SOME EMPIRICAL RESULTS FOR THE NUMBER OF BOOTSTRAP REPLICATIONS

Lorenc Ekonomi[1], Edlira Donefski[2], Eljona Milo[3], Lorena Margo[4]

[1]Department of Mathematics, University of Korca, ALBANIA, lorencekonomi@yahoo.co.uk
[2]Department of Mathematics, University of Korca, ALBANIA, edadonefski@yahoo.com
[3]Department of Mathematics, University of Korca, ALBANIA, eljonamilo@yahoo.com
[4]Department of Mathematics, University of Korca, ALBANIA, lorena.margo@yahoo.com

Abstract

Given a set of data $X=(X_1,\ldots,X_n)$ and a statistic T(X), a key statistical question is about the distribution of T(X). The bootstrap is a general technique which gives estimates of this distribution for any X and T by substituting raw computing power for analytical expertise. The computing, a Monte Carlo calculation of an expectation, can be quite lengthy, especially in problems where T is itself a complex computation. Typical problems require 50-200 bootstrap replications to estimate a standard error and 1000-2000 replications to compute a bootstrap confidence interval. These numbers assume that the bootstrap estimation is done in the most obvious way. Various computational and probabilistic methods have been suggested to reduce the number of replications required. The classical importance sampling estimate is well-suited for variance reduction in rare event applications. It fails in many other applications. The ratio and regression estimates, well-known in sampling theory, succeed in many of these cases. In our work we have done various simulations in linear models to determine the needed number of the bootstrap replications. We have calculated bootstrap estimation for OLS estimators in linear regression when errors are homoscedastic or heteroscedastic. In this case we have used the bootstrap with residual resampling and bootstrap with vector resampling. We have concluded that the needed number of bootstrap replications is about 300-500.

Keywords: *bootstrap replication, standard error, linear regression, Monte Carlo.*

1. Introduction

Given a set of data $X=(X_1,\ldots,X_n)$ and a statistic T(X), a key statistical question is "What is the behavior (distribution) of T(X)"? The answer to, or even the ability to answer, that question often determines our choice of a statistic T. The bootstrap (Efron, (1979, 1982)) is a general technique which gives estimates of this distribution for any X and T by substituting raw computing power for analytical expertise. The computing, a Monte Carlo calculation of an expectation, can be quite lengthy, especially in problems where T is itself a complex computation. Such T is often the very ones where the bootstrap technique is most welcome, since they represent cases for which theoretical attacks are hopeless.

For those with a finite computer budget two questions immediately arise "How many Monte Carlo trials are necessary to achieve a sufficiently accurate answer?" and "Can a better accuracy/trial ratio be obtained using some modified calculation?"

Typical problems require 50-200 bootstrap replications to estimate a standard error and 1000-2000 replications to compute a bootstrap confidence interval. These numbers assume that the bootstrap estimation is done in the most obvious way. Various computational and probabilistic methods have been suggested to reduce the number of replications required. The promise of such methods is not only a reduction of the computational burden, but also a deeper understanding of the bootstrap. Various methods of improved bootstrap for reducing the number of bootstrap replications appear in Johns (1988), Graham et al. (1990), Ogbonmwan and Wynn (1986), Therneau (1983), Hesterberg (1988), Efron (1987, 1990), DiCiccio and Efron (1990). But, these methods fail in some cases in regression estimates (Therneau (1983)).

In our work we have done various simulations in linear models to determine the needed number of the bootstrap replications. We have calculated bootstrap estimation for standard errors in linear regression when errors are homoscedastic or heteroscedastic. In this case we have used the ordinary bootstrap and block bootstrap. We have determined empirically that the adequate estimation of variance or standard error in linear regression require 300-500 bootstrap replications. In Section 2 we have made the notion of bootstrap estimation of the variance and its Monte Carlo approximation. In this section we have discussed about the errors occurred in bootstrap estimation. In Section 3 we have given bootstrap with residual resampling and vector resampling in linear regression. In Section 4 we have shown the simulations results in linear regression with homoscedastic and heteroscedastic errors.

