Monte Carlo Simulation: exploring the usage of clustered standard errors when using instrumental variable analysis

Susanne Schaftenaar

3 June 2019

1 Assignment motivation

The main purpose of using instrumental variable analysis is to avoid problems with omitted variable bias. This problem especially occurs when having unobserved confounders that are not represented in a model (Angrist & Pischke, 2009: 115-116). Using instruments offer a solution. An IV-analysis is a two-step procedure. The first stage estimates the effect of the instrument Z on the main independent variable X1. In the second stage, the predicted values of X1 from the first stage are used to estimate the effect on the outcome variable Y. An instrument Z must thus be correlated with the main causal variable X1, or the "treatment" variable. However, it also should be uncorrelated to any other causes of the outcome variable Y. In short, the IV-analysis rest on two important assumptions: the instrument Z must have a clear effect on the main independent variable X1 (which is testable), and second, the only reason for the relationship between Y and Z is the first stage. The latter is usually referred to as the exclusion criterion. If both these assumptions hold, the instrumental variable analysis avoids problems with omitted variable bias.

This assignment focuses on the usage of IV-analysis with panel data. In a working paper, I use IV analysis to reevaluate the effect of gender inequality indicators on armed conflict onset. The instruments, geo-climatic suitability for plough-positive cereals or plough-negative cereals are on the country-level. The gender inequality, conflict onset, and control variables are all available on the country-year level (panel data). The data are thus clustered by design.

Clustered data can be seen as a special form of omitted variable bias where the cluster(s) account for within-group variation. This can lead to bias and inefficiency. The question posed in this assignment is whether one must account for clustered data in an IV-framework. The main purpuse of IV-analysis is to overcome issues with omitted variable bias. If clustered data is a form of omitted variable bias, does this mean that the IV-analysis accounts for this by design? Many IV-analyses include design features accounting for clustered data though, such as fixed effects (Alesina et al, 2013) or clustered robust standard errors (Miguel et al, 2004). In addition, Abedie et al (2017) suggest to cluster the standard errors if a) the sample is not randomly drawn or b) the assignment of the treatment is clustered. In my working paper, the instruments can be seen as the assignment mechanism. Subsequently, the assignment of the treatment is clustered (i.e. country-level instruments assigns values at the country-year level). Does their advice hold in an IV-framework? The main issue when unnecessarily clustering standard errors, is that one might commit to type-II error. The researcher would end up with too conservative standard errors (Abedie et al, 2017). However,

if not acounting for clustered data where one should, this could lead to overconfidence and, as a consequence, type-I error.

This assignment uses Monte-Carlo simulations to assess which solution to use in my paper. Should I proceed with using a) unadjusted standard errors, b) clustered standard errors, or c) neither work (i.e. I don't only have an efficiency problem, but also biased estimates). Below, I present the procedure.

2 Assignment procedure

The true DGP in this assignment is as follows:

$$y_i = a_i + \beta_1 X1 + \beta_2 W + u_i \tag{1}$$

The true DGP contains clustered data. I generate a_i from a random uniform distribution U(-1,1). I create 100 intercepts (with 100 unique IDs). These intercepts each have 50 nested observations per intercept. This set-up: 100 IDs/ intercepts, with 50 observations aims to mimic a country-year set-up. I have 100 countries and 50 years within each country. I set the true values as follows: $\beta_1 = -2$, $\beta_2 = 1.5$.

I construct three different instruments. Z1 mimics the instruments in my working paper. The instrument is clustered: I draw the instrument from a random normal distribution(0, 1). I create 100 ids (to approximate "countries") with 50 nested observations (to approximate "country-years"). The instrument values are fixed by ID. Z1 is uncorrelated with W. I also base X1_A on Z1 (see below) to satisfy the two assumptions of an IV-analysis (Z should be correlated with X and not with any of the other causes of Y). Z2 is constructed as follows: Z2 = Z1 + a_i . It is similar to Z1, but I also correlate it with a_i. Z2 is thus not only a clustered assignment mechanism: it potentially opens an alternative pathway to Y by being correlated to the intercept, violating the exclusion criterion. Z3 is drawn from the random normal distribution and varies per "country-year" and is thus not clustered by id.

