UPPSALA
UNIVERSITET

A COMPARISON OF SOME ESTIMATION METHODS FOR HANDLING OMITTED VARIABLES : A SIMULATION STUDY

Submitted by
Philomina Amartey

A thesis submitted to the Department of Statistics in partial fulfillment of the requirements for a two-year Master of Arts degree in Statistics in the Faculty of Social Sciences

Supervisor
Philip Fowler

Spring, 2020

ABSTRACT

Omitted variable problem is a primary statistical challenge in various observational studies. Failure to control for the omitted variable bias in any regression analysis can alter the efficiency of results obtained. The purpose of this study is to compare the performance of four estimation methods (Proxy variable, Instrumental Variable, Fixed Effect, First Difference) in controlling the omitted variable problem when they are varying with time, constant over time and slightly varying with time. Results from the Monte Carlo study showed that, the prefect proxy variable estimator performed better than the other models under all three cases. The instrument Variable estimator performed better than the Fixed Effect and First Difference estimator except in the case when the omitted variable is constant over time. Also, the Fixed Effect performed better than First Difference estimator when the omitted variable is time-invariant and vice versa when the omitted is slightly varying with time.

Keywords: Panel Data, Fixed Effect, Proxy, Instrumental Variable, First Difference.

Contents

1 Introduction

Having all the required theoretical assumptions met, the Ordinary Least Square (OLS) estimator consistently estimates the coefficients of any linear regression model. One common notable reason for the inconsistency in the OLS estimator is when an explanatory variable in the model correlates with another variable whose data is omitted from the observed data yet affects the outcome.

The problem of omitted variables **(OV)** in research analysis cuts across most scientific disciplines but mostly in the social sciences when dealing with observational data. The omitted variable problem is an issue when we would like to control for one or more additional variables but, usually because of data unavailability, they are not included in a regression model (Wooldridge, 2010). This produces a bias in the coefficients estimated in a regression model which makes the problem of omitted variables one of the most serious problems in regression analysis. Most variables are usually omitted from a regression model because of unavailability of data, inability of it being measured, the impossibility to model how the omitted variables interact with the included variables, and the insufficient knowledge about the influence of the omitted variables (Leightner and Inoue, 2007). Amidst these reasons, researchers have over the years sought to find solutions or methods in handling them as evidenced by quite a number of published papers and articles. In recent studies, economists have come up with improved but complex techniques in handling this omitted variable problem especially in relation to the data used such as the panel data methods (Leightner and Inoue, 2012).

Panel data is a data set collected over a period of time for the same cross section units or individuals. One fundamental motive for using panel data is to solve the omitted variable problem (Wooldridge, 2010, pp. 281). Detailed explanations on panel data estimation methods and its practical applications have been covered in most econometric literature and textbooks (see eg. Greene, 2003 and Baltagi, 2001). The most frequently discussed panel data estimators are the Random effects, Fixed effects (FE), First Difference (FD) among others. Some studies (eg. Suparman, 2015, Arellano, 2003 and Wooldridge, 2010) have shown that these estimators work better when the omitted variables are not varying with time and the traditional methods of proxies and instrumental variables works better with cross section data. All these estimators function well when all their strong assumptions are met (Leightner & Inoue, 2012). Inspite of numerous publications made on the topic of omitted variables, much has not been done on comparing some estimation methods that corrects the omitted variable bias in panel data.

The purpose of this study is to compare the performance of four different approaches of solving the omitted variable problem using panel data through a simulation study. The four methods evaluated are the; Proxy variable, Instrumental Variable, Fixed Effect and First Difference methods. The motivation behind this study is to explore which of these four estimation methods will perform better in the case when the omitted variable is varying with time, constant over time and slightly varying with time.

The rest of this thesis is structured as follows. The next section gives a literature review of the study which summarizes research works that have been carried out on the topic of the OV problem. Section 3 describes the estimation methods, data generation process and the simulation settings. The results from the study is outlined in Section 4 while the conclusions drawn are in Section 5.

2 Literature review

There have been many studies carried out on the omitted variable problem since this has been a rising issue in most statistical and economic research works. Thus many techniques and ways of solving this problem has emerged over the years and have been documented in journals, articles and books (Bou and Satorra, 2017). Although there are numerous studies on the topic of the omitted variable problem with panel data, this project mainly focuses on comparing some estimation methods that have been used to solve the omitted variable problem. This section also reviews literature that is relevant to the study.

A Monte Carlo simulation study was conducted by Suparman (2015) in which time-varying omitted variables were controlled for using the methods of latent fixed effects regression, demeaning, first order differencing, autoregression and constrained autoregression (Suparman et al., 2014). From the bias, standard error and mean squared errors of the estimators, they concluded that the constrained autoregression performs better in controlling for time-variant omitted variables than the other estimation methods considered in the study.

Also, Beccarini (2010) demonstrated a method of handling omitted variables bias in panel data by regime-switching regression approach with some Monte Carlo simulations. The Monte Carlo simulations were performed under three different cases; the first case being that the omitted variable is not autocorrelated but binary, the second case for which the omitted variable is autocorrelated and binary and the last case were the autocorrelated omitted variable is defined on a real line. He further gave an empirical verification based on Fisher's equation and thus concluded that the omitted variable bias can be corrected using the regime-switching approach.