2. The bootstrap and importance sampling

Given a statistic T and a data sample $X=(X_1,\dots,X_n)$ from F, the actual variance of T(X) is

$$var_F T(X)=E_F(T-E_F T(X)^2. \qquad (1)$$

Since F is in practice unknown, this number is, of course, unobtainable. The bootstrap estimates simply replace F with the empirical distribution function $\hat{F}: \text{mass} \frac{1}{n}$ at x_i, i=1,...,n. Let $X^* = (x_1^*,..., x_n^*)$ is a data sample from this distribution. This is a bootstrap sample. The bootstrap variance of T(X) is

$$var_{BOOT} T(X) = E_{\hat{F}}\left(T\left(X^*\right) - E_{\hat{F}} T\left(X^*\right)\right)^2. \qquad (2)$$

If we have B bootstrap samples $X_1^*,..., X_B^*$, the Monte Carlo approximation for the variance is

$$\hat{v} = \frac{1}{B-1}\sum_{b=1}^{B}\left(T\left(X^{*b}\right) - T\left(X^{*(.)}\right)\right)^2, \qquad (3)$$

where $T\left(X^{*(.)}\right) = \frac{1}{B}\sum_{b=1}^{B} T\left(X^{*b}\right)$.

There are thus two sources of error in the bootstrap estimate of variance:

a) $var_{BOOT} T(X) \neq var_F T(X)$, because $\hat{F} \neq F$,

b) $\hat{v} \neq var_{BOOT} T(X)$, because $B < \infty$.

The error of the second type, which is the concern of this paper, can be very important in the bootstrap. We can, therefore find a series of efficient bootstrap techniques intended to reduce or quantify some of these errors. From a wider perspective we can find methods like: the centering method of Efron (1990), the linear bootstrap introduced by Davison et al. (1986), the control function estimates, discussed by Therneau (1988), the balanced bootstrap by Davison et al. (1986) and Graham et al. (1990), the accelerated procedures outlined by Ogbonmwan and Wynn (1986).

Other methods include the Monte Carlo device of importance sampling. Importance sampling in Monte Carlo simulation is the process of estimating a distribution using observations from a different distribution. Importance sampling has been very successful as a variance reduction technique in rare event applications. It can also be applied in many other applications, as a variance reduction technique, as a means of solving problems that are otherwise intractable, or for analyzing the performance or a physical process under multiple input distribution using a single data set observations, as in the response surface estimation or in the analysis of robust estimates. Introduced by Therneau (1988), in the context of the bootstrap estimation, it has been used by Johns (1988) in a quantile problem, by Hinkley and Shi (1989) in a double bootstrap problem, and has been widely reviewed by Hesterberg (1988).

The classical importance sampling estimate is well-suited for variance reduction in rare event applications. It fails in many other applications. The ratio and regression estimates, well-known in sampling theory, succeed in many of these cases. To avoid this problem, we have determined by empiric method the adequate number of bootstrap replication in linear regression.

3. Bootstrap methods in linear models

Bootstrap methods in linear models were first considered by Efron (1979, 1982) and then have been examined in greater depth by Freedman (1991), Freedman and Peters (1984), Hinkley (1988), Wu (1986), Moulton and Zeger (1991). These methods may be used for estimating the variability of estimators and are particularly useful in situations with small sample sizes. Freedman (1991) showed that the bootstrap approximation of the least squares estimates is valid. Let we see some aspects about bootstrapping of a linear model.

Let us take the model

$$y_i = x_i^T \beta + e_i, \ i=1,...,n, \qquad (4)$$

where x_i^T, i=1,...,n is a kx1 fixed or random vector, β is a kx1 vector of unknown parameters, e_i, i=1,...,n are errors with mean zero and variance σ^2. Writing $Y=(y_1,...,y_n)^T$, $e=(e_1,...,e_n)^T$, $X=[x_1,...,x_n]^T$, the model (1) can be written as below

$$Y=X\beta+e, \ E(e)=0, \ var(e)=\sigma^2. \qquad (5)$$

The ordinary OLS for β is given by $\hat{\beta}=\left(X^T X\right)^{-1} X^T Y$. Let us describe two methods for bootstrapping the given linear model.

3.1 Residual resampling

Resampling of residuals requires that x_i^T, i=1,...,n is a kx1 fixed vector, e_i, i=1,...,n are independent and identically random variables, so the errors are homoscedastic. Let us have the residuals vector $r = y - X\hat{\beta}$. We construct the empirical distribution function $\hat{F}: \text{mass} \frac{1}{n}$ at r_i, i=1,...,n, where r_i, i=1,...,n are the elements of the residual vector.

