Low sample size and regression: A Monte Carlo approach

Riveros Gavilanes, John Michael

Corporación Centro de Interés Público y Justicia -CIPJUS-

17 November 2019

JOURNAL of Applied Economic Sciences

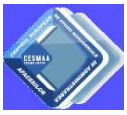
JAES

Journal of Applied Economic Sciences

Volume XV
Issue 1(67) Spring 2020

ISSN-L 1843 - 6110
ISSN 2393 - 5162

Editorial Board



European Research Center of Managerial Studies in Business Administration
Email: jaes_secretary@yahoo.com
Web: http://cesmaa.org/Extras/JAES

Low Sample Size and Regression: A Monte Carlo Approach

John Michael RIVEROS GAVILANES
Corporation Center of Public Affairs and Justice, Economic Research, Bogotá
Universidad Colegio Mayor de Cundinamarca, Bogotá, Colombia
jmriveros@unicolmayor.edu.co jmrg2992@hotmail.com

Article's history:

Received 19 January 2020; *Received in revised form* 5 February 2020; *Accepted* 6 March, 2020;
Published 30 March 2020.

Suggested citation:

Riveros Gavilanes, J.M. 2020. Low sample size and regression: A Monte Carlo approach. *Journal of Applied Economic Sciences*, Volume XV, Spring, 1(67): 22-44. DOI: https://doi.org/10.14505/jaes.v15.1(67).02

Abstract:

This article performs simulations with different small samples considering the regression techniques of OLS, Jackknife, Bootstrap, Lasso and Robust Regression in order to establish the best approach in terms of lower bias and statistical significance with a pre-specified data generating process (DGP). The methodology consists of a DGP with 5 variables and 1 constant parameter which was regressed among the simulations with a set of random normally distributed variables considering sample sizes of 6, 10, 20 and 500. Using the expected values discriminated by each sample size, the accuracy of the estimators was calculated in terms of the relative bias for each technique. The results indicate that Jackknife approach is more suitable for lower sample sizes while the Bootstrap approach reported to be sensitive for the lower sample sizes indicating that it might not be suitable for establishing statistically significant relationships in the regressions. The Monte Carlo simulations also reflected that when a significant relationship is found in small samples, this relationship will also tend to remain significant when the sample size is increased.

Keywords: small sample size; statistical significance; regression; simulations; bias.

JEL Classification: C15; C19; C63.

Introduction

One situation that might happen while we're trying to analyze different types of data and make empirical inferences over a phenomenon is that we may have a low (or reduced) number of observations. This is usually associated with the lack of confidence in the estimations, especially when we're opting for the regression analysis in the multivariate framework. A possible answer to avoid this problem is to perform descriptive statistics and proceed with the deduction patterns, however, it could be asked: Are we really sure that our estimations are unreliable? Do they really lack of confidence? These are usual questions in the context of quantitative analysis when we're regressing a model in the presence of low observations. Naturally, the literature supports this idea from different perspectives, as an example Bujang, Sa'at, and Tg Abu Bakar Sidik (2017) studies state that in order to obtain coefficients closer to the population parameters we need around 300 observations to be sure they're reliable.

But if our phenomenon has not been studied (or documented) properly in order to obtain a significant number of observations, should we discard immediately the multiple regression technique to analyze it? The aim of this paper is to provide evidence that regression can have consistent estimates of the coefficients even when we're dealing with a low number of observations. The methodology consists mainly of the use of Monte Carlo simulations derived from a linear data generating process (DGP) to perform conclusions about the bias of the estimated coefficients in the regression framework with a different number of observations. The estimation techniques involve ordinary least squares (OLS), Jackknife, Bootstrap, Robust Regression, and Lasso approaches.

1. Research background

The sample size can be classified in general terms depending on the number of observations, as it can be found in the study of Mason and Perreault (1991), a sample size of 30 observations or lesser is considerate small, samples around 150 observations can be considered as moderate and finally, samples bigger than 250 or 300 are tagged as large. One interesting problem that arises in small samples is relative to the statistical inferences, in fact, "using a sample smaller than the ideal increases the chance of assuming as true a false premise" (Faber and Fonseca 2014). This implies considering the two types of errors in statistical hypothesis testing, the type I and II errors. In simple words, the first type of error refers that our null hypothesis H_0 (relative to a specific proposition) is true but we reject it, while the second type of error refers when our H_0 is false but we don't reject it.

Small sample size and incorrect inferences in the parameters' significance tests are studied by Colquhoun (2014) indicating that a p-value lesser than 5% might not be statistically significant since the results are derived from "underpowered statistical inferences". From this, the risk of using a small size would be the possible type I error in the regression framework.

More from this idea can be found in another study of Forstmeier, Wagenmakers, and Parker (2017) where the problem of false-positive findings can be derived from a decreased sample size and incorrect p-values. Also, the problem of statistical inferences is correlated with the replication procedure, in other words, the last two types of errors seem to be sensitive to the number of replications in a way that the results derived from one inference might not match the result of a similar exercise concerning a similar set of data. This is a fair point in the analysis, the number of replications might affect the statistical inference and the overall converge rate to the population parameters of the estimations, therefore it should be taken into account. This idea leads to a basic statement: as we increase the number of replications of an experiment, we're getting closer and closer to the expected behavior of the population parameters in the inference. These authors also make a valid point regarding some underlying assumptions of the estimations, for example, autocorrelation, correct specifications, no omitted variables in general. In this case, small sample size inferences can be harmful where also the ordinary least squares assumptions are not satisfied.

A remarkable study performed by Holmes Finch and Hernandez Finch (2017) starts by analyzing tools like Lasso, Elastic net, Ridge regression and the Bayesian approach regarding the situation when we got high dimensional multivariate data relative to an even bigger number of variables. In this case, the number of independent variables may be close or equal to the sample size, yielding in unstable coefficients and standard errors (these ones are needed to the formulation of the hypothesis testing procedure) (Bühlmann, Van De Geer 2011). The result of these experiments tends to demonstrate that the regularization methods, in particular, the Ridge regression approach were more accurate in terms to control bias and type I errors produced in the estimations with low sample data for multiple regression analysis. Speed (1994) establishes a contribution to the solution of the problem of low sample size in the regression framework, considering sample reuse validation techniques. These techniques refer to the Jackknife and Bootstrap approaches related to the multiple regression estimation.

An important statement of this author is: *Researchers should note that the overwhelming case is that reduction in sample size is far more likely to reduce the likelihood of finding any significant relationships than to increase it. This is due to the way that sample size affects test power. The researcher sets the level of type I error (the probability of accepting a hypothesis when false in reality) in any test, normally at 0-05, and critical values calculated for the given size of sample. Small sample sizes are no more likely to result in wrongfully claiming a relationship exists than is the case for larger samples*. (Speed 1994, 91)

This interpretation is indeed useful since it states that low sample relationships are more likely to be found when the sample size increases over the experiments. In fact, there is some literature that also critiques the role of large samples in the estimations, arguing that anything becomes significant. Within this idea, we can find the study of Lin, Lucas Jr. and Shmueli (2013) where they affirm that as the sample size is increasing, the p-value starts to decrease drastically to 0, which could lead to statistically significant results which are not sensitive over the regression analysis. Meanwhile, a low sample size is more sensitive to the correlation between the variables (this implies sensibilization to the changes too) leading to think that large sample sizes might find significant results when it's just an overwhelming product of the power of the sample without accurately indicating real (or strong) relationships among the variables. In fact, Faber and Fonseca (2014) appoints that samples cannot be either too big or too small in order to perform statistical inferences.

Up to this point, we're facing problems on both sides of the sample size, too much can be misleading and insensitive to true relationships among the variables (which can be especially the case of the regression analysis) and on the opposite when we got a little sample size, we might have results that are inconsistent across replications driving to errors of type 1.

2. Methodology

The main idea of the methodology is to perform Monte Carlo approximations across different types of estimations which involves OLS, Jackknife, Bootstrap, Lasso and Robust Regression, assuming a multivariate data generating process in a linear form as it follows:

$$y_i = \alpha + \gamma x_{1,i} + \delta x_{2,i} + \theta x_{3,i} + \vartheta x_{4,i} + \varphi x_{5,i} + u_i \quad (1)$$

Equation (1) is calibrated setting the population parameters α, γ, δ, θ, ϑ, φ as all equal to 10 for the *i* observations. The independent variables are x_j with $j = \{1,2,3,4,5\}$ and the residuals are expressed in u_i.The objective is to identify which of the estimation types suits better in terms of accuracy of the estimators. In this case, across the simulations it is assumed that:

$$x_j \sim N(0,1)$$
$$u_i \sim N(0,1) \quad (2)$$

From (1) we're setting the number of replications to 10, 100 and 500 while the number of observations would be set at first to 6 in order to induce on purpose the micronumerosity phenomenon and see how the estimators react to this problem, the next size of observations across replications are set to 10, 20 and 500. There's no need to test for a higher number of observations since empirical literature has established that the overall significance and unbiasedness are influenced by a large sample size. The relative bias of the estimators among the coefficients would be expressed as a relative difference from the population parameter, following a general idea that:

$$Bias = \left| \frac{\beta_j - \overline{\beta_j}}{\beta_j} \right| \quad (3)$$

where: β_j represents the true parameter associated to the x_j variable contained in equation (1) and $\overline{\beta_j}$ represents the estimated parameters in the regressions.

The overall bias $O.B.$ can be expressed in terms of expected values as it follows:

$$O.B. = \left| \frac{k - E(\overline{\beta_j})}{k} \right| \quad (4)$$

where: the mean value of the estimated parameters would be our expected value $E(\overline{\beta_j})$ of the coefficients by each type of regression; k represents the expected value of the true parameters considering that all of the population parameters in the DGP are set to 10 therefore $k = 10$.

In equation (4), the bias would be expressed in terms of percentage comparing the true parameters with the mean of the estimated parameters by the regressions as a relative difference, indicating that 0 would be closer to a perfect match with the absence of bias.

In order to see the changes in the statistical significance of the coefficients, single Monte Carlo simulations would be presented in the usual regression output for each type of estimation (OLS, Jackknife, Bootstrap, Lasso and Robust Regression) with the different sizes in the observations as mentioned before, then the bias results are presented for each type of estimation discriminated by the size of the sample and the number of replications. For the overall calculus, simulations and results, the statistical software Stata 16 was used (StataCorp, 2019).

3. Results

3.1. Statistical significance

The OLS simulation practiced, establish that the pattern of statistical significance for all estimators will remain as long as the sample size is increasing, the special case of micronumerosity tend to disrupt the statistical significance as expected, but the yielding estimators seems to be closer to the DGP.