Du et al. (2018) also used some common empirical strategies to control for omitted variables in a sample linked administrative labour market data. They conducted their analysis on exemplary wage regressions and labor market transitions and related the results of cross sectional analysis with those of panel analysis to investigate the extent additional cross sectional variables explain the variation in unobserved individual time invariant effects. Their findings suggested that additional cross sectional variables control for considerably less relevant information than fixed effects in panel analysis. Also from their study, they concluded that unobserved effects panel data models with a restricted regressor set are found to control far more information than cross sectional analysis with an extended variable set.

The question of how omitted variable bias can be eliminated when the strong assumptions of using the proxy and instrumental variables are not met motivated Leightner and Inoue (2012) to

come up with a new analytical technique named Reiterative Truncated Projected Least Squares (RTPLS). In their paper, they explained that the RTPLS produces reduced form estimations while minimizing the influence of omitted variables and the researcher would not have to know what the omitted variables are, neither its relationship to an instrument, proxy, or the dependent variable, nor finding an appropriate instrument or proxy. With results from Monte Carlo Simulations, they concluded that the RTPLS produces less bias than OLS when there are omitted variables that interact with included variables. Also, they referred their found technique as a recent one and may require more studies to improve it.

This paper combines knowledge from the aforementioned research studies with the aim of controlling the omitted variable problem by a comparative study of four well-known estimation methods.

3 Methodology

3.1 Model

Consider the panel data regression model for the study :

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + u_{it} + \varepsilon_{it} \tag{1}$$

where, for each individual *i* at time point *t*,

$$\begin{aligned} y_{it} &= \text{values of the dependent variable} \\ x_{it} &= \text{values of time-varying independent observed variables} \\ u_{it} &= \text{the omitted or unobserved variable} \\ \varepsilon_{it} &= \text{the random error term , } \varepsilon_{it} \sim N(0, \sigma_{\varepsilon}^2) \\ \beta_0, \beta_1, \beta_2, \beta_3 &= \text{constant values.} \end{aligned}$$

We would like to estimate β_1 which is the effect of the observed explanatory variable x_{1it} on the outcome y_{it}, holding all the others constant and this can easily be done with the OLS estimator on condition that data for all variables are at hand. Now, suppose variable u_{it} is omitted from the equation (1) due to unavailability of data but correlates with the variable of interest and also a determinant of the dependent variable y_{it}, then using the OLS method to estimate the coefficient of x_{1it} will give rise to the OV bias. In this study, four estimation techniques (Proxy variable, Instrument Variable, Fixed Effect and First Difference) are used to control for the OV bias and their results are compared.

3.2 Estimation Methods

There are several estimation methods (see eg. Wooldridge, 2010; Greene, 2003 and Baltagi, 2001) that are used in controlling for omitted variable bias but for the purposes of this study, only these will be discussed.

3.2.1 Proxy Variable approach

According to Judd and Kenny (1981), "Proxy variables are variables that are included in an analysis because they represent at least in part the omitted theoretical construct that is the major interest". Hence the omitted variables bias can be at least corrected, if a proxy variable is

available for the unobserved variable. According to Wooldridge (2010), a good proxy variable for the omitted or unobserved variable must fulfill these requirements;

- The first is that the proxy variable should be redundant in the equation model (1). This means that the proxy variable is not required for explaining the dependent variable once the explanatory and omitted variables have been controlled for.

- The second requirement of a good proxy variable is that the correlation between the omitted variable and each explanatory variable, should be zero once we take out the influence of the proxy. This requirement needs the proxy to be closely enough related to the omitted variable so that once it is included in the model equation (1), the explanatory variables are not partially correlated with the omitted variable.

It is quite simple for a proxy variable to satisfy the first requirement but not the second requirement. A proxy is therefore described as imperfect when it doesn't meet the second requirement. In practice, researchers work with imperfect proxy variables (Leightner & Inoue, 2012).

3.2.2 Instrumental Variable method

An instrumental variable is a variable used in a regression analysis that involves an endogenous variable (one that correlates with the error term) to describe the true correlation between the dependent variable and the explanatory variable of interest (Stephanie, 2016). The method of instrumental variables (IV) provides an answer to the problem of an endogenous explanatory variable (Wooldridge, 2010, p. 89).

To have a clearer understanding of this approach, consider the model in equation (1) where the variable of interest x_{1it} correlates with the omitted variable u_{it}, hence x_{1it} is an endogenous variable. There is therefore the need to find a suitable instrument, z_{it} that satisfies the following requirements;

- It must be irrelevant in the structural model in equation (1)

- It must be uncorrelated with the omitted variable, u_{it} ie. $Cov(z_{it}, u_{it}) = 0$ and

- It must be highly correlated with the endogenous variable.

When z_{it} attains all these requirements, it is described as an instrumental variable for x_{1it}. Anytime an instrumental variable does not fully fulfil the third requirement, it is termed as a

weak instrument. The two-stage least squares (2SLS) method is one of the most efficient way to estimate models with instrumental variables (Wooldridge, 2010). As the name presupposes, this method is carried out in a two-step procedure: First, regress the endogenous variables on the instruments and obtain fitted values. Afterwards, run the OLS regression of the dependent variable on the fitted endogenous variables and the other explanatory variables. For easier implementation of this estimation method, a package AER (Kleiber & Zeileis, 2020) from the R (RStudio Team, 2019) software will be used.