We draw randomly B bootstrap samples from $\hat{F}$. So, we have the vectors r_i^{*b}, b=1,...,B. We calculate $Y^{*b} = X\hat{\beta} + r^{*b}$, b=1,…,B and then obtain $\hat{\beta}^{*b} = \left(X^T X\right)^{-1} X^T Y^{*b}$, b=1,...,B. The Monte Carlo approximations for the covariance matrix of $\hat{\beta}$ is

$$\text{var}_r^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^{*b} - \hat{\beta}^{*(.)}\right)\left(\hat{\beta}^{*b} - \hat{\beta}^{*(.)}\right)^T, \qquad (6)$$

where $\hat{\beta}^{*(.)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{*b}$.

3.2 Vector resampling

Turn now to the correlation method, when the matrix X is nor fixed, but is random. Here, we find, in general some dependence between errors and the matrix X. this case is inappropriate to resample the residuals. We construct the empirical distribution function $\hat{F}: \text{mass}\frac{1}{n}$ at $\left(y_i, x_i^T\right)$, i=1,...,n. We draw randomly B bootstrap samples from $\hat{F}$ and have $\left(y_i^{*b}, x_i^{*bT}\right)$, b=1,...,B. Then $\hat{\beta}^{*b} = \left(X^{*bT} X^{*b}\right)^{-1} X^{*bT} Y^{*b}$. Then we use

$$\text{var}_{xy}^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^{*b} - \hat{\beta}^{*(.)}\right)\left(\hat{\beta}^{*b} - \hat{\beta}^{*(.)}\right)^T, \qquad (7)$$

where $\hat{\beta}^{*(.)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{*b}$ to take Monte Carlo approximation for the covariance matrix of $\hat{\beta}$.

4. Simulation results

Simulation 4.1 Let we have the model

$$y_i = x_i^T\beta + u_i = 10.0 + 0.4x_{2i} + 0.6x_{3i} + u_i,\ i=1, ..., 20, \qquad (8)$$

u_i=0.25e_i, e_i is a random variable with normal standard distribution N(0,1) and the matrix X of observations is known (Judge et al. (1988)). In this case the errors are homoscedastic. After calculations we find that the true standard deviations of OLS estimators for parameters β are respectively 0.2025, 0.0833, 0.1245. In the following tables we can see the bootstrap approximations $\text{var}_r^*, \text{var}_{xy}^*$ of standard deviations of $\hat{\beta}_i$, i=1,2,3 for different values of bootstrap replications.

B	100	500	1000	1200
β_1	0.2096	0.2058	0.2116	0.2088
β_2	0.0838	0.0871	0.0867	0.0857
β_3	0.1290	0.1270	0.1320	0.1307

Table 1. The bootstrap approximations var_r^* of OLS estimator standard deviations for different values of bootstrap replications.

B	100	500	1000	1200
β_1	0.2072	0.2265	0.2290	0.2147
β_2	0.1002	0.1078	0.1131	0.1056
β_3	0.1354	0.1361	0.1349	0.1325

Table 2. The bootstrap approximations var^*_{xy} of OLS estimator standard deviations for different values of bootstrap replications.

From these results we concluded that the adequate number of bootstrap replications in linear regression is about 300-500 bootstrap replications. In the following simulations we have done 500 replications for the bootstrap with residual resampling and 300 replications for the bootstrap with vector resampling.

To study the variability of bootstrap approximations, we calculated for 100 different simulations the quantity $\frac{1}{100}\sum_{i=1}^{100}\frac{\text{approx.boot}_i - \text{true value}}{|\text{true value}|}$. In Table 3 we see a good variability in estimations results of the OLS standard deviations.

	var^*_r	var^*_{xy}
β_1	-0.03	-0.05
β_2	-0.03	-0.01
β_3	-0.03	-0.03

Table 3. The variability of the bootstrap approximations of OLS estimator standard deviation.

In Table 4 we see the variability for the approximation bootstrap of the covariance OLS estimators matrix of unknown parameters β. The symbol (i,j) shows the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$.

	(1,1)	(1,2)	(1,3)	(2,2)	(2,3)	(3,3)
var^*_r	-0.03	0.00	-0.04	-0.04	-0.11	-0.04
var^*_{xy}	-0.04	0.03	-0.06	0.01	0.44	-0.03

Table 4. The variability of the bootstrap approximations of OLS estimator covariance matrix.

Simulation 4.2 Now, let suppose that the errors are heteroscedastic in the form $var(u_i) = 0.0625(1 + x^2_{2i} + x^2_{3i})$, i=1,...,20. In the following tables we can see the standard deviations and the covariance matrix of OLS estimators for unknown parameter β.

	var^*_r	var^*_{xy}
β_1	0.13	0.03
β_2	0.00	-0.01
β_3	-0.01	-0.02

Table 5. The variability of the bootstrap approximations of OLS estimator standard deviations.