Table 1. OLS Monte Carlo simulation with different sizes

VARIABLES	(1)	(2)	(3)	(4)
	Y	Y	Y	Y
x1	9.549	9.288***	9.915***	9.991***
	(0)	(0.440)	(0.208)	(0.0468)
x2	10.36	9.915***	10.01***	9.961***
	(0)	(0.491)	(0.200)	(0.0499)
x3	8.952	9.709***	10.27***	9.979***
	(0)	(0.362)	(0.211)	(0.0457)
x4	10.66	10.44***	9.977***	10.04***
	(0)	(0.295)	(0.207)	(0.0453)
x5	10.70	9.233***	10.59***	10.02***
	(0)	(0.530)	(0.289)	(0.0506)
Constant	8.902	9.979***	10.10***	9.997***

VARIABLES	(1)	(2)	(3)	(4)
	Y	Y	Y	Y
	(0)	(0.394)	(0.225)	(0.0463)
Observations	6	10	20	500
R-squared	1.000	0.999	0.999	0.998
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

Source: Own construction

As an interesting thing to consider, the R^2 values changes when we estimate the DGP with 20 observations, to a lower accuracy (but still closer to 1) in the context of 500 observations, this is proof that the property of consistency among the OLS estimator is achievable (and of course all classical assumptions of the linear regression model are also satisfied). This tends to indicate that the affirmation of Speed (1994) regarding to the relationships found in small sample sizes tend to remain as the size of the sample increases.

Going further with the jackknife estimation, it can be observed that it cannot be computed in the presence of perfect micronumerosity, leading to the impossibility to even approach to get a result from observed coefficients, among the statistical significance it also remains across the different sample sizes, suggesting the same result from OLS.

Table 2 Jackknife estimation with different sample size

VARIABLES	(1)	(2)	(3)	(4)
	Y	Y	Y	Y
x1	-	9.733***	10.25***	10.00***
	-	(0.470)	(0.358)	(0.0445)
x2	-	9.892***	9.891***	9.926***
	-	(0.296)	(0.380)	(0.0454)
x3	-	10.42***	10.33***	10.02***
	-	(0.667)	(0.296)	(0.0445)
x4	-	11.04***	10.09***	9.977***
	-	(0.523)	(0.403)	(0.0476)
x5	-	10.29***	9.627***	10.03***
	-	(0.784)	(0.401)	(0.0434)
Constant	-	9.454***	9.830***	10.04***
	-	(0.433)	(0.282)	(0.0462)
Observations	6	10	20	500
R-squared	-	0.999	0.998	0.998
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

Source: Own construction

The goodness of fit of the model tends to be reduced as the sample size is increased considering this type of estimation, we can also see that the coefficients vary from the ones estimated via OLS. The bootstrap estimation is presented in the Table 3 and display results a little bit different from the OLS and the jackknife, in the induced model with micronumerosity the coefficients can be computed, however, standard errors cannot be estimated.

Table 3. Bootstrap estimation with different sizes

VARIABLES	(1)	(2)	(3)	(4)
	Y	Y	Y	Y
x1	10.06	10.26**	9.651***	10.01***
	(0)	(4.026)	(0.214)	(0.0512)
x2	9.375	10.67**	10.23***	9.958***
	(0)	(4.408)	(0.264)	(0.0359)
x3	10.85	9.744***	10.40***	10.02***
	(0)	(2.750)	(0.200)	(0.0379)
x4	10.24	10.27	9.718***	10.05***
	(0)	(8.483)	(0.177)	(0.0422)
x5	10.68	10.15***	9.959***	10.01***
	(0)	(3.871)	(0.238)	(0.0479)
Constant	11.34	10.50***	10.04***	9.993***

VARIABLES	(1)	(2)	(3)	(4)
	Y	Y	Y	Y
	(0)	(2.564)	(0.286)	(0.0396)
Observations	6	10	20	500
R-squared	1.000	0.993	0.999	0.998
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

Source: Own construction

According to the Monte Carlo experiment with the bootstrap technique, it can be seen that as the sample size is increasing, the statistical significance will also be increased. The variables x_1, x_2 and x_4 demonstrate this situation, where for example with n = 10, for x_4 there wasn't a statistically significant relationship with y in the regression model. Then as soon as we increased the sample size to n = 20 the variable turned to be significant, the similar case can be observed with x_1 and x_2 where they only were significant at a 5% with n = 10. Then with n = 20 they become significant at 1%, indicating that the bootstrap approach is sensitive to the number of observations regarding the coefficient hypothesis testing. This might suggest is not a good idea to perform this technique with a low sample size since it might discard a real relationship among the variables.

Following with the Lasso regression, micronumerosity doesn't allow the estimation of the coefficients. And the overall statistical significance remains equal across regressions with different sample sizes. This result indicates that estimations are consistent across models using the right variables with the specific function formal equally to the DGP.

Table 4. Lasso estimations with different sizes

VARIABLES	(1)	(2)	(3)	(4)
	Y	Y	Y	Y
x1	-	11.18***	10.09***	10.03***
	-	(0.832)	(0.242)	(0.0453)
x2	-	9.605***	9.785***	9.913***
	-	(0.376)	(0.272)	(0.0452)
x3	-	10.70***	9.866***	10.07***
	-	(0.551)	(0.264)	(0.0442)
x4	-	9.441***	9.585***	9.873***
	-	(0.355)	(0.235)	(0.0421)
x5	-	10.17***	9.861***	9.984***
	-	(0.397)	(0.336)	(0.0433)
Observations	6	10	20	500
Robust standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

Source: Own construction

It can be noted that Lasso regression omits the constant parameter in this single exercise, but the highest possible goodness of fit has been selected according to the variables. Thus, the statistical significance of the estimators prevails across the models with different sample sizes. However, it is necessary to appoint that Lasso regression doesn't look directly at the p-values or the standard errors since its sole objective is to isolate a model where the predictions become more suitable according to the data (StataCorp, 2019).

The robust regression estimates are similar to the ones done with Lasso and jackknife in terms that the model cannot be estimated when micronumerosity is present. The other results related to the statistical significance of the estimators indicate that when we're in the context of short samples, the relationships remain significant as the number of observations increase.

Table 5 Robust regression with different sizes

VARIABLES	(1)	(2)	(3)	(4)
	Y	Y	Y	Y
x1	-	9.866***	9.883***	10.02***
	-	(0.443)	(0.282)	(0.0442)
x2	-	11.04***	9.560***	9.978***
	-	(0.618)	(0.255)	(0.0460)
x3	-	10.15***	10.35***	10.02***
	-	(0.611)	(0.290)	(0.0415)
x4	-	9.315***	10.16***	9.972***
	-	(1.361)	(0.333)	(0.0441)
x5	-	10.88***	10.25***	9.963***
	-	(0.649)	(0.215)	(0.0430)
Constant	-	10.58***	9.949***	10.02***
	-	(0.935)	(0.226)	(0.0444)
Observations	6	9	20	500
R-squared	-	1.000	0.999	0.998
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

Source: Own construction

An interesting thing to appoint is that as long as we're having a large sample regarding our regressions, the goodness of fit tends to be somewhat reduced across estimations. This led to confirm the conclusion that R^2 is sensitive to the number of observations among the sample.

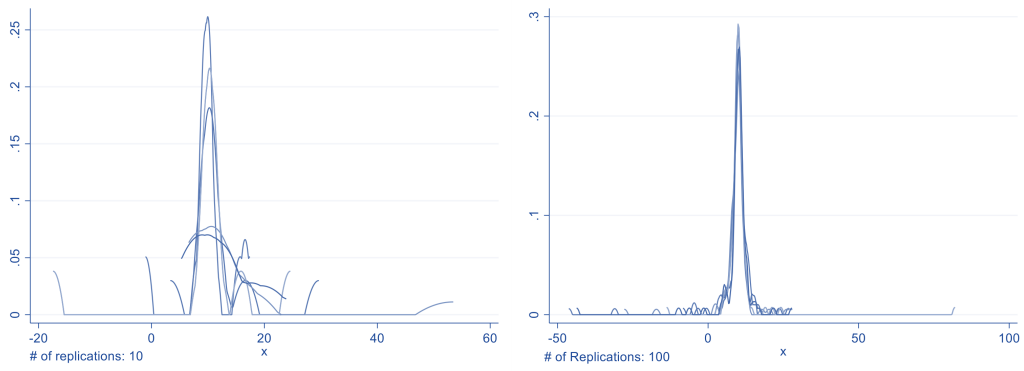
3.2. Bias behavior of the parameters

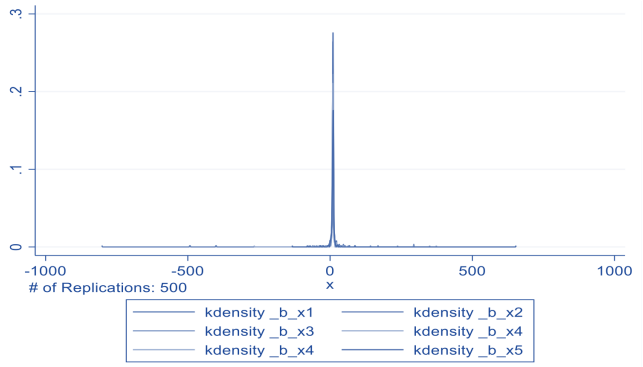
This section consists of the results for each type of estimation (OLS, jackknife, bootstrap, lasso and robust regression) referring to the distributions across replications for the coefficients, kernel densities were used for each coefficient of the different x variables in order to provide analysis regarding the importance of the number of replications.

3.2.1. Ordinary Least Squares

Considering a number of 6 observations, the coefficients for each variable tend to be somewhat unstable when the number of replications is low, meaning that in the presence of micronumerosity, the estimators are less likely to be trustable. As replications are increased to 100 and 500, the estimators seem to converge to their true value of 10, the situation clearly implies that across regressions with normally distributed data, as long as we replicate enough times the experiments, the expected value seems to be close to our DGP, it should be noted that OLS estimators stills covers some extreme values which could be affecting the consistency across replications, as we can see it in the graphical pattern in Figure1.

Figure 1. OLS - Distributions of the coefficients with n=6





Source: Own construction

These results proofs evidence that under micronumerosity, OLS estimates are unstable so it should be avoided at all cost. Considering the 500 replications for the 6 observations regression with OLS, the descriptive statistics for each coefficient reflects an undeniable reality. The minimum and maximum values are out of scale regarding to our DPG where each coefficient is equal to 10, even when the mean value is somewhat closer, the results yield unstable.

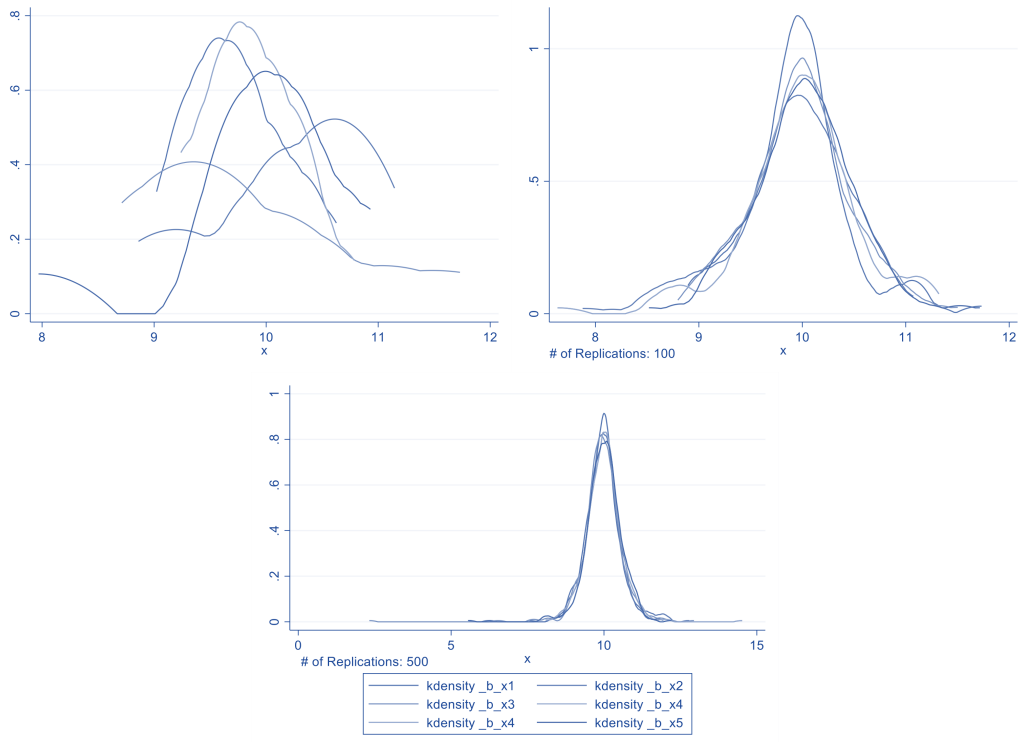
Table 6. OLS Descriptive Statistics with n=6

Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.78	13.529	-86.594	243.693
_b_x2	500	13.807	102.362	-588.518	2199.746
_b_x3	500	7.444	49.54	-1044.307	199.705
_b_x4	500	10.553	28.372	-306.405	526.281
_b_x5	500	4.668	62.136	-1043.826	62.116
_b_cons	500	5.83	92.71	-2015.365	188.439

Source: Own construction.

