Constructing statistical inference about determination coefficient R^2 in regression analysis by using bootstrap

Dr. Mahmoud Farouk El-Said

Abstract

In this paper, the bootstrap method is used to consider the estimate of the standard errors of R^2 and $\overline{R}^2$ which are measures of their precision. It is shown that the bootstrap estimation of standard errors are accurate estimates of the exact ones.

1. Introduction

In regression analysis, the coefficient of determination R^2 and the adjusted coefficient of determination $\overline{R}^2$ are usually referred to as measures of goodness of fit. Thus, the sampling properties of R^2 and $\overline{R}^2$ have been examined by many researchers. The reason is that since the distributions of R^2 and $\overline{R}^2$ are complex and depend on unknown parameters, it is difficult to calculate precision. When the measure of precision is not presented, we cannot say that since a value of R^2 is large, a value of a parent coefficient of determination ϕ is also large. When ϕ is not substantially large than zero, the model is poorly specified.

In this paper, using the bootstrap method, we consider the estimate of the standard errors of R^2 and $\overline{R}^2$ which are measures of their precision. In Section 2 the model and estimators are presented, and in Section 3 the bootstrap procedure is shown. In Section 4 we carry out some experiments to examine the sampling performances of the bootstrap estimates of R^2 and $\overline{R}^2$.

It is shown that the bootstrap standard errors are considerably accurate estimates of the exact ones.

2. Model and estimators

Consider the linear regression model,

$$\mathrm{y} = \ell\alpha + \mathrm{x}\beta + \varepsilon \ , \quad \text{where } \varepsilon \sim N(0, \sigma^2 I_n) \quad .....(1)$$

Where:

I_n : identity matrix of order *n*.

The rank of matrix X is (k-1) of observations on non-stochastic independent variables, α is an intercept term , β is $(\mathrm{k}-1)\times 1$ vector of regression coefficients, and ε is an $(\mathrm{n}\times 1)$ vector of normal error terms. We assume that all the independent variables are measured as deviations from their sample means.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} \qquad \underset{(n\times 1)}{\ell} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_{\mathrm{k-1}} \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{x}=\begin{pmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1(k-1)} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2(k-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & .... & .... & x_{n(k-1)} \end{pmatrix} \qquad \overline{\mathrm{X}}=\begin{pmatrix} \overline{\mathrm{X}}_1 \\ \overline{\mathrm{X}}_2 \\ \vdots \\ \vdots \\ \overline{\mathrm{X}}_{k-1} \end{pmatrix}$$

$$x_{ij} = \mathrm{X}_{\mathrm{ij}} - \overline{\mathrm{X}}_j \quad \text{for i =1,…,n and j =1,…..,(k-1)}$$

$$\mathrm{E}(\varepsilon) = \mathrm{o} \quad , \qquad \mathrm{E}(\varepsilon\varepsilon') = \sigma^2 I_n$$

Since $\mathbf{x}'\ell = o$, the ordinary least square (OLS) estimators of α and β are *a* and *b* as follows :

$$a = \ell' \mathbf{y} / n \qquad \text{..... (2)}$$

$$b = \mathrm{S}^{-1}\mathbf{x}'\mathbf{y} \qquad \text{.....(3)}$$

where $\mathrm{S} = X'X$. Let e be an $n \times 1$ vector of the OLS residuals (i.e. $e = y - l\bar{y} - Xb$). Then, using $y'y = b'Sb + e'e$, the sample coefficient of determination is written as

$$R^2 = 1 - \frac{e'e}{y'y - n\bar{y}^2} \qquad \text{..... (4)}$$

Also, the adjusted R^2 is

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}(1-R^2) \quad \text{..... (5)}$$

If we define a formally general estimator as

$$R_h^2 = hR^2 + (1-h) \quad \text{..... (6)}$$

where $h \geq 1$, then R_h^2 reduces to R^2 when $h = 1$, and to $\overline{R}^2$ when $h = (n-1)/(n-k)$. As is shown in [10], the distribution function of R_h^2 is

$$F(c) = P(R_h^2 < c) = \sum_{i=0}^{\infty} w_i(\lambda) I_{c^*}\left(\tfrac{k-1}{2} + i, \tfrac{n-k}{2}\right) \quad \text{..... (7)}$$

where $w_i(\lambda) = \exp(-\lambda/2)(\lambda/2)^i / i!$, $\lambda = \beta' S \beta / \sigma^2$, $c^* = (c + h - 1)/h$, and $I_x(a,b)$ is the incomplete beta function ratio. Also, the formula for the m^{th} moment of R_h^2 is

$$E\left[(R_h^2)^m\right] = \sum_{r=0}^{m} m c_r h^r (1-h)^{m-r} \sum_{i=0}^{\infty} w_i(\lambda) \times \frac{\Gamma[(n-1)/2+i]\Gamma[(k-1)/2+r+i]}{\Gamma[(n-1)/2+r+i]\Gamma[(k-1)/2+i]} \quad \text{..... (8)}$$

As stated above, the formulas for the distribution functions and the moments of R^2 and $\overline{R}^2$ are obtained by putting $h = 1$ and $h = (n-1)/(n-k)$ in Eqs. (7) and (8) , respectively.