3.2.3 Fixed Effect method

The subject of fixed effect estimation has mostly been considered with panel data analysis (Greene, 2003). Suppose the omitted variable is constant over time, then equation (1) will be redefined as;

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + u_i + \varepsilon_{it} \tag{2}$$

A fixed effects regression is an estimation technique that involves averaging equation (2) over time and subtracting from each variable then estimating the resulting transformed model by Ordinary Least Squares (Arellano, 2003). This procedure, known as within transformation, gives the chance to leave out the unobserved component and consistently estimate the coefficient of the variable of interest. Analytically, the FE model is then defined as:

$$\tilde{y}_{it} = \boldsymbol{\beta}\tilde{\mathbf{x}}_{it} + \tilde{\varepsilon}_{it} \tag{3}$$

where $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\tilde{\mathbf{x}}_{it} = x_{it} - \bar{x}_i$ and $\tilde{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$ with $\bar{y}_i$, $\bar{x}_i$, $\bar{\varepsilon}_{it}$ as the individual averages over time. This time demeaning of the original equation (2) has removed the individual effect (which is also the omitted variable) u_i. This is because in this case the omitted variable, u_i is constant over time hence time demeaning of the original equation (1) eliminates u_i. The fixed effect estimator is the pooled OLS estimator from the regression of $\tilde{y}_{it}$ on $\tilde{\mathbf{x}}_{it}$. It is also called the within estimator because it uses the time variation within each individual unit. Since the time demeaning has been carried out, the fixed effect estimator can be simply computed. As stated by Stock and Watson (2007, p. 289-290), "The key insight is that if the unobserved variable does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics".

According to Wooldridge (2010), these assumptions must hold for the fixed effect estimator to be consistent;

- Strict exogeneity of the explanatory variables conditional on u_i :

$$E(\varepsilon_{it}|x_i, u_i) = 0$$

- Full rank :

$$rank(\mathbf{X}) = rank(\mathbf{X'X}) = K$$

  where K is the number of explanatory variables.

- Homoscedasticity:

$$E(\varepsilon_{it}^2|x_i, u_i) = \sigma_u^2$$

Inasmuch as the FE is an efficient estimator, there are however some shortcomings of this estimator; the within transformation does not allow one to include time-invariant independent variables in the regression, because they get eliminated similarly to the fixed unobserved component. Also, standard errors for fixed effects coefficients are mostly higher than those for other methods, especially when the explanatory variable has little variation over time (Allison, 2009).

3.2.4 First Difference method

The first difference method is another approach to solve the OV problem. Let's consider the equations below with the omitted variable constant over time and $t = 1, ..., T$:

$$y_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + u_i + \varepsilon_{it} \tag{4}$$

$$y_{it-1} = \beta_1 x_{1it-1} + \beta_2 x_{2it-1} + \beta_3 x_{3it-1} + u_i + \varepsilon_{it-1} \tag{5}$$

Differencing these two equations removes the time-invariant variable u_i resulting in:

$$\Delta y_{it} = \beta_1 \Delta x_{1it} + \beta_2 \Delta x_{2it} + \beta_3 \Delta x_{3it} + \Delta \varepsilon_{it} \tag{6}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\Delta x_{it} = x_{it} - x_{i,t-1}$, and $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$.
This is called the first difference regression model. A pooled OLS estimator from the regression on Δy_i on Δx_{1i} is referred to as the fixed difference estimator (Wooldridge, 2010, p. 316). For the FD estimator to be consistent, these assumptions must be valid:

- Strict exogeneity of the explanatory variables :

$$E(\Delta\varepsilon_{it}|\Delta x_{it}) = 0$$

- Rank condition of the FD estimator must be equal to the number of explanatory variables in the model, *K*:

$$rank(\Sigma_{t=2}^{T}\mathbf{\Delta x_{it}}'\mathbf{\Delta x_{it}}) = K$$

- the first differences of the random errors, ε_{it} are serially uncorrelated (and have constant variance) :

$$E(\mathbf{\Delta}\varepsilon_{\mathbf{it}}'\mathbf{\Delta}\varepsilon_{\mathbf{it}}|\Delta x_{it}, u_i) = \sigma_{\varepsilon}^2$$

The FD estimator is easier to implement and produces identical estimates as the FE estimator when $t = 2$ but can also be applied when the time periods are more than two. One limitation of the FD estimator is that it cannot be used when the explanatory variable of interest is time-constant and imprecise when the explanatory variable changes little over time (Hill et al., 2019).

3.3 Simulation Design

3.3.1 Data Generating Process (DGP)

We begin the simulation study by generating data based on the model in equation (3). An illustration of how the data are generated for this study is shown in the Figure 1 below :

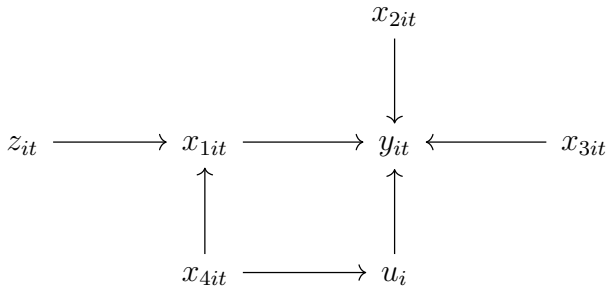

Figure 1: A diagram of the DGP

First, the explanatory variables x_{2it}, x_{3it}, proxy variable x_{4it}, instrumental variable z_{it} and random errors, *r* and ε are generated to follow a multivariate random normal distribution with

a zero mean vector and a covariance matrix;

$$\sum = \begin{array}{c} \\ x_{2it} \\ x_{3it} \\ x_{4it} \\ z_{it} \\ r_{it} \\ \varepsilon_{it} \end{array} \begin{array}{c} \begin{array}{cccccc} x_{2it} & x_{3it} & x_{4it} & z_{it} & r_{it} & \varepsilon_{it} \end{array} \\ \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 10 \end{bmatrix} \end{array}$$