Now considering the number of observations as 10, the following pattern of distributions can be found in Figure 2.

Figure 2. OLS - Distributions of the coefficients with *n* = 10



Source: Own construction.

There is a quick and stable rate of convergence relative to the distributions of the estimators for each variable which is depicted across replications. The distributions tend to be normal as the simulation number increase, leading to the true value of the estimators for all x variables and the constant term. The descriptive statistics are shown ahead in Table 7 considering 500 hundred replications of the Monte Carlo simulations with n=10 observations.

Table 7. OLS Descriptive Statistics n=10

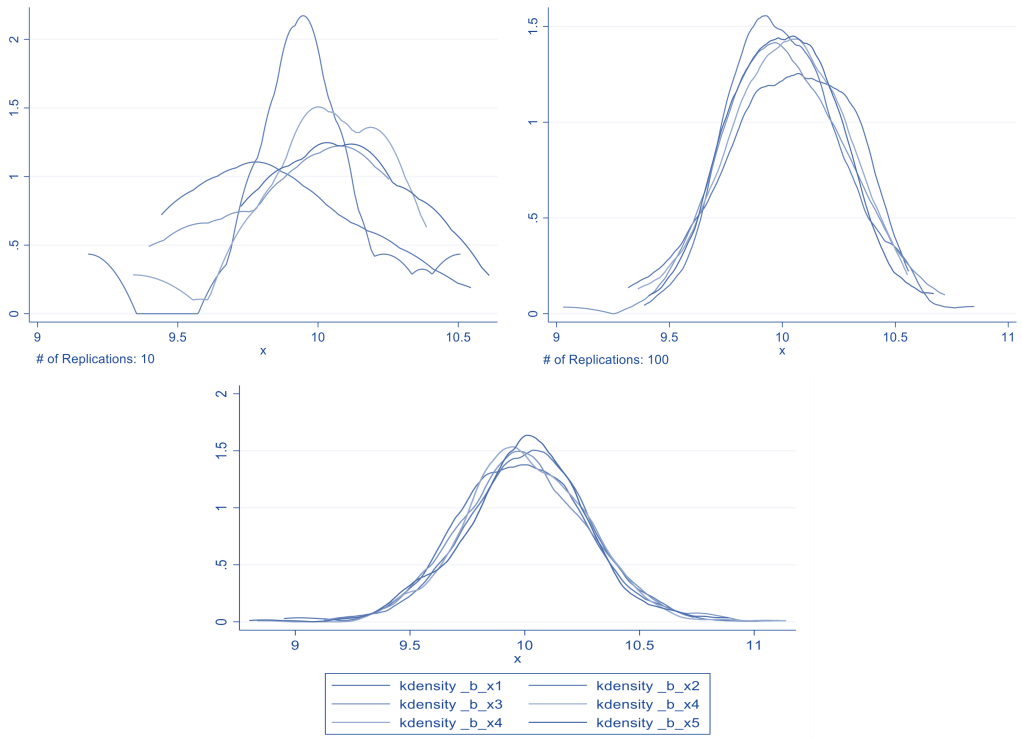
Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.937	.575	5.552	11.942
_b_x2	500	9.994	.599	7.56	12.789
_b_x3	500	9.957	.688	2.325	12.944
_b_x4	500	10.007	.583	8.089	14.52
_b_x5	500	9.999	.582	5.572	12.208
_b_cons	500	10.002	.535	7.253	12.053

Source: Own construction.

We can see that the minimum and maximum values for the 500 hundred replications with n = 10 tends to be more stable than when n = 6 which is the micronumerosity simulation. In this case the mean values are also more accurate in terms to approach to the data generating process of equation (1).

Now considering the number of observations to 20, the pattern of the distributions for each parameter is shown ahead in Figure 3, indicating a possibly significant difference from the n = 10 exercise because the shape of the curves for each distribution are different.

Figure 3. OLS - Distributions of the coefficients with n = 20



Source: Own construction

The range of the distribution is more accurate (from 9 to 11 in the x axis) for all replications with 20 observations, this tends to indicate that the precision of the estimates is increasing as expected. However, the shape of the curve is somewhat different but stills relies over 10. Which is a sign of the consistency and unbiasedness property of the estimator. The descriptive statistics in Table 8 from the 500-replication exercise within this number of observations reflects a good precision of the estimators.

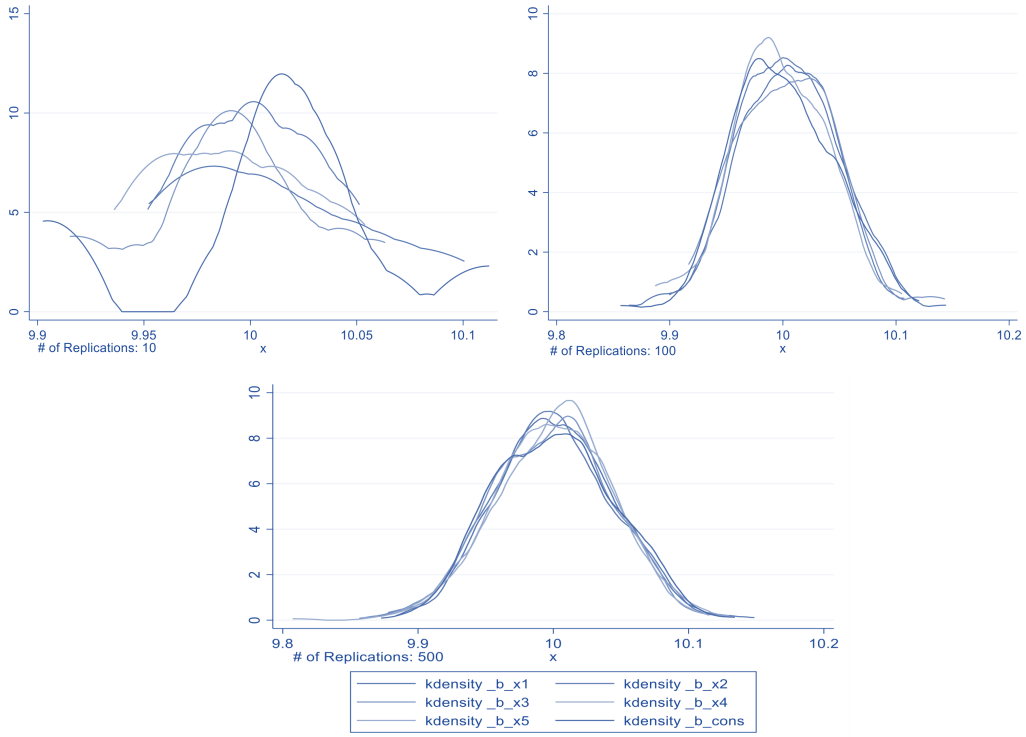
Table 8. OLS Descriptive statistics with n = 20

Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.979	.274	8.95	10.736
_b_x2	500	10.018	.269	9.145	11.104
_b_x3	500	10.003	.284	9.126	11.082
_b_x4	500	10.007	.268	8.834	11.139
_b_x5	500	10.002	.275	8.8	10.889
_b_cons	500	9.998	.271	9.262	10.837

Source: Own construction

Finally, as a comparing exercise, we're setting the number of observations to 500 in order to understand the behavior of the coefficients' distribution as it is shown in Figure 4.

Figure 4. OLS - Distributions of the coefficients with n = 500



Source: Own construction

As expected, the higher number of observations tend to have a faster converging rate to the true value of the parameters than the other simulations with lesser observations, the accuracy of the regressions are shown in the descriptive statistics ahead in Table 9.

Table 9 OLS Descriptive Statistics n= 500

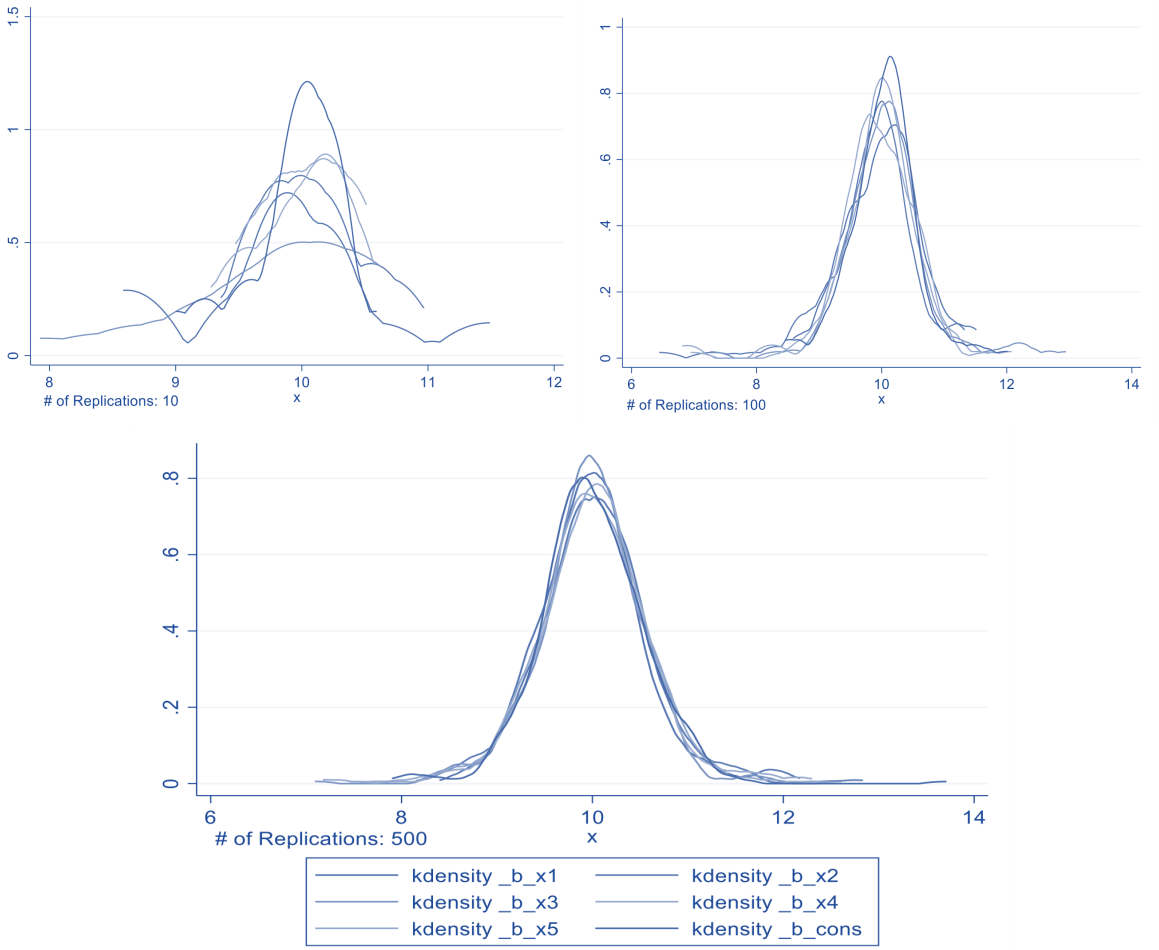
Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	10.002	.043	9.883	10.149
_b_x2	500	9.999	.044	9.857	10.112
_b_x3	500	9.998	.043	9.878	10.115
_b_x4	500	10.001	.044	9.86	10.13
_b_x5	500	10.001	.044	9.807	10.132
_b_cons	500	10.001	.044	9.873	10.134

Source: Own construction.

