Explaining Bootstraps and Robustness

Tony Lancaster,
Brown University, Providence RI 02912[1]

Abstract

In this note we consider several versions of the bootstrap and argue that it can be helpful in explaining and thinking about such procedures to use an explicit representation of the random resampling process. To illustrate the point we give such explicit representations and use them to produce some results about bootstrapping linear models that are, apparently, not widely known, at least in the econometric literature. Among these are a demonstration of the equivalence, to order n^{-1} of the covariance matrix of the bootstrap distribution of the least squares estimator and the Eicker(1967)/White(1980) heteroscedasticity robust covariance matrix estimate. The method also shows the precise relations between an Efron(1979) bootstrap procedure and the Bayesian bootstrap of Rubin(1981)

KeyWords
heteroscedasticity; Bayes; Least Squares

1 INTRODUCTION

The bootstrap is usually explained algorithmically, as a set of computational instructions. (This description seems to apply to books, e.g. Efron and Tibshirani(1993); survey articles, e.g. Horowitz(2001); and to textbooks, e.g. Wooldridge(2002).) In the case of the Efron nonparametric bootstrap the algorithm would be something like

1. Randomly sample your data, with replacement, n times.
2. Compute the statistic of interest using your new data
3. Repeat steps 1 and 2 B times
4. Calculate the standard deviation of the B values of the statistic.

Justification for the algorithm would then be provided by explaining that the empirical distribution of the data, say F_n is an approximation to the true but unknown distribution F and that repeated sampling from F_n is approximately the same as repeated sampling from F which, in turn, is what is required to

[1]Tony Lancaster is Professor of Economics at Brown University, Providence RI 02912 (email: Tony_Lancaster@brown.edu)

calculate a repeated sampling distribution. Further justification would be provided by study of the exact and asymptotic properties of the random variable calculated according to 4.

This way of explaining the bootstrap is different from the way in which we normally explain statistical procedures. For example, in explaining least squares estimation we would not write down an algorithm for calculation and then try to give some intuition for what the algorithm does. Instead we write down elements of a statistical model in which the sources of random variation are explicitly denoted and identified. We then show that a way of estimating a parameter is by minimizing a sum of squares and that this method can have desirable properties. Lastly, instructions for calculation of the least squares estimate are provided.

We might call the approach described in the first paragraph an *algorithmic* explanation and that of the second paragraph an *explicit* explanation. Both approaches are, of course, valid. But a question is which is the more helpful. In this note we shall give explicit descriptions of three bootstrap procedures and show how a familiar and elementary calculation, the delta method, leads easily to some interesting results and connections.

2 An Efron Bootstrap.

Consider first a bootstrap procedure, due to Efron(1979), in which we resample rows of the data matrix.

Let Z, an $n \times m$ matrix, contain your data, where n is the number of individuals in your sample; let $t = t(Z)$ be a statistic whose value depends (only) upon the data; and let $v(1 \times n)$ be a n dimensional multinomial random variable. The row vector v contains $n-1$ zeroes and 1 one. If the 1 is in the j'th position in the vector it means "cell" j has been selected. The process of selecting a row from the matrix Z then has the explicit representation

$$vZ$$

and if the multinomial distribution is such that the probabilities of the the n cells are the same, and thus equal to $1/n$, the operation vZ is an explicit representation of randomly selecting a row of Z..

To represent the operation of randomly sampling n times with replacement we can use n independent, equal probability, multinomial variates represented by n vectors $v_{1.}, v_{2.}....v_{n.}$ and assemble these as rows of an $n \times n$ matrix V.

$$V = \begin{bmatrix} v_{1.} \\ v_{2.} \\ . \\ . \\ v_{n.} \end{bmatrix},$$

The matrix multiplication VZ produces another $n \times m$ matrix Z^*.

$$Z^* = VZ.$$

Z^* is a bootstrap replication of the data.