Since the formulas for the moments and the distribution functions of R^2 and $\overline{R}^2$ include unknown parameter (i.e. $\lambda = \beta' S \beta / \sigma^2$) , we cannot calculate the moments of R^2 and $\overline{R}^2$ numerically in practical situations. However, if we use the

bootstrap method, we can evaluate them numerically. In the next section, we consider the bootstrap procedure to estimate the moments of R^2 and $\bar{R}^2$.

3. Algorithms of bootstrap

The Algorithms of bootstrap is as follows.

1. Given the data on y and X, we first calculate the OLS estimates of α and β: $a = \bar{y}$ and $b = S^{-1}X'y$. Using *a* and *b*, we calculate the residual vector: $e = y - la - Xb$

2. The case of the parametric bootstrap: using *e*, we obtain the OLS estimate of $\sigma^2 : \hat{\sigma}^2 = e'e/(n-k)$. Generating n random numbers, e_1^*, e_2^*, . . . e_n^*, from $N(la + Xb, \hat{\sigma}^2 I_n)$, we obtain the bootstrap sample of $y : y^* = la + Xb + e^*$

where $e^* = (e_1^*, e_2^*, ...e_n^*)^t$. Then, the bootstrap estimates of α and β are obtained : $a^* = \bar{y}^*$ and $b^* = S^{-1}X'y^*$, where $\bar{y}^*$ is the sample mean of y^*. Also, the bootstrap residual vector is $e_B^* = y^* - la^* + Xb^*$.

3. Using the bootstrap sample of y and residuals, the bootstrap estimates of R^2 and $\bar{R}^2$ are obtained:

$$R^{*2} = 1 - \frac{e_B^{*'} e_B^*}{y^{*'} y^* - n\bar{y}^{*2}} \quad \text{..... (9)}$$

$$\bar{R}^{*2} = 1 - \frac{n-1}{n-k}(1 - R^{*2}) \quad \text{..... (10)}$$

4. Repeating Steps 2. and 3. *B* times, we obtain the *B* estimates of R^2 and $\bar{R}^2$. Denoting the bootstrap estimates of R^2

and $\overline{R}^2$ obtained in the i^{th} iteration as $R^{*2}_{(i)}$ and $\overline{R}^{*2}_{(i)}$, the bootstrap estimates of R^2 and $\overline{R}^2$ are $\hat{R}^{*2} = \frac{1}{B}\sum_{i=1}^{B} R^{*\,2}_{(i)}$, $\hat{\overline{R}}^{*2} = \frac{1}{B}\sum_{i=1}^{B} \overline{R}^{*\,2}_{(i)}$,

and their standard errors are

$$Se(\hat{R}^{*2}) = \left[\frac{1}{B-1}\sum_{i=1}^{B}\left(R^{*\,2}_{(i)} - \hat{R}^{*2}\right)^2\right]^{1/2} \quad \text{..... (11)}$$

$$Se(\hat{\overline{R}}^{*2}) = \left[\frac{1}{B-1}\sum_{i=1}^{B}\left(\overline{R}^{*\,2}_{(i)} - \hat{\overline{R}}^{*2}\right)^2\right]^{1/2} \quad \text{..... (12)}$$

4. Results of applying bootstrap

Table (4-1) Mean and standard error of R^2 by the bootstrap

k	n	Φ	μ (exact)	σ (exact)	μ (parametric)	σ (parametric)
3	30	0.1470788	0.1456158	0.0083246	0.1729385	0.0008901
3	30	0.3159046	0.3144416	0.0210435	0.3417643	0.0136090
3	30	0.4847303	0.4832673	0.0752869	0.5105900	0.0678524
3	30	0.6535561	0.6520931	0.2202477	0.6794158	0.2128132
3	30	0.8223818	0.8209188	0.5246154	0.8482416	0.5171809
3	50	0.1495100	0.1480500	0.0083800	0.1753700	0.0009400
3	50	0.3207600	0.3193000	0.0218300	0.3466200	0.0144000
3	50	0.4920200	0.4905600	0.0792500	0.5178800	0.0718100
3	50	0.6632700	0.6618100	0.2326900	0.6891300	0.2252600
3	50	0.8345300	0.8330700	0.5548900	0.8603900	0.5474600
3	80	0.1508960	0.1494330	0.0084060	0.1767560	0.0009710
3	80	0.3235390	0.3220760	0.0223020	0.3493990	0.0148670
3	80	0.4961820	0.4947190	0.0815850	0.5220420	0.0741500
3	80	0.6688250	0.6673620	0.2400400	0.6946850	0.2326050
3	80	0.8414680	0.8400050	0.5727720	0.8673280	0.5653370