3.2.2. Jackknife

This type of estimation cannot be performed in the presence of perfect micronumerosity, so distribution analysis cannot be done with the case of 6 observations. Moving ahead with 10 observations, the behavior of the distributions of the parameters according to different replications are shown in Figure 5.

Figure 5. Jackknife - Distributions of the coefficients with n = 10



Source: Own construction.

It appears that the range of the different parameters' distributions in the case of 100 replications is higher than the rest of the simulations considering 10 observations, something particular but yet over the long-run not important since the mean value of all replications stills converge to the true value. The shape of the distributions cannot be established as better from the OLS, since the range varies widely. From this, descriptive statistics in Table 10 would be useful.

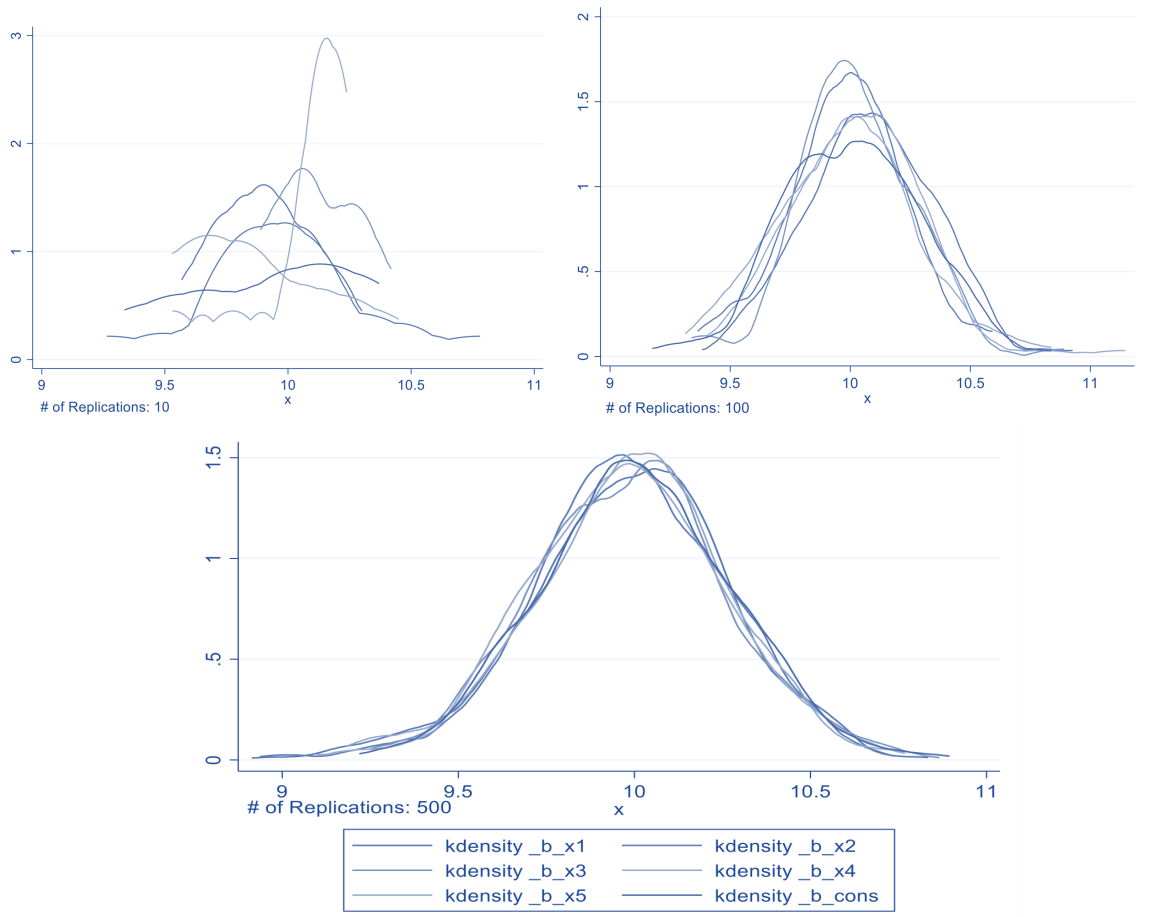
Table 10 Jackknife descriptive statistics n = 10

Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.999	.554	8.398	12.178
_b_x2	500	10.024	.594	8.152	12.835
_b_x3	500	9.987	.554	7.092	12.61
_b_x4	500	10.009	.63	7.176	12.303
_b_x5	500	10.001	.577	7.451	12.494
_b_cons	500	9.997	.565	7.9	13.709

Source: Own construction.

The expected value of the parameters is more accurate in the jackknife simulations than it is with the OLS, also the standard deviation tends to be lower for the jackknife approach. Considering now a sample size of 20 observations, the following pattern can be observed in Figure 6.

Figure 6. Jackknife - distributions of the coefficients with n = 20



Source: Own construction.

Jackknife estimation seems to be more unstable with a lower number of replications considering n=20, however we're not sure yet if it's more suitable than OLS by the graphic interpretation, looking at the descriptive statistics in Table 11 we can have a better approximation.

Table 11 Jackknife descriptive statistics n=20

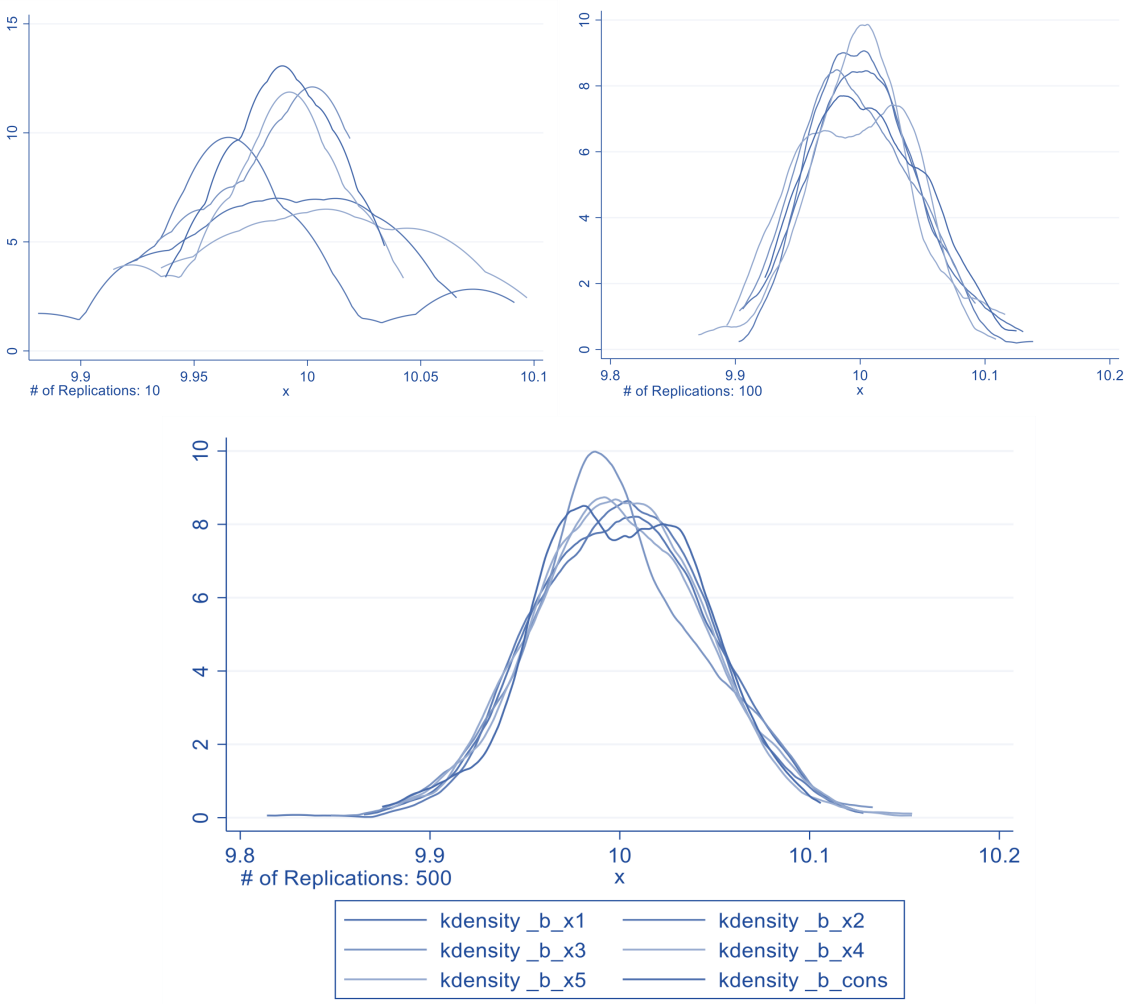
Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.985	.284	8.913	10.834
_b_x2	500	9.986	.272	8.936	10.74
_b_x3	500	9.988	.273	9.249	10.867
_b_x4	500	9.978	.274	9.097	10.769
_b_x5	500	9.99	.276	9.181	10.766
_b_cons	500	10.006	.274	9.219	10.895

Source: Own construction

The estimations with jackknife seem to be pretty close to the ones performed with OLS at this number of observations, however OLS seems to have the advantage to be more stable with lesser replications than Jackknife does and the expected value with n = 20 of the estimators is closer to the DGP for OLS than it is for jackknife.

Finally, with 500 observations the pattern is shown in Figure 7, it is noted that jackknife has the counterpart to require a higher and significant time of computing during the estimations.

Figure 7. Jackknife - distributions of the coefficients with n = 500



Source: Own construction.

The jackknife distribution with n=500 seems to converge somewhat equal to the OLS estimations. If we analyze the statistics relative to the OLS for the same number of observations, we'll find that the OLS performs better in terms of the standard deviation and minimum and maximum values closer to 10.

