Linear and Non-Linear Panel IV Estimators and the Control Function Approach – A Simulation Based Investigation of Estimator Performance

Richard Winter*

2022-06-08

Abstract

In this brief I investigate the performance of the Control Function method used to identify a causal effect in the presence of endogeneity of the treatment variable. The methods will be illustrated for linear and non-linear panel models, and in a simulation exercise the Control Function method will be compared for PPML and Log-Linear Panel OLS.

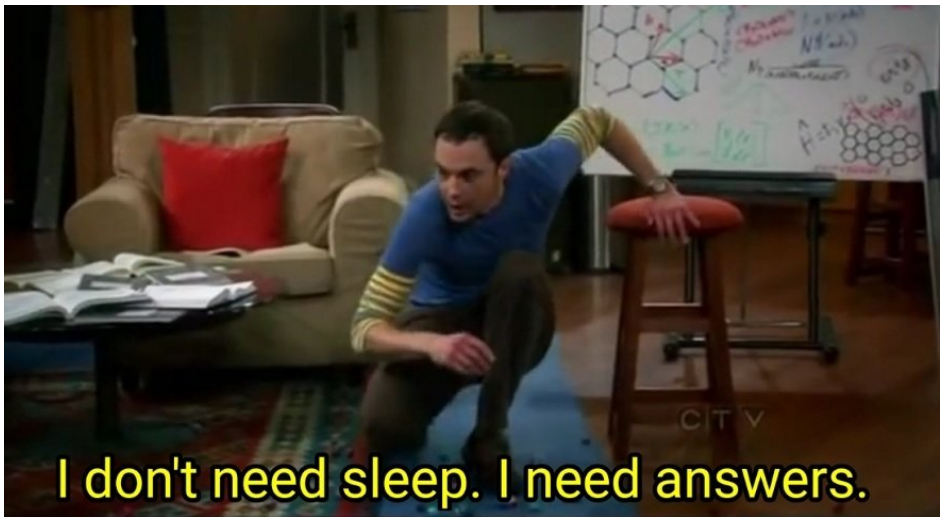


Figure 1: The Hunt for Truth

*University of Mannheim, rwinter@uni-mannheim.de

1 Introduction

This paper evaluates the performance of several approaches to account for omitted variable bias and endogeneity in linear and non-linear panel models. To this end, three data generating processes will be simulated and estimated.

Linear Panel Data Model

To start off, we take a closer look at a linear panel data model with time-invariant individual effects c_i, time fixed effects λ_t and an endogenous regressor x_{it}. The model thus can be specified as follows:

$$y_{it} = x_{it} + c_i + \lambda_t + \varepsilon_{it}, \tag{1}$$

where $Cov(x_{it}, c_i) \neq 0$, $Cov(c_i\varepsilon_{it}) \neq 0$, $Cov(x_{it}\lambda_t) \neq 0$, $Cov(\lambda_t\varepsilon_{it}) \neq 0$ and $Cov(x_{it}\varepsilon_{it}) \neq 0$.

Hence, we are confronted with omitted variable bias, two-way error clustering as well as endogeneity of the explanatory variable even after controlling for individual and time fixed effects, as the model features time-varying shocks that are related both to the regressor as well as the outcome of interest.

To identify the causal effect of x_{it} on y_{it}, we have an exogenous instrument z_{it}, that is uncorrelated with all components of the composite error term ε_{it}.

Poisson Panel Data Model

For our second process, we simulate a Poisson distributed outcome variable with conditional mean

$$E[y_{it}|x, \gamma_i, \delta_t] = \exp\{\beta_1 x_{it} + \gamma_i + \delta_t \ + \log(n_{it})\} \tag{2}$$

where the fixed effects are analogous to the linear panel specification. The offset n can be interpreted as the population size in period t.

Zero-Inflated Poisson Panel Data Model

The final model we consider is a zero-inflated Poisson model given by

$$E[y_{it}|x, \gamma_i, \delta_t] = \begin{cases} 0 & \text{with probability } \pi \\ \exp\{\beta_1 x_{it} + \gamma_i + \delta_t + \log(n_{it})\} & \text{with probability } (1-\pi), \end{cases} \tag{3}$$

with otherwise identical structure as the regular Poisson model.

