FAST DOUBLE BOOTSTRAP TESTS OF NONNESTED LINEAR REGRESSION MODELS

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

and

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

Key Words and Phrases: nonnested test; bootstrap test; J test.

JEL Classification: C12, C15, C20.

ABSTRACT

It has been shown in previous work that bootstrapping the J test for nonnested linear regression models dramatically improves its finite-sample performance. We provide evidence that a more sophisticated bootstrap procedure, which we call the fast double bootstrap, produces a very substantial further improvement in cases where the ordinary bootstrap does not work as well as it might. This FDB procedure is only about twice as expensive as the usual single bootstrap.

Final Version, January, 2001.

1. INTRODUCTION

The J test proposed by Davidson and MacKinnon (1981) is the most widely-used procedure for testing nonnested regression models; see McAleer (1995). Its popularity stems from the fact that it is conceptually simple and easy to compute. However, its finite-sample distribution can be very far from the standard normal distribution that it follows asymptotically. As a consequence, it often overrejects severely. A natural way to improve the finite-sample properties of the test is to bootstrap it, as Fan and Li (1995) and Godfrey (1998) were among the first to suggest.

In Davidson and MacKinnon (2002), we developed a theoretical approach which enabled us to show precisely what determines the finite-sample distribution of the J test. By using our theoretical results to design simulation experiments, we showed that, in most cases, the J test will perform very reliably in finite samples if it is bootstrapped. However, there can still be some cases in which the bootstrapped J test rejects noticeably more often than it should.

In Davidson and MacKinnon (2001), we developed a simple and inexpensive technique for computing bootstrap P values, which we called the "fast double bootstrap," or FDB. Like the double bootstrap proposed by Beran (1988), it leads to a theoretical improvement in the performance of bootstrap tests. Unlike the double bootstrap, it requires only about twice as much computation as the ordinary, or single, bootstrap. Although the FDB is not as widely applicable as the double bootstrap, it can be applied to a broad range of econometric tests, including the J test and other nonnested tests. In this paper, we develop the FDB J test and demonstrate, by means of simulations, that it works extraordinarily well.[1]

In the next section, we briefly describe the J test and discuss some standard ways in which it can be bootstrapped. In Section 3, we describe how the fast double bootstrap can be used to make the J test more reliable than the ordinary, or single, bootstrap J test. Then, in Section 4, we present some simulation results on the performance of the single bootstrap and FDB J tests. Section 5 concludes the paper.

2. THE J TEST

Consider two nonnested, linear regression models with IID errors:

$$H_1: \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma_1^2 \mathbf{I}), \text{ and}$$

$$H_2: \quad \boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{v}, \quad \boldsymbol{v} \sim \text{IID}(\boldsymbol{0}, \sigma_2^2 \mathbf{I}),$$

[1] Davidson and MacKinnon (2001) contains limited simulation results for FDB tests on the mean of a lognormal distribution and tests for omitted variables in a probit model. It does not discuss the application of the FDB to nonnested tests. The FDB is not discussed at all in Davidson and MacKinnon (2002).

where $\boldsymbol{y}$, $\boldsymbol{u}$, and $\boldsymbol{v}$ are n–vectors, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are, respectively, $n \times k$ and $n \times l$ matrices of regressors, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are, respectively, a k–vector and an l–vector of unknown parameters. The J statistic for testing H_1 is the ordinary t statistic for $\alpha = 0$ in the regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \alpha \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y} + \text{residuals}, \tag{1}$$

where $\boldsymbol{P}_{\boldsymbol{Z}} \equiv \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, so that $\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y}$ is the vector of fitted values from OLS estimation of H_2. Asymptotically, under regularity conditions that are given in Davidson and MacKinnon (1981), the J statistic is distributed as $N(0,1)$ under the null hypothesis H_1. In practice, the $t(n-k-1)$ distribution is often used for finite-sample inference, although there is, in general, no formal justification for doing so.

The J statistic for testing H_1 can be written as

$$\hat{J} = \frac{\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}}{\acute{s}^2 (\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y})^{1/2}}, \tag{2}$$

where $\acute{s}$ is the usual estimated standard error from regression (1), and $\boldsymbol{M}_{\boldsymbol{X}}$ is the projection matrix $\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Since $\hat{J}$ depends on the regressor matrices only through the projections $\boldsymbol{M}_{\boldsymbol{X}}$ and $\boldsymbol{P}_{\boldsymbol{Z}}$, it is invariant to any changes in $\boldsymbol{X}$ and $\boldsymbol{Z}$ that do not change the subspaces spanned by the columns of these matrices.

