Biometrika (1988), **75**, 2, *pp.* 370–3
Printed in Great Britain

A note on bootstrapping the correlation coefficient

By G. A. YOUNG
Statistical Laboratory, University of Cambridge, Cambridge, CB2 1SB, U.K.

Summary

Smoothed bootstrap estimation of the sampling standard deviation of the variance-stabilized correlation coefficient is reconsidered. An approximation to the mean squared error of the bootstrap estimator is obtained and an empirical procedure for choosing the degree of smoothing in the bootstrap estimation is presented. Performance of the procedure is examined in a simulation study.

Some key words: Bootstrap; Computer algebra; Correlation coefficient.

1. Introduction

Silverman & Young (1987) discuss use of the smoothed bootstrap for estimation of the sampling standard deviation of the variance-stabilized correlation coefficient in bivariate samples and examine, from a theoretical viewpoint, circumstances when some smoothing is advantageous in the bootstrap estimation. In the present paper the analytical tools used in that investigation are applied to estimate the mean squared error of the bootstrap estimator and to define a procedure for empirical choice of the degree of smoothing. The procedure is illustrated in a small simulation study.

2. Bootstrapping the correlation coefficient

Use of the bootstrap in estimation of the sampling properties of the correlation coefficient has been discussed previously by, for example, Efron (1982), Dolker, Halperin & Divgi (1982), Schluchter & Forsythe (1986) and Silverman & Young (1987). When interest is in comparing the correlation parameters for a number of independent populations, it is important to be able to attach a standard error to a point estimate of such a parameter. The bootstrap and smoothed bootstrap are convenient tools for such standard error estimation. The smoothed bootstrap may be a substantially more accurate estimation procedure than the standard, unsmoothed, bootstrap; see Efron (1982, Table 5.2). The aim is to have some means of choosing, from the sample data itself, an appropriate degree of smoothing.

Details of the smoothed bootstrap procedure based on a kernel density estimator with the same variance structure as the sample data are given by Silverman & Young (1987). Given data with variance matrix V, the procedure requires choice of a symmetric probability density function K, with unit variance matrix, as kernel and a nonnegative smoothing parameter h. The case $h = 0{\cdot}0$ corresponds to the standard bootstrap.

Following Silverman & Young (1987), let F_0 be a bivariate distribution with mean zero and correlation ρ. Let r be the correlation coefficient based on a sample of n independent observations from F_0 and let $z = \tanh^{-1} r$: the quantity to be estimated is $\alpha_n(F_0) = \{\mathrm{var}_{F_0}(z)\}^{\frac{1}{2}}$. Let

$$\alpha(F_0) = \left[\frac{\rho^2}{(1-\rho^2)^2}\left\{\frac{\mu_{22}}{\mu_{11}^2} + \frac{1}{4}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}}\right) - \left(\frac{\mu_{31}}{\mu_{11}\mu_{20}} + \frac{\mu_{13}}{\mu_{11}\mu_{02}}\right)\right\}\right]^{\frac{1}{2}},$$

where $\mu_{ij} = \int x_1^i x_2^j \, dF_0(x)$. As regards choice of degree of smoothing, estimation of $\alpha_n(F_0)$ is approximately equivalent to that of $\alpha(F_0)$, since $\alpha_n(F_0) = n^{-\frac{1}{2}}\alpha(F_0) + O(n^{-3/2})$.

Let $X = (X_1, X_2)$ be a random vector with distribution F_0 and set

$$S = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{\frac{1}{2}},$$

where $\sigma_1^2 = \mathrm{var}_{F_0}(X_1)$, $\sigma_2^2 = \mathrm{var}_{F_0}(X_2)$ and $\rho = \mathrm{corr}_{F_0}(X_1, X_2)$. Consider the transformation $Y = (Y_1, Y_2) = S^{-1}X$ and let $m_{ij} = E_{F^*}(Y_1^i Y_2^j)$, where F^* denotes the distribution of Y induced by that of X. Define the function $a_S(u)$ by

$$a_S(u) = \beta_0\{u_1^2u_2^2 + \tfrac{1}{2}\rho(u_1^2 - u_2^2)(m_{31} - m_{13}) - m_{22}(u_1^2 + u_2^2) - (m_{13} + m_{31})u_1u_2\}, \qquad (2{\cdot}1)$$

with $\beta_0 = \{2\alpha(F_0)\}^{-1}$.

The mean squared error of the smoothed bootstrap estimator of $\alpha(F_0)$ is, for any choice of smoothing parameter h, to $O_p(n^{-1})$, the same as that of the linear estimator

$$\hat{A}(F_0) = n^{-1} \sum_{i=1}^{n} w^*(X_i) \qquad (2{\cdot}2)$$

of $A(F_0) = \int a(t)\, dF_0(t)$. Here X_i denotes the ith data point and

$$w^*(x) = \int a\{(1+h^2)^{-\frac{1}{2}}(x + hV^{\frac{1}{2}}\xi)\}K(\xi)\, d\xi,$$

with $a(Su) = a_S(u)$. Assume henceforth that V is the fixed variance matrix of F_0.