2 Example Simulation and Estimation

2.1 Simulated Data

In this section, we simulate our three data processes one time each as an example of how estimation is carried out. We focus here on the point estimates and just note that the standard errors for the control function approaches are incorrect and should be computed via bootstrap procedures. However, for the purposes of this illustration the point estimates will suffice.

We simulate a dataset of 500 individuals observed over 10 time periods. The regressor x and the composite error term ε_{it} is simulated by

$$x_{it} = z_{it} + \mu_i + \nu_t + \eta_{it} \tag{4}$$
$$e_{it} = \zeta_i + \kappa_t + \phi_{it} + \vartheta_{it} \tag{5}$$

Table 1: Descriptive Statistics

	Mean	SD	Median	P25	P75	Min	Max
x	−1.12	1.77	−1.15	−2.33	0.11	−7.46	4.85
z	0.00	1.01	0.01	−0.71	0.69	−3.51	3.23
pop	3080.44	1709.87	3021.49	1627.04	4441.65	100.51	7993.24
y	−2.75	3.01	−2.69	−4.71	−0.69	−12.84	7.00
y_p	4025.23	30 179.56	165.00	26.00	1069.00	0	1241307
y_p0	3580.39	29 469.77	109.00	11.00	794.25	0.00	1 239 482.00
y_pr	3.03	28.00	0.07	0.01	0.50	0.00	1090.49
y_p0r	2.59	24.65	0.04	0.00	0.37	0.00	1089.97

Note:
Summary Statistics for the explanatory variables and outcomes for the three processes. Statistics are based on 500 individuals observed for 10 periods, yielding a sample size of 5000.

where by construction components with equal subscripts are correlated and have non-zero means, apart from ϑ_{it}, which is a white noise error term, and the instrument, which is constructed independently of the error components thereby satisfying the exclusion restriction. The true effect of x on the outcome is 1 for all processes. Table 1 presents descriptive statistics of the simulated data. Figure 2 depicts the distribution of the dependent variables in the linear model (a), Poisson count model (b) and zero-inflated Poisson count model (c), whereas 3 presents scatter plots of the respective dependent variables with the endogenous regressor.

2.2 Estimation

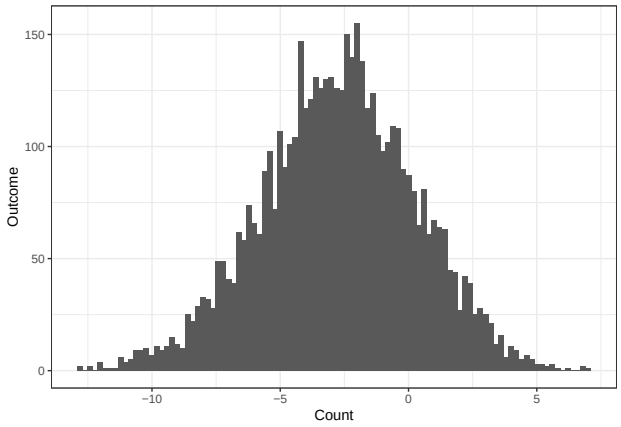
2.2.1 Linear Model

We begin by estimating the linear model using OLS fixed effects, instrumental variables and the Control Function method. The results are presented in Table 2. The way we set up the data generating process, all omitted variables are positively correlated with the endogenous regressor. Hence, as expected, the slope coefficient in the pooled bivariate regression overstates the true effect of x on y substantially, the upward-biased point estimate from Model 1 is 1.530. The inclusion of individual and time fixed effects in Models 2 and 3 decrease the upward bias substantially and yield coefficient estimates of 1.513 and 1.168 respectively. Models 4 and 5 implement 2SLS and the Control Function approach to account for the remaining endogeneity in the regressor. In the linear model, the two approaches are equivalent, which is confirmed by the identical coefficient estimates of 0.992.

