Bootstrapping Regression Models With BLUS Residuals*

Michèle Grenier† Christian Léger†

CRM-2559

August 1998

*This paper is part of the Master's thesis of the first author under the supervision of the second author. We thank NSERC and FCAR for their financial support. We would also like to thank a referee and an Associate Editor for their comments which led to an improved presentation of the simulation results.

†Département de mathématiques et de statistique, Université de Montréal

Abstract

To bootstrap a regression problem, pairs of response and explanatory variables or residuals can be resampled, according to whether we believe that the explanatory variables are random or fixed. In the latter case, different residuals have been proposed in the literature, including the ordinary residuals (Efron, 1979), standardized residuals (Bickel and Freedman, 1983) and studentized residuals (Weber, 1984). Freedman (1981) has shown that the bootstrap from ordinary residuals is asymptotically valid when the number of cases increases and the number of variables is fixed while Weber (1984) has done the same for studentized residuals. Bickel and Freedman (1983) have shown the asymptotic validity for ordinary residuals when the number of variables as well as the number of cases increase provided that the ratio of the two converges to 0 at an appropriate rate. In this paper, we introduce the use of Best Linear Unbiased Scaled (BLUS) residuals in bootstrapping regression models. The main advantage of the BLUS residuals, introduced in Theil (1965), is that they are uncorrelated. The main disadvantage is that only $n-p$ residuals can be computed for a regression problem with n cases and p variables. The asymptotic results of Freedman (1981) and Bickel and Freedman (1983) for the ordinary (and standardized) residuals are generalized to the BLUS residuals. A small simulation study shows that even though only $n-p$ residuals are available, bootstrapping BLUS residuals is as good in small samples, and sometimes better, than bootstrapping from standardized or studentized residuals.

Key words and phrases: BLUS residuals, standardized residuals, studentized residuals, bootstrap, regression

AMS 1991 subject classifications: Primary 62G09, 62J05; secondary 62G20.

Résumé

Pour appliquer le bootstrap en régression, on peut soit rééchantillonner conjointement les variables réponse et explicatives ou encore rééchantillonner des résidus selon que l'on pense que les variables explicatives sont aléatoires ou fixes. Dans ce dernier cas, plusieurs sortes de résidus ont été présentés dans la littérature, notamment les résidus ordinaires (Efron, 1979), les résidus standardisés (Bickel et Freedman, 1983) et les résidus studentisés (Weber, 1984). Freedman (1981) a démontré que l'utilisation des résidus ordinaires dans le bootstrap en régression est asymptotiquement valide lorsque le nombre d'unités d'observation augmente alors que le nombre de variables explicatives est fixe. Bickel et Freedman (1983) ont démontré la même validité asymptotique lorsque le nombre de variables explicatives augmente en même temps que le nombre d'unités d'observation en autant que le rapport des deux converge vers 0 à un taux approprié. Dans cet article, nous introduisons l'utilisation des résidus BLUS (*Best Linear Scaled Unbiased*) dans le bootstrap en régression. Le principal avantage des résidus BLUS, introduits par Theil (1965), est qu'ils sont non corrélés. Toutefois, on ne peut calculer que $n-p$ résidus dans un problème de régression avec n unités d'observation et p variables explicatives. Les résultats asymptotiques de Freedman (1981) et Bickel et Freedman (1983) pour les résidus ordinaires (et standardisés) sont généralisés aux résidus BLUS. Une simulation démontre que bien que seulement $n-p$ résidus soient disponibles, le bootstrap à partir des résidus BLUS fait aussi bien et parfois mieux dans les petits échantillons que le bootstrap à partir des résidus standardisés ou studentisés.

1 Introduction

The bootstrap algorithm of Efron (1979) for independently and identically distributed (i.i.d.) random variables is well known. To estimate the distribution of a statistic computed on i.i.d. data, the statistic is computed on bootstrap samples generated by resampling with replacement from the original data. In a linear regression model, the vector of dependent observations is the sum of the regression mean vector, which is a linear combination of the independent (fixed) observations, and a vector of i.i.d. errors from the distribution F. To apply the bootstrap to this model, the regression mean and the distribution F must be estimated. The regression mean is usually estimated using the least squares estimate of the regression coefficients. The bootstrap errors cannot be generated by resampling the original errors as they are not observed. Instead, residuals are used as proxies.

To that effect, many different residuals have been used. Efron (1979) used the least squares residuals centered at their mean and Freedman (1981) has shown the asymptotic validity of the bootstrap with these residuals. In that case, the variance of the bootstrap errors is the maximum likelihood estimate of the variance of F which is biased downwards. Bickel and Freedman (1983) suggested a global solution which consists of dividing the least squares residuals by $n-p$ where n is the number of observations and p is the number of independent variables. Weber (1984), noting that the residuals do not all have the same variance, suggested a local solution consisting of modifying each residual so that they all have equal variance. In either case, the resulting residuals remain dependent, just like the ordinary residuals.

In this paper, we consider using centered BLUS residuals, which are a linear transformation of the original residuals. These residuals, introduced by Theil (1965), have equal variance and are uncorrelated, (they are even independent if the original errors are normally distributed). The price to pay is that only $n-p$ BLUS residuals can be computed. So there are p fewer residuals to resample with replacement from.

The paper is organized as follows. In Section 2, we describe the BLUS residuals. In Section 3, we show that the bootstrap distribution of the regression coefficients from BLUS residuals is asymptotically valid for fixed p as well as for increasing p, generalizing the results of Freedman (1981) and Bickel and Freedman (1983). A small sample simulation is presented in Section 4 which shows that, at least for the cases which we tried, our method is almost always as good or better than those of Efron, Bickel and Freedman, and Weber. So even in small samples, resampling from fewer uncorrelated residuals does not hurt. Concluding remarks are in Section 5 while an Appendix contains the proofs of the results in Section 3.