Table (4-2) Mean and standard error of R^2 by the bootstrap

k	n	Φ	μ (exact)	σ (exact)	μ (parametric)	σ (parametric)
6	30	0.1716070	0.1701440	0.0089480	0.1974670	0.0015140
6	30	0.3649610	0.3634980	0.0307140	0.3908210	0.0232790
6	30	0.5583150	0.5568520	0.1237230	0.5841750	0.1162890
6	30	0.7516690	0.7502060	0.3725160	0.7775290	0.3650820
6	30	0.9450230	0.9435600	0.8951750	0.9708830	0.8877410
6	50	0.1742160	0.1727530	0.0090300	0.2000760	0.0015960
6	50	0.3701800	0.3687170	0.0319830	0.3960400	0.0245480
6	50	0.5661430	0.5646800	0.1300860	0.5920030	0.1226510
6	50	0.7621070	0.7606440	0.3925240	0.7879660	0.3850900
6	50	0.9580700	0.9566070	0.9438770	0.9839300	0.9364420
6	80	0.1757100	0.1742470	0.0090780	0.2015700	0.0016440
6	80	0.3731670	0.3717040	0.0327320	0.3990270	0.0252980
6	80	0.5706250	0.5691620	0.1338430	0.5964840	0.1264090
6	80	0.7680820	0.7666190	0.4043410	0.7939410	0.3969070
6	80	0.9655390	0.9640760	0.9726410	0.9913990	0.9652070

When n = 30, 50, 80 and Φ = 0.147, 0.315, 0.484, 0.653, 0.822. The experiments are executed on a personal computer, using the statistical package MATHCAD.

As stated above, the number of replications in the experiments is 1000. The results for k = 3 and 6 are shown in Tables (4-1) and (4-2). Tables (4-1) and (4-2) show the parametric bootstrap means of R^2 (denoted as " μ ") and the

standard errors (denoted as " σ ") , where 'Exact' shows that the mean and standard error are calculated based on the exact formula for the moment given in Eq.(8).

Table (4-3) Mean and standard error of $\overline{R}^2$ by the bootstrap

k	n	Φ	μ (exact)	σ (exact)	μ (parametric)	σ (parametric)
3	30	0.1370790	0.1356160	0.0073250	0.1728380	0.0008900
3	30	0.3059050	0.3044420	0.0200430	0.3416640	0.0136090
3	30	0.4747300	0.4732670	0.0742870	0.5104900	0.0678520
3	30	0.6435560	0.6420930	0.2192480	0.6793160	0.2128130
3	30	0.8123820	0.8109190	0.5236150	0.8481420	0.5171810
3	50	0.1395080	0.1380450	0.0073760	0.1752680	0.0009410
3	50	0.3107640	0.3093010	0.0208350	0.3465230	0.0144000
3	50	0.4820190	0.4805560	0.0782470	0.5177790	0.0718120
3	50	0.6532740	0.6518110	0.2316920	0.6890340	0.2252570
3	50	0.8245300	0.8230670	0.5538920	0.8602890	0.5474580
3	80	0.1408960	0.1394330	0.0074060	0.1766560	0.0009710
3	80	0.3135390	0.3120760	0.0213020	0.3492990	0.0148670
3	80	0.4861820	0.4847190	0.0805850	0.5219420	0.0741500
3	80	0.6588250	0.6573620	0.2390400	0.6945850	0.2326050
3	80	0.8314680	0.8300050	0.5717720	0.8672280	0.5653370

