Discussion on Brad Efron's paper: Model selection, estimation, and bootstrap smoothing

Dimitris N. Politis
Department of Mathematics
University of California—San Diego
La Jolla, CA 92093-0112, USA
email: dpolitis@ucsd.edu

1 Which bootstrap?

Professor Brad Efron, a pioneer in the re-casting of modern statistics in its current computer-intensive framework, has given us another important and thought-provoking piece of work. To discuss it, consider the standard additive regression model

$$Y_j = \mu_p(\underline{x}_j) + \varepsilon_j \quad \text{for } j = 1, \ldots, n \tag{1}$$

where $Y_1, \ldots, Y_n$ are the data, ε_j are the errors assumed i.i.d. $(0, \sigma^2)$, and $\underline{x}_j$ is a length p vector of explanatory (predictor) variables associated with the observation Y_j. The function $\mu_p(\cdot)$ is unknown but assumed to belong to a certain class of functions that is either finite-dimensional or not. For simplicity, let us focus on the simple case where $\mu_p(\cdot)$ is affine in its arguments, i.e., $\mu_p(\underline{x}_j) = \beta_0 + \underline{x}_j' \underline{\beta}_p$ with $\underline{\beta}_p = (\beta_1, \ldots, \beta_p)'$. Also for simplicity assume that the p coordinates of $\underline{x}_j$ are ranked in terms of their importance so that model selection is tantamount to choosing the order p; this is the case with the polynomial regression example of the cholesterol data.

In the above, the regressor $\underline{x}_j$ is most often thought of as deterministic, and $\mu_p(\underline{x}_j)$ has the interpretation of expected value of the response Y_j associated with regressor $\underline{x}_j$. But if the regressors are random, then Efron's set-up where the pairs

$$(Y_j, \underline{x}_j) \text{ for } j = 1, \ldots, n \text{ are i.i.d.} \tag{2}$$

is appropriate. In that case, eq. (1) still applies by defining $\mu_p(\underline{x}_j)$ to be the (theoretical) orthogonal projection of Y_j onto the linear span of the elements of $\underline{x}_j$ (plus a constant), and letting eq. (1) serve as the definition of ε_j which would then be uncorrelated with $\underline{x}_j$. Of course, under joint normality of $(Y_j, \underline{x}_j)$, the projection $\mu_p(\underline{x}_j)$ would equal the conditional

expectation $E(Y_j|\underline{x}_j)$ since the latter would be affine as a function of $\underline{x}_j$; in this case the ε_j would be normal as well, and independent of $\underline{x}_j$.

In the cholesterol example, it is obvious that even though the (transformed) compliance variable may be normally distributed, when raised to the power of two or higher it will not be. However, the assumption of normality is innocuous if it is just used as a trick to derive the orthogonal projection via the simple formula for Gaussian conditional expectations. Note that it is possible to avoid the assumption of normality but still retain the linearity of $E(Y_j|\underline{x}_j)$. A simple way of doing that is to assume that eq. (1) is true with $\mu_p(\underline{x}_j) = \beta_0 + \underline{x}_j'\underline{\beta}_p$ and ε_j being i.i.d. $(0, \sigma^2)$ as before, coupled with the 'exogeneity' assumption that $\underline{x}_1, \ldots, \underline{x}_n$ are i.i.d. and independent of $\{\varepsilon_1, \ldots, \varepsilon_n\}$.

The two aforementioned viewpoints on the same scatterplot motivate the two most popular bootstrap methods for homoscedastic regression, namely the *residual bootstrap* and the *pairs bootstrap*. The latter is a straightforward implication of eq. (2). By contrast, the residual bootstrap keeps the $\underline{x}_j$ fixed, and creates pseudo-data using eq. (1), i.e., letting

$$Y_j^* = \hat{\beta}_0 + \underline{x}_j'\hat{\underline{\beta}}_p + \varepsilon_j^* \text{ for } j = 1, \ldots, n \tag{3}$$

where $\hat{\beta}_0, \hat{\underline{\beta}}_p$ are the Least Squares (LS) estimators of $\beta_0, \underline{\beta}_p$, and ε_j^* is a random draw from the set of fitted residuals $\{e_1, \ldots, e_n\}$ with $e_j = Y_j - \hat{\beta}_0 - \underline{x}_j'\hat{\underline{\beta}}_p$. Bose and Chatterjee (2002) review and compare several different resampling methods for linear regression including the pairs and residual bootstraps.