Table 12 Jackknife Descriptive Statistics n=500

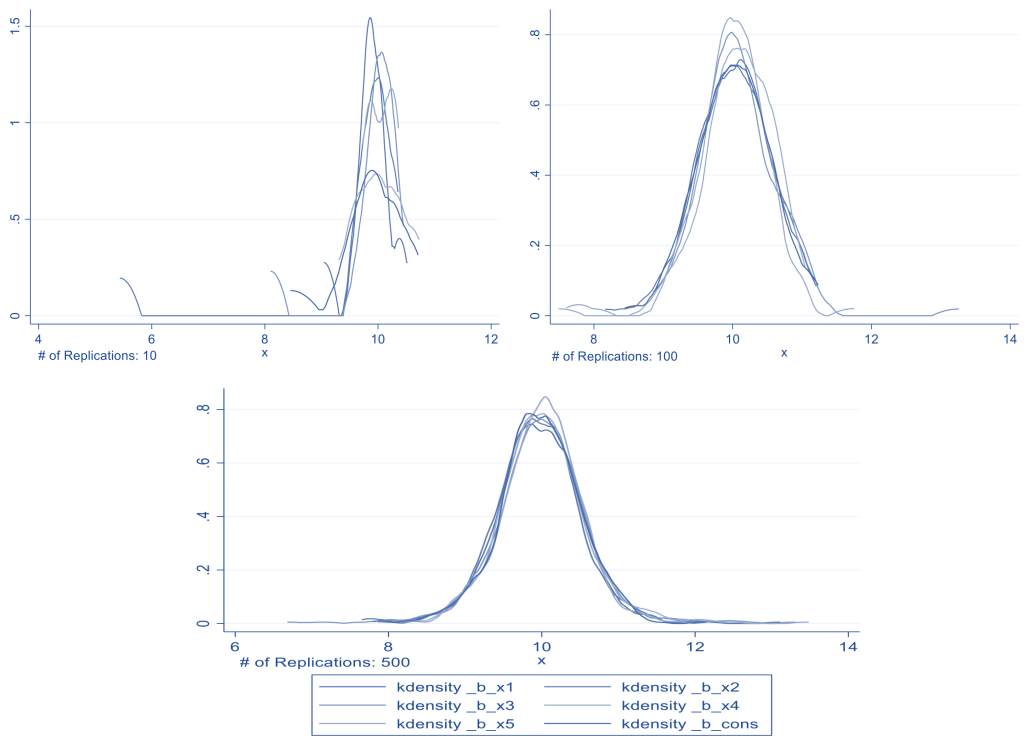
Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	10.002	.045	9.865	10.126
_b_x2	500	10.001	.044	9.814	10.128
_b_x3	500	9.999	.046	9.848	10.133
_b_x4	500	10	.043	9.879	10.154
_b_x5	500	10	.045	9.875	10.154
_b_cons	500	10	.043	9.875	10.106

Source: Own construction.

3.2.3. Bootstrap

Similar to the Jackknife approach, bootstrap estimation cannot be performed if the number of observations is 6, so it is not allowed perfect micronumerosity. Moving to the analysis with n = 10 we can observe the following patterns of the parameters via bootstrap in Figure 8.

Figure 8. Bootstrap - Distributions of the Coefficients with n = 10



Source: Own construction.

The pattern related to the lowest replications (10) tends to be unstable with the bootstrap technique with n=10, but as it gets more replications the parameters converge to their true value. The descriptive statistics are presented ahead in Table 13 indicating a similar behavior to the Jackknife technique.

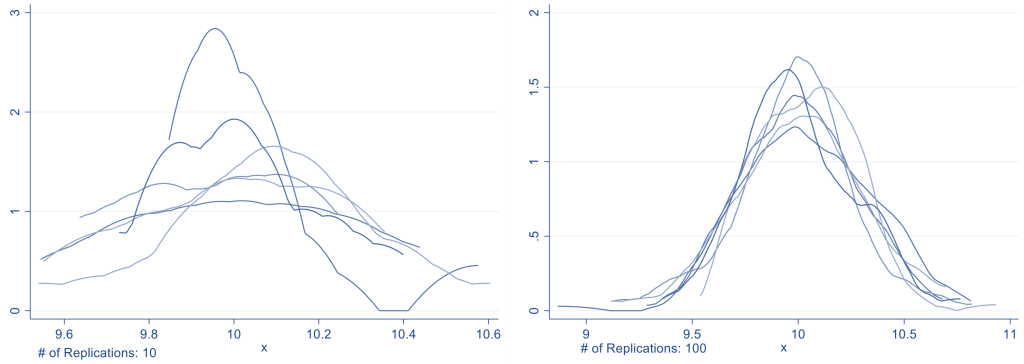
Table 13. Bootstrap Descriptive Statistics n=10

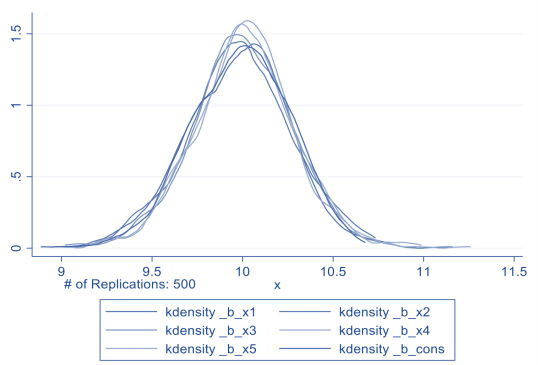
Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.955	.588	7.653	13.115
_b_x2	500	9.968	.552	7.915	12.409
_b_x3	500	9.995	.633	6.682	13.314
_b_x4	500	10.014	.558	7.865	12.28
_b_x5	500	10.015	.556	7.969	13.482
_b_cons	500	9.98	.532	7.794	12.132

Source: Own construction.

Moving to n=20, the patterns relative to the lesser replications tend to be more stable than with n=10, indicating a sensitive behavior of the bootstrap with lower samples, however stills yielding results similar to OLS and Jackknife.

Figure 9. Bootstrap - Distributions of the Coefficients with n=20





Source: Own construction.

As happens with the jackknife, the bootstrap in Figure 9 seems to have variations for each distribution of each variable when the replication number is set to 100, however the distributions converge as OLS and jackknife in the case of bootstrap when replications are set to 500.

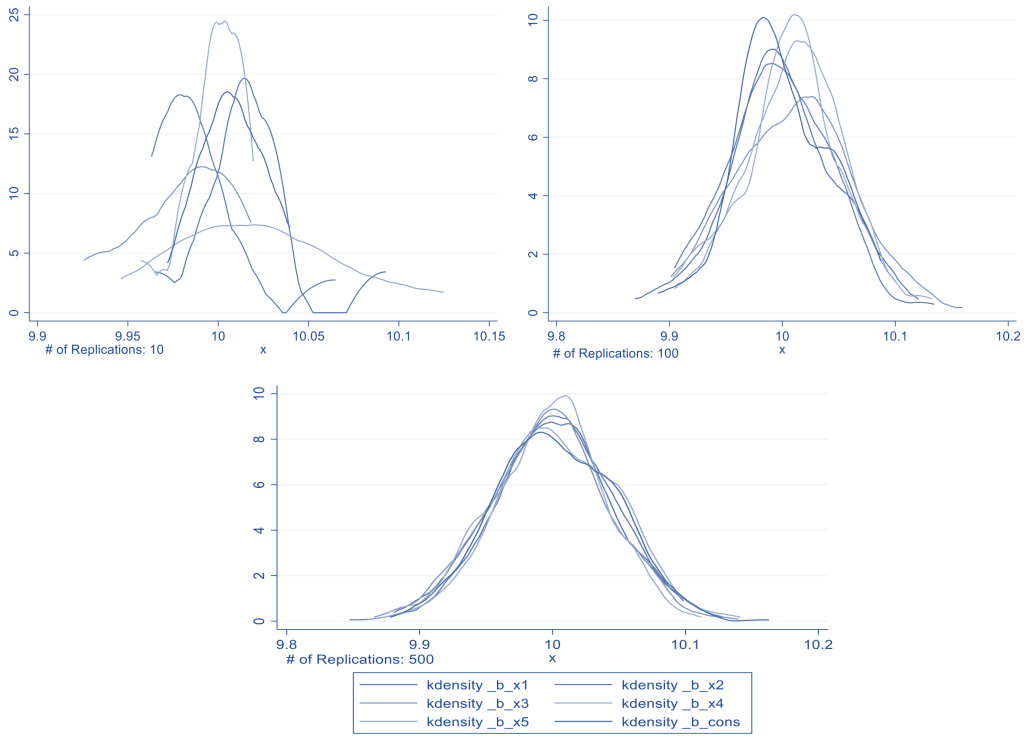
Table 14 Bootstrap Descriptive Statistics n=20

Estimated Parameter	Replications	Mean	Standard Deviation	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.977	.288	8.93	11.161
_b_x2	500	9.986	.288	9.015	10.734
_b_x3	500	9.998	.277	9.019	10.957
_b_x4	500	10.013	.279	9.185	11.261
_b_x5	500	10.009	.275	8.953	10.988
_b_cons	500	9.988	.273	8.885	10.678

Source: Own construction

Going further with the bootstrap technique and using n=500 observations, the graphical pattern in Figure 10 indicates some better adjustment regarding the lower replications compared to n=10 and n=20.

Figure 10. Bootstrap - Distributions of the Coefficients with n=500



Source: Own construction.

The pattern of the distributions among the coefficients when the number of replications is set to 500 tends to be more different from the OLS and the Jackknife estimations, which might suggest that bootstrap performs different distributions for each estimator even when the OLS and jackknife tend to converge the distribution for all estimators with the same number of n=500 observations. According to the descriptive statistics in Table 15, bootstraps seems to be as efficient as OLS and Jackknife specially because of the mean value of the coefficients, it's stills as accurate relative to the expected value of the estimators in comparison.