Using (2·2), the mean squared error of $\hat{A}(F_0)$ may, after the transformation, be expressed as

$$\textsc{mse}\{\hat{A}(F_0)\} = [E_{F^*}\{w^{**}(Y)\} - A(F_0)]^2 + n^{-1}\,\mathrm{var}_{F^*}\{w^{**}(Y)\}, \qquad (2{\cdot}3)$$

where

$$w^{**}(y) = \int a_S\{(1+h^2)^{-\frac{1}{2}}(y + hS^{-1}V^{\frac{1}{2}}\xi)\}K(\xi)\, d\xi.$$

It is easily shown from (2·1) that for the Gaussian kernel K

$$\begin{aligned} w^{**}(y) = \beta_0\{&c_1^4y_1^2y_2^2 + c_1^2c_2^2y_1^2 + c_1^2c_2^2y_2^2 + c_2^4 - m_{22}c_1^2(y_1^2 + y_2^2) \\ &+ \tfrac{1}{2}\rho(m_{31} - m_{13})c_1^2(y_1^2 - y_2^2) - 2m_{22}c_2^2 - (m_{13} + m_{31})c_1^2y_1y_2\}, \qquad (2{\cdot}4) \end{aligned}$$

where $c_1 = (1+h^2)^{-\frac{1}{2}}$ and $c_2 = hc_1$.

Use of the computer algebraic manipulation package REDUCE, combining of (2·3) and (2·4), shows that, apart from the multiplicative factor β_0^2, not depending on h, the mean squared error of $\hat{A}(F_0)$ has the form

$$\begin{aligned} \textsc{mse}\{\hat{A}(F_0)\} = \tfrac{1}{4}n^{-1}(1+h^2)^{-4}[&\gamma_0 + 2(\gamma_0 + \gamma_2 + \gamma_3)h^2 + \{\gamma_0 + 4(1 - n^{-1})\gamma_1 + \gamma_2 + 2\gamma_3\}h^4 \\ &+ 4\gamma_1h^6 + \gamma_1h^8], \qquad (2{\cdot}5) \end{aligned}$$

where

$$\begin{aligned} \gamma_0 = &-4\rho(m_{13}^3 + m_{31}^3) + \rho^2(m_{04} + m_{40})(m_{13} - m_{31})^2 + 2m_{22}(6 - \rho^2)(m_{13}^2 + m_{31}^2) \\ &+ 4\rho(m_{13}^2m_{31} + m_{31}^2m_{13}) + 4\rho(m_{13} - m_{31})(m_{40} - m_{04})(m_{22} - 1) \\ &- 8m_{33}(m_{13} + m_{31}) + 4(\rho^2 + 6)m_{22}m_{13}m_{31} + 4m_{22}^2(2m_{22} - 1) \\ &+ 4m_{22}(m_{22} - 2)(m_{04} + m_{40}) + 4m_{04}m_{40}, \end{aligned}$$

$$\gamma_1 = 4n(m_{22} - 1)^2, \quad \gamma_2 = 4(m_{04} + m_{40} - m_{04}m_{40}) - 12m_{22}^2 + 8m_{22},$$

$$\gamma_3 = 4m_{33}(m_{13} + m_{31}) - 4(m_{13} + m_{31})^2.$$

3. Empirical smoothing procedure

Now suppose data $X_1, \ldots, X_n$, from some completely unspecified distribution F_0, are given. Let $\bar{X}$ be the mean of the X_i and let $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, $\hat{\rho}$ be estimates of the marginal variances and correlation of F_0.

Set

$$\hat{S} = \begin{bmatrix} \hat{\sigma}_1 & 0 \\ 0 & \hat{\sigma}_2 \end{bmatrix} \begin{bmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{bmatrix}^{\frac{1}{2}}$$

and let $Y_i = (Y_{1i}, Y_{2i}) = \hat{S}^{-1}(X_i - \bar{X})$ $(i = 1, \ldots, n)$. Compute the sample moments $\hat{m}_{ij} = n^{-1} \Sigma\, Y_{1m}^i Y_{2m}^j$. Substitution of these moments, and $\hat{\rho}$, into (2·5) yields a function of the smoothing parameter h which serves as an estimate of the mean squared error of the bootstrap estimator. The proposed procedure is that the choice of h for the bootstrap estimation itself should minimize this estimated mean squared error. The minimization is easily performed numerically: the REDUCE package will automatically produce the FORTRAN code for evaluation of the mean squared error function.