	(1,1)	(1,2)	(1,3)	(2,2)	(2,3)	(3,3)
var_r^*	0.33	0.31	0.11	0.03	-0.20	0.02
var_{xy}^*	0.11	0.12	0.05	0.03	-0.21	0.00

Table 6. The variability of the bootstrap approximations of OLS estimation variance

Simulation 4.3 We have the model (8) and the errors are of the form $u_i = \rho u_{i-1} + e_i$, where e_i has normal distribution N(0,0.0625) and $|\rho| < 1$. In the following tables we can see the standard deviations and the covariance matrix of OLS estimators for unknown parameter β.

◦	0.999	0.99	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20
(1,1)	-0.99	-0.92	-0.69	-0.53	-0.39	-0.35	-0.32	-0.28	-0.22	-0.15	-0.08
(1,2)	-0.73	-0.62	-0.26	0.02	0.27	0.36	0.38	0.36	0.30	0.22	0.13
(1,3)	-2.80	-2.47	-1.45	-0.78	-0.23	-0.03	0.05	0.07	0.06	0.03	0.00
(2,2)	1.22	1.10	0.69	0.38	0.11	0.03	0.01	0.03	0.06	0.09	0.12
(2,3)	-1.20	-1.18	-1.13	-1.10	-1.14	-1.21	-1.29	-1.37	-1.49	-1.72	-2.43
(3,3)	2.96	2.72	1.89	1.28	0.72	0.49	0.36	0.29	0.24	0.20	0.16
◦	0.10	0.08	0.05	0.001	0.00	-0.00	-0.05	-0.10	-0.20	-0.30	-0.40
(1,1)	-0.02	-0.01	0.01	0.03	0.03	0.03	0.05	0.06	0.07	0.05	0.00
(1,2)	0.05	0.03	0.00	-0.04	-0.04	-0.04	-0.08	-0.12	-0.20	-0.27	-0.33
(1,3)	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.02	0.02	0.08	0.16
(2,2)	0.12	0.12	0.11	0.10	0.10	0.10	0.08	0.05	-0.03	-0.13	-0.24
(2,3)	-177	-5.67	-1.71	-0.35	-0.34	-0.32	0.12	0.36	0.60	0.73	0.81
(3,3)	0.11	0.09	0.07	0.04	0.04	0.04	0.00	-0.04	-0.13	-0.24	-0.35
◦	-0.50	-0.60	-0.70	-0.80	-0.90	-0.95	-0.99	-0.99			
(1,1)	-0.08	-0.17	-0.26	-0.35	-0.41	-0.48	-0.78	-0.97			
(1,2)	-0.36	-0.37	-0.32	-0.24	-0.46	-10.3	-1.22	-1.01			
(1,3)	0.26	0.36	0.44	0.51	0.59	0.68	0.90	0.99			
(2,2)	-0.37	-0.50	-0.62	-0.71	-0.77	-0.77	-0.82	-0.96			
(2,3)	0.87	0.91	0.94	0.96	0.97	0.97	0.98	0.99			
(3,3)	-0.47	-0.57	-0.66	-0.73	-0.79	-0.83	-0.93	-0.99			

Table 7. The variability of the bootstrap approximations var_r^* of OLS estimator covariance matrix.

◦	0.999	0.99	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20
(1,1)	-0.98	-0.87	-0.53	-0.32	-0.21	-0.22	-0.24	-0.23	-0.19	-0.11	-0.02
(1,2)	-2.10	-1.88	-1.16	-0.65	-0.15	0.08	0.19	0.21	0.16	0.04	-0.10
(1,3)	-4.02	-3.48	-1.87	-0.88	-0.15	0.09	0.20	0.24	0.24	0.21	0.18
(2,2)	1.84	1.70	1.18	0.78	0.44	0.35	0.37	0.46	0.58	0.69	0.76
(2,3)	-0.25	-0.33	-0.57	-0.73	-0.95	-1.27	-1.71	-2.26	-3.00	-4.19	-7.24
(3,3)	3.08	2.78	1.75	0.99	0.38	0.25	0.24	0.28	0.32	0.34	0.33
◦	0.10	0.08	0.05	0.00	0.00	-0.00	-0.05	-0.10	-0.20	-0.30	-0.40
(1,1)	0.07	0.09	0.12	0.16	0.16	0.16	0.19	0.22	0.25	0.24	0.17
(1,2)	-0.29	-0.33	-0.39	-0.49	-0.49	-0.49	-0.60	-0.70	-0.88	-1.04	-1.13