Table 15 Bootstrap Descriptive Statistics n=500

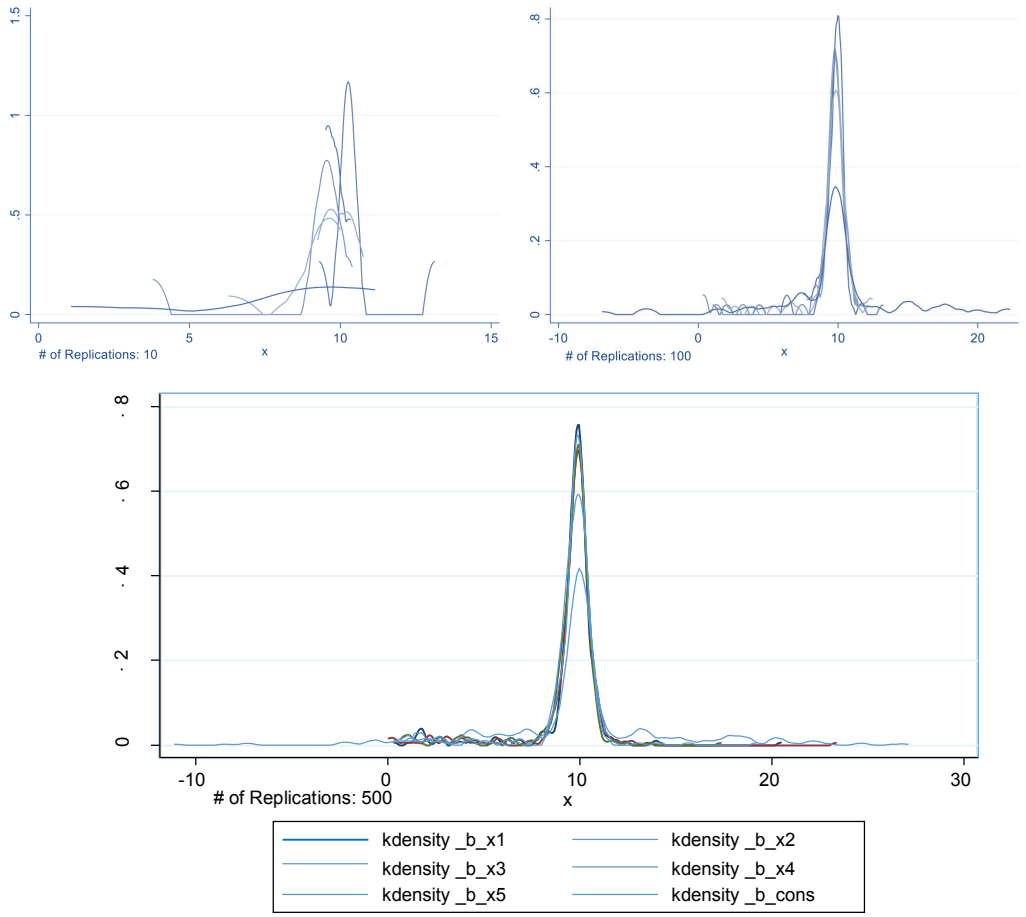
Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.998	.044	9.881	10.099
_b_x2	500	10.003	.044	9.847	10.128
_b_x3	500	9.997	.045	9.866	10.14
_b_x4	500	9.996	.042	9.875	10.112
_b_x5	500	10.005	.045	9.861	10.141
_b_cons	500	10.002	.045	9.878	10.163

Source: Own construction.

3.2.4. Lasso regression

As mentioned before, lasso cannot compute the model when the number of observations is equal to 6, so we're going straight to the analysis with 10 observations, the graphical pattern is shown ahead in Figure 11.

Figure 11. Lasso - Distributions of the Coefficients with n=10



Source: Own construction.

The figure suggest that the distributions are different for each variable across replications, in that case the constant coefficient remains with difference ranges when its converging to the true parameter. The descriptive statistics in Table 16 suggest that from the 500 simulations some of them failed and were just covering up to 307,

317, 318 replications, the constant term was the only one which remained across the simulations, however even when the mean value it's somewhat accurate, the minimum and maximum values are varying more in the coefficients associated with the x variables.

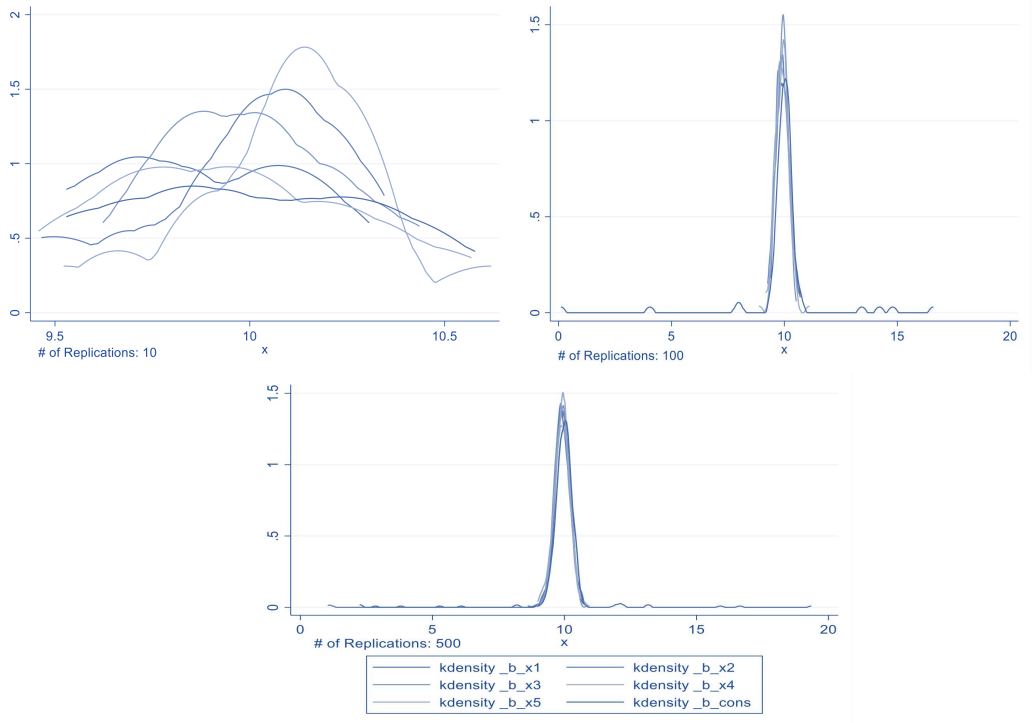
Table 16 Lasso Descriptive Statistics n=10

Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	317	9.484	1.963	.287	20.51
_b_x2	317	9.444	2.142	.014	23.37
_b_x3	314	9.371	1.921	.657	17.363
_b_x4	318	9.382	2.063	.29	19.233
_b_x5	307	9.457	1.841	.894	12.664
_b_cons	500	10.108	4.509	-11.102	27.153

Source: Own construction.

Proceeding with lasso estimations with n=20 we watch the graphical pattern associated to the distribution of the parameters as it follows in Figure 12.

Figure 12. Lasso - Distributions of the Coefficients with n=20



Source: Own construction

According to the distributions, the estimators associated to the different variables seem to behave over a wide range during the simulations with n=20 observations. Relying in the descriptive statistics in Table 17, we can find a significant range regarding the x_1 variable and the constant term in the regression. Also, some simulations failed to accomplish the main total of 500, which tends to indicate that lasso approach is sensitive to the number of replications and therefore, the overall range of the estimators differs across replications.

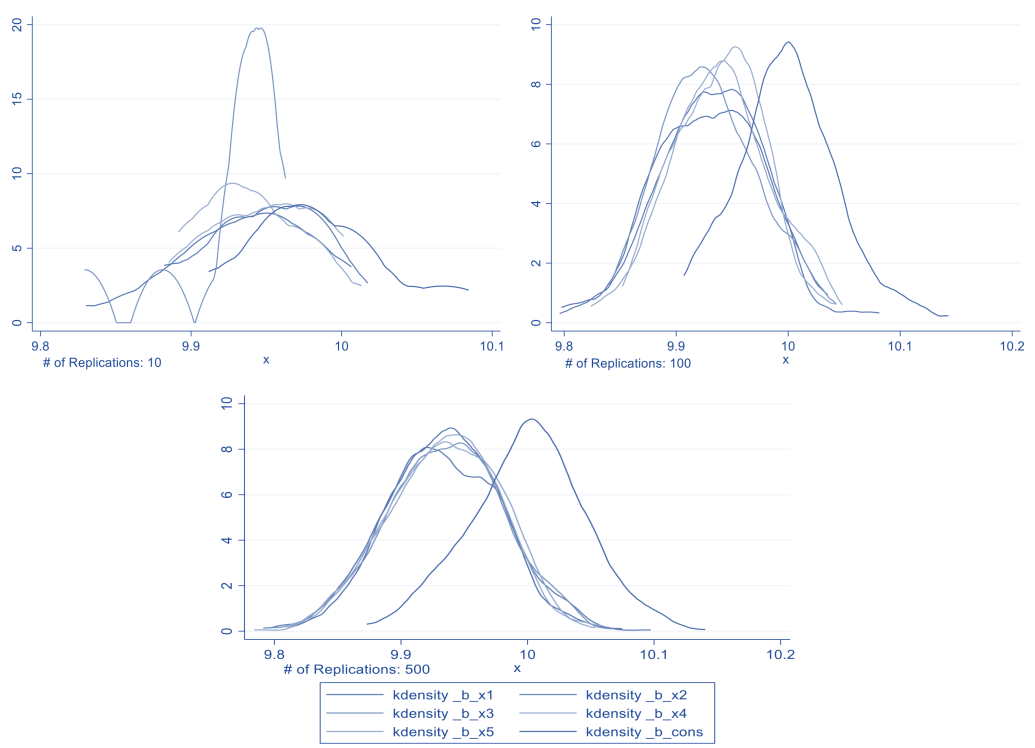
Table 17 Lasso Descriptive Statistics n=20

Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	476	9.889	.635	1.037	10.743
_b_x2	474	9.908	.296	8.608	10.746
_b_x3	474	9.904	.276	8.965	10.901
_b_x4	474	9.903	.31	8.986	10.962
_b_x5	474	9.917	.272	8.957	10.609
_b_cons	500	9.98	1.002	2.243	19.351

Source: Own construction.

The descriptive statistics tend to indicate some instability of the lasso regression with n=10 and 20, which would be judge in overall with the 500 observations simulations. Proceeding with the analysis with n=500 simulations, the graphical pattern is shown ahead in Figure 13.

Figure 13. Lasso - Distributions of the Coefficients with n=500



Source: Own construction.

The distribution seems not to converge to the exact value of the DGP, lasso regression also seems to perform a different distribution relative to the other x variables and the constant coefficient. This doesn't mean Lasso regression is inconsistent, since it's close to 10, however is not as consistent as other estimations are. The descriptive statistics in Table 18 of the estimated parameters, tends to confirm this idea since the expected value of the estimators is not as close to the other types of estimations, also it tends to have a standard deviation a little bit higher than the others.

Table 18 Lasso Descriptive Statistics n=500

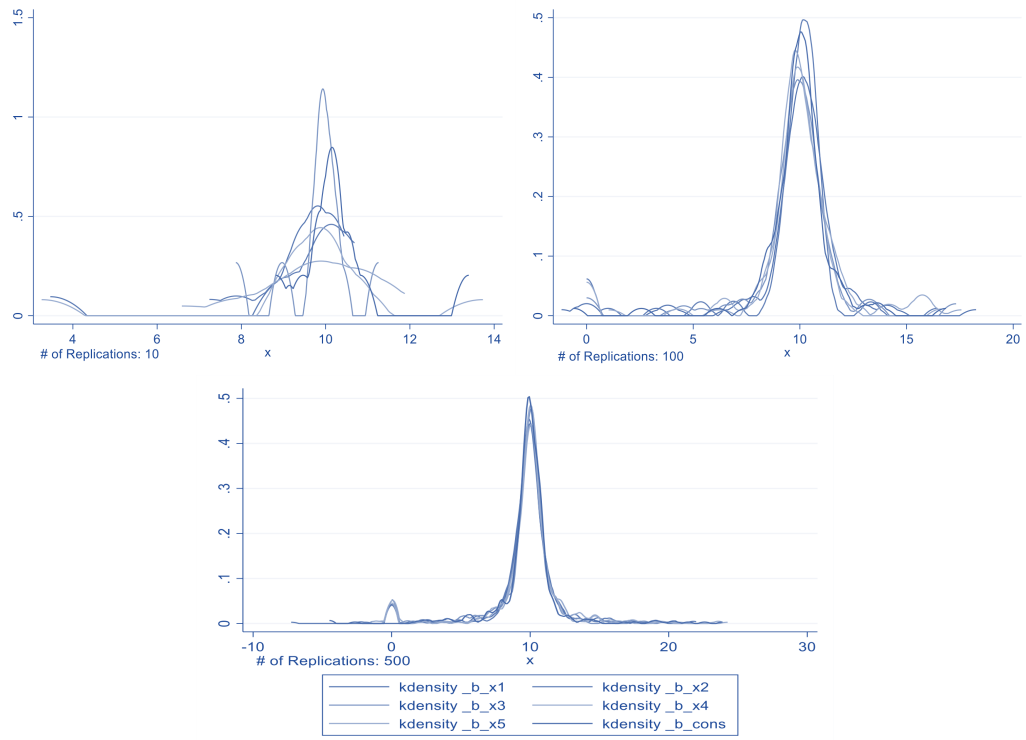
Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.934	.043	9.81	10.075
_b_x2	500	9.934	.047	9.791	10.097
_b_x3	500	9.936	.046	9.795	10.067
_b_x4	500	9.934	.044	9.784	10.059
_b_x5	500	9.935	.044	9.807	10.054
_b_cons	500	9.999	.046	9.873	10.14

Source: Own construction.

