PP 421: Problem Set 4

Luke Chen

Acknowledgements: Ibrahim Khan

May 19, 2020

Problem 1

Table 1: Regression Results

	No Cluster	Cluster
treat	222.2**	222.2**
	(68.85)	(66.97)
Constant	-99.06*	-99.06*
	(47.92)	(40.59)
Observations	900	900
Adjusted R^2	0.010	0.010

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results of the OLS regression of y_i on $treat_i$ are presented in Table 1. The results show that the coefficient estimates are identical between the case with cluster-robust standard error calculation and non-cluster-robust SE calculation. This is true by construction since the cluster-robusting procedure only affects the estimation of the variance-covariance matrix and does not influence the regression itself. Meanwhile, the standard errors change very slightly (go down) upon clustering. This is because we constructed the standard errors to be independently and identically distributed (homoskedastic), meaning that there is no within-cluster standard error correlations by construction. Therefore, even after clustering by the group variable, the calculation for the standard error turns out almost identical to the homoskedastic case. However, by chance, the error terms are such that clustering by group causes the standard error to decrease slightly in our example.

Problem 2

a)

Table 2: Regression Results

	(1) No Cluster	(2) Cluster
treat	16413.1***	16413.1***
	(416.0)	(707.6)
Constant	8134.0***	8134.0***
	(299.3)	(513.2)
Observations	900	900
Adjusted R^2	0.634	0.634

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

b) The regression results of the regular OLS and the OLS with cluster-robust SE are reported in Table 2. The coefficients again do not change as expected. However, the standard errors this time are significantly larger in the cluster-robust case. This makes sense since the errors are in fact correlated within clusters and so we cannot treat all 900 observations as i.i.d. Instead, we have to calculate the standard errors, essentially treating each group as an observation. Therefore, the standard error increases since we end up dividing the sample variance by a smaller amount.

c) If this were a real empirical project, I would report the cluster-robust standard errors since the errors are in fact clustered at the group level. Additionally, the number of groups is large while the size of each group is small and so we know that the clustered SEs perform relatively well. While the non-cluster-robust standard errors appear more attractive because they suggest more precision in our estimate, they underestimate the SEs because they falsely assume errors are i.i.d

In a more realistic case where the errors are unobserved, use of the cluster-robust standard errors would depend on whether there was any theoretical reason to do so. For example, if I was looking at individuals randomly sampled from around the country, I may cluster at the state or county level if there is a high possibility for some common shocks to affect their outcomes and thus cluster their standard errors. On the other hand, if we could be reasonably sure that the errors are in fact homoskedastic (perhaps in a laboratory setting) then I would use the non-cluster-robust errors (though I might use heteroskedastic-robust errors instead of homoskedastic ones).

d) In Stata, I manually bootstrapped the standard error for $\hat{\beta}$ using the *bsample* command to generate a bootstrap sample from the original sample and running the simple OLS regression to get an estimate $\hat{\beta}_{b_i}$ for $i = 1, 2, ..., 1000$. The results of one such bootstrap regression are presented in Table 3.

Table 3: Example Bootstrap Results

	(1)
	y
treat	16412.6***
	(402.8)
Constant	8517.9***
	(288.3)
Observations	900
Adjusted R^2	0.649

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

After getting all of $\hat{\beta}_{b_1}...\hat{\beta}_{b_{1000}}$, I used the formula

$$\hat{V}(\hat{\beta}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\beta} - \bar{\hat{\beta}})^2$$

where in this case $B = 1000$. This procedure yielded a standard error of **412.5** which is quite close to the parametrically estimated homoskedastic standard error of 416.0

e) Table 4 shows some example results from one clustered bootstrap sample.