These variables are independently distributed and generated to vary randomly over time across each individual *i*. Next, the omitted variable is generated as a linear function of the proxy, x_4it and the random error, r_{it} with a constant term as stated in Wooldridge (2010) . The omitted variables are generated under three settings;

1. Time varying: With the time varying, u_i is generated to be randomly varying over time for each individual unit and this brings about a within-unit variation in the values of the OV. It is defined as :

$$u_{it} = \theta_0 + \theta_1 x_{4it} + r_{it}$$

2. Time-invariant: Under the time-invariant setting, the omitted variable is made constant over time for each individual unit in that an individual has the same random constant OV value for all four time periods. This brings about a between-unit variation in the values of the OV. The values of proxy variable were also made constant over time under this setting. It is mathematically defined as :

$$u_i = \theta_0 + \theta_1 x_{4i} + r_i$$

3. Slightly varying with time: Here, each individual is made to have the same OV value for the first two time points and a different but same OV value for the next two. The individual has the same error term, r_{it} for the first two time points and a different but same value for the next two. This causes both between-unit and within-unit variation in the OV values generated. There can be many possible ways of defining the term "slightly varying with time" but the choice of defining it this way is to investigate how sensitive the FE and FD models are to this setting.

The θ_0 and θ_1 have constant values of 0.2 and 0.5 respectively in all cases. Also, the variable of interest x_{1it} is generated as a linear function of the instrumental variable z_{it}, the proxy variable x_{4it} and some random noise, ϵ_{it} with a constant term :

$$x_{it} = 0.1 + \beta_z z_{it} + \beta_u x_{4it} + \epsilon_{it}$$

where $\epsilon_{it} \sim N(0, 1)$. The β_z and β_u have constant values of 0.7 and 0.8 respectively. The weak instrument was made to have a low correlation with x_{1it} by reducing β_z to 0.1 and this violates the third requirement of a strong instrument (see section 3.2.2). Also for the imperfect proxy, the random error term,r_{it} in the OV function was made to correlate with the variable of interest, x_{1it} by including a $0.2r_{it}$ term in the x_{it} function (Wooldridge, 2010). Given the explanatory variables, the omitted variable, and the error term, ε_{it} the dependent variable is generated according to the true model;

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + u_i + \varepsilon_{it} \tag{7}$$

The values of β_0, β_1, β_2 and β_3 are 0.5, 0.5, 0.3 and 0.3 respectively.

3.3.2 Simulation method

A number of replications, $R = 10,000$ was chosen for the study. For each replication, sample sizes of 100, 500 and 1000 with a fixed time point of four periods was conducted. In each replications, data sets were generated according to the DGP and carried out differently for all three settings of the omitted variables. For each generated data set in each replication, these five models were estimated.

1. Omitted model: In this model, the dependent variable y_{it}, was regressed on only the explanatory variables x_{1it}, x_{2it}, x_{3it} and the estimated coefficient of x_{1it} is stored in each replication. The omitted model is termed as model omit in the table of results. This model reveals the severity of the bias caused by the OV.

2. Proxy model: For the proxy model, an OLS regression was run on the dependent variable, y_{it} and the explanatory variables including the proxy x_{4it}.

3. Instrumental variable model: In this model, the 2SLS method (discussed in section 3.2.2) was used. Here, an OLS regression was run on y_{it} and the explanatory variables, with the variable z_{it} as an instrument for x_{1it} and variables x_{2it}, x_{3it} serving as their own instrumental variables.

4. Fixed effect model: Here, an OLS regression was run on the demeaned variables $\tilde{y}_{it}$ and $\tilde{x}_{1it}$, $\tilde{x}_{2it}$ $\tilde{x}_{3it}$ as stated in section 3.2.3.

5. First difference model: Just as the FE model, the OLS regression was run on the differenced data as discussed in section 3.2.3.

It was essential to investigate how these methods will perform in the presence of a weak IV and an imperfect proxy therefore two simulation settings were conducted under the case of the time-variant OV and time-invariant OV and one simulation under the case of slightly time-variant OV. For the time-invariant and time-variant OV, one simulation setting was for the perfect proxy and strong IV methods and the other simulation for the imperfect proxy and a weak IV methods. The motivation for doing this was that, generally in applied or observational research works, researchers mostly work with weak instruments and poor proxies (Leightner & Inoue, 2007) so it is needful to evaluate the performance of these model under different circumstances.

These methods were estimated using commands from packages in the R software. The *plm* function from the plm (Croissant & Millo, 2008) package was used to estimate the FE and FD models and *ivreg* function from the AER (Kleiber & Zeileis, 2020) package for the IV model. The estimates from the 10000 replications for each estimation method are stored. The performance of these methods were measured by means of the bias, standard deviations (SD) and mead square errors (MSE) of the estimated coefficient, $\hat{\beta}_1$.

4 Results and Discussion

After implementing the methods of estimation on the data generated as discussed, the results are outlined in tables below. There are two tables each for when the omitted variable is randomly varying with time and constant over time and only one for the slightly varying with time.