A statistic $t(Z^*) = t(VZ)$ calculated from a bootstrap replication has a bootstrap distribution determined solely by that of V. The properties of the bootstrap distribution depend on those of the random matrix V and upon the data Z.

Example 1 ***The Linear Model***

In the linear model $y = X\beta + \varepsilon$ with $E(X'\varepsilon) = 0$, the data matrix Z is $Z = (y : X)$ and the bootstrap replication is

$$Z^* = VZ = (Vy : VX).$$

As an example of a statistic whose bootstrap distribution is to be studied consider the least squares estimate $b = (X'X)^{-1}Xy$. A bootstrap replication of this estimate is found by replacing y by Vy and X by VX leading to

$$\begin{aligned} \beta^* &= (X'V'VX)^{-1}X'V'Vy \\ &= (X'WX)^{-1}X'Wy, \quad \textit{for} \quad W = V'V. \end{aligned} \tag{1}$$

The matrix W has a typical off-diagonal element equal to $\Sigma_k v_{ki}v_{kj}$ for $i \neq j$. This is identically zero because, for every k, $v_{ki}v_{kj} \equiv 0$ because all elements of any vector $v_{k\cdot}$ are either zero or one and there is only one 1. It follows that W is a diagonal matrix and that any bootstrap replication of β^* is a weighted least squares estimate. The weights are the diagonal elements of W, of which a typical one is $\Sigma_k v_{ki}^2$. But this measures the number of times n independent multinomial variates have a one in the k'th position Equivalently, it is the number of successes in n Bernoulli trials with probability of a success – $v_{ik} = 1$ – equal to $1/n$, so $W_{kk} \sim B(n, 1/n)$ with expectation 1 and variance $(n-1)/n$. This implies, in particular, that

$$E(W) = I_n.$$

The rather simple weighted least squares representation of a bootstrap replication of the least squares estimate, and the almost equally simple properties of the random matrix W lead to an easy study of the properties of the bootstrap distribution. For example, consider a delta method calculation of the mean and variance of this distribution. Think of β^* as a function of the n vector w which contains the n (binomial) diagonal elements of W. So $\beta^* = \beta^*(w)$ where the expected value of w is $\mathbf{1}$, a vector of n ones. Taking the Taylor series expansion of β^* up to the second term gives

$$\begin{aligned} \beta^*(w) &= \beta^*(\mathbf{1}) + \left[\frac{\partial \beta^*}{\partial w'}\right]_{w=1}(w - \mathbf{1}) \\ &= b + \left[e' \otimes (X'X)^{-1}X'\right]\frac{\partial vec W}{\partial w'}(w - \mathbf{1}) \end{aligned}$$

where $e = y - Xb$ is the least squares residual vector. The covariance matrix of $\beta^* - b$ depends on that of w and, using the properties described above, this is

$$E(w - \mathbf{1})(w - \mathbf{1})' = I_n - \frac{1}{n}J_n$$

where $\acute{J}_n$ is an $n \times n$ matrix of ones. Using this fact we find, after a little algebra, that

$$V(\beta^* - b) = (X'X)^{-1}X'EX(X'X)^{-1} \quad \text{where} \quad E = diag\{e_i^2\}.$$

Thus we conclude that the delta method approximate moments of β^* are

$$E(\beta^*) = b; \quad V(\beta^*) = (X'X)^{-1}X'EX(X'X)^{-1},$$

where b is the least squares estimate and $(X'X)^{-1}X'EX(X'X)^{-1}$ is the Eicker/White heteroscedasticity robust least squares covariance matrix estimate. (In the econometric textbooks, if robust covariance matrix estimates and bootstrapping are both mentioned the connection between these methods is rarely, if ever, mentioned.)

These remarks are subject to the qualification that, with positive probability in this resampling scheme, the matrix $X'WX$ will be singular and β^* not defined. For example, a bootstrap replication can give n identical rows for X^* with positive probability and in this case $X'WX$ will have rank one. So, strictly, we must consider the bootstrap distribution that we are examining as subject to the condition that realizations of V for which $X'WX$ is singular are discarded. Such a restriction will slightly alter the moment results that we are giving.