3.2.5. Robust regression

The last type of estimation we're analyzing is the robust regression, this one cannot be estimated with n=6 observations (the perfect micronumerosity case) so we're going straight forward to set n=10 observations and perform the graphical distribution patterns.

Figure 14. Robust Regression - Distributions of the Coefficients with n=10



Source: Own construction

With 500 simulations and n=10, Stata calculated 484 replications, the rest of the remaining replications failed in the maximization process. There are some appoints to make here, first: the range of the distribution with n=10 observations across replications is way too high in comparison OLS, Jackknife, Bootstrap or Lasso types of estimations, second: some of the distributions of some variables tend to have spikes closer to the value of 0 indicating that a significant number of times, the robust regression adjusted some coefficients as 0. According to the descriptive statistics in Table 19, the mean value of the coefficients tends to converge better than Lasso, however Jackknife and Bootstrap perform better with this set of observations.

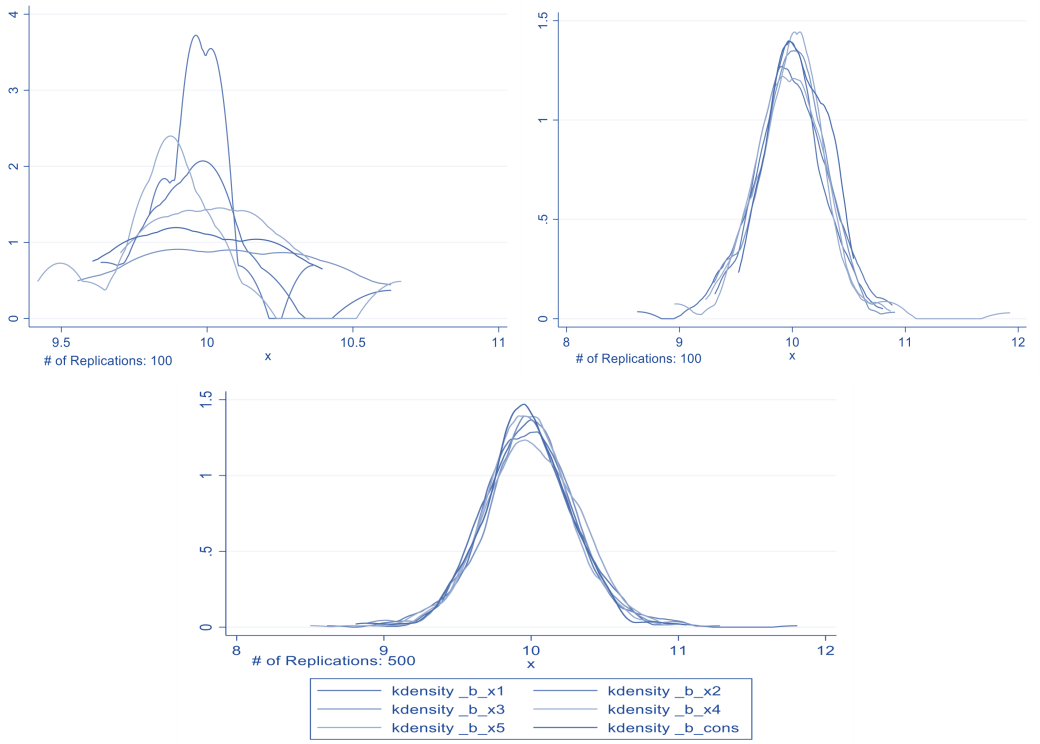
Table 19 Robust Regression Descriptive Statistics n=10

Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	484	9.704	2.682	-7.252	20.553
_b_x2	484	9.732	2.689	-2.542	21.981
_b_x3	484	9.494	2.57	0	19.703
_b_x4	484	9.794	2.661	-.849	23.795
_b_x5	484	9.694	2.949	-1.525	24.251
_b_cons	484	9.898	2.408	-4.489	23.887

Source: Own construction.

Moving forward and setting n=20 observations, we can observe that the graphical pattern of the distributions for each estimator of each variable is going more accurate with the robust regression technique, however no significant changes can be concluded from the other types of estimations.

Figure 15. Robust Regression - Distributions of the Coefficients with n=20



Source: Own construction.

The behavior with n=20 observations is far better than with n=10, also these results are consistent with a lesser range over the estimators. The mean value of the estimators is getting closer to 10 as we increased the number of replications.

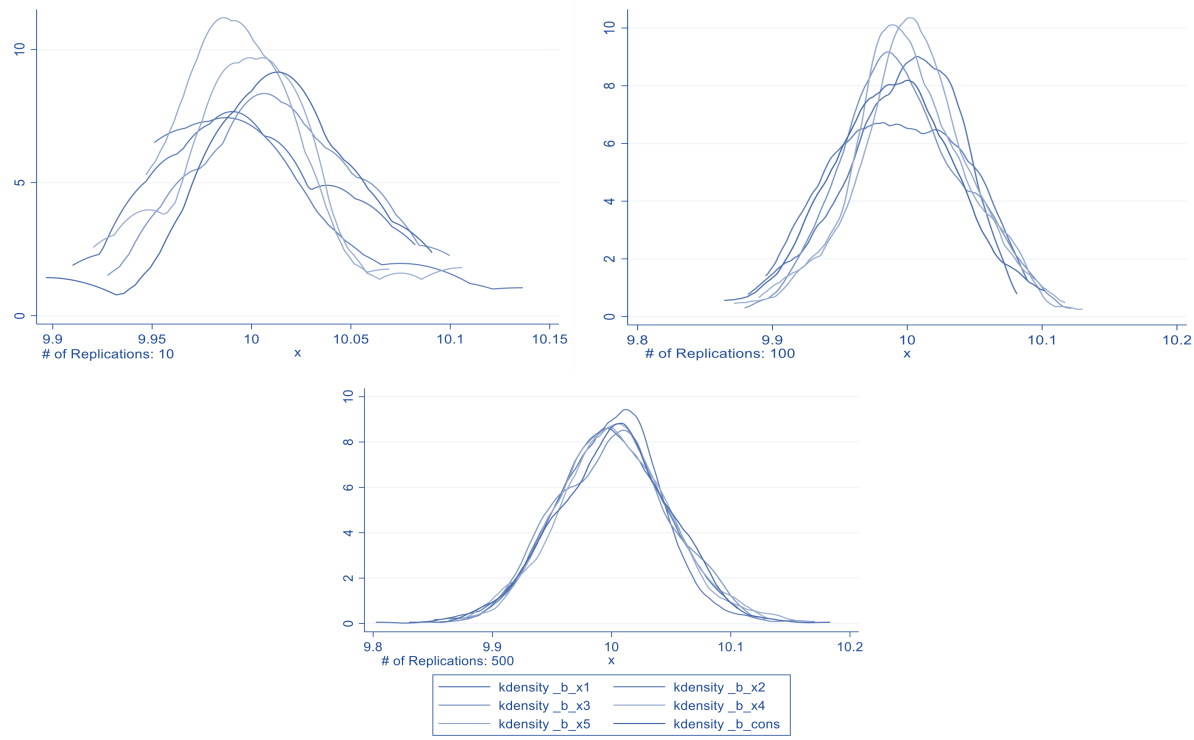
Table 20 Robust Regression Descriptive Statistics n=20

Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.989	.313	8.915	11.808
_b_x2	500	9.981	.324	8.613	11.054
_b_x3	500	9.998	.32	8.888	11.283
_b_x4	500	9.974	.3	8.499	11.044
_b_x5	500	10.003	.317	8.958	10.864
_b_cons	500	9.972	.3	8.808	11.054

Source: Own construction.

Setting the final simulations with n=500, the results of the distributions with kernel densities are shown ahead in Figure 16. The pattern tends to indicate a convergence to the true value of the parameter as the number of replications are increased, also with the descriptive statistics in Table 21, the mean value is closer to 10, leading to think that robust regression is also a good option in large samples.

Figure 16. Robust Regression - Distributions of the Coefficients with n=500



Source: Own construction.

Table 21 Robust Regression Descriptive Statistics n=500

Estimated Parameter	Replications	Mean	Standard Deviation.	Minimum Value of the Parameter	Maximum Value of the Parameter
_b_x1	500	9.998	.046	9.802	10.106
_b_x2	500	9.997	.043	9.851	10.129
_b_x3	500	10.002	.048	9.856	10.17
_b_x4	500	10.002	.046	9.872	10.132
_b_x5	500	10.002	.047	9.864	10.154
_b_cons	500	10.002	.048	9.83	10.183

Source: Own construction

3.3. Comparing the estimations

In order to synthesize the previous part, we can discriminate the results by the number of observations (from the lowest) and the descriptive statistics for the coefficients, in this order of ideas the mean value of the whole estimators across simulations would be our reference point, standard deviation as lower it is the better, and the minimum and maximum values closer to 10 would be ranked.

Table 22 Comparison between Estimations, n=10

Estimation Type n=10	Expected Value of the Estimators	Expected Std. Deviation	Expected Minimum Value	Expected Maximum Value
OLS	9,98266667	0,59366667	6,0585	12,7426667
Jackknife	10,0028333	0,579	7,69483333	12,6881667
Bootstrap	9,98783333	0,56983333	7,64633333	12,7886667
Lasso	9,541	2,4065	-1,49333333	20,0488333
Robust Regression	9,71933333	2,65983333	-2,77616667	22,3616667
Best Option	Jackknife	Bootstrap	Jackknife	Jackknife

Source: Own construction

According to Table 22, when we're considering a sample size with n=10 observations in the context of a 6-coefficient estimation in the regression models, the best option is the jackknife estimation technique. It should be

noted that the number of freedom degrees in the residuals for this case is equal to 4. It is expected that when this number gets higher, we might have more accurate estimators from the other techniques.

Table 23 Comparison between Estimations, n=20

Estimation Type n=20	Expected Value of the Estimators	Expected Std. Deviation	Expected Minimum Value	Expected Maximum Value
OLS	10,0011667	0,2735	9,0195	10,9645
Jackknife	9,98883333	0,2755	9,09916667	10,8118333
Bootstrap	9,99516667	0,28	8,99783333	10,9631667
Lasso	9,91683333	0,46516667	6,466	12,2186667
Robust Regression	9,98616667	0,31233333	8,78016667	11,1845
Best Option	OLS	OLS/Jackknife	Jackknife	Bootstrap

Source: Own construction.