In order to bootstrap the J test, we need to generate B bootstrap samples from a DGP that approximates what H_1 would be if it had actually generated the data. The natural choice for $\boldsymbol{\beta}$ is the OLS estimator $\hat{\boldsymbol{\beta}}$. Then the j^{th} bootstrap sample will be

$$\boldsymbol{y}_j^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{u}_j^*, \tag{3}$$

where the elements of $\boldsymbol{u}_j^*$ can be simulated in various ways. One possibility would be to draw them from the $N(0, s^2)$ distribution, where s is the standard error of the regression H_1 estimated by OLS. However, this approach is based on the assumption that the error terms are normally distributed, which may be uncomfortably strong. Since we are using the bootstrap, it is natural to generate $\boldsymbol{u}^*$ by resampling from the vector

$$\tilde{\boldsymbol{u}} \equiv \left(\frac{n}{n-k}\right)^{1/2} \hat{\boldsymbol{u}}. \tag{4}$$

Here we have rescaled the residuals so that the average squared residual has expectation σ_1^2. In many applications of the bootstrap, this step is omitted. However, as we will see in Section 4, in the case of the J test it is very

important to resample from the rescaled residuals (4) rather than from the ordinary residuals $\hat{\boldsymbol{u}}$ when k/n is not small. There are other, more complicated, ways to rescale the residuals, which take into account the different leverage of each observation, but we have not observed any advantage to using them in this application.

In writing (3), we have implicitly assumed that all the regressors are exogenous. If there are lagged dependent variables, it will be necessary to generate the elements of the vector $\boldsymbol{y}_j^*$ recursively, using the value of the dependent variable in observation 0 to begin the recursion. In this case, the regressor matrix $\boldsymbol{X}$ will be different for every bootstrap sample, as will the regressor matrix $\boldsymbol{Z}$ if it too includes a lagged dependent variable. Of course, when it is H_1 that is being tested, the lagged dependent variable that appears in $\boldsymbol{Z}$ for the bootstrap samples must be generated from H_1.

Ideally, B should be a reasonably large number, and it should be chosen so that $\alpha(B+1)$ is an integer for all levels α of interest. One common choice is $B = 999$, although for a test as easy to compute as the J test, it might be reasonable to use an even larger number, such as 4999 or even 9999; see Davidson and MacKinnon (2000) for a practical way to choose B endogenously so as to minimize simulation error.

For the j^{th} bootstrap sample, $j = 1, \ldots, B$, a bootstrap test statistic J_j^* is computed in exactly the same way as $\hat{J}$ was computed from the original data. Then we can compute a bootstrap P value by the formula

$$\hat{p}^*(\hat{J}) = \frac{1}{B} \sum_{j=1}^{B} I(J_j^* \geq \hat{J}), \tag{5}$$

where $I(\cdot)$ is an indicator function, equal to 1 if its argument is true and equal to 0 otherwise. This assumes that the test is a one-tailed test which rejects in the upper tail, as is usually the case with the J test and other nonnested tests. For a two-tailed test, the indicator function in (5) would become $I(|J_j^*| \geq |\hat{J}|)$.

Since the statistic $\hat{J}$ is not pivotal, a bootstrap test based upon it will not be exact. The problem is that the true P value depends on the unknown true distribution of $\hat{J}$, while the bootstrap P value (5) is based on the distribution of the bootstrap statistics J_j^*, which depends on the bootstrap DGP. These two distributions will differ whenever a test statistic is not pivotal and the parameter estimates used in the bootstrap DGP differ from the true values of the parameters. However, as Beran (1988) showed, provided the test statistic is asymptotically pivotal, the bootstrap P value will converge to the true P value, as the sample size increases, at a rate faster than the asymptotic P value converges to it.

In Davidson and MacKinnon (2002), we derived an expression for $\hat{J}$ as a function of various random variables and quantities that depend on $\boldsymbol{X}$,

$\boldsymbol{Z}$, and the parameters of H_1, under the assumptions that the error terms are normally distributed and the regressor matrices are exogenous. This expression is rather complicated, but a key determinant of the finite-sample distribution of $\hat{J}$ turned out to be the quantity

$$\|\boldsymbol{\theta}\|^2 \equiv \|\boldsymbol{M_X P_Z X\beta}\|^2/\sigma_1^2. \tag{6}$$

The numerator of (6) is the squared length of the part of $\boldsymbol{X\beta}$ that is explained by $\boldsymbol{Z}$, projected off $\boldsymbol{X}$. The denominator is just the variance of the error terms. Other things being equal, the larger is $\|\boldsymbol{\theta}\|^2$, the closer the finite-sample distribution of $\hat{J}$ will be to the standard normal distribution. Although this result depends on the normality assumption, simulation results in Davidson and MacKinnon (2002) suggest that (6) affects the finite-sample distribution of $\hat{J}$ in the same way even when this assumption does not hold.