2 BLUS Residuals

Consider the following regression model with fixed independent variables:

$$y = X\beta + \epsilon, \tag{1}$$

where y is the n-vector of the dependent variable, X is the $n \times p$ full rank matrix of fixed independent variables, β is the p-vector of unknown regression coefficients and ϵ is the vector of n i.i.d. random variables from distribution F with mean 0 and finite variance σ^2. The least squares estimate of β is $\hat{\beta} = (X'X)^{-1}X'y$ and the predicted values are $\hat{y} = Hy$ where $H = X(X'X)^{-1}X$ is the hat matrix. The ordinary residuals are $\hat{\epsilon} = y - \hat{y} = My$, where $M = I - H$. It is well known that the ordinary residuals have mean 0 and variance $\sigma^2 M$ and are correlated.

If A is an $(n-p) \times n$ matrix such that $AX = 0$ and $AA' = I_{n-p}$ where I_n is the identity matrix of order n, then $\hat{\epsilon}^B = Ay = A\epsilon$ will have mean 0 and variance $\sigma^2 I_{n-p}$. Such residuals are LUS residuals, i.e., Linear, Unbiased with a Scalar covariance matrix. Theil (1965) introduced BLUS residuals, i.e., Best LUS residuals, by finding the matrix A satisfying the previous two conditions and which minimizes the quadratic error

$$E\{(A\epsilon - J\epsilon)'(A\epsilon - J\epsilon)\},$$

where $J = [0\ I_{n-p}]$ with 0 an $(n-p) \times p$ null matrix. So BLUS residuals are the best estimates (in a quadratic sense) of the last $n-p$ true errors. Focusing on the last $n-p$ errors is without loss of generality

as we could simply reorder the observations. If the distribution of the errors is normal then the BLUS residuals $\hat{\epsilon}^B$ are i.i.d.. Otherwise, they are uncorrelated.

The optimal matrix A was found by Theil (1965). Let

$$X = \begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \text{ and } M = \begin{bmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{bmatrix}, \tag{2}$$

where X_0 is $p \times p$ nonsingular, X_1 is $(n-p) \times p$ and M is similarly partitioned. Also let $M_{11} = PDP'$ where D is the diagonal matrix of the eigenvalues of M_{11} and P is its matrix of eigenvectors. Then

$$A = [-PD^{1/2}P'X_1X_0 \quad PD^{1/2}P']. \tag{3}$$

Consequently the BLUS residuals can easily be computed in computing environments such as S-Plus.

3 Asymptotic Validity of BLUS Residuals Bootstrap in Regression

Consider the following regression model.

Model 1

a) *The observations are $y_{(n)} = X_{(n)}\beta + \epsilon_{(n)}$, where $y_{(n)}$ is the n-vector of the dependent variable, β is the p-vector of unknown parameters, and $\epsilon_{(n)}$ is the n-vector of unobservable random errors.*

b) *The $n \times p$ design matrix $X_{(n)}$ is fixed and of full rank $p \leq n$.*

c) *The errors $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are independent and identically distributed from an unknown distribution F with mean 0 and finite variance σ^2.*

d) *$n^{-1}X_{(n)}^t X_{(n)}$ converges elementwise to a positive definite matrix V.*

Let $\hat{F}_n$ be an estimate of the distribution of the errors ϵ. Bootstrap observations are computed from the least squares estimator $\hat{\beta}$ and from ϵ^*, a sample of n i.i.d. bootstrap errors from $\hat{F}_n$, as follows:

$$y^* = X\hat{\beta} + \epsilon^*. \tag{4}$$

Different $\hat{F}_n$ lead to different bootstrap estimators. Efron (1979) used the empirical distribution function of the centered (ordinary) residuals and Bickel and Freedman (1981) showed the asymptotic validity of the procedure. In small samples, these bootstrap observations are underdispersed because the bootstrap variance of ϵ^* is $[(n-p)/n]\hat{\sigma}^2$ where $\hat{\sigma}^2$ is the unbiased estimate of σ^2 based on the residual sums of squares. Bickel and Freedman (1983) suggested multiplying the residuals by $[n/(n-p)]^{1/2}$ and to apply the bootstrap from these (centered) residuals. Weber (1984) suggested instead to use the residuals $\hat{e}_i^W = \hat{e}_i(1-h_{ii})^{-1/2}$, where $\hat{e}_i$ is the i^{th} (ordinary) residual and h_{ii} is the i^{th} diagonal element of the hat matrix H and then to bootstrap from the centered residuals $\hat{e}_i^W - n^{-1}\sum_{j=1}^n \hat{e}_j^W$. Note that the variance of each residual $\hat{e}_i^W$ is identical, although they are correlated.

In this paper, we consider using the BLUS residuals. Let $\hat{\epsilon}^B = Ay$ be the BLUS residuals where A is given by the equations (2) and (3). Let $\hat{F}_n^B$ be the empirical distribution function of the centered BLUS residuals, i.e., of $\hat{\epsilon}_i^B - (n-p)^{-1}\sum_{j=1}^{n-p} \hat{\epsilon}_j^B$. Then the bootstrap algorithm of (4) is applied with $\hat{F}_n = \hat{F}_n^B$. Note that n bootstrap errors are chosen with replacement from the $n-p$ BLUS residuals. With p fixed and n large, resampling from $n-p$ rather than n residuals should not be a problem, but the small sample behavior might be affected. This is studied in the next section.