Efron uses the pairs bootstrap in the paper, and I do not think this is simply a matter of taste. Doing the residual bootstrap presupposes a choice of the order p, i.e., model selection. Suppose $\hat{p}$ is a data-based selector of the order p; then the residual bootstrap would generate data from a model with dimension $\hat{p}$, and model selection procedures when applied in the bootstrap world would disproportionally often select the same $\hat{p}$ again. To elaborate, suppose that with this type of data your favorite model selection procedure, say Mallows C_p, would select $\hat{p} = k$ with sampling probability p_k, i.e., $100p_k\%$ of such scatterplots would result into $\hat{p} = k$. The sampling probability p_k would not be well captured/replicated when pseudo-scatterplots are generated by residual bootstrap that always uses the order $\hat{p}$ (say $\hat{p} = 3$) that was chosen based on the original data.

The question arises: can we still employ a residual bootstrap in such a case where model selection is also involved? The answer appears to be yes but it may be quite more cumbersome. To start with, one can use the pairs bootstrap (or other considerations) in order to estimate the aforementioned sampling probability p_k. The important thing here is not to underestimate model order; so one can probably afford to be slightly less parsimonious at the stage of estimating p_k. Then, use a two-step residual bootstrap: first generate the order, say p^*, using the discrete distribution that puts mass p_k on the number k, and then

generate a pseudo-scatterplot via a residual bootstrap based on order p^*. The collection of many bootstrap scatterplots generated this way should reflect well the variability associated with model selection. If the regressors are not ranked, i.e., the models are not nested, then one may associate sampling probability p_k to candidate model k, and modify the above two-step procedure accordingly.

2 Estimation and prediction

Efron focuses on the linear combination $\mu_p(\underline{x}) = \beta_0 + \underline{x}'\underline{\beta}_p$ as the parameter of interest. As previously mentioned, $\mu_p(\underline{x})$ has the interpretation of the mean response when the regressor vector takes the value $\underline{x}$. As such, it is a quantity that has precise meaning for *all* models considered; indeed, all models should be able to capture such a quantity regardless of whether individual β–parameters are zeroed out or not. Interestingly, $\mu_p(\underline{x})$ has an additional interpretation: it is the L_2–optimal (linear) *predictor* of the future response Y_{n+1} that is associated with a regression vector $\underline{x}_{n+1}$ that is equal to $\underline{x}$.

Estimation and prediction often go hand-by-hand. It is not a coincidence that popular model selection methods such as Mallows C_p or Cross-Validation rank models in terms of their predictive ability. On the other hand, prediction is typically conducted using an estimated model which implies a preliminary step of model-fitting. Since fitting a model gives the practitioner the ability to predict future responses one can ask if the converse is also true. The answer is yes: if one is able to predict the future response that is associated with *any* regressor value $\underline{x}$, then an implied model-fitting is taking place as the curve explaining/predicting Y on the basis of $\underline{x}$ is being constructed.

But how can one predict without a model? The *Model-Free (MF) Prediction Principle* of Politis (2013) substitutes the notion of *transformation* in place of a model, and places the emphasis on observable quantities, i.e., current and future data, as opposed to unobservable model parameters and estimates thereof. To briefly state it, consider the vector of responses $\underline{Y}_m = (Y_1, \ldots, Y_m)'$ where Y_j is associated with regressor $\underline{x}_j$; the latter can be assumed deterministic for the time being. Thus, $\underline{Y}_n$ contains the already observed responses while $\underline{Y}_{n+1}$ contains $\underline{Y}_n$ plus the future (yet unobserved) response Y_{n+1} associated with regressor value $\underline{x}_{n+1}$.

If the Y_is were i.i.d., then prediction would be trivial: the L_2—optimal predictor of Y_{n+1} would simply be given by the common mean of the Y_is, totally disregarding the regressor value $\underline{x}_{n+1}$. Since the Y_is are not i.i.d., the Model-Free Prediction Principle amounts to using the structure of the problem—that also utilizes the regressors—in order to find an *invertible transformation* H_m that can map the non-i.i.d. vector $\underline{Y}_m$ to a vector $\underline{\epsilon}_m = (\epsilon_1, \ldots, \epsilon_m)'$ that has i.i.d. components; here m could be taken equal to either n or $n+1$ as needed.