3 A BAYESIAN BOOTSTRAP

If we view the rows of the data matrix, Z, as realizations of independent multinomial variates on, say, $L+1$ points of support with probabilities $\{\pi_l\}$, summing to one for $l = 0, 1, 2, ..L$ we have a likelihood for the data. In this model the data provided by any one agent is a vector of $k+1$ numbers, a row of Z, and this vector is an element of a set of $L+1$ such vectors. The application to the regression model interprets the $\{z_i\}$ as $\{y_i, x_i\}$, that is, as rows of the data matrix $\{y : X\}$. This model does not restrict the conditional distribution of y_i given x_i and in particular y need not have linear regression on x, nor need the variance of y given x be independent of x. Thus it permits both non-linearity and heteroscedasticty. The most substantive restriction of the model is that the observations $\{y_i, x_i\}$ must be independent. It thus does not apply to models with autocorrelated errors or regressors which are lagged dependent variables.

The Bayesian bootstrap, so named by its originator Rubin(1981), assigns a specific prior to the vector of probabilities $\pi = \{\pi_l\}$ The method proceeds by selecting a parameter of interest, say θ, defined as a functional of the data distribution, whose components are π and the points of support of z say $\{z^l\}$. The prior for π together with the multinomial likelihood of the data enables computation of the posterior distribution of π. This in turn, because θ is a function of π, enables computation of the posterior distribution of θ.

Specifically, the computation is as follows. The multinomial likelihood for the data provided by agent i is $\ell_i(\pi) = \Pi_{l=0}^{L}\pi_l^{j_l}$ where j_l is the indicator of the

event $z_i = z^l$. Multiplying n such terms gives the likelihood for the whole data set

$$\ell(\pi) = \Pi_{l=0}^{L} \pi_l^{n_l} \quad \text{where } n_l \text{ is the number of times } \{z_i\} = z^l.$$

The natural conjugate prior for π is the multivariate beta, or dirichlet, with kernel

$$p(\pi) \propto \Pi_{l=0}^{L} \pi_l^{\nu_l - 1}$$

in which the $L+1$ parameters $\{\nu_l\}$ are positive and, of course, $\Sigma_{l=0}^{L} \pi_l = 1$. It follows from Bayes' theorem that the kernel of the posterior density of π is

$$p(\pi|z) \propto \Pi_{l=0}^{L} \pi_l^{n_l + \nu_l - 1} \tag{2}$$

This is again a dirichlet distribution.

For any parameter $\theta = \theta(\pi)$ we can calculate its posterior distribution by sampling from (2) and forming θ for each realization of π. By repeating this calculation many times an abitrarily accurate approximation to the posterior distribution of θ may be formed.

Sampling from (2) may be accomplished by simulating $L+1$ independent gamma variates with shape parameters equal to $(n_l + \nu_l)$ and unit scale parameters. Call these $\{g_l\}$ and form

$$\pi_l = \frac{g_l}{\Sigma_{j=0}^{L} g_j} \quad \text{for } l = 1, 2, ...L. \tag{3}$$

Then π is dirichlet distributed (with $\pi_0 = 1 - \Sigma_{j=1}^{L} \pi_j$).

The remaining issue is the choice of the $\{\nu_l\}$ and in the Bayesian bootstrap these are set to zero giving an improper version of the dirichlet prior. The effect of this is to produce (effectively) zero realizations for points in the support of z that have $n_l = 0$, that is, that were not observed in the sample. This in turn means that the relevant points of support are the distinct values of z that appear in the data, and, in particular, if all z values in the data are distinct then expectations such as $\Sigma_{l=0}^{L} y_l \pi_l$ may be equivalently written as $\Sigma_{i=1}^{n} y_i \pi_i$.