We now show that the bootstrap from BLUS residuals is asymptotically valid. Let

$$y^*(B) = X\hat{\beta} + \epsilon^*(B), \tag{5}$$

where $\epsilon^*(B)$ is distributed according to $\hat{F}_n^B$, be a bootstrap regression sample from BLUS residuals. The bootstrap least squares estimate of β is

$$\hat{\beta}^*(B) = (X'X)^{-1}X'y^*(B), \tag{6}$$

and an estimate of the bootstrap variance is

$$[\hat{\sigma}_n^*(B)]^2 = n^{-1}\sum_{i=1}^{n}\left(\hat{\epsilon}_i^*(B) - n^{-1}\sum_{j=1}^{n}\hat{\epsilon}_j^*(B)\right)^2, \tag{7}$$

where $\hat{\epsilon}_j^*(B)$ is the j^{th} bootstrap residual, i.e., the j^{th} element of the vector $y^*(B) - X\hat{\beta}^*(B)$. This estimate has a small bias which is not important in this (asymptotic) treatment. Let $J_n(F)$ be the distribution of $\sqrt{n}(\hat{\beta}-\beta)$ where F is the distribution of the true errors ϵ so that $J_n(\hat{F}_n^B)$ is the distribution of $\sqrt{n}(\hat{\beta}^*(B)-\hat{\beta})$ (conditional on y). Also let $K_n(F)$ be the distribution of $(X'X)^{1/2}(\hat{\beta}-\beta)/\sigma$ so that $K_n(\hat{F}_n^B)$ is the conditional law of $(X'X)^{1/2}(\hat{\beta}^*(B)-\hat{\beta})/\hat{\sigma}^*$.

Under the conditions of Model 1, it is well known that $J_n(F)$ converges weakly to a multivariate normal with mean 0 and covariance matrix $\sigma^2 V$ and that $K_n(F)$ converges to a multivariate normal with mean 0 and covariance I_p. To show that the bootstrap is asymptotically valid, we need to show that the bootstrap distributions have the same asymptotic distribution in probability. The method of proof follows Freedman (1981) who used Mallow's metric to metrize weak convergence. Mallow's metric between two p-dimensional distributions G_1 and G_2 is defined as $d_\alpha^p(G_1, G_2) = \inf E[\|\upsilon - \tau\|^\alpha]^{1/\alpha}$ where $\|\cdot\|$ is the p-dimensional Euclidean norm and the infimum is taken with respect to all joint distributions of random variables υ and τ for which their marginal distributions are G_1 and G_2, respectively. (Whenever $p = 1$, we will omit the superscript.)

■ Suppose Model 1. Consider the BLUS bootstrap estimators of (6) and (7). Then

1. $d_2^p\{J_n(\hat{F}_n^B), J_n(F)\}^2 \xrightarrow{P} 0$;
2. $d_1\{\hat{\sigma}_n^*(B), \sigma\} \xrightarrow{P} 0$;
3. $d_2^p\{K_n(\hat{F}_n^B), K_n(F)\}^2 \xrightarrow{P} 0$.

Consider now the case where the number of independent variables p increases with the number of observations n. The case of bootstrapping from the ordinary (centered) residuals has been studied by Bickel and Freedman (1983). We generalize their results to the bootstrap from BLUS residuals. We begin with the distribution of normalized contrasts. Let c be a $p \times 1$ vector such that $c'(X'X)^{-1}c = 1$. We are interested in the distribution of the scalar $c'(\hat{\beta} - \beta)$ which is normalized to have variance σ^2. This can be used to construct a confidence interval for a coefficient of β. Let $\Psi_{npc}(F)$ be the distribution of $c'(\hat{\beta} - \beta)$ when the errors ϵ are distributed according to F. When p increases with n we need to modify Model 1 as follows.

Model 2 *Conditions a), b), and c) are as in Model 1, but condition d) is modified to*
d) The matrix $X_{(n)}^t X_{(n)}$ is positive definite.

Let s^2 be the usual estimate of the variance σ^2 given by

$$s^2 = (n-p)^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2, \tag{8}$$

and let $s^{*2}{}_{(\mathrm{B})}$ be the bootstrap estimate computed from the least squares bootstrap residuals when the bootstrap errors come from the centered BLUS residuals and defined as

$$s^{*2}{}_{(\mathrm{B})} = (n-p)^{-1}\sum_{i=1}^{n}(\hat{\epsilon}^*{}_{(\mathrm{B})})^2 \tag{9}$$

The next Theorem justifies the use of the bootstrap based on BLUS residuals for normalized contrasts as long as $p/n \to 0$. It is a generalization of Theorem 1.2 of Bickel and Freedman (1983).

■ Suppose Model 2. We condition with respect to $y_1, y_2, \ldots, y_n$. We suppose that $n \to \infty$ and $p/n \to 0$. Consider the standard deviation estimates s and $s^{*(\mathrm{B})}$ given by (8) and (9). Then

1. The d_2 distance between the conditional distribution of $c'(\hat{\beta}^{*(\mathrm{B})} - \hat{\beta})$ and the distribution of $c'(\hat{\beta} - \beta)$ converges to 0 in probability, uniformly in c.

2. The conditional distribution of $s^{*(\mathrm{B})}$ converges weakly to the point mass at σ, in probability.

3. The d_2 distance between the conditional distribution of $c'(\hat{\beta}^{*(\mathrm{B})} - \hat{\beta})/s^{*(\mathrm{B})}$ and the distribution of $c'(\hat{\beta} - \beta)/s$ converges to 0 in probability, uniformly in c.

Next, we look at the whole p-dimensional distribution. Let $\Psi_{np}(F)$ be the distribution of $(X'X)^{1/2}(\hat{\beta} - \beta)$ when the errors ϵ are distributed according to F.