(1,3)	0.13	0.12	0.11	0.09	0.09	0.09	0.08	0.07	0.07	0.10	0.16
(2,2)	0.79	0.78	0.78	0.76	0.76	0.75	0.72	0.67	0.53	0.35	0.14
(2,3)	-712	-25.0	-9.11	-3.65	-3.59	-3.54	-1.71	-0.76	0.20	0.65	0.88
(3,3)	0.28	0.26	0.24	0.19	0.19	0.19	0.14	0.08	-0.04	-0.18	-0.31
°	-0.50	-0.60	-0.70	-0.80	-0.90	-0.95	-0.99	-0.99			
(1,1)	0.06	-0.08	-0.21	-0.31	-0.36	-0.41	-0.74	-0.96			
(1,2)	-1.15	-1.08	-0.93	-0.76	-1.03	-14.5	-1.31	-1.02			
(1,3)	0.24	0.33	0.41	0.47	0.53	0.63	0.88	0.98			
(2,2)	-0.08	-0.29	-0.48	-0.62	-0.68	-0.68	-0.76	-0.95			
(2,3)	1.00	1.05	1.07	1.06	1.02	1.00	0.99	0.99			
(3,3)	-0.44	-0.56	-0.66	-0.73	-0.77	-0.80	-0.92	-0.99			

Table 8. The variability of the bootstrap approximations var^{*}_{xy} of OLS estimator covariance matrix.

From Tables 7 and 8 we see bad estimations for the covariance between $\hat{\beta}_2$ and $\hat{\beta}_3$. This happened because the true value of this parameter is very small in absolute value.

5. Reference

1. Davison, A.C., Hinkley, D.V., Schechtman, E. (1986). Efficient bootstrap simulation. *Biometrika* 73, 555-566.
2. DiCiccio, T. and Efron, B. (1992). More accurate confidence intervals in exponential families. *Biometrika* 79, 231-245.
3. Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1-26
4. Efron, B. (1982). The jackknife, the bootstrap and other sesampling plans. *CBMS-NSF Regional Conference Series in Applied Mathematics 38*. Society for Industrial and Applied Mathematics (SIAM) Philadelphia.
5. Efron, B. (1987). Better bootstrap confidence intervals. *JASA* 82, 171-185.
6. Efron, B. (1990). More efficient bootstrap computations. *JASA 85*, 79-89.
7. Efron, B. and Gong, G. (1993). A leisury look at the bootstrap, the jackknife and cross-validation. *Amer. Statist*. 37, 36-48
8. Freedman, D.A. (1991). Bootstrapping regression models. *Ann. Statist.* 9, 1218-1228.
9. Freedman, D.A. and Peters, S.C. (1984). Bootstrapping a regression equation: some empirical results. *J. Amer. Statist. Assoc*. 79, 97-106.
10. Graham, R.L., Hinkley, D.V., John, P.W.M., and Shi, S. (1990). Balanced design of bootstrap simulations. *J. R. Statist.. Soc. B* 52, 185-202.
11. Hesterberg, T. C. (1988). Advances in importance sampling. *PhD Thesis*. Stanford University.
12. Hinkley, D.V. (1988). Bootstrap methods. *J. Roy. Statist. Ser*. B 50, 321-337.
13. Hinkley, D.V. and Shi, S. (1989). Importance sampling and the nested bootstrap. *Biometrika* 76, 435-446.
14. Johns, M.V. (1988). Importance sampling for bootstrap confidence intervals. *JASA* 83, 709-714.
15. Judge, G.G., Carter Hill, R., Griffiths, W.E., Lutkepohl, H., Lee, T. (1988). *Introduction to the Theory and Practice of Econometrics* (Second ed.), John Wiley&Sons, New York.

16. Moulton, H.L., Zeger, L.S. (1991). Bootstrapping generalized linear models. *Comput. Data Anal.* 11, 53-63.
17. Ogbonmwan, S.M., and Wynn, H.P. (1988). Resampling generated likelihoods. *In Statistical Decision Theory and Related Topics IV*. Editors S. Gupta and J. Berger, New York. Springer-Verlag, 137-147.
18. Therneau, T. (1988). Variance reduction techniques for the bootstrap. *PhD Thesis*. Stanford University.
19. Wu C.F.J. (1986) Jackknife, bootstrap and other resampling plans methods in regression analyses. *Ann. Statist.* 14, 1261-1295.