Example 2 ***The Linear Model and the Bayesian Bootstrap***

To see the Bayesian bootstrap in action and to study analytically the distribution consider an application to the linear model.

Let $z_j = (y_j \ x_j)$ where x_j is the j'th row of X. Define the functional β by the condition that

$$EX'(y - X\beta) = 0.$$

Thus,

$$\beta = [E(X'X)]^{-1} E(X'y)$$

where a typical element of $E(X'X)$ is $\Sigma_{i=1}^{n} x_{il} x_{im} \pi_i$ and a typical element of $E(X'y)$ is $\Sigma_{i=1}^{n} x_{il} y_i \pi_i$. (Defined in this way β is the coefficient vector in the linear projection of y on X.) Thus we can write β as

$$\beta = (X'PX)^{-1} X'Py$$

where $P = diag\{\pi_i\}$. But the $\{\pi_i\}$ can be represented as in (3) so we can write

$$\beta^* = (X'GX)^{-1}X'Gy \tag{4}$$

where G is an $n \times n$ diagonal matrix with elements that are independent gamma(1), or unit exponential, variates because when $v_l = 0$ and $n_l = 1$ the random variables $\{g_i\}$ are unit exponentials.

As G varies from realization to realization so does β and this variation is the Bayesian bootstrap (posterior) distribution of β. Note that, just as in the Efron bootstrap, realizations of β are equivalent to calculations of a weighted least squares estimate whose weight matrix, G, is on average, equal to the identity matrix. This is because the mean of a unit exponential variate is one, so $E(G) = I_n$. In fact the differences between (4) and (1) are rather small. Both W and G are diagonal matrices whose diagonal elements are non-negative random variables. The $\{w_i\}$ are binomial variates with means 1 and variances $(n-1)/n$; the diagonal elements of G, $\{g_i\}$ are exponential variates with means 1 and variances 1.The Bayesian boostrap can be thought of as a smoothed version of the Efron bootstrap in which every row of the data matrix appears in every bootstrap replication but different rows receive different weights in each recalculation. (It thus avoids the difficulty with the Efron boostrap noted earlier that, with some positive probability, β^* will not exist. The Bayesian bootstrap β^* exists with probability one.)

As in the Efron bootstrap the delta method can be used to find the approximate mean and variance of the posterior or bootstrap distribution, giving the expansion of $\beta^*(g)$ about the expectation of g, which is $\mathbf{1}$.

$$\beta^* = b + \left[e' \otimes (X'X)^{-1}X'\right] \frac{\partial vecG}{\partial g'}(g - \mathbf{1}).$$

A calculation similar to that for the Efron bootstrap, though somewhat simpler, then gives the approximate moments

$$E(\beta^*) = b; \qquad V(\beta^*) = (X'X)^{-1}X'EX(X'X)^{-1}$$

which are identical to those of the Efron bootstrap.

The Efron and Bayesian boostrap distributions are identical, to this order of approximation. Both have a mean equal to the least squares estimate and both have the Eicker/White heteroscedasticity robust covariance matrix.

4 THE EFRON RESIDUAL BOOTSTRAP

As Efron and Tibshirani note, "bootstrapping is not a uniquely defined concept" and there is a second Efron bootstrap which derives from resampling the model residuals. In a model with covariates in X and dependent variable in y model residuals take some estimate of the relation between y and X, and form a residual vector, e. The vector e is resampled as e^* and the corresponding y vector, say y^*, is calculated. Using y^*, X the statistic of interest is calculated

and the procedure repeated B times. Again the simplest context is the linear model and the explicit representation is given in the following:

Example 3 ***Bootstrapping Residuals in the Linear Model***

Residuals are easy to define in the linear model, in particular $e = y - Xb$ *provides the least squares residual vector. In the same way as for the earlier bootstraps a randomly resampled residual vector is* $e^* = Ve$. *Then the implied* y *vector is*

$$y^* = Xb + e^*.$$

A bootstrap replication of the least squares estimate is then

$$\begin{aligned} \beta^* &= (X'X)^{-1}X'y^* \\ &= (X'X)^{-1}X'Xb + (X'X)^{-1}X'Ve \\ &= b + (X'X)^{-1}X'Ve. \end{aligned}$$