From Table 1, a comparison of these methods based on the omitted variables randomly varying with time was done. It is observed that the estimates for the proxy model and IV model (both having a value of 0.499) performed better than that of the FE and FD model of values 0.686 and 0.687 respectively. As a result, the FE and FD model is as biased as the omitted model when the OV is varying with time when the sample sizes increases. The standard deviations of all the models keep decreasing as the sample size increases and the IV model has the largest SD among the methods for all three sample sizes. The SD for the IV model is 10 percent more than the proxy and FE methods. Also the MSE slowly decreases as the sample size increases for the FE and FD models but with the proxy and IV model, it decreases rapidly as the sample sizes increases. Although the proxy and IV produced unbiased estimates, the MSE for the IV is the highest and that of the proxy is the lowest. The time invariant methods performed poorly compared to the others and the FE model did slightly better than the FD model as expected (Suparman, 2015).

Table 1: Simulation results with time-variant omitted variables.

		$\beta_1 = 0.5$, R =10000				
		Model Omit	**Proxy**	**IV**	**FE**	**FD**
n=100	Estimate	0.687	0.499	0.499	0.686	0.687
	Bias	0.187	-0.001	-0.001	0.186	0.187
	Std. Dev.	0.115	0.135	0.242	0.132	0.152
	MSE	0.049	0.019	0.060	0.052	0.059
n=500	Estimate	0.687	0.499	0.499	0.686	0.688
	Bias	0.187	-0.001	-0.001	0.186	0.188
	Std. Dev.	0.052	0.061	0.107	0.039	0.062
	MSE	0.038	0.004	0.012	0.039	0.040
n=1000	Estimate	0.688	0.500	0.500	0.687	0.687
	Bias	0.188	0.000	0.000	0.187	0.187
	Std. Dev.	0.037	0.043	0.076	0.042	0.048
	MSE	0.037	0.002	0.006	0.037	0.038

A graphical representation of how these methods performed based on the values of their MSE's is illustrated in the Figure 2 below. The plot further reveals how well the proxy and IV methods performed as compared to the other methods since a lower MSE indicates a better performance of an estimator.

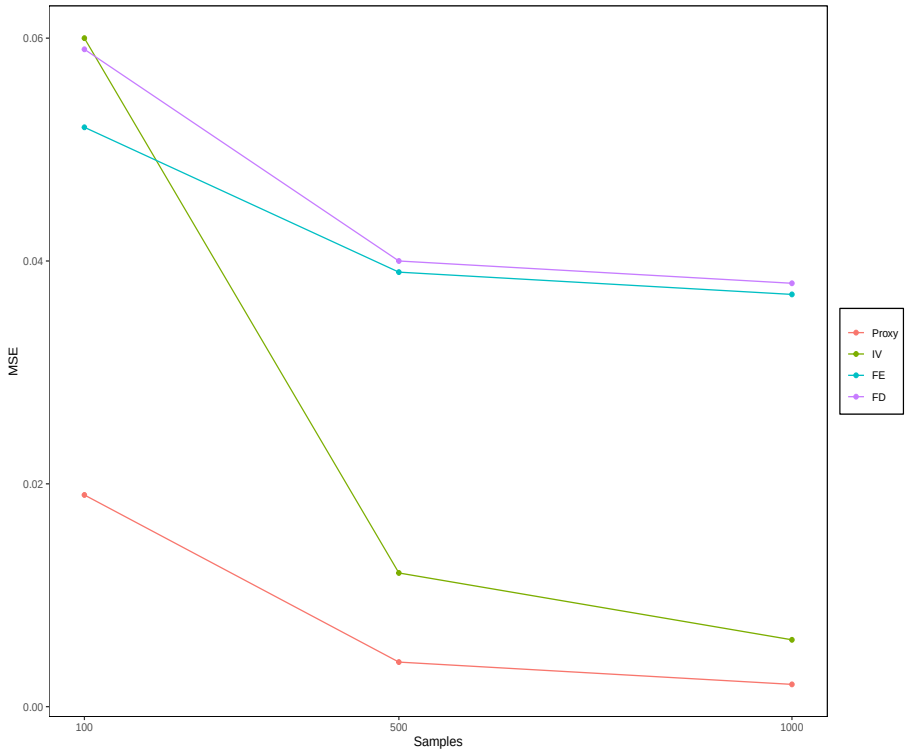

Figure 2: Time-Variant OV

In Table 2, the instance of the IV being a weak instrument for the variable of interest and the proxy being imperfect while the omitted variable is still varying with time was investigated. It is observed that all the four methods produced biased estimates with the weak IV model being highly biased and imperfect proxy the least biased estimator. The SD and MSE values of all the methods keeps decreasing with an increasing sample size just as in Table 1. The weak IV model gets less biased as the sample size increases to 500 and gets even better when sample size is 1000. The SD and the MSE for the weak IV model when the sample size is 100 are enormously larger than the others but spontaneously gets smaller with increasing sample size. These findings confirms the documented literature on weak IV models and their biases with small sample size (see Wooldridge, 2013). The FE and FD models performed worse in this case as in Table 1. Although the imperfect proxy produced biased estimates, it obtained the lowest MSE for all sample sizes as compared to the others. It can be deduced that the imperfect proxy performed better than the rest in this case and shows how efficient the proxy method can be even when the proxy is an imperfect one under an OV time-variant setting.

Taking a closer look at Table 2, one key observation is that the FE and FD models performed worse when the OV is varying with time, even in the midst of weak IV and imperfect proxy.

Table 2: Simulation results with time-variant omitted variables.