Table (4-4) Mean and standard error of $\overline{R}^2$ by the bootstrap

k	n	Φ	μ (exact)	σ (exact)	μ (parametric)	σ (parametric)
6	30	0.1616070	0.1601440	0.0079480	0.1973670	0.0015140
6	30	0.3549610	0.3534980	0.0297140	0.3907210	0.0232790
6	30	0.5483150	0.5468520	0.1227230	0.5840750	0.1162890
6	30	0.7416690	0.7402060	0.3715160	0.7774290	0.3650820
6	30	0.9350230	0.9335600	0.8941750	0.9707830	0.8877410
6	50	0.1642160	0.1627530	0.0080300	0.1999760	0.0015960
6	50	0.3601800	0.3587170	0.0309830	0.3959400	0.0245480
6	50	0.5561430	0.5546800	0.1290860	0.5919030	0.1226510
6	50	0.7521070	0.7506440	0.3915240	0.7878660	0.3850900
6	50	0.9480700	0.9466070	0.9428770	0.9838300	0.9364420
6	80	0.1657100	0.1642500	0.0080800	0.2014700	0.0016400
6	80	0.3631700	0.3617000	0.0317300	0.3989300	0.0253000
6	80	0.5606200	0.5591600	0.1328400	0.5963800	0.1264100
6	80	0.7580800	0.7566200	0.4033400	0.7938400	0.3969100
6	80	0.9555400	0.9540800	0.9716400	0.9913000	0.9652100

When n = 30, 50, 80 and Φ = 0.137, 0.305, 0.474, 0.643, 0.812. The experiments are executed on a personal computer, using the statistical package MATHCAD.

As stated above, the number of replications in the experiments is 1000. The results for k = 3 and 6 are shown in Tables (4-3) and (4-4). Tables (4-3) and (4-4) show the parametric bootstrap means of R^2 (denoted as " μ ") and the standard errors (denoted as " σ ") , where ‘Exact’ shows that the

mean and standard error are calculated based on the exact formula for the moment given in Eq.(8).

5. Recommendations

a) we can apply non parametric bootstrapping method to estimate determination coefficients R^2 and $\overline{R}^2$ for multiple regression.

b) We can , too , apply bootstrapping method to estimate determination coefficients R^2 and $\overline{R}^2$ for seemingly unrelated regression model**.**

References

1. Adkins, L.C., 1992. Finite sample moments of a bootstrap estimator of the James-Stein rule. Econometric Rev. 11, 173-193.
2. Brownstone, D., 1990. Bootstrapping improved estimators for linear regression models. J. Econometrics 44, 171-187.
3. Carrodus, M.L., Giles, D.E.A., 1992. The exact distribution of R2 when regression disturbances are auto correlated. Econ. Lett. 38, 375-380.
4. Chi, X.E., Judge, G.G., 1985. On assessing the precision of Stein's estimator. Econ. Lett. 18, 143-148.
5. Cramer, J.S., 1987. Mean and variance of R^2 in small and moderate samples. J. Econometrics 35, 253-266.
6. Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1-26.
7. Helland, I.S., 1987. On the interpretation and use of R^2 in regression analysis. Biometrics 43, 61-69.
8. James, W., Stein, C., 1961. Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press, Berkeley, pp. 361-379.
9. Meepagala, G., 1992. The small sample properties of R^2 in a misspecified regression model with stochastic regressors. Econ. Lett. 40, 1- 6.
10. Ohtani, K., 1994. The density functions of R^2 and R^2, and their risk performance under asymmetric loss in misspecified linear regression models. Econ. Modell. 11, 463-471.

11. Ohtani, K., Hasegawa, H., 1993. On small sample properties of R^2 in a linear regression model with multivariate t errors and proxy variables. Econometric Theory 9, 504-515.

12. Press, S.J., Zellner, A., 1978. Posterior distribution for the multiple correlation coefficient with fixed regressors. J. Econometrics 8, 307-321.

13. Rencher, A.C., Pun, F.C., 1980. Inflation of R^2 in best subset regression. Technometrics 22, 49-53.

14. Srivastava, A.K., Ullah, A., 1995. The coefficient of determination and its adjusted version in linear regression models. Econometric Rev. 14, 229-240.

15. Wu, C.F.J., 1986. Jackknife, bootstrap and other resembling methods in regression analysis. Ann. Stat. 14, 1261-1295.

16. Yi, G., 1991. Estimating the variability of the Stein estimator by bootstrap. Econ. Lett. 37, 293-298.

الملخص

في هذا البحث يقوم الباحث بتقدير تقريبي للمتوسطات والأخطاء المعيارية لمعامل التحديد ومعامل التحديد المصحح وذلك باستخدام طريقة البوتستراب والتي تعتبر أحد طرق إعادة المعاينة ، وقد تبين أن المتوسطات والأخطاء المعيارية المقدرة باستخدام طريقة البوتستراب هي تقديرات دقيقة إذا ما قورنت بالمتوسطات والأخطاء المعيارية الناتجة عن استخدام طرق أخرى.