Unlike the first two bootstraps this is linear in its random component, V *in this case, and its bootstrap distribution is easily calculated. In particular, from the fact that the rows of* V *are independent multinomials with equal probabilities it follows that*

$$E(V) = \frac{1}{n}J_n \quad \text{and} \quad E(Vee'V) = \left(\frac{\Sigma_i e_i^2}{n}\right) I_n = s^2 I_n. \tag{5}$$

(The second result here requires that $\Sigma_i e_i = 0$ *so the model must contain an intercept.) These results imply that*

$$E(\beta^*) = b \quad \text{and} \quad V(\beta^*) = s^2(X'X)^{-1}.$$

So, as is well known, the covariance matrix of the Efron residual bootstrap distribution is identicial to the standard least squares covariance matrix and the distribution is not robust to heteroscedasticity.

5 CONCLUSIONS

We have argued that there is some merit, both pedagogical and for research, to giving a more explicit presentation of bootstraps. Results on the properties of the bootstrap distribution can be easier to understand and to explain. To illustrate this we considered the linear model and described how one might explain the bootstrap distribution in that simple context.

This exposition lead to the easy derivation of several results. The first is that the bootstrap that resamples rows of the data matrix has a covariance matrix equal to the Eicker/White heteroscedasticity robust covariance matrix estimate – to $O(1/n)$. It is well known that this bootstrap is much less dependent on model assumptions than the bootstrap that resamples residuals and it is, perhaps, not surprising that the procedure turns out to be robust against failure of homoscedasticity (Textbook discussions of bootstrapping and robustness with

which I am familiar typically see no connection between them and they are discussed in widely separated parts of the book.).

The bootstrap that resamples data rows is equivalent, to $O(1/n)$, to the posterior distribution corresponding to a multinomial likelihood and an improper dirichlet prior. This is a particular case of a general result due to Lo(1987) in which the large sample equivalence of the Efron and Bayesian bootstraps is proved. One implication of this equivalence is the following. It is sometimes said that one ought to resample residuals – the third method we considered in the examples – because resampling data rows does not appear to correspond to a repeated sampling distribution in which X is held fixed. But resampling rows is equivalent to a Bayes procedure which is, of course, conditional on the entire data set, both y and X. This fact suggests that this argument in favour of resampling residuals is dubious.

Finally we should note this paper provides yet another situation in which Bayesian and frequentist procedures give numerically very similar answers but the philosophical interpretation of these answers is very different. The Bayesian bootstrap shows exactly the posterior uncertainty about the coefficients in a linear projection of y on X in a specific (multinomial/dirichlet) model. The (resampling rows) frequentist boostrap estimates the uncertainty in the least squares estimate of the coefficients in a possibly heteroscedastic linear model. These are quite different objectives.

6 References

Chesher A. D., and I. Jewitt (1987), "The bias of a heteroskedasticity consistent covariance matrix estimator", *Econometrica*, 55, 1217-1222.

Efron B. (1979) "Bootstrap methods: another look at the jacknife", *Annals of Statistics*, 7, 1-26.

Efron B. and Tibshirani R. J., (1993) *An introduction to the bootstrap*, Chapman and Hall, London.

Eicker, F., (1967) "Limit theorems for regressions with unequal and dependent errors", *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1, 59-82, Berkeley: University of California Press.

Horowitz, J. L. (2001) "The bootstrap", in *Handbook of Econometrics*, v. 5. eds J. J. Heckman and E. Leamer. North-Holland.

Lo, A. Y., (1987) "A large sample study of the Bayesian bootstrap", *Annals of Statistics*, 15, 360-375.

Rubin D. (1981) "The Bayesian bootstrap", *Annals of Statistic*s, 1981, 9, 130-134.