		$\beta_1 = 0.5$, R =10000				
		Model Omit	**Imp. Proxy**	**Weak IV**	**FE**	**FD**
n=100	Estimate	0.853	0.688	2.228	0.853	0.853
	Bias	0.353	0.188	1.728	0.353	0.353
	Std. Dev.	0.129	0.163	188.507	0.148	0.171
	MSE	0.141	0.062	35534.275	0.147	0.154
n=500	Estimate	0.855	0.691	0.464	0.855	0.853
	Bias	0.355	0.191	-0.036	0.355	0.353
	Std. Dev.	0.057	0.072	1.239	0.066	0.076
	MSE	0.129	0.042	1.537	0.131	0.132
n=1000	Estimate	0.855	0.690	0.484	0.855	0.855
	Bias	0.355	0.190	-0.016	0.355	0.355
	Std. Dev.	0.040	0.072	0.577	0.047	0.053
	MSE	0.127	0.042	0.333	0.128	0.129

The next table (Table 3), reports the case when the OV is constant over time for each individual. Here, the FE and FD estimators performed equally well as the proxy and IV estimators. Though both FE and FD methods give the same unbiased estimates, the values of SD and MSE for the FE estimator are lower than that of FD. This is as a result of the errors in the model not serially correlated making the FE estimator more efficient than the FD (Wooldridge, 2013, pp. 487). This result of the FE and FD methods producing same estimates for time-invariant omitted variables is an evidence that both models are consistent with fixed T and $N \to \infty$ (Wooldridge, 2013, pp. 487).

Just as previous tables, the values of the SD and MSE keeps decreasing as the sample size increases for all methods with the IV having the highest values of SD and MSE.

Table 3: Simulation results with time-invariant omitted variables.

		$\beta_1 = 0.5$, R =10000				
		Model Omit	**Proxy**	**IV**	**FE**	**FD**
n=100	Estimate	0.686	0.500	0.496	0.500	0.500
	Bias	0.186	0.000	-0.004	0.000	0.000
	Std. Dev.	0.120	0.135	0.242	0.149	0.171
	MSE	0.049	0.018	0.058	0.022	0.029
n=500	Estimate	0.686	0.499	0.500	0.498	0.499
	Bias	0.186	-0.001	0.000	-0.002	-0.001
	Std. Dev.	0.054	0.062	0.107	0.067	0.077
	MSE	0.038	0.004	0.011	0.004	0.006
n=1000	Estimate	0.687	0.500	0.499	0.500	0.500
	Bias	0.187	0.000	-0.001	0.000	0.000
	Std. Dev.	0.038	0.044	0.076	0.048	0.055
	MSE	0.036	0.002	0.006	0.002	0.003

A graphical representation of how these methods performed based on the values of their MSE's are illustrated in the f Figure 3 below. It also reveals the performance of the FE and FD estimators now that the OV is time-invariant.

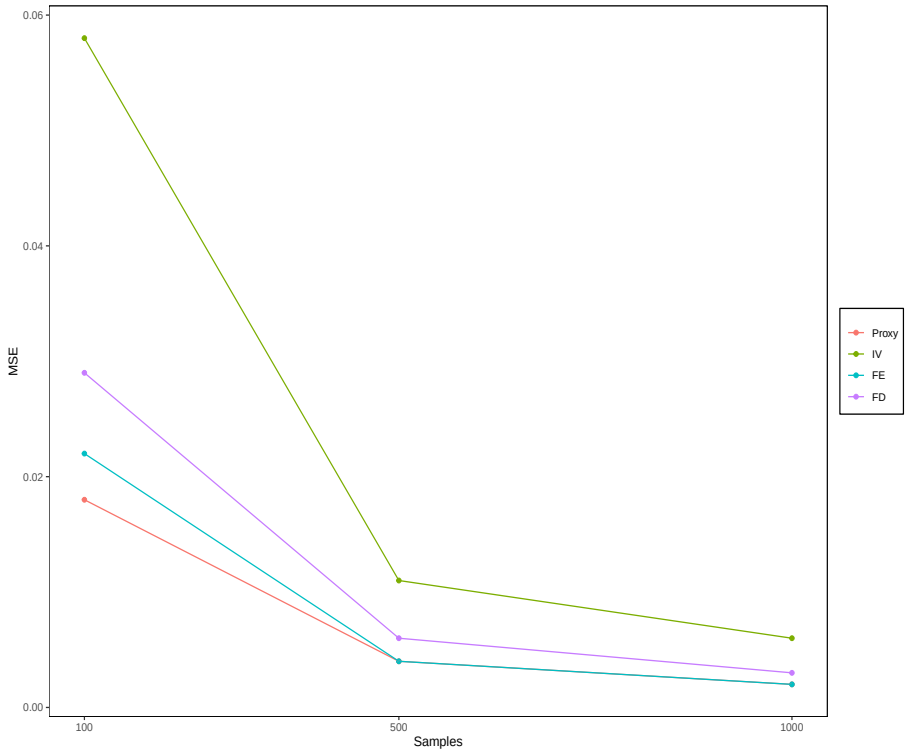

Figure 3: Time-Invariant OV

Here in Table 4, the FE and FD estimators still did better while the imperfect proxy and weak IV did worse but the weak IV estimator produced a high biased estimates with higher SD and MSE values making the weak IV a very poor estimator of time-invariant OV. The weak IV also showed a similar trend of an extremely large value of SD and MSE when the sample size is 100 but quite better (still the highest) when the sample size is 1000 just as in Table 2. This provides evidence to what Baltagi (2007, pp. 263) said about weak instruments producing biased estimates even with large samples since the 2SLS estimator is sensitive to weak instruments. The good performance of the FE and FD models when the OV is constant over time confirms the theoretical intuition of these models working well when the OV is time-invariant (see Baltagi, 2001, Greene, 2003).

Table 4: Simulation results with time-invariant omitted variables.