■ Consider Model 2. If $p \to \infty$ while $p^2/n \to 0$, and $E\{d_2(F_n, F)^2\} = o(1/p)$, then

$$d_2[\Psi_{np}(F), \Psi_{np}(\hat{F}_n^B)] \xrightarrow{P} 0.$$

Remark 1 Weber (1984) introduced a bootstrap method based on "studentized" residuals $\hat{\epsilon}_i^W = \hat{\epsilon}_i(1 - h_{ii})^{-1/2}$ where h_{ii} is the i^{th} diagonal element of the hat matrix. He showed the equivalent of part 1 of Theorem 3. It can be shown that Parts 2 and 3, as well as Theorems 3 and 3 are also valid when the resampling is done using Weber's method.

Remark 2 Bickel and Freedman (1983) discuss a relationship between the tails of the distribution F and the assumption that $E\{d_2(F_n, F)^2\} = o(1/p)$.

4 Simulations

We have compared the small sample behavior of four different bootstrap methods in a simulation. Each method consists of resampling from one of the four different sets of residuals. The first three methods are described at the beginning of Section 3 and will be referred to as the methods [o], [BF], and [W] for the centered ordinary residuals, the centered standardized residuals of Bickel and Freedman (1983), and the centered studentized residuals introduced by Weber (1984). The fourth method (method [B]) consists of resampling from the $n - p$ centered BLUS residuals.

The observations in the simulation are generated from equation (1). The value of β is arbitrarily set to a vector of 1's. We consider $n \times p$ design matrices X of size $n = 20, 40, 60$ and $p = 4, 8, 16$. To make the different cases as comparable as possible, we have constructed design matrices such that the $p \times p$ matrix $S = (X'X)^{-1}$ is defined by $S_{ij} = \rho^{|i-j|}$. This ensures that the covariance matrix of $\hat{\beta}$ is the same for the different values of n when p is fixed, while keeping the same covariance structure as p changes. Moreover the variance of each least squares coefficient is σ^2. The value of ρ was arbitrarily set at 0.6. The QR decomposition was used to construct the fixed X matrices, the R matrix being determined by S while the Q matrix was computed from the first p columns of the $n \times 16$ Q matrix of the QR decomposition of an $n \times 16$ matrix of i.i.d. $N(0, 1)$ random variables.

The errors in the regression model were generated either from a $N(0, 4)$ or a Contaminated Normal distribution where with probability .9 the observation is $N(0, 1)$ and with probability .1, it is $N(0, 9)$. For each distribution, 1000 regression data sets were generated, and for each of them 100 bootstrap data sets of each method were generated in Splus.

For each data set and each bootstrap method, bootstrap estimates of the variance of $\hat{\beta}_i$, $i = 1, \ldots, p$ are computed by the usual formula, i.e., the sample variance of the 100 bootstrap estimates $\hat{\beta}_i^*$'s. The mean

squared error of the 1000 bootstrap estimates of the variance of $\hat{\beta}_i$ are computed for each $i = 1, \ldots, p$. Figures 4.1 and 4.3 show the average (over the p coefficients) mean squared errors (MSE) of the bootstrap estimates of the variance of $\hat{\beta}$ for each of the four bootstrap methods for the Normal and Contaminated Normal errors, respectively. The lines at the left are for $n = 20$ with $p = 16$, 8, or 4, while the lines in the center and at the right are for $n = 40$ and 60 respectively. Each linetype corresponds to a bootstrap method as explained in the legend.

We begin with the results for the normal distribution. >From Figure 4.1, we see that the average MSE of the variance of the regression coefficients decreases as $n - p$ increases for a given n as well as when n increases for a fixed p for all four methods. The behavior of the methods [BF], [W], and [B] are very similar, especially as n increases. On the other hand, the method [o] has a larger average MSE unless p is quite small. This is basically due to the large bias of [o] when p is large compared to n. Figure 4.2 shows boxplots of the 1000 bootstrap estimates of the variance of each least squares regression estimate for each of the four bootstrap methods when $n = 20$ and $p = 8$. This case is similar to all other cases. The bias of the ordinary bootstrap is obvious and its smaller variance is unable to compensate for its large bias.

The bias in estimating σ^2 can be computed exactly for the four methods. The bias of the BLUS bootstrap estimate of σ^2 is $\sigma^2/(n-p)$. For the other three bootstrap methods, the bias depends on the hat matrix although a simplification exists when a constant is included in the model (which is not the case in this simulation) for two of the methods. If a constant is included, then the bias of the ordinary bootstrap is $p\sigma^2/n$ whereas the standardized bootstrap estimate of Bickel and Freedman is unbiased. The bias of the studentized bootstrap estimate of σ^2 depends on the hat matrix, even in the case where an intercept is included in the model. But, in all the cases that we have studied, the bias is always extremely small.

To actually compare the different average MSE in Figures 4.1 and 4.3, we have performed two-sample paired t-tests when comparing the four bootstrap methods for a given n and p (since they are computed from the same original data sets) and two-sample t-tests when comparing any two methods for distinct pairs of n and p. For that reason, it is not simple to report standard errors for the estimates. For instance, for $n-p = 44$, the [BF] and [W] average MSE estimates are 1.011 and 1.021 with 95% confidence intervals of $[0.938, 1.082]$ and $[0.946, 1.094]$, respectively. But a paired two-sample 95% confidence interval for the difference between the two estimates is $[-0.0187, -0.0009]$ leading to a statistically significant difference!