4. Simulation

Performance of the procedure for choice of smoothing was studied for data sets of three sizes, $n = 10, 20, 50$, generated from three bivariate distributions: Gaussian, log normal and t. Random observations from the latter two distributions can be generated via the first. If $Y = (Y_1, Y_2) \sim N_2(\mu, \Omega)$ and $u_1 = e^{Y_1}$, $u_2 = e^{Y_2}$, then $u = (u_1, u_2)$ has a bivariate log-normal distribution. If $Y \sim N_2(\mu, \Omega)$ and $W \sim \chi_k^2$, independently of Y, and $u_1 = Y_1/(W/k)^{\frac{1}{2}}$, $u_2 = Y_2/(W/k)^{\frac{1}{2}}$, then u has a bivariate t distribution on k degrees of freedom. All simulations were based on the bivariate Gaussian distribution of mean zero, unit variances and correlation $\frac{1}{2}$. The NAG subroutine library was used to generate univariate variables and a linear transformation method used to produce bivariate Gaussian samples. The study included the t distribution on (i) 10 degrees of freedom, and (ii) 3 degrees of freedom: in the former but not the latter case all the moments of (2·5) exist. Table 1 shows, for each combination of distribution and sample size, the root mean squared error of the estimators of $\alpha_n(F_0)$ obtained over 500 replications, for both the standard bootstrap and the smoothed bootstrap with automatic choice of h. All bootstrap estimators were calculated on the basis of 200 resamples. All figures in Table 1 have standard errors in the range 0·0002 to 0·0008.

In this study, where choice of h was by numerical minimization, using a NAG routine, of the estimated mean squared error function, automatic smoothing added about 8% to the computational load in calculating the bootstrap estimator for sample size 10. For sample size 50, the increase in computational load is only some 2%.

In small bivariate samples computational degeneracies of the type discussed by Dolker et al. (1982) and Schluchter & Forsythe (1986) may occur, when standard bootstrap samples have 2 or fewer distinct points. For small sample sizes, the smoothed bootstrap will prevent such

Table 1. *Estimates of root mean squared error of bootstrap estimates of $\alpha_n(F_0)$; sample size, n; standard and automatic choice of smoothing parameter h*

		Distribution			
				t	t
n	h	Gaussian	Log normal	(10 d.f.)	(3 d.f.)
10	Standard	0·0927	0·1397	0·1092	0·1639
	Automatic	0·0336	0·1468	0·0548	0·1636
20	Standard	0·0467	0·1156	0·0603	0·1436
	Automatic	0·0252	0·1147	0·0399	0·1458
50	Standard	0·0198	0·0845	0·0322	0·1179
	Automatic	0·0141	0·0882	0·0278	0·1241

degeneracies. However, since our purpose is to compare the standard and smoothed bootstrap procedures in circumstances where they are genuine competitors, we take sample size 10 as the smallest where no smoothing is needed.

5. Discussion

When using the smoothed bootstrap to estimate the standard deviation of the transformed correlation coefficient, the optimal degree of smoothing will depend on the underlying distribution and it is important to have some empirical procedure for choosing the amount of smoothing applied. The linear estimator (2·2) and computer algebra yield a closed form approximation to the mean squared error of the smoothed bootstrap estimator, valid for all h. This expression furnishes a data-driven procedure for choosing the degree of smoothing in general bivariate samples which costs little computationally compared to the bootstrap estimation itself.

In the situation where the underlying distribution is Gaussian or nearly Gaussian, use of the procedure can lead to substantially more accurate estimation, particularly for smaller sample sizes. In these cases, the improvements due to smoothing decrease with the sample size. In the situation where the underlying distribution is far from Gaussian, smoothing is less effective and may indeed lead to less accurate estimation.

Detailed study of the simulation results shows that the smoothed bootstrap works by reducing the variance of the estimator at the expense of increasing the bias. Smoothing is only really effective in circumstances where this variance dominates the squared bias of the bootstrap estimator, as in the Gaussian case for smaller sample sizes. When the bias is large, as in the cases of the log normal distribution and t distribution on 3 degrees of freedom, smoothing has little effect on the bootstrap estimation. If the bootstrap is to be applied to data samples which look very non-Gaussian, it would presumably be advisable to choose the degree of smoothing with reference to some fitted parametric family of distributions, though the question arises as to whether we would really be interested in correlation parameters in such circumstances.

References

Dolker, M., Halperin, S. & Divgi, D. R. (1982). Problems with bootstrapping Pearson correlations in very small bivariate samples. *Psychometrika* **47**, 529-30.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: SIAM.

Schluchter, M. D. & Forsythe, A. B. (1986). A caveat on the use of a revised bootstrap algorithm. *Psychometrika* **51**, 603-5.

Silverman, B. W. & Young, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika* **74**, 469-79.

[*Received July* 1987. *Revised October* 1987]