I further generate three versions of X1 (A, B, C). First I generate xstar and W (my confounders). Xstar serves as the basis for X1A, B, and C. Xstar and W are correlated with one another. In a normal OLS regression, we thus need to include W to correctly model the DGP, otherwise there would be omitted variable bias. In the IV-analysis, this should be solved by design (i.e. not by including W in the model, but by instrumenting X by Z).

X1_A is constructed as follows: X1_A = xstar + Z1. Z1 is thus a valid instrument: it is correlated with my main independent variable X1_A, but uncorrelated with any of the other causes (W) of Y. Using Z1 to instrument X1_A should thus give unbiased estimates, since the exclusion criterion is not violated. Z1 is fixed by id though and might affect the standard errors.

X1_B is constructed as follows: X1_B = xstar + Z2. Z2 is thus not a valid instrument: it is correlated with my main independent variable X1_B, but also correlated with any of the other causes of Y: the intercept a_i. Using Z2 to instrument X1_B should thus give biased estimates (the inclusion criterion is potentially violated). It is also likely to affect the standard errors by having a clustered instrument.

X1_C is constructed as follows: X1_C = xstar + Z3. Z3 is thus a valid instrument: it is correlated with my main independent variable X1_C, but uncorrelated with any of the other causes of Y. Using Z3 to instrument X1_C should thus give unbiased estimates (the exclusion criterion is not violated) and should not affect the standard errors, since this is not a clustered instrument.

Since the values of X1_A, X1_B and X1_C vary since they are based on different instruments, I estimate three Ys. Note u_i = N(0,1) for all.

$$y1_i = a_i + \beta_1 X1\_A + \beta_2 W + u_i \tag{2}$$

$$y2_i = a_i + \beta_1 X1\_B + \beta_2 W + u_i \tag{3}$$

$$y3_i = a_i + \beta_1 X1\_C + \beta_2 W + u_i \tag{4}$$

In PART I I estimate four normal OLS/ fixed effects OLS models. In PART II I estimate different IV-analyses. I compare PART I and PART II to show when and why it is useful to employ IV-estimation (all models are Two Stage Least Squares (2SLS)). PART II also incorporates bootstrapping (1000x) to provide a robustness test for the unadjusted and adjusted standard errors. I thus draw the following realisations:

PART I

1. an OLS: $Y \sim X1$
2. fixed effects OLS: $Y \sim X1$
3. an OLS: $Y \sim X1 + W$
4. fixed effects OLS: $Y \sim X1 + W$

PART II

1. IV-estimation: $Y1 \sim X1\_A|Z1$
2. IV-estimation: $Y1 \sim X1\_A|Z1$ with clustered standard errors.
3. IV-estimation: $Y1 \sim X1\_A|Z1$ bootstrapped
4. IV-estimation: $Y2 \sim X1\_B|Z2$
5. IV-estimation: $Y2 \sim X1\_B|Z2$ with clustered standard errors.
6. IV-estimation: $Y2 \sim X1\_B|Z2$ bootstrapped
7. IV-estimation: $Y3 \sim X1\_C|Z3$
8. IV-estimation: $Y3 \sim X1\_C|Z3$ with clustered standard errors.
9. IV-estimation: $Y3 \sim X1\_C|Z3$ bootstrapped

I then proceed with the Monte Carlo (MC) simulation. I am mainly interested in whether I should cluster the standard errors in the IV-analysis. I thus generalise PART II in the MC experiment. I run 1000 simulations that include all models from PART II.