Let us now compare the BLUS average MSE to the other three methods as illustrated in Figure 4.1. All tests were performed at the 5% level. The BLUS estimate is statistically different from the other three estimates for $n - p = 4$ and different from the [o] and [W] estimates for $n - p = 12$. The 95% confidence interval for the difference between the BLUS estimate and the [BF] estimate is of length 0.27. For $n - p = 16$, the BLUS estimate is not statistically different from the other three and the largest confidence interval for the difference between the average MSE for the BLUS and another estimate is of length 0.38. For $n-p$ equals 24, 32, 44, and 52, the ordinary estimate and the BLUS estimate have average MSE statistically different, but not for $n-p$ equal to 36 or 56. The largest 95% confidence interval between the BLUS estimate and the [BF] or [W] estimate for $n - p$ equal to 24, 32, 44, and 52 is of length .12, .11, .07, and .10, respectively. The largest 95% confidence interval between the BLUS estimate and the other three estimates for $n - p$ equal to 36 or 56 is of length .15 and .12. Overall, when the observations are normal, resampling BLUS centered residuals seems to give the best results; it is sometimes significantly better than the other methods and never worse. Even when very few BLUS residuals are available, the BLUS bootstrap method is not performing worse than the others.

The results for the contaminated normal distribution are a little bit different, as shown in Figure 4.3. Even though the average MSE decreases as n increases for fixed p, it need not decrease for fixed n as p decreases. We do not have any heuristic explanation for that. But note that when p changes, the dimension of the covariance matrix of $\hat{\beta}$ changes and even though we have tried to make the different covariance matrices as comparable as possible by keeping the same model of autocovariance of order 1, the covariance matrices are *not* directly comparable. Moreover, the variance of the bootstrap estimate of σ^2 involves fourth moments. As for the BLUS method, it continues to behave similarly to the [BF] and [W] methods. It is significantly different from them for $n - p = 4$ with the BLUS method doing better. In all other cases, they are not significantly different with the largest confidence interval for the difference

between the average MSE estimates of the BLUS and the [BF] or [W] methods being of length .15, .28, .05, .06, .05, .04, and .04, as $n-p$ increases from 12 to 56. The BLUS method is significantly worse than the ordinary bootstrap for $n-p$ equal to 16, 36, and 52, and significantly better for $n-p=44$. In all other cases, the difference between the [B] and [o] methods is not significantly different with confidence intervals of length .60, .33, .20, .10, and .04 for $n-p$ equal to 4, 12, 24, 32, and 56, respectively. Overall, the bias induced by resampling from the centered ordinary residuals is largely compensated by the smaller variance of this method in the case of the contaminated normal distribution which has heavy tails, leading in some cases to a better performance than the other methods. On the other hand, the less biased BLUS method performs similarly to the other two less biased methods and it never performs badly, even when few BLUS residuals are available.

5 Conclusion

The bootstrap can be applied in many different ways in a regression problem. The two main approaches consist of resampling pairs or resampling residuals and depend on whether we consider the X matrix to be random or fixed, e.g., Efron and Tibshirani (1993). But if one resamples residuals, which ones should be resampled? In this paper, we have introduced the use of the $n-p$ uncorrelated BLUS residuals. We have shown the asymptotic validity of their use. Moreover, in small samples they do as well, and sometimes better, than the standardized or studentized residuals of Bickel and Freedman (1981,83) and Weber (1984), respectively, In fact, even when only four BLUS residuals are available compared to 20 for the ordinary, standardized, or studentized residuals, the BLUS bootstrap gives bootstrap estimates of the variance of the regression coefficients with smaller mean squared error.

Appendix

To prove Theorem 3, we first need some notation and lemmas. Recall that F is the distribution function of the (true) errors and that $\hat{F}_n^B$ is the empirical distribution function (e.d.f.) of the centered BLUS residuals. Let F_n be the e.d.f. of the true errors $\epsilon_1,\ldots,\epsilon_n$. Let $\tilde{F}_n$ be the e.d.f. of the uncentered ordinary residuals while $\tilde{F}_n^B$ is the e.d.f. of the uncentered BLUS residuals.

Lemma 1 $d_2(\tilde{F}_n^B,\tilde{F}_n)^2 \xrightarrow{P} 0$.

Proof: Consider the following random variables. Let $\upsilon=\hat{\epsilon}_i$ with probability $1/n,\ i=i,\ldots,n$. If $\upsilon=\hat{\epsilon}_i$ for $i\geq(p+1)$, let $\tau=\hat{\epsilon}_{i-p}^B$, otherwise let $\tau=\hat{\epsilon}_i^B$ with probability $1/(n-p)$ for $i=1,\ldots,n-p$. In other words,

$$\tau=\begin{cases}\hat{\epsilon}_{i-p}^B & \text{if } \upsilon=\hat{\epsilon}_i, i=p+1,p+2,\ldots,n\\ \hat{\epsilon}_{U(1,\ldots,n-p)}^B & \text{if } \upsilon=\hat{\epsilon}_j, j=1,2,\ldots,p\end{cases}$$

where $U(1,\ldots,n-p)$ is a uniform random variable on the integers $1,\ldots,n-p$. Clearly, the marginal distributions of υ and τ are $\tilde{F}_n$ and $\tilde{F}_n^B$, respectively. By definition, $d_2(\tilde{F}_n^B,\tilde{F}_n)^2\leq E\left[(\upsilon-\tau)^2\,|\,y_1,y_2,\ldots,y_n\right]$.

Now

$$E\left[(\upsilon-\tau)^2\,|\,y_1,y_2,\ldots,y_n\right]=\frac{1}{n}\sum_{i=1}^{n-p}(\hat{\epsilon}_{i+p}-\hat{\epsilon}_i^B)^2+\frac{1}{n(n-p)}\sum_{i=1}^{p}\sum_{j=1}^{n-p}(\hat{\epsilon}_i-\hat{\epsilon}_j^B)^2. \tag{10}$$

Using Markov's inequality, we only need to prove that the expected value of the right hand side of (10) converges to 0 to prove the lemma.