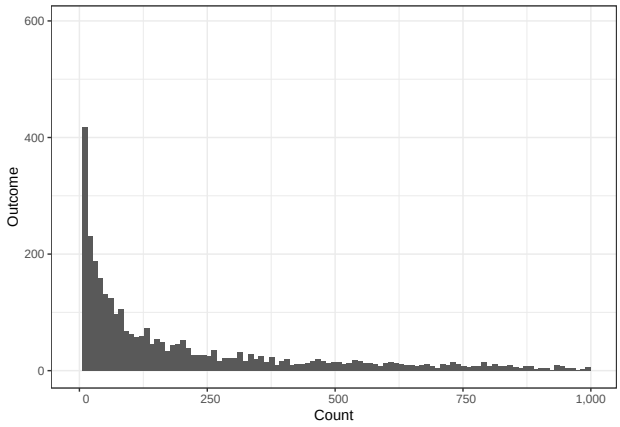
As we can see, for the case of a continuous dependent variable and a linear conditional mean specification, both instrumental variable methods can account for several sources of endogenity and precisely estimate the causal effect of interest.

2.2.2 Poisson distribution

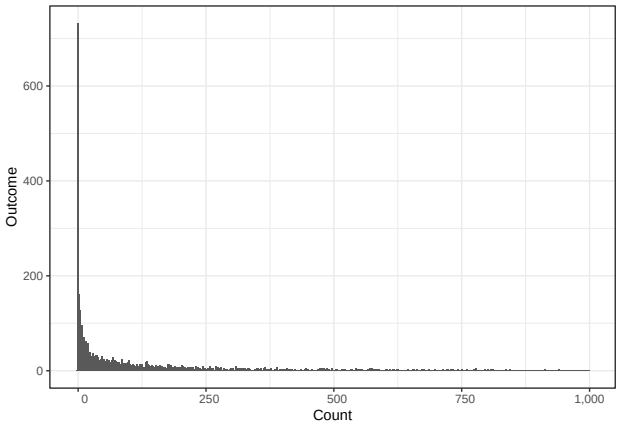
Next, we estimate the Poisson process using PPML, OLS fixed effects, panel IV and the Control Function method. Table 3 illustrates the results. Models 1 through 4 estimate variations of the PPML model, where to the baseline Model 1 with no controls successively individual (Model 2) and time fixed effects (Model 3) are added. Model 4, which constitutes our preferred specification, estimates the Control Function approach by including the estimated residuals from the first stage into the structural equation. The coefficient estimate on the regressor is with NA pretty close to the true value 1. Models 5-10 estimate variations of OLS regressions. In Models 5-9 the dependent variable is the log-transformed count $\log(y_{it}+1)$ with the ad-hoc addition of one to incorporate zero counts into the estimation. As for the PPML regressions, fixed effects are successively