		$\beta_1 = 0.5$, R =10000				
		Model Omit	**Imp. Proxy**	**Weak IV**	**FE**	**FD**
n=100	Estimate	0.852	0.686	0.363	0.496	0.496
	Bias	0.352	0.186	-0.137	-0.004	-0.004
	Std. Dev.	0.137	0.165	39.726	0.184	0.211
	MSE	0.142	0.062	1578.046	0.034	0.045
n=500	Estimate	0.855	0.690	0.671	0.501	0.501
	Bias	0.355	0.190	0.171	0.001	0.001
	Std. Dev.	0.061	0.073	38.540	0.081	0.094
	MSE	0.130	0.042	1485.226	0.007	0.009
n=1000	Estimate	0.855	0.690	0.486	0.500	0.500
	Bias	0.355	0.190	-0.014	0.000	0.000
	Std. Dev.	0.043	0.052	0.570	0.058	0.066
	MSE	0.128	0.039	0.325	0.003	0.004

Table 5, contains results of the case where the omitted variable is slightly varying with time. Here, the proxy and IV model have unbiased estimates with their SD and MSE decreasing as the sample size increases. Both FE and FD methods produced biased estimates but the the FD has a 1.8 less bias than the FE. The MSE for the FD is also lower than FE for all the samples which has never been the case for the previous tables. Hence a conclusion of the FD estimator performing better than the FE when OV is slightly time-invariant can be made. The rationale for the FD performing better than the FE method in this case could be as a result of how the slightly varying OV was defined in the data generation process. The reason could be that since the FD has to do with differencing data while the FE has to do with time-demeaning, it is highly probable the FD method will perform better in a data where an individual has the same value for two different time points. Also, the data generation process for the slightly time varying OV was carried out in another way where each individual was made to have same constant values for three time points and different for the last time point. Even with that, the FD still performed better than the FE model. These tables are available also upon requests. It can then be said that, the FD estimator is less sensitive to time variation in the OV than the FE is.

Table 5: Simulation results from slightly time-variant omitted variables.

		$\beta_1 = 0.5$, $\mathbf{R} = 10000$				
		Model Omit	**Proxy**	**IV**	**FE**	**FD**
n=100	Estimate	0.688	0.499	0.499	0.639	0.578
	Bias	0.188	-0.001	-0.001	0.139	0.078
	Std. Dev.	0.117	0.138	0.244	0.140	0.163
	MSE	0.049	0.019	0.060	0.038	0.033
n=500	Estimate	0.688	0.500	0.500	0.639	0.578
	Bias	0.188	0.000	0.000	0.139	0.078
	Std. Dev.	0.052	0.061	0.109	0.062	0.073
	MSE	0.038	0.004	0.012	0.023	0.012
n=1000	Estimate	0.688	0.500	0.500	0.639	0.578
	Bias	0.188	0.000	0.000	0.139	0.078
	Std. Dev.	0.037	0.043	0.075	0.044	0.051
	MSE	0.037	0.002	0.006	0.021	0.009

A graphical representation of how these methods performed based on the MSE's are illustrated in the Figure 4 below. It reveals how better the FD estimator performed compared to the FE model when the OV is slightly time varying.

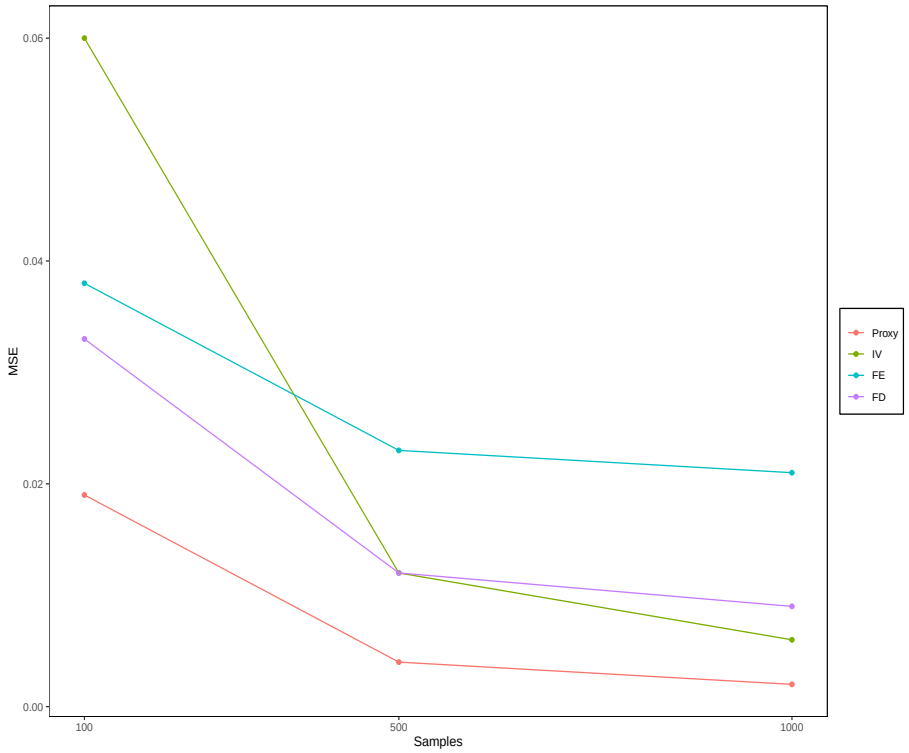

Figure 4: Slightly Time-Invariant OV

From the discussions on all tables of results, not much was said about the omitted model (model omit) because it serves a control for the other four models ie. to verify how biased or unbiased the four estimation methods are in solving OV problem. Also, the model omit produced the expected biased estimate under all circumstances, as evident in the results. There are no table of results for weak IV model and imperfect model for the case of slightly varying OV because the main point was to explore how sensitive the FE and FD model were to very small changes in the OV in respect to its variation with time. Also the aspect of time points increasing from $T = 4$ to $T = 8$ to $T = 16$ for sample size 100, 500 and 1000 respectively was looked at and the trend in results were not any different to when only the sample sizes were increasing. The tables for these results are available upon request.