Beginning with the first term in (10), we have

$$\begin{aligned}E\left[n^{-1}\sum_{i=1}^{n-p}(\hat{\epsilon}_{i+p}-\hat{\epsilon}_i^B)^2\right] &= n^{-1}E\left[(J^t\hat{\epsilon}-\hat{\epsilon}^B)^t(J^t\hat{\epsilon}-\hat{\epsilon}^B)\right]\\ &= n^{-1}E\left[(J^tM\epsilon-A\epsilon)^t(J^tM\epsilon-A\epsilon)\right]\\ &= n^{-1}trace\left[(A-MJ)(A-MJ)^tE(\epsilon\epsilon^t)\right]\\ &= n^{-1}\sigma^2trace\left\{(A^tA)+[(MJ)^tMJ]-(A^tMJ)-[(MJ)^tA]\right\}\end{aligned}$$

Note that $trace(A^tA) = trace(I_{n-p}) = (n-p)$ and $trace[(MJ)^tMJ] = trace(M_{11}) < (n-p)$. Since $AX = 0$, $MA = A$ and so $trace(A^tMJ) = trace(J^tA) = trace(A_1)$. Recall that $A_1 = PD^{1/2}P^t$ (equation (3)). Thus, $traceA_1 = traceD^{1/2}$. Let $d_i,\ i = 1, 2, \ldots, n-p$, be the diagonal elements of D. Theil (1965) has shown that D contains $(n-2p)$ eigenvalues equal to 1 and p eigenvalues less than 1. Hence,

$$\begin{aligned} E\left[n^{-1}\sum_{i=1}^{n-p}(\hat{\epsilon}_{i+p} - \hat{\epsilon}_i^B)^2\right] &< n^{-1}\sigma^2\left[2(n-p) - 2traceA_1\right] \\ &< \frac{2p\sigma^2}{n}\,. \end{aligned} \tag{11}$$

The expected value of the second term in (10) is:

$$\begin{aligned} E\left[\frac{1}{n(n-p)} \sum_{i=1}^{p}\sum_{j=1}^{n-p}(\hat{\epsilon}_i - \hat{\epsilon}_j^B)^2\right] &= \frac{1}{n(n-p)}\sum_{i=1}^{p}\sum_{j=1}^{n-p}E(\hat{\epsilon}_i - \hat{\epsilon}_j^B)^2 \\ &= \frac{1}{n(n-p)}\left[(n-p)\sum_{i=1}^{p}E(\hat{\epsilon}_i^2) + p\sum_{j=1}^{n-p}E(\hat{\epsilon}_j^B)^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{n-p}E(\hat{\epsilon}_i\hat{\epsilon}_j^B)\right]. \end{aligned} \tag{12}$$

Note that $E(\hat{\epsilon}_i^2) = (1-h_{ii})\sigma^2 < \sigma^2$ and $E(\hat{\epsilon}_j^B)^2 = \sigma^2$. Also let A_j be the j^{th} column of the matrix A and a_{ij} be its $(i,j)^{\text{th}}$ element and let M_i be the i^{th} row of M. With $MA = A$ we have

$$E(\hat{\epsilon}_i\hat{\epsilon}_j^B) = E(M_i\epsilon\epsilon^tA_j) = \sigma^2M_iA_j = \sigma^2a_{ij} \tag{13}$$

Since $A^tA = I$, $|a_{ij}| \leq 1$ for all i, j. Thus,

$$|E(\hat{\epsilon}_i\hat{\epsilon}_j^B)| < \sigma^2.$$

So, equation (12) is bounded by $4p\sigma^2/n$ and

$$\begin{aligned} E[d_2(\tilde{F}_n^B, \tilde{F}_n)^2] &\leq E\left[E\left((\upsilon - \tau)^2|y_1, y_2, \cdots, y_n\right)\right] \\ &< \frac{6p\sigma^2}{n}. \end{aligned} \tag{14}$$

Hence, as $n\to\infty$, $d_2(\tilde{F}_n^B, \tilde{F}_n)^2 \xrightarrow{P} 0$.

Lemma 2 $d_2(\hat{F}_n^B, \tilde{F}_n^B)^2 \xrightarrow{P} 0$.

Proof: Using Lemma 8.8 of Bickel and Freedman (1981), we have

$$\begin{aligned} d_2(\hat{F}_n^B, \tilde{F}_n^B)^2 &= d_2(\hat{F}_n^B, \tilde{F}_n^B - \hat{\mu}_n^B)^2 + \|\hat{\mu}_n^B\|^2 \\ &= d_2(\hat{F}_n^B, \hat{F}_n^B)^2 + \|\hat{\mu}_n^B\|^2 \\ &= \|\hat{\mu}_n^B\|^2 \\ &= \left(\sum_{i=1}^{n-p}\frac{\hat{\epsilon}_i^B}{n-p}\right)^2, \end{aligned}$$

where $\hat{\mu}_n^B = (n-p)^{-1}\sum_{i=1}^{n-p}\hat{\epsilon}_i^B$. Its expected value is

$$E[d_2(\hat{F}_n^B, \tilde{F}_n^B)^2] = \frac{1}{(n-p)^2}\sum_{i=1}^{n-p}Var(\hat{\epsilon}_i^B) = \frac{\sigma^2}{n-p} \tag{15}$$

and so $d_2(\hat{F}_n^B, \tilde{F}_n^B)^2 \xrightarrow{P} 0$.

Proof of Theorem 3, part 1: Using Theorem 2.1 of Freedman (1981), $d_2^p\{J_n(\hat{F}_n^B), J_n(F)\}^2 \leq n \cdot trace(X'X)^{-1} \cdot d_2(\hat{F}_n^B, F)^2$. Moreover, by assumption d) of Model 1, $n \cdot trace(X'X)^{-1}$ converges to a constant. So proving that $d_2(\hat{F}_n^B, F)^2 \xrightarrow{P} 0$ is sufficient.