Letting H_m^{-1} denote the inverse transformation, we would then have $\underline{\epsilon}_m = H_m(\underline{Y}_m)$ and $\underline{Y}_m = H_m^{-1}(\underline{\epsilon}_m)$, i.e.,

$$\underline{Y}_m \stackrel{H_m}{\longmapsto} \underline{\epsilon}_m \quad \text{and} \quad \underline{\epsilon}_m \stackrel{H_m^{-1}}{\longmapsto} \underline{Y}_m. \tag{4}$$

If the practitioner is successful in implementing the MF procedure, i.e., in identifying the transformation H_m to be used, then the prediction problem is reduced to the trivial one of predicting i.i.d. variables. To see why, note that eq. (4) with $m = n + 1$ yields $\underline{Y}_{n+1} = H_{n+1}^{-1}(\underline{\epsilon}_{n+1}) = H_{n+1}^{-1}(\underline{\epsilon}_n, \epsilon_{n+1})$. But $\underline{\epsilon}_n$ can be treated as known given the data $\underline{Y}_n$; just use eq. (4) with $m = n$. Since the unobserved Y_{n+1} is just the $(n+1)^{th}$ coordinate of vector $\underline{Y}_{n+1}$, the former can also be expressed as a function of the unobserved ϵ_{n+1}. Finally, note that predicting a function, say $g(\cdot)$, of an i.i.d. sequence $\epsilon_1, \ldots, \epsilon_n, \epsilon_{n+1}$ is straightforward because $g(\epsilon_1), \ldots, g(\epsilon_n), g(\epsilon_{n+1})$ is simply another sequence of i.i.d. random variables.

Under regularity conditions, such a transformation H_m always exists although it is not unique. The challenge to the skills and expertise of the statistician is to be able to devise and estimate a workable such transformation for the problem at hand; see Politis (2013) for a complete treatment of the regression paradigm. Note, however, that having mapped our data onto the i.i.d. variables $\epsilon_1, \ldots, \epsilon_n$, a *Model-Free bootstrap* scheme readily presents itself, namely: (a) generate bootstrap variables $\epsilon_1^*, \ldots, \epsilon_n^*$ by random drawing (without replacement) from the set $\{\epsilon_1, \ldots, \epsilon_n\}$, and (b) generate a pseudo-response vector $\underline{Y}_n^* = \hat{H}_n^{-1}(\underline{\epsilon}_n^*)$ where $\underline{\epsilon}_n^* = (\epsilon_1^*, \ldots, \epsilon_n^*)'$ and $\hat{H}_n^{-1}$ is the estimated (inverse) transformation.

The MF bootstrap can be viewed as an extension of the residual bootstrap to settings where a model is not available. To see why, note that if the additive model (1) is actually available, then the transformation H_n can be readily estimated by first estimating $\mu_p(\cdot)$. For example, constructing the fitted residuals $e_j = Y_j - \hat{\beta}_0 - \underline{x}_j' \hat{\underline{\beta}}_p$ can be viewed as a transformation of the $\underline{Y}_n$ data towards (approximate) i.i.d.–ness; recall that the residuals are approximately i.i.d. being proxies for the true errors.

However, this is not the only possible transformation; for instance, one can define $\epsilon_j = Y_j - \hat{\beta}_0^{(j)} - \underline{x}_j' \hat{\underline{\beta}}_p^{(j)}$ where $\hat{\beta}_0^{(j)}, \hat{\underline{\beta}}_p^{(j)}$ are the LS estimates obtained from the *delete–one* dataset $\{(Y_t, \underline{x}_t)$ for $t = 1, \ldots, n$ but with $t \neq j\}$. In the above, the ϵ_j are nothing more than the *predictive* residuals that are typically used in Cross-Validation; see e.g. Geisser (1993) and the references therein. Politis (2013) gives an argument based on the Model-Free Prediction Principle that favors using the predictive (as opposed to the fitted) residuals for resampling; doing so appears to partially correct the under-coverage of bootstrap prediction intervals noticed early on by Efron (1983) and Stine (1985).