3 Results

3.1 OLS vs 2SLS

Table 1-3 show the results of the OLS regressions. Model 1 is a bivariate model. Model 2 is a bivariate model with fixed effects, Model 3 is the complete multivariate model without fixed effects, and Model 4 is the complete multivariate model with fixed effects. Table 1 estimates Y1, Table 2 estimates Y2, Table 3 estimates Y3. To refresh the memory: $\beta_1 = -2$, $\beta_2 = 1.5$. The bivariate models return, as expected, biased estimates, since they do not include W. Models 2, the fixed effects bivariate, returns slightly more biased estimates than model 1. This is surprising, since that model should in fact be less biased as it accounts for the fixed intercepts. Models 3 and 4 return estimates of B1 and B2 that are close to its true values. In Table 1 and 2 the fixed effects multivariate model is slightly closer to the true values. In Table 3 the model 3 actually returns a slightly closer estimate to the true value for B1. Generally speaking, the standard errors for the bivariate models are similar to one another. The standard errors for the multivariate models are also similar. Overall model 4 has the highest adjusted R2 value suggesting that this is the best model. Thus, when accounting for omitted variable bias (confounders and clustered data), an OLS would function perfectly well.

Table 1:

	Dependent variable:			
	Y1			
	(1)	(2)	(3)	(4)
X1_A	-1.786^{***}	-1.647^{***}	-2.020^{***}	-2.011^{***}
	(0.014)	(0.017)	(0.009)	(0.011)
W			1.483^{***}	1.479^{***}
			(0.017)	(0.015)
Constant	-1.874^{***}	-2.801^{***}	-0.102^{***}	-0.970^{***}
	(0.038)	(0.239)	(0.031)	(0.141)
Observations	5,000	5,000	5,000	5,000
R^2	0.757	0.792	0.904	0.929
Adjusted R^2	0.757	0.788	0.904	0.928
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2:

	Dependent variable:			
	Y2			
	(1)	(2)	(3)	(4)
X1_B	-1.686^{***}	-1.629^{***}	-1.902^{***}	-1.996^{***}
	(0.014)	(0.017)	(0.009)	(0.011)
W			1.440^{***}	1.491^{***}
			(0.017)	(0.015)
Constant	-1.679^{***}	-2.492^{***}	0.053^{*}	-0.934^{***}
	(0.037)	(0.242)	(0.031)	(0.142)
Observations	5,000	5,000	5,000	5,000
R^2	0.751	0.778	0.898	0.924
Adjusted R^2	0.751	0.773	0.898	0.923
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3:

	Dependent variable:			
	Y3			
	(1)	(2)	(3)	(4)
X1_C	-1.743^{***}	-1.738^{***}	-1.995^{***}	-1.990^{***}
	(0.015)	(0.014)	(0.010)	(0.009)
W			1.478^{***}	1.482^{***}
			(0.017)	(0.015)
Constant	-1.771^{***}	-2.553^{***}	-0.058^{*}	-0.965^{***}
	(0.039)	(0.245)	(0.032)	(0.143)
Observations	5,000	5,000	5,000	5,000
R^2	0.735	0.769	0.891	0.922
Adjusted R^2	0.735	0.764	0.891	0.921
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4 shows the results of the 2SLS IV-analyses. Model 1 and 2 show the estimates when using Z1. Model 2 has clustered standard errors by id. Model 3 and 4 show the estimates when using Z2. Model 4 has clustered standard errors by id. Model 5 and 6 show the estimates when using Z3. Model 6 has clustered standard errors by id.

First, Z1 and Z3 function as valid instruments: the estimates are unbiased and close to B1's true value of -2. However, when using robust standard errors for the IV-model estimates by Z1 (model 2), the standard errors almost triple in size compared to using unadjusted standard errors (model 1). Z1 provides a clustered assignment of values of X1_A. This is not the case for the model estimated by Z3 (unclustered instrument), where the standard errors are almost the same. Finally, Z2 violates the exclusion criterion by being correlated with the intercept a_i. It returns a biased estimate. The standard errors also more than double when using clustered standard errors.