Furthermore, the proxy and IV model obtained unbiased estimates under all three cases with exceptions of when the proxy is imperfect and the IV is a weak one.

5 Conclusion

Omitted variables can really affect the accuracy of results when it is not tackled in observational studies. The main purpose of this study is to compare four estimation methods of solving the OV problem when the OV is randomly varying with time, constant over time and slightly varying with time through a Monte Carlo simulation study. Four methods were considered and these were; Proxy variable, Instrumental Variable, Fixed effect and First differencing methods. From the results of this study, the proxy variable method performed best under all three cases on condition that the proxy is perfect. The IV estimator performed better than the FE and FD methods except in the case when the OV is time-invariant. Also, a weak IV estimator can cause a huge bias with an enormous MSE when the sample size is small. Further, the FE method performed better than the FD when the OV is constant over time and vice versa when the OV is slightly varying with time making the FE more sensitive to any small variation from the time-invariant OV.

Usually in practice, it is difficult to get a perfect proxy and good IV but from the study, it would be advisable to use an imperfect proxy rather than a weak instrument when the sample size is small and the OV is varying with time. In addition, anytime the OV is constant over time with an increasing sample size and fixed time point, the FE is recommendable. Also if any real data has its OV slightly varying with time as generated in this study, then FD model can be chosen over the FE model.

For further studies on this topic, more estimation methods like the RTPLS can be compared with these four in order to evaluate their performances in solving the OV problem. Otherwise, a comparison of this kind can be carried out with real data to confirm conclusions drawn on the performance of these estimation methods from this study.

References

Allison, P. D. (2009). *Fixed effects regression models* (Vol. no. 07-160.). Los Angeles, SAGE.

Arellano, M. (2003). *Panel data econometrics*. New York, Oxford University Press.

Baltagi, B. H. (2001). *Econometric analysis of panel data* (2.). New York, Wiley.

Baltagi, B. H. (2007). *Econometrics*. New York, Springer Science & Business Media.

Beccarini, A. (2010). Eliminating the omitted variable bias by a regime-switching approach. *Journal of Applied Statistics*, 37(1), 57–75. https://doi.org/10.1080/02664760902914474

Bou, J. C., & Satorra, A. (2017). Univariate Versus Multivariate Modeling of Panel Data: Model Specification and Goodness-of-Fit Testing. *Organizational Research Methods*, 21(1), 150–196.

Croissant, Y., & Millo, G. (2008). Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software*, 27(1), 1–43. https://doi.org/10.18637/jss.v027.i02

Du, S., Homrighausen, P., & Wilke, R. A. (2018). *On Omitted Variables, Proxies and Unobserved Effects in Analysis of Administrative Labour Market Data* (tech. rep.). The Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

Greene, W. H. (2003). *Econometric analysis* (5th ed). Upper Saddle River, N.J, Prentice Hall.

Hill, T. D., Davis, A. P., Roos, J. M., & French, M. T. (2019). Limitations of Fixed-Effects Models for Panel Data. *Sociological Perspectives*, (Journal Article), 73112141986378. https://doi.org/10.1177/0731121419863785

Judd, C. M., & Kenny, D. A. (1981). *Estimating the Effects of Social Intervention*. CUP Archive.

Kleiber, C., & Zeileis, A. (2020). AER: Applied Econometrics with R. Retrieved May 13, 2020, from https://CRAN.R-project.org/package=AER

Leightner, J. E., & Inoue, T. (2007). Tackling the omitted variables problem without the strong assumptions of proxies. *European Journal of Operational Research*, 178(3), 819–840. https://doi.org/10.1016/j.ejor.2006.02.022

Leightner, J. E., & Inoue, T. (2012). Solving the Omitted Variables Problem of Regression Analysis Using the Relative Vertical Position of Observations. *Advances in Decision Sciences*, 2012, 1–25. https://doi.org/10.1155/2012/728980

RStudio Team. (2019). R: A Language and Environment for Statistical Computing. *Vienna, Austria*, R Foundation for Statistical Computing. https://www.R-project.org/

Stephanie. (2016). Instrumental Variable: Definition & Overview. Retrieved May 19, 2020, from https://www.statisticshowto.com/instrumental-variable/

Stock, J. H., & Watson, M. W. (2007). *Introduction to econometrics* (2nd ed.). Boston, MA, Pearson Addison Wesley.

Suparman, Y. (2015). *Controlling omitted variables and measurement errors by means of constrained autoregression and structural equation modeling* (Doctoral dissertation). University of Groningen.

Suparman, Y., Folmer, H., & Oud, J. (2014). Hedonic price models with omitted variables and measurement errors: A constrained autoregression-structural equation modelling approach with application to urban Indonesia. *Journal of Geographical Systems*, 16(1), 49–70. https://doi.org/10.1007/s10109-013-0186-3

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, United States, MIT Press.

Wooldridge, J. M. (2013). *Introductory econometrics : A modern approach* (5th ed., international ed). Mason, Ohio, South-Western Cengage Learning.