In any case, when model selection is also involved, i.e., when the number p of regressors to be used in the transformation H_n is up for debate, the analogy between the residual bootstrap and the MF bootstrap suggests that a similar trick as the one suggested at

the end of last section may be helpful. To elaborate, one can use a two-step resampling procedure: (a) generate the model order, say p^*, using some estimated distribution (say p_k), and then generate a pseudo-scatterplot via the MF bootstrap based on an estimated transformation $\hat{H}_n$ that uses p^* regressors.

Nevertheless, there is nothing to stop the MF practitioner from using the pairs bootstrap in this setting; this could be done just to obtain an estimate of the sampling distribution p_k needed above, or in order to carry out the complete task of capturing the variability of an estimator that includes the model selection step. But using the pairs bootstrap is associated with an assumption that the regressors $\underline{x}_j$ are random, and furthermore that the pairs $(Y_1, \underline{x}_1), (Y_2, \underline{x}_2), \ldots$ are i.i.d. as in eq. (2). In the case of random regressors, the Model-Free Prediction Principle can be simply re-stated by conditioning on the regressor values. In other words, the transformation H_m of eq. (4) would be constructed conditionally on the values $\{\underline{x}_1, \ldots, \underline{x}_m\}$, and the goal of the MF practitioner is to render the transformed variables $\epsilon_1, \ldots, \epsilon_m$ as close to i.i.d. as possible conditionally on $\{\underline{x}_1, \ldots, \underline{x}_m\}$.

3 Models vs. transformations: a reconciliation

The Model-Free (MF) approach can form the basis for a complete statistical inference that includes point estimators and predictors in addition to confidence and prediction intervals without assuming an additive model such as (1); see Politis (2013, 2014) for details. Interestingly, however, when an additive model is known to hold true, there is no discrepancy if one adheres to the MF approach, i.e., tries to find a transformation towards "i.i.d.-ness".

To see why, let us assume eq. (1) with $\mu_p(\underline{x}_j) = \beta_0 + \underline{x}_j' \underline{\beta}_p$. The essence of this model—as far as MF prediction is concerned—is that the variables $\epsilon_j \equiv Y_j - \underline{x}_j' \underline{\beta}_p$ are i.i.d. albeit with (possibly) non-zero mean β_0. Thus, a candidate transformation to 'i.i.d.–ness' may be constructed by letting $r_j = Y_j - \underline{x}_j' \hat{\underline{\beta}}_p$ where $\hat{\underline{\beta}}_p$ is a candidate vector. The MF principle now mandates choosing $\hat{\underline{\beta}}_p$ with the objective of having the r_js become as close to i.i.d. as possible. However, under the stated regression model, the r_js would be i.i.d. if only their first moment was properly adjusted.

To elaborate, a homoscedastic regression model such as (1) implies that all central moments of order two or higher are constant; the only non-i.i.d. feature is in the first moment. So, in this case, the MF principle suggests choosing $\hat{\underline{\beta}}_p$ in such a way as to make $r_1, \ldots, r_n$ have (approximately) the *same* first moment. Noting that the first moment—if it is common—would be naturally approximated by the empirical value $\hat{r} = n^{-1} \sum_{i=1}^{n} r_i$, we can use a *subsampling* construction to make this happen.

To fix ideas, assume for simplicity that $p = 1$, and that the univariate design points

$x_1, ..., x_n$ are sorted in ascending order. Then compute the overlapping block means

$$\bar{r}_{k,b} = b^{-1} \sum_{j=k}^{k+b-1} r_j \quad \text{for} \quad k = 1, \ldots, q \tag{5}$$

where b is the block size, and $q = n - b + 1$ is the number of available blocks.

Note that $\bar{r}_{k,b}$ is an estimate of the first moment of the r_is found in the kth block. In order to achieve the target requirement that all $r_1, \ldots, r_n$ have first moment that is the same (and thus approximately equal to $\hat{r}$), the MF practitioner may

$$\text{choose } \hat{\beta}_1 \text{ that minimizes } LS(b) = \sum_{k=1}^{q} (\bar{r}_{k,b} - \hat{r})^2 \text{ or } L1(b) = \sum_{k=1}^{q} |\bar{r}_{k,b} - \hat{r}| \tag{6}$$

according to whether an L_2 or L_1 loss criterion is preferred.