To conclude, the multivariate OLS models (especially those with fixed effects) would function as well as the IV-models when the IV-assumptions hold: the estimates of B1 are very close to its true value. If there are unobserved confounders and the assumptions of IV-estimation hold, the IV-models would thus be a good alternative. However, the IV-models with a clustered assignment mechanism (i.e. Z1 and Z2) return differently sized standard errors when using robust clustered standard errors. This is not the case for Z3 where clustering the standard errors makes no difference.

What about when we bootstrap the standard errors? Would they come closer to the unadjusted or the clustered standard errors? Figure 1 and 2 show that the bootstrapped standard errors and the clustered standard errors overlap. The unadjusted standard errors would have led to overconfidence. This suggests that clustering the standard errors is the preferable choice.

Figure 3 is puzzling: the unadjusted and clustered standard errors overlap, but the bootstrapped standard errors continue to be larger. Likely, it is wise to continue bootstrapping the standard errors when estimating the real data and use them instead of the unadjusted standard errors. I will further explore this in the MC-analysis.

Table 4:

	Dependent variable:					
	Y1		Y2		Y3	
	instrumental variable	*coefficient test*	*instrumental variable*	*coefficient test*	*instrumental variable*	*coefficient test*
	(1)	(2)	(3)	(4)	(5)	(6)
X1_A	-2.023^{***}	-2.023^{***}				
	(0.024)	(0.061)				
X1_B			-1.783^{***}	-1.783^{***}		
			(0.022)	(0.050)		
X1_C					-1.984^{***}	-1.984^{***}
					(0.026)	(0.025)
Constant	-2.332^{***}	-2.332^{***}	-1.870^{***}	-1.870^{***}	-2.253^{***}	-2.253^{***}
	(0.054)	(0.138)	(0.050)	(0.122)	(0.059)	(0.078)
Observations	5,000		5,000		5,000	
R^2	0.743		0.748		0.721	
Adjusted R^2	0.743		0.748		0.721	
Note:						*p<0.1; **p<0.05; ***p<0.01

Figure 1: density plot of IV model 1: comparing unadjusted standard errors, clustered standard errors and bootstrapped standard errors

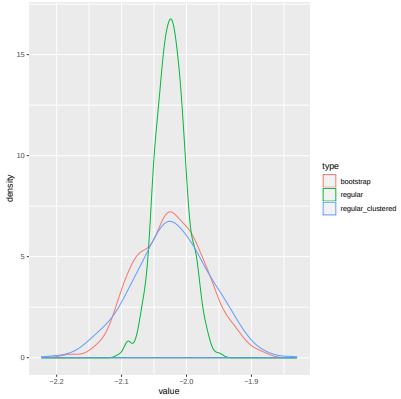


Figure 2: density plot of IV model 2: comparing unadjusted standard errors, clustered standard errors and bootstrapped standard errors

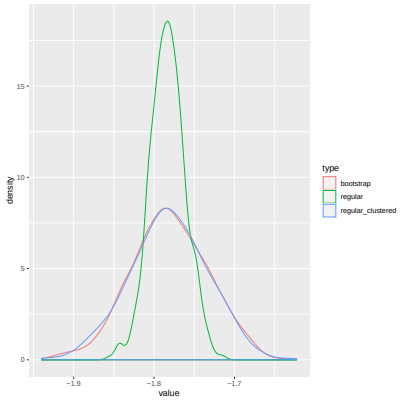
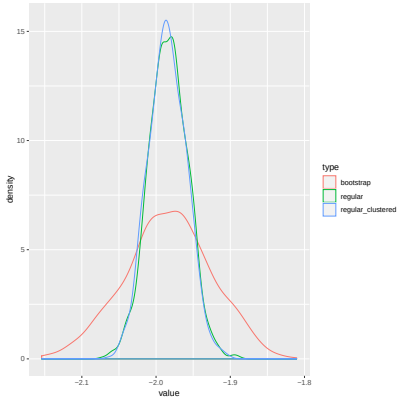


Figure 3: density plot of IV model 3: comparing unadjusted standard errors, clustered standard errors and bootstrapped standard errors



3.2 MC- experiment

I generalize the experiment with MC simulations. As mentioned in the procedure, I only focus on the IV-models (2SLS) in this part of the assignment. I run the MC-simulation 1000 times. I estimate the three IV models with unadjusted standard errors and clustered standard errors. I also include the bootstrapped IV-models in the simulation.