Noting that $A \leq B + C$ implies $A^2 \leq 2B^2 + 2C^2$, then the triangle inequality implies that $d_2(F, G)^2 \leq 2d_2(F, H)^2 + 2d_2(G, H)^2$. Using this result repeatedly, we have that

$$\begin{aligned} d_2(\hat{F}_n^B, F)^2 &\leq 2d_2(\hat{F}_n^B, F_n)^2 + 2d_2(F_n, F)^2 \\ &\leq 4d_2(\hat{F}_n^B, \tilde{F}_n^B)^2 + 4d_2(\tilde{F}_n^B, F_n)^2 + 2d_2(F_n, F)^2 \\ &\leq 4d_2(\hat{F}_n^B, \tilde{F}_n^B)^2 + 8d_2(\tilde{F}_n^B, \tilde{F}_n)^2 + 8d_2(\tilde{F}_n, F_n)^2 + 2d_2(F_n, F)^2 \quad (16) \\ &\xrightarrow{P} 0, \quad (17) \end{aligned}$$

since all terms in (16) converge to 0 in probability, the first one by Lemma 2, the second one by Lemma 1, the third one by Lemma 2.1 of Freedman (1981) and the last one by Lemma 8.4 of Bickel and Freedman (1981).

Proof of Theorem 3, part 2: Let's define the following quantities. First let's recall the standard deviation of the bootstrap residuals $\hat{\epsilon}_i^*(B)$ given in (7),

$$\hat{\sigma}_n^{*(\mathrm{B})} = \left[n^{-1} \sum_{i=1}^{n} \left(\hat{\epsilon}_i^{*(\mathrm{B})} - n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i^{*(\mathrm{B})} \right)^2 \right]^{1/2}.$$

Consider the standard deviation of the bootstrap errors $\epsilon_i^*(B)$,

$$\sigma_n^{*(\mathrm{B})} = \left[n^{-1} \sum_{i=1}^{n} \left(\epsilon_i^{*(\mathrm{B})} - n^{-1} \sum_{i=1}^{n} \epsilon_i^{*(\mathrm{B})} \right)^2 \right]^{1/2}.$$

For the purpose of the proof we introduce the following quantity which is a function of the original (unobserved) errors,

$$\sigma_n = \left[n^{-1} \sum_{i=1}^{n} \left(\epsilon_i - n^{-1} \sum_{i=1}^{n} \epsilon_i \right)^2 \right]^{1/2}.$$

We first show that $d_1(\hat{\sigma}_n^{*(\mathrm{B})}, \sigma_n^{*(\mathrm{B})}) \to 0\, a.s.$ before showing that $d_1(\sigma_n^{*(\mathrm{B})}, \sigma_n) \xrightarrow{P} 0$. This will imply $d_1(\hat{\sigma}_n^{*(\mathrm{B})}, \sigma_n) \xrightarrow{P} 0$ and since $\sigma_n \to \sigma\, a.s.$, we will conclude that $d_1(\hat{\sigma}_n^{*(\mathrm{B})}, \sigma) \xrightarrow{P} 0$.

By definition,

$$d_1(\hat{\sigma}_n^{*(\mathrm{B})}, \sigma_n^{*(\mathrm{B})})^2 = inf E[\|U - V\| \,\big|\, y_1, y_2, \cdots, y_n]^2$$

where the infimum is taken with respect to all bivariate laws (U, V) whose marginal distributions are those of $\hat{\sigma}_n^{*(\mathrm{B})}$ and $\sigma_n^{*(\mathrm{B})}$, respectively. Writing the conditional moment as $E^*(\cdot)$ and computing $\hat{\sigma}_n^{*(\mathrm{B})}$ and $\sigma_n^{*(\mathrm{B})}$ from the same bootstrap errors $\epsilon_i^{*(\mathrm{B})}$, we have:

$$\begin{aligned} d_1(\hat{\sigma}_n^{*(\mathrm{B})}, \sigma_n^{*(\mathrm{B})})^2 &\leq E^*[(\hat{\sigma}_n^{*(\mathrm{B})} - \sigma_n^{*(\mathrm{B})})^2] \\ &\leq E^*[n^{-1} \sum_{i=1}^{n} \{\hat{\epsilon}_i^{*(\mathrm{B})} - \epsilon_i^{*(\mathrm{B})}\}^2] \quad (18) \\ &= (p/n)(\hat{\sigma}_n^B)^2, \quad (19) \end{aligned}$$

where inequality (18) comes from lemma 2.7 of Freedman (1981) and $\hat{\sigma}_n^B$ is the standard deviation of the BLUS residuals, i.e.,

$$\hat{\sigma}_n^B = \left[\frac{1}{(n-p)} \sum_{i=1}^{n} \left(\hat{\epsilon}_i^B - \frac{1}{(n-p)} \sum_{i=1}^{n} \hat{\epsilon}_i^B \right)^2 \right]^{1/2}.$$

Using properties of BLUS residuals, it is easy to show that $E[(\hat{\sigma}_n^B)^2] = [(n-p-1)/(n-p)]\sigma^2$. Using Markov's inequality, we show that $(p/n)(\hat{\sigma}_n^B)^2 \xrightarrow{P} 0$ and conclude that $d_1(\hat{\sigma}^*_{n(\mathrm{B})}, \sigma^*_{n(\mathrm{B})}) \xrightarrow{P} 0$.