(a) Linear Panel Model

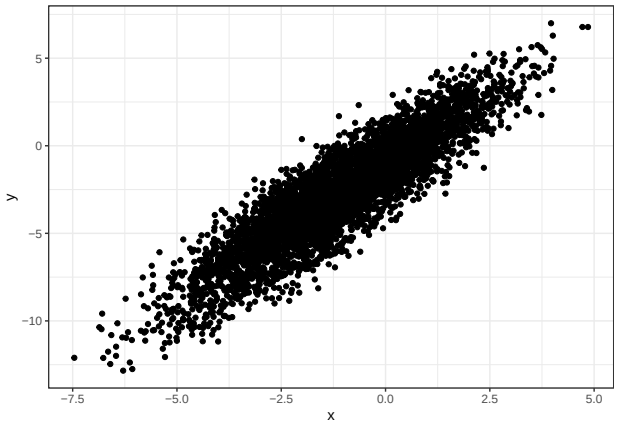


(b) Poisson Panel Model

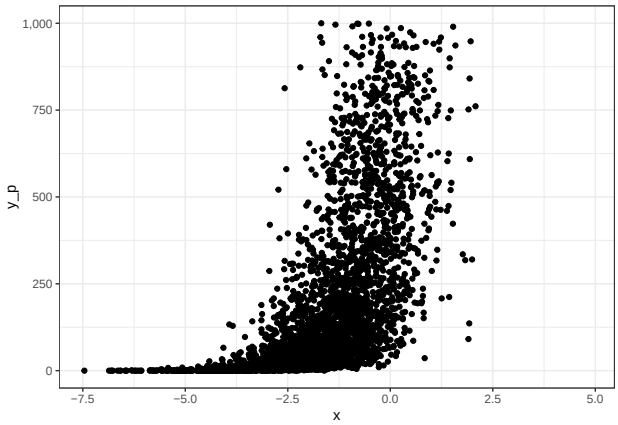


(c) Zero-Inflated Poisson Model

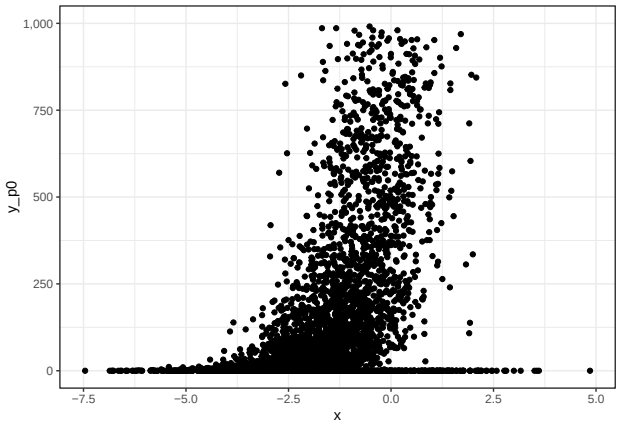
Figure 2: Distribution of Outcome Variables



(a) Linear Panel Model



(b) Poisson Panel Model



(c) Zero-Inflated Poisson Model

Figure 3: Joint distribution of Outcome and Endogenous Regressor

Table 2: Regression Results Linear Process

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	−1.043				
	(0.028)				
x	1.530	1.513	1.168		0.992
	(0.010)	(0.010)	(0.009)		(0.012)
fit_x				0.992	
				(0.013)	
e_xf					0.389
					(0.017)
Num.Obs.	5000	5000	5000	5000	5000
R2	0.815				
R2 Adj.	0.815				
RMSE	1.29	1.16	0.77	0.80	0.73
Std.Errors	by: id	by: id	by: id	by: id	by: id
FE: id		X	X	X	X
FE: t			X	X	X

Note:
This table presents estimation results for the linear panel data process. Model 1 estimates a pooled regression without fixed effects. Models 2 and 3 successively add individual and time fixed effects. Model 4 employs the 2SLS estimator. Model 5 estimates the causal effect using the Control Function method.

added to the specification and finally the two IV methods are employed. The coefficient estimates in the models accounting for all sources of endogeneity and OVB is NA, which is close but somewhat off the true value. Model 10 presents results from regressing the rate y_{it}/pop_{it}, as is sometimes done in the literature. The coefficient estimate is 2.412 and completely off target.

2.2.3 Zero-Inflated Poisson Distribution

Finally, we repeat the analysis of the preceding section for the zero-inflated Poisson process. The results are depicted in Table 4. The basic model structure is equivalent to the one in Table 3, and we see similar patterns for the PPML models. Interestingly, the introduction of a larger zero count mass introduces a downward bias in the estimation. The estimated coefficients are consistently lower when compared to the respective models in Table 3. For the pooled and fixed effects regressions, the upward and downward biases almost exactly cancel out, whereas the instrumental variable approaches, which correctly account for the endogeneity of the regressor, are now downward biased and with 0.807 substantially below the true value 1.