The means of the models in the MC-simulation are very close to those in the realizations above. IV-model 1 and IV-model 3 are very close to the true value -2 and are unbiased, whilst model 2 (violating the exclusion criteria) is biased. I have calculated the Mean Squared Error (MSE) for all models. IV model 3 has the lowest MSE at 0.0007. IV model 1 has an MSE of 0.004, and IV model 2 an MSE of 0.065. The MSE is similar for all models when bootstrapped. In figure 4 (IV model 1), figure 5 (IV model 2), figure 6 (IV model 3) show that IV-model 1 and 3 return unbiased estimates, whilts IV-model 2 returns biased estimates. IV-model 3's estimates are tighter than IV-model 1's estimates, which confirm the MSE scores presented above.

Finally, I calculate the coverage probabilities and construct three coverage plots. Figure 7 shows that the IV model with unadjusted standard errors includes the true parameter in 60-65 % of the simulations. This is less than the expected 95% and means that the estimated standard errors are too small on average. The models with clustered standard errors and bootstrapped standard errors include the true parameter in almost 95% of the cases, which means that these methods perform well when estimating the standard errors. The coverage plot for IV-model 2 returns the true estimate in 0% of the cases for all models. This since the estimate is biased. Finally, IV-model 3 provides insight into the puzzling results found in the regression models above where the bootstrapped errors were -surprisingly- larger than either the unadjusted and clustered standard errors. The unadjusted and clustered standard errors perform well: 95% of the simulations returned the true parameter -2. However, the bootstrapped models return always return the true value. The standard errors calculated by the bootstrapping are, in fact, too large.