Table 4: Example Clustered Bootstrap Results

	(1)
	y
treat	16451.7***
	(419.7)
Constant	7309.3***
	(306.8)
Observations	900
Adjusted R^2	0.631

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This time, I re-did the procedure as outlined in the previous part (2d) but letting the *bsample* command take the argument *cluster(group)*, letting us create bootstrap by sampling with replacement from the groups instead of individuals. One run-through of the boostrap procedure yielded a standard error of **705.3** which is again very close to the parametrically estimated standard error of 707.6

f) To visualize how the number of bootstrap samples used affects the accuracy and precision of our bootstrapped standard error estimates, I ran 100 iterations of the bootstrapping procedure

using different numbers of bootstrap samples (B). The results are shown in Figure 1. We can see that there are quite large gains in precision as B rises from 5 to 30 and the gains become more modest afterwards. Therefore, we might consider 30-50 samples to be the lower bound on what we would consider a "large enough" number of bootstrap samples.

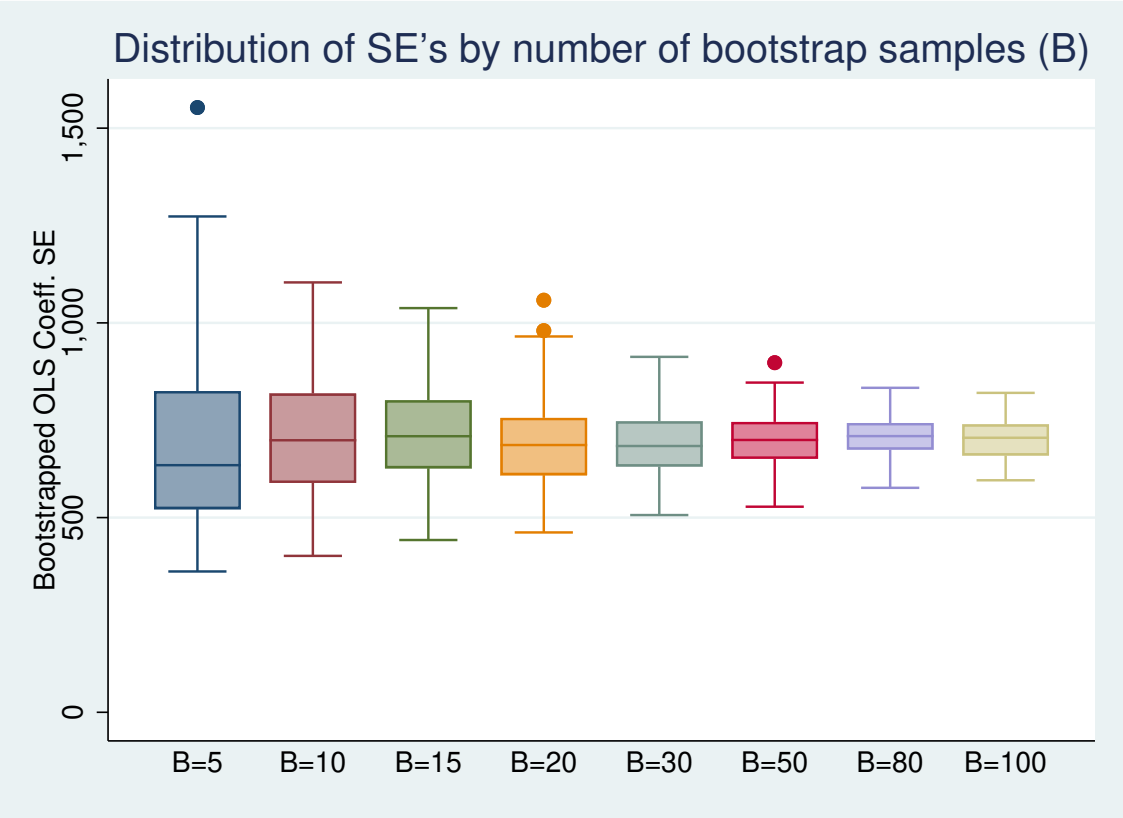

Figure 1: The "box graph" showing the mean standard errors and bootstrapped sample variances of standard errors for the dataset $simulated1_2020$ using clustered bootstrap samples.