White H., (1980) "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", 1980, *Econometrica*, 48, 817-838.

Wooldridge, J., W. (2002) *Econometric analysis of cross-section and panel data*, Cambridge, MA. MIT Press, 2002.

7 Appendix

We have used the Taylor series expansion of $\beta^*(g)$ or $\beta^*(w)$ several times and since this may not be obvious to readers we here sketch the derivation.

Consider $\beta^*(g) = (X'GX)^{-1}X'Gy$ where G is a diagonal matrix with diagonal elements equal to $\{g_i\}$ contained in a n vector g. Then $\beta^*(1) = (X'X)^{-1}X'y = b$ and

$$\beta^* = b + \frac{\partial (X'GX)^{-1}X'Gy}{\partial g'}(g - \mathbf{1}).$$

Next,

$$\begin{aligned}
\frac{\partial (X'GX)^{-1}X'Gy}{\partial g'} &= \frac{\partial vec(X'GX)^{-1}X'Gy}{\partial g'} \\
&= (1 \otimes (X'GX)^{-1})\frac{\partial vecX'Gy}{\partial g'} + (y'GX \otimes I_k)\frac{\partial vec(X'GX)^{-1}}{\partial g'}, \\
(\text{using } \frac{\partial vecAB}{\partial g'}) &= (I_q \otimes A)\frac{\partial vecB}{\partial g'} + (B' \otimes I_n)\frac{\partial vecA}{\partial g'} \text{ when } A \text{ is } n \times p \text{ and } B \text{ is } p \times q).
\end{aligned}$$

Next we use

$$\frac{\partial vecA^{-1}}{\partial vec(A)'} = -(A^{-1})' \otimes A^{-1} \text{ and } \frac{\partial vec(X'GX)}{\partial g'} = (X' \otimes X')\frac{\partial vecG}{\partial g'}$$

to get

$$\frac{\partial vec(X'GX)^{-1}}{\partial g'} = -(X'GX)^{-1} \otimes (X'GX)^{-1}X' \otimes X'\frac{\partial vecG}{\partial g'}.$$

This then leads to

$$\begin{aligned}
\frac{\partial (X'GX)^{-1}X'Gy}{\partial g'} &= (1 \otimes (X'GX)^{-1})(y' \otimes X')\frac{\partial vecG}{\partial g'} \\
&\quad -(y'GX \otimes I_k)((X'GX)^{-1} \otimes (X'GX)^{-1})(X' \otimes X')\frac{\partial vecG}{\partial g'} \\
&= \left\{y' \otimes (X'GX)^{-1}X' - (y'GX(X'GX)^{-1} \otimes (X'GX)^{-1})(X' \otimes X')\right\}\frac{\partial vecG}{\partial g'} \\
&= \left\{y' \otimes (X'GX)^{-1}X' - y'GX(X'GX)^{-1}X' \otimes (X'GX)^{-1}X'\right\}\frac{\partial vecG}{\partial g'} \\
&= \left\{(y' - y'GX(X'GX)^{-1}X') \otimes (X'GX)^{-1}X'\right\}\frac{\partial vecG}{\partial g'} \\
&= \left\{(y - X\beta^*)' \otimes (X'GX)^{-1}X'\right\}\frac{\partial vecG}{\partial g'} \\
&= \left\{\varepsilon^{*'} \otimes (X'GX)^{-1}X'\right\}\frac{\partial vecG}{\partial g'},
\end{aligned}$$

and at $g = \mathbf{1}$ $\varepsilon^* = e = y - Xb$ this reduces to

$$\frac{\partial \beta^*}{\partial g'} = (e' \otimes (X'X)^{-1}X')\frac{\partial vecG}{\partial g'}.$$

Thus

$$\beta^* = b + (e' \otimes (X'X)^{-1}X')\frac{\partial vecG}{\partial g'}(g - \mathbf{1})$$

as required.