Figure 4: density plot B1 IV model 1: normal and bootstrapped models

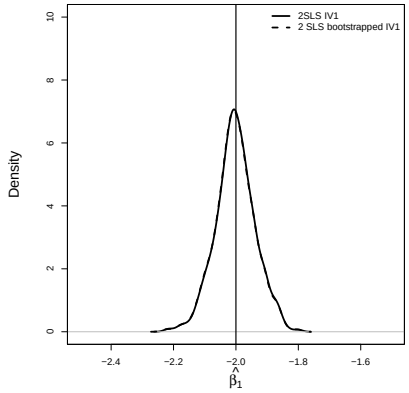


Figure 5: density plot B1 IV model 2: normal and bootstrapped models

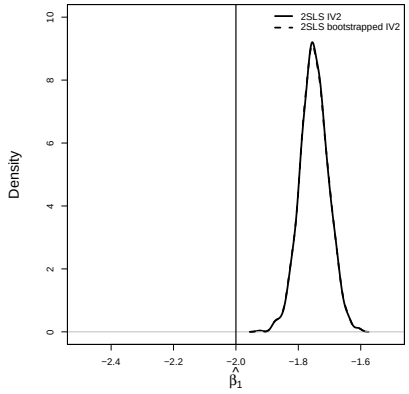


Figure 6: density plot B1 IV model 3: normal and bootstrapped models

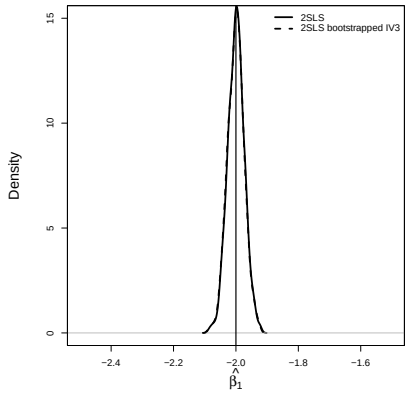


Figure 7: Coverage plot B1 IV model 1

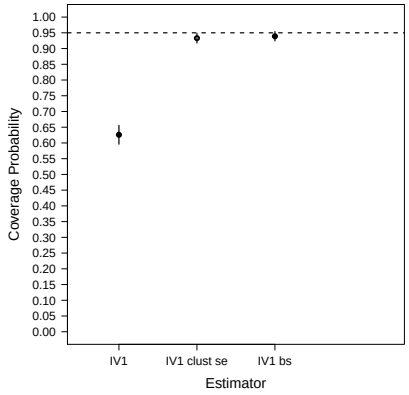


Figure 8: Coverage plot B1 IV model 2

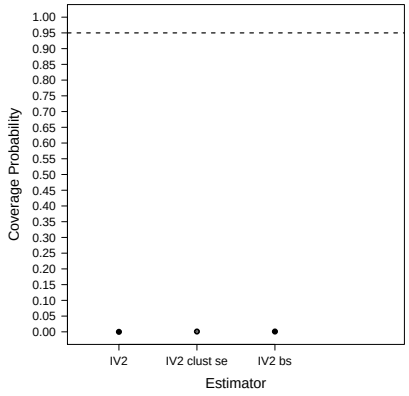
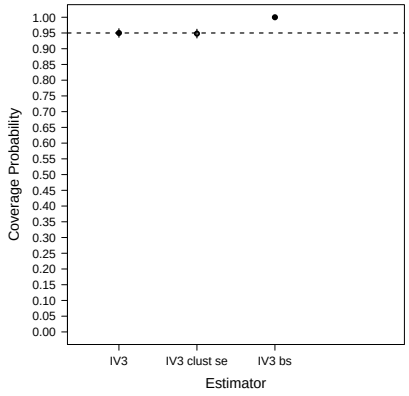


Figure 9: Coverage plot B1 IV model 3



4 Conclusion

In this assignment, I set out to explore whether I need to cluster standard errors in an IV-analysis. The findings suggest that I should. If the instrument is clustered and thus assigns the treatment in a clustered manner, the standard errors need to be adjusted. However, if the instrument is not clustered, unadjusted standard errors can be used. Clustering the standard errors, would, however, not lead to inflated standard errors in this case and appears harmless. Finally, if the instrument is correlated to the intercepts, then the exclusion assumption is violated and clustering the standard errors would not remedy this. In that case, one should likely a) return to an OLS model, or b) model this in the IV-model. To solve this is beyond the scope of the assignment (since I do not believe this is a problem within my working paper).

5 Implementation in working paper (extra)

Based on the MC-simulation, I have proceeded with clustered standard errors in my working paper. Figures 10 and 11 compared the unadjusted, clustered, and bootstrapped standard errors of the real data. They behave very similarly to the simulations: the bootstrapped and the clustered standard errors overlap, whilst the unadjusted standard errors are too small and would lead to overconfidence.

Figure 10: 2SLS estimation of the effects of women's labour force participation on armed conflict onset. The figure makes a comparison of the unadjusted, clustered, and bootstrapped standard errors using real data

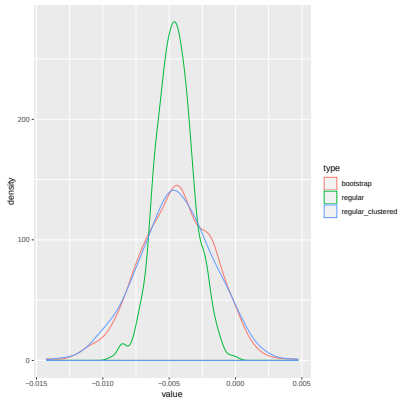


Figure 11: 2SLS estimation of the effects of women's political empowerment on armed conflict onset. The figure makes a comparison of the unadjusted, clustered, and bootstrapped standard errors using real data