3 Simulation of Control Function Approach

In this section, we simulate the zero-inflated Poisson process 500 times and examine the distribution of estimates for the PPML Control Function approach (Table 4, Model 4) in comparison to the Log-linear Panel Control Function method (Table @ref(tab:reg-zipois, Model 9). In order to make an informed choice about which estimator performs better, we provide histograms of the simulated coefficient estimates. Furthermore, we compute a comparative measure for estimator performance, namely, the Root Mean Squared Error (RMSE). The RMSE for a parameter θ is defined as

Table 3: Regression Results Poisson Process

	PPML				OLS					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
(Intercept)	−0.482				6.592					
	(0.101)				(0.032)					
x	1.397	1.465	1.148	1.023	1.304	1.247	1.099		0.930	
	(0.064)	(0.061)	(0.048)	(0.062)	(0.011)	(0.011)	(0.011)		(0.014)	
e_xf				0.337					0.374	
				(0.075)					(0.019)	
fit_x								0.930		2.412
								(0.014)		(0.506)
Num.Obs.	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000
R2					0.785					
R2 Adj.					0.785					
RMSE	27 843.39	10 393.68	8612.68	8160.81	1.21	1.05	0.82	0.85	0.78	25.42
Std.Errors	by: id	by: id	by: id	by: id	by: id	by: id	by: id	by: id	by: id	by: id
FE: id		X	X	X		X	X	X	X	X
FE: t			X	X			X	X	X	X

Note:
This table shows estimation results for the Poisson model. The first three models are estimated using PPML while models 4, 5 and 6 employ linear models using the log transformation of the dependent variable. Model 1 estimates a pooled Poisson model, Model 2 and 3 add individual and time fixed effects, whereas Model 4 estimates the Control Function approach using first stage residuals in the second stage. Model 5 estimates a pooled model with the log transformation as dependent variable

Table 4: Regression Results Zero-Inflated Poisson Process

	PPML				OLS					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
(Intercept)	−0.577				5.881					
	(0.110)				(0.050)					
x	1.387	1.461	1.155	1.044	1.135	1.083	0.952		0.807	
	(0.070)	(0.070)	(0.054)	(0.070)	(0.019)	(0.021)	(0.024)		(0.030)	
e_xf				0.303					0.322	
				(0.101)					(0.042)	
fit_x								0.807		2.034
								(0.030)		(0.421)
Num.Obs.	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000
R2					0.482					
R2 Adj.					0.482					
RMSE	27 190.02	10 261.11	8941.41	8569.16	2.09	1.93	1.83	1.84	1.82	22.48
Std.Errors	by: id	by: id	by: id	by: id	by: id	by: id	by: id	by: id	by: id	by: id
FE: id		X	X	X		X	X	X	X	X
FE: t			X	X			X	X	X	X

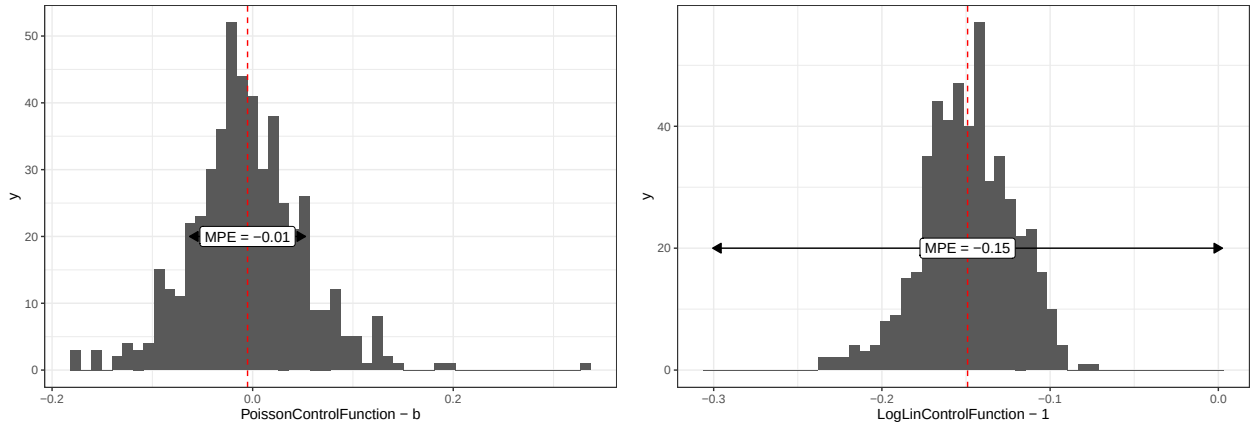
Note:
This table shows estimation results for the Poisson model. The first three models are estimated using PPML while models 4, 5 and 6 employ linear models using the log transformation of the dependent variable. Model 1 estimates a pooled Poisson model, Model 2 and 3 add individual and time fixed effects, whereas Model 4 estimates the Control Function approach using first stage residuals in the second stage. Model 5 estimates a pooled model with the log transformation as dependent variable