Instead of $\hat{r}$, we could equally use the mean of means, i.e., $\bar{\bar{r}} = q^{-1} \sum_{k=1}^{q} \bar{r}_{k,b}$ as the centering value in eq. (6). If $b = 1$, then $\hat{r} = \bar{\bar{r}}$; if $b > 1$, then $\hat{r} = \bar{\bar{r}} + O_P(b/n)$ so the difference is negligible provided b is small as compared to n. Recall that in the typical application of subsampling for variance or distribution estimation, it is suggested to take the block size b to be large (but still of smaller order than n); this is for the purpose of making the subsample statistics $\bar{r}_{k,b}$ have asymptotically the same distribution as the statistic $\hat{r}$ computed from the full sample; see e.g. Politis, Romano and Wolf (1999).

Nevertheless, it is not crucial in our current setting that each of the $\bar{r}_{k,b}$ have asymptotically the same distribution as $\hat{r}$. What is important is that all the $\bar{r}_{k,b}$ (for $k = 1, \ldots, q$) have approximately the same distribution whatever that may be. Therefore, it is not necessary in eq. (6) to use a large value for b. Even the value $b = 1$ is acceptable, in which case we have:

$$\frac{d}{d\hat{\beta}_1} LS(1) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{where } \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \text{ and } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

In other words, the MF fitting procedure (6) with L_2 loss and $b = 1$ is re-assuringly *identical* to the usual Least Squares estimator! Note that the r_is serve as proxies for the unobservable ε_is which have expected value β_0 under model (1). Hence, β_0 is naturally estimated by the sample mean of the r_is, i.e.,

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 x_i) = \bar{Y} - \hat{\beta}_1 \bar{x}$$

which is again the Least Squares estimator.

Minimizing $LS(b)$ with $b > 1$ gives a more robust way of doing Least Squares in which the effect of potential outliers is diminished by the local averaging of b neighboring values; details

are omitted due to lack of space. Similarly to the above, minimizing $L1(1)$ is equivalent to L_1 regression, whereas minimizing $L1(b)$ with $b > 1$ gives additional robustness.

Finally, let us revisit the general case of model (1) with $\mu_p(\underline{x}_j) = \beta_0 + \underline{x}_j'\underline{\beta}_p$. When $p > 1$, the regressors $\underline{x}_j$ can not be sorted in ascending order. One could instead use a local-averaging or nearest-neighbor technique to compute the subsample means. But no such trick is needed in the most interesting case of $b = 1$ since the quantities $LS(1)$ and $L1(1)$ are unequivocally defined as

$$LS(1) = \sum_{k=1}^{n}(r_k - \hat{r})^2 \text{ and } L1(1) = \sum_{k=1}^{n}|r_k - \hat{r}|. \tag{7}$$

It is now easy to see that the MF practitioner that chooses the β's in order to minimize $LS(1)$ or $L1(1)$, is effectively doing Least Squares or L_1 regression respectively. Hence, when an additive model is available, there is no discrepancy between the MF approach and traditional model fitting. Nevertheless, the MF approach can still lend some insights such as the aforementioned use of predictive residuals in connection with the model-based residual bootstrap.

References

[1] Bose, A. and Chatterjee, S. (2002). Comparison of bootstrap and jackknife variance estimators in linear regression: second order results, *Statist. Sinica*, vol. 12, pp. 575-598.

[2] Efron, B. (1983), Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78, 316-331.

[3] Geisser, S. (1993). *Predictive Inference: An Introduction*, Chapman and Hall, New York.

[4] Politis, D.N. (2013). Model-free model-fitting and predictive distributions (with Discussion), *Test*, vol. 22, no. 2, pp. 183-250.

[5] Politis D.N. (2014). Bootstrap confidence intervals in nonparametric regression without an additive model, *Topics in Nonparametric Statistics: Proceedings of the First Conference of the International Society for NonParametric Statistics,* M.G. Akritas, S.N. Lahiri and D.N. Politis (Eds.), Springer, New York, 2014.

[6] Politis, D.N., Romano, J.P. and Wolf, M. (1999), *Subsampling,* Springer, New York.

[7] Stine, R.A. (1985). Bootstrap prediction intervals for regression. *J. Amer. Statist. Assoc.*, 80, 1026-1031.