Next we show that $d_1({\sigma_n^*}^2_{(\mathrm{B})}, \sigma_n^2) \xrightarrow{P} 0$. Note that

$$\begin{aligned} d_1[{\sigma_n^*}^2_{(\mathrm{B})}, \sigma_n^2] &\leq d_1\left[{\sigma_n^*}^2_{(\mathrm{B})}, n^{-1}\sum_{i=1}^{n} \epsilon_i^{*2}{}_{(\mathrm{B})} \right] + d_1\left[n^{-1}\sum_{i=1}^{n} \epsilon_i^{*2}{}_{(\mathrm{B})}, n^{-1}\sum_{i=1}^{n} \epsilon_i^2 \right] \\ &+ d_1\left[\sigma_n^2, n^{-1}\sum_{i=1}^{n} \epsilon_i^2 \right]. \end{aligned}$$

First,

$$\begin{aligned} d_1\left[{\sigma_n^*}^2_{(\mathrm{B})}, n^{-1}\sum_{i=1}^{n} \epsilon_i^{*2}{}_{(\mathrm{B})} \right] &\leq E\left[\left| n^{-1}\sum_{i=1}^{n} \epsilon_i^*{}_{(\mathrm{B})}{}^2 - \{n^{-1}\sum_{i=1}^{n} \epsilon_i^*{}_{(\mathrm{B})}\}^2 - n^{-1}\sum_{i=1}^{n} \epsilon_i^{*2}{}_{(\mathrm{B})} \right| \right] \\ &= (\hat{\sigma}_n^B)^2/n \xrightarrow{P} 0, \end{aligned}$$

since we have shown that $\hat{\sigma}_n^B/n \xrightarrow{P} 0$. Arguing as in the proof of Theorem 2.2 of Freedman (1981) and using results of Bickel and Freedman (1981), we can show that

$$\begin{aligned} d_1\left[n^{-1}\sum_{i=1}^{n} \epsilon_i^{*2}{}_{(\mathrm{B})}, n^{-1}\sum_{i=1}^{n} \epsilon_i^2 \right] &\leq \sum_{i=1}^{n} d_1[\epsilon_i^{*2}{}_{(\mathrm{B})}/n, \epsilon_i^2/n] \\ &= d_1(\epsilon_i^{*2}{}_{(\mathrm{B})}, \epsilon_i^2) \xrightarrow{P} 0, \end{aligned} \tag{20}$$

using Lemma 8.5 of Bickel and Freedman (1981) and since we have shown that $d_2(\hat{F}_n^B, F) \xrightarrow{P} 0$. Finally,

$$\begin{aligned} d_1\left[\sigma_n^2, n^{-1}\sum_{i=1}^{n} \epsilon_i^2 \right] &\leq E\left[\left| n^{-1}\epsilon_i^2 - \{n^{-1}\sum_{i=1}^{n} \epsilon_i\}^2 - n^{-1}\sum_{i=1}^{n} \epsilon_i \right| \right] \\ &= \sigma^2/n \to 0. \end{aligned}$$

Hence we have shown that $d_1[{\sigma_n^*}^2_{(\mathrm{B})}, \sigma_n^2] \xrightarrow{P} 0$. Using Lemma 8.5 of Bickel and Freedman (1981) one more time with $\phi(\sigma^2) = (\sigma^2)^{1/2}$, we get $d_1[\sigma^*_{n(\mathrm{B})}, \sigma_n] \xrightarrow{P} 0$.

Clearly $\sigma_n \to \sigma$ *p.p.* so that the proof of part 2 is complete.

Proof of Theorem 3, part 3: This is an immediate consequence of parts 1 and 2.

Proof of Theorem 3: We begin with part 1. Using Theorem 1.1 (b) of Bickel and Freedman (1983),

$$d_2[\Psi_{npc}(\hat{F}_n^B), \Psi_{npc}(F)]^2 \leq d_2(F, \hat{F}_n^B)^2. \tag{21}$$

Using equation (16), the right hand side of (21) is bounded by 4 terms. The fourth one does not depend on p and converges in probability to 0 by Lemma 8.4 of Bickel and Freedman (1981). The expected value of the first three terms can be bounded above using equations (15), and (14), as well as Lemma 2.1 of

Freedman (1981). In all three cases, the expected value is $O(p/n)$ and converges to 0 provided $p/n \to 0$, hence the result.

The proof of part 2 follows essentially from the proof of the corresponding part of Theorem 3, in particular equation (20), and the argument of the previous paragraph. Part 3 is an immediate consequence of parts 2 and 3.

Proof of Theorem 3: >From Theorem 1.1 (a) of Bickel and Freedman (1983),

$$\begin{aligned} d_2[\Psi_{np}(F), \Psi_{np}(\hat{F}_n^B)]^2 &\leq p \cdot d_2(F, \hat{F}_n^B)^2 \\ &\leq 2p \cdot [d_2(F, F_n)^2 + d_2(F_n, F_n^B)^2]. \end{aligned} \tag{22}$$

This is like the decomposition of equation (16), where the first three terms of (16) are bounding the first term of (22) while the fourth term of (16) and the second term of (22) are identical. Since the expected value of the first three terms of (16) is $O(p/n)$ and $E[d_2(F, F_n)^2]$ is $o(1/p)$ by hypothesis, then $d_2[\Psi_{np}(F), \Psi_{np}(\hat{F}_n^B)]^2 \to 0$ in probability provided $p \to \infty$ and $p^2/n \to 0$.

References

Bickel, P.J., and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statis.,* **9**, 1196–1217.

Bickel, P.J., and Freedman, D.A. (1983). Bootstrapping regression models with many parameters. *A Festschrift for Erich L. Lehman*, P.J. Bickel, K.A. Doksum, and J.L. Hodges eds, Wadsworth, Belmont (Calif.), 28–48.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.,* **7**, 1–26.

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

Freedman, D.A. (1981). Bootstrapping regression models. *Ann. Statist.,* **9**, 1218–1228.

Theil, H. (1965). The analysis of disturbances in regression analysis. *J. Amer. Statist. Assoc.,* **60**, 1067–1079.

Weber, N.C. (1984). On resampling techniques for regression models. *Statist. Probab. Lett.,* **2**, 275–278.