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E[(\hat{\theta} - \theta)^2]}, \tag{6}$$

which we approximate by its sample counterpart $RMSE(\hat{\theta}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta)^2}$, where $\hat{\theta}_m$ denotes the coefficient estimate for θ of simulation m. The RMSE is the standard deviation of the prediction errors and hence gives an indication of how far off the estimates are from the true value.

3.1 Large N, Small T

The sample is based on 1000 individuals and 10 time periods. Figure 4 illustrates the distribution of estimates for the 500 simulations. The two panels show histograms with 50 bins each. The RMSE of the estimates from the PPML model is 0.06, whereas the RMSE for the log-linear model is 0.15, which is 162.43% larger. Hence, the PPML model performs better than the log-linear specification.



(a) Poisson Control Function

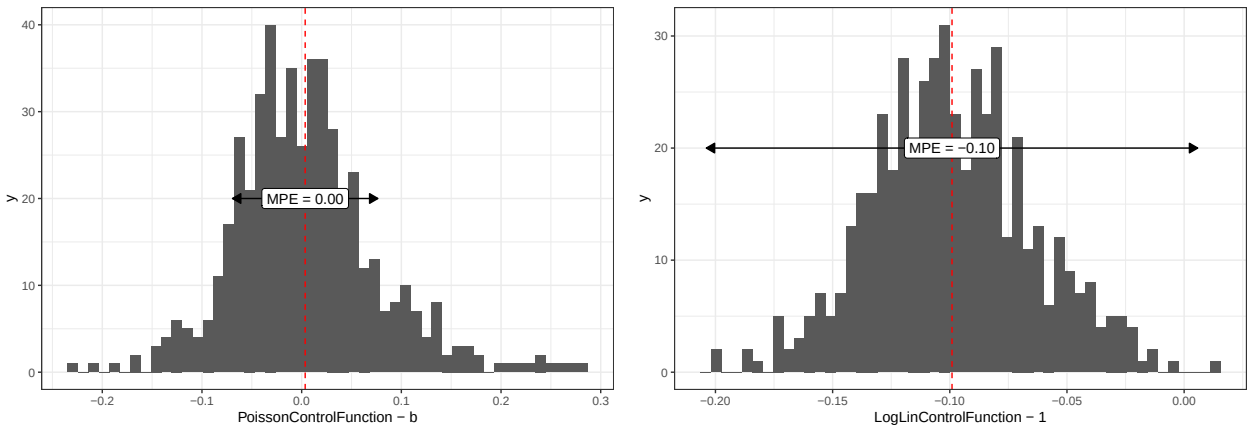
(b) Log-linear Control Function

Figure 4: Distribution of Prediction Error for Simulation 1

3.2 Small N, Large T

Sample is based on 10 individuals and 1000 time periods.

The sample is based on 10 individuals and 1000 time periods. Figure 5 illustrates the distribution of estimates for the 500 simulations. The two panels show histograms with 50 bins each. The RMSE of the estimates from the PPML model is 0.07, whereas the RMSE for the log-linear model is 0.10, which is 43.77% larger. Hence, the PPML model performs better than the log-linear specification.



(a) Poisson Control Function

(b) Log-linear Control Function

Figure 5: Distribution of Prediction Error for Simulation 2