When the number of observations is increased to n=20 and the degrees of freedom is higher to a value of 14, the OLS performs quite better in the expected value of the coefficients according to Table 23, meanwhile we got a draw with OLS and jackknife in the case for the lesser expected value of the standard deviation. It is noted that the jackknife approach has better performance regarding the minimum expected value of the estimator.

Table 24 Comparison between Estimations, n=500

Estimation Type n=500	Expected Value of the Estimators	Expected Std. Deviation	Expected Minimum Value	Expected Maximum Value
OLS	10,0003333	0,04366667	9,85966667	10,1286667
Jackknife	10,0003333	0,04433333	9,85933333	10,1335
Bootstrap	10,0001667	0,04416667	9,868	10,1305
Lasso	9,94533333	0,045	9,81	10,082
Robust Regression	10,0005	0,04633333	9,84583333	10,1456667
Best Option	Bootstrap	OLS	Bootstrap	Lasso

Source: Own construction.

In Table 24, when our sample size is sufficiently large (n=500), the bootstrap technique performs better than OLS, Jackknife, Lasso or Robust regression, although OLS tends to have a lesser expected deviation than the rest. Over this stage, since samples sizes are large, there are sufficient arguments to prefer one method over another but this selection needs to be accounted for specific contexts, for example, robust regression wasn't scored as the best in any of these statistics but it would be extremely useful when we got outliers in the sample, where in such case the OLS estimator fails to account for them (Adepoju & Olaomi, 2012). In fact, a new development performed by Mishra (2008) of the robust approach can be more useful in the presence of outliers.

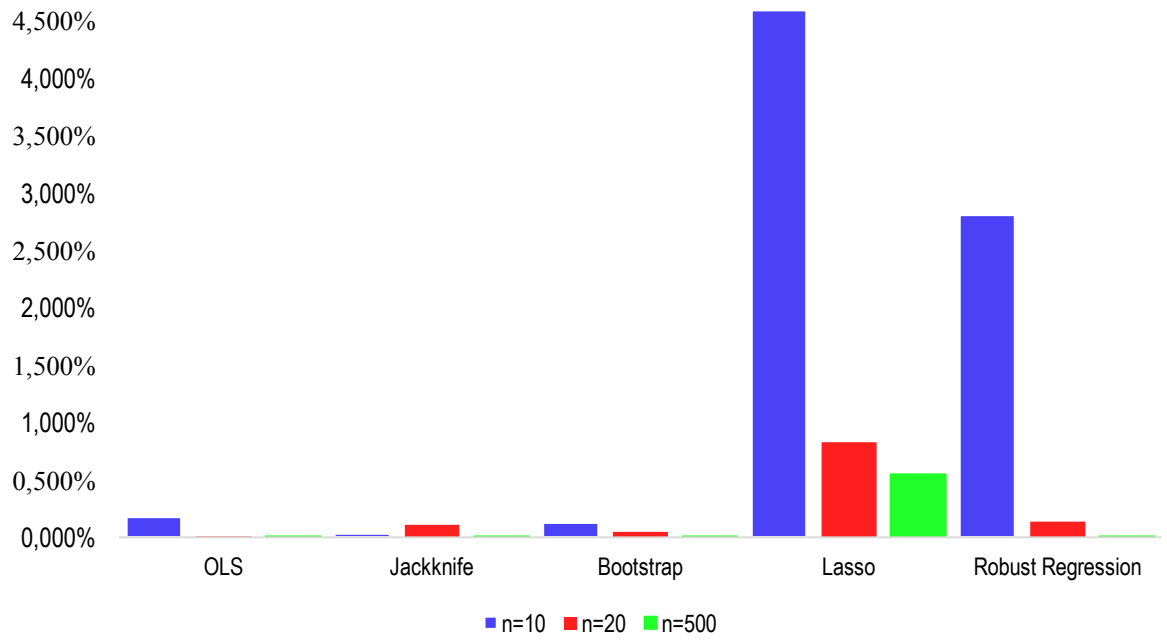
We need to remember that this analysis was performed with random variables that followed a distribution of $N \sim (0,1)$ and the main interest was to analyze the estimations for low samples (the perfect micronumerosity case n=6 and the others with n=10, n=20, n=500 with replications of 10, 100 and 500 simulations). The DGP in equation (1) was also established to be a cross-sectional type of data, so no autoregressive problems or incorrect specifications were assumed for the types of estimations.

The relative bias analysis in Figure 17, with n=10 observations, suggest that Lasso regression performs the worse bias value, reaching a score of deviation of 4.59% calculated from the expected value of the estimators compared with the true parameter, followed by the robust regression with a value of 2.807%. Bootstrap and OLS perform far better than these types of regressions with respective scores of 0.173% and 0.122%, the lower bias was obtained with the jackknife approach with a bias of 0.028%.

Moving to the sample size of n=20 observations, Lasso and robust regression performs also the worse value of the relative bias, respectively with values of 0.832% and 0.138%. Jackknife now turns to be in third place with a relative bias of 0.112% while bootstrap has a value of 0.048% and the OLS with a score of 0.012% indicating a lesser bias.

Finally, when the sample size is large (n=500), Lasso remains with the worse score in the relative bias with a value of 0.547%, meanwhile robust regression has a score of 0.005% of relative bias against the DGP. OLS and Jackknife have the same relative bias with a score of 0.003%. The best performance in terms of relative bias in this case was obtained with Bootstrap with a score of 0.002% of relative bias among the simulations, the proceeding graph summarizes this result.

Figure 17. Relative Bias for each estimation type by sample size



Source: Own construction.

Conclusion

This paper performed over 1500 simulations distributed among different sample sizes (n=6, n=10, n=20 and n=500) with a linear Data Generating Process in order to regress a model with six coefficients and five variables, these variables were normally distributed with zero mean and variance of one, the estimations types for the regressions across simulations were the approaches of OLS, Jackknife, Bootstrap, Lasso and Robust Regression.

The statistical significance of the coefficients across the models tends to follow the pattern described by Speed (1994) where a significant relationship is found in a small sample also will prevail when the sample size gets bigger. However, the Bootstrap approach seems to be sensitive to the sample size, since with n=10 observations it didn't present a significant relationship for one variable which was part of the DGP, suggesting that Bootstrap might discard significant relationships of certain variables during the regressions with a small sample size. As soon as the sample size increased to n=20, the bootstrap approach presented significant relationships with a 5% significance level, and with a larger sample size, the statistical significance was of 1%. On the other hand, OLS, Jackknife, Lasso and Robust regression performed well in terms of the statistical significance of the coefficients for all the variables in the DGP across the Monte Carlo simulations with different sample size.

Comparing the results with n=10 observations, the best estimation type was performed with the Jackknife approach, since the expected value of the coefficients was the best in terms to be closer to the true value of the DGP, also this approach suggests a lesser relative bias across the replications for the coefficients with this sample size. Bootstrap on the other hand with this sample size had the lowest expected standard deviation. In this case, it is confirmed that Speed (1994) was right in affirming that Jackknife and Bootstrap techniques are more suitable in small samples, however the drawback of the bootstrap approach is the sensitiveness in the statistical significance of the coefficients. According to these results, the jackknife approach seems to be more suitable for lower sample analysis.

In the case of n=20 observations, OLS obtained the best score regarding the accuracy of the estimators across simulations, as a reference for this, the relative bias was the lowest among the other types of estimations. In terms of the expected standard deviation, OLS matched the jackknife approach, but the minimum expected value of the estimators across replications of the jackknife was closer to the true value instead of the OLS regressions.

In the final simulations with n=500 observations, Bootstrap approach performed better than the rest of the estimation's types in terms of the accuracy of the estimator, a relative bias of 0.002% regarding from the true parameter was calculated with this approach. Also, the minimum expected value of the estimators was closer from this approach than the others, suggesting that bootstrap might be more appropriate for large samples.

According to the last results and as it is suggested by Speed (1994), researchers should perform also jackknife and bootstrap approaches when they're analyzing relationships from a set of variables in the multivariate regression framework, this in order to obtain more accurate estimations. However, the statistical significance might not be a good idea to be checked with the bootstrap approach since from this study, it was proved that its sensitive to the size of the samples and might induce to errors of type 1 more easily. Jackknife approach seems to be the most reliable method to perform correct inferences when the sample size is small.

Acknowledgments:

This research was facilitated by the Corporation Center of Public Affairs and Justice (Corporación Centro de Interés Público y Justicia -CIPJUS-). Special thanks to Jeisson Riveros Gavilanes, Oscar Betancurt and Luisa Tirado León for the support provided during the development of this research.

References

[1] Adepoju, A.A., and Olaomi, J.O. 2012. Evaluation of small sample estimators of outliers infested simultaneous equation model: A Monte Carlo approach. *Journal of Applied Economic Sciences*, Volume VII, Spring, 1(19): 8-16. Available at: http://cesmaa.org/Docs/JAES_Spring1(19)_2012.pdf

[2] Bühlmann, P., and Van De Geer, S. 2011. *Statistics for high-dimensional data: Methods, theory and applications*. Berlin: Springer-Verlag. ISBN: 978-3-642-20192-9, 556 pp.

[3] Bujang, M., Sa'at, N., and Tg Abu Bakar Sidik, T. 2017. Determination of minimum sample size requirement for multiple linear regression and analysis of covariance based on experimental and non-experimental studies. *Epidemiology Biostatistics and Public Health*, 14(3): 1-9. DOI: https://doi.org/10.2427/12117

[4] Colquhoun, D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1(3). DOI: https://doi.org/10.1098/rsos.140216

[5] Faber, J., and Fonseca, L. 2014. How sample size influences research outcomes. *Dental Press Journal Orthod*, 19(4): 27-29. July-Aug. DOI: http://dx.doi.org/10.1590/2176-9451.19.4.027-029.ebo

[6] Forstmeier, W., Wagenmakers, E., and Parker, T. 2017. Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews*, 92(4): 1941-1968. DOI: https://doi.org/10.1111/brv.12315

[7] Holmes Finch, W., Hernandez Finch, M. 2017. Multivariate regression with small samples: A comparison of estimation methods. *General Linear Model Journal*, 43(1): 16-30. DOI: http://dx.doi.org/10.31523/glmj.043001.002

[8] Lin, M., Lucas Jr, H.C., and Shmueli, G. 2013. Research commentary - too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24(4): 883-1167. DOI: https://doi.org/10.1287/isre.2013.0480

[9] Mason, C.H., and Perreault, W.J. 1991. Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, 28(3): 268-280. DOI: https://doi.org/10.2307/3172863

[10] Misha, S.K. 2008. A new method of robust linear regression analysis: Some Monte Carlo experiments. *Journal of Applied Economic Sciences*, Volume III, Fall, 3(5): 261-268.

[11] Speed, R. 1994. Regression type techniques and small samples: A guide to good practice. *Journal of Marketing Management,* Volume X, 1(3): 89-104. DOI: https://doi.org/10.1080/0267257X.1994.9964262

*** Stata Corp. 2019. Stata lasso reference manual release 16. College Station, Texas: Stata Press Publication.

JOURNAL of Applied Economic Sciences

ISSN 2393 – 5162 ISSN - L 1843-6110