3. THE FAST DOUBLE BOOTSTRAP

The fast double bootstrap of Davidson and MacKinnon (2001) can be thought of as an approximation to the double bootstrap of Beran (1988). It involves calculating two different bootstrap statistics for each replication. These are based on two different bootstrap datasets drawn from two different bootstrap DGPs. In the case of the J test, the first bootstrap DGP is the one already described, in which, for the j^{th} replication, a bootstrap sample $\boldsymbol{y}_j^*$ is drawn from (3), with the error terms obtained by resampling from the vector $\tilde{\boldsymbol{u}}$ defined in (4). This vector is used to compute a statistic J_j^*. For the FDB procedure, these data are also used to compute estimates $\hat{\boldsymbol{\beta}}_j^*$ and rescaled residuals $\tilde{\boldsymbol{u}}_j^*$, which characterize the second bootstrap DGP. A second bootstrap sample $\boldsymbol{y}_j^{**}$ is then drawn from this DGP in precisely the same way as the first bootstrap sample was drawn from the DGP characterized by $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$, and this sample is used to compute a statistic J_j^{**}.

The fast double bootstrap P value is easily calculated from the actual test statistic $\hat{J}$, the B first-level bootstrap test statistics J_j^*, and the B second-level bootstrap test statistics J_j^{**}. We first calculate the single-bootstrap P value $\hat{p}^*$ using expression (5). Next, we calculate the $1-\hat{p}^*$ quantile of the J_j^{**}, denoted by $\hat{Q}^*(1-\hat{p}^*)$ and defined implicitly by the equation

$$\frac{1}{B}\sum_{j=1}^{B} I\big(J_j^{**} > \hat{Q}^*(1-\hat{p}^*)\big) = \hat{p}^*. \tag{7}$$

Of course, for finite B, there will be a range of values of $Q^*(1-\hat{p}^*)$ that satisfy (7), and we will need to choose one of them in a somewhat arbitrary manner. Then the FDB P value is

$$\hat{p}^{**} = \frac{1}{B}\sum_{j=1}^{B} I\big(J_j^{*} > \hat{Q}^*(1-\hat{p}^*)\big). \tag{8}$$

Thus, instead of seeing how often the bootstrap test statistics are more extreme than the actual test statistic, we see how often they are more extreme than the $1 - \hat{p}^*$ quantile of the J_j^{**}.

The intuition behind this procedure is as follows. Suppose, for concreteness, that the J_j^{**} tend to be less extreme than the J_j^*. This suggests that the J_j^* will tend to be less extreme than they would be if they were drawn from the true unknown DGP rather than from the bootstrap DGP. Therefore, the ordinary bootstrap P value will be too small, and the bootstrap test will overreject. In this situation, $\hat{Q}(1 - \hat{p}^*)$ will be less extreme than $\hat{J}$ itself, and $\hat{p}^{**}$ will consequently be larger than $\hat{p}^*$. Thus it appears that using $\hat{p}^{**}$ instead of $\hat{p}^*$ will be a step in the right direction.

The properties of $\hat{p}^{**}$, not specialized to the J test, were studied in Davidson and MacKinnon (2001). It is valid under quite weak conditions, but it can be expected to be more accurate than $\hat{p}^*$ only when the bootstrap DGP is asymptotically independent of the test statistic. Since the quantities on which the bootstrap DGP depends (namely, $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$) are either efficient estimates under the null hypothesis that $\alpha = 0$ or functions of those estimates, the J test statistic must be asymptotically independent of them; see Davidson and MacKinnon (1999). Therefore, the theory suggests that, for the J test, $\hat{p}^{**}$ will be more accurate than $\hat{p}^*$. Simulation results to be presented in the next section strongly support this conjecture.

4. EVIDENCE FROM MONTE CARLO EXPERIMENTS

In this section, we present some simulation results for models deliberately constructed to make the bootstrap J test work relatively poorly. Our results should not be considered at all typical for J tests computed using real data. We consider a pair of linear regression models of the form

$$H_1: \quad y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \delta_1 y_{t-1} + u_t \tag{9}$$

$$H_2: \quad y_t = \boldsymbol{Z}_t\boldsymbol{\gamma} + \delta_2 y_{t-1} + v_t, \tag{10}$$

where the error terms for H_1, which is assumed to have generated the data, are $t(5)$ rescaled to have variance σ_1^2. The first elements of $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$ are unity, and their dimensions are $k - 1$ and $l - 1$, respectively. In the first set of experiments, the components of $\boldsymbol{X}_t$, except for the constant term, were distributed as $N(0, 1)$. Each component of $\boldsymbol{Z}_t$ was also normally distributed and correlated with the corresponding component of $\boldsymbol{X}_t$, with correlation ρ. When $k > l$ or $l > k$, any extra components of $\boldsymbol{X}_t$ or $\boldsymbol{Z}_t$ were uncorrelated with everything else. We experimented with the choice of δ_1, σ_1, ρ, k, and l. We found that the values of δ_1 and ρ had relatively little effect on the performance of the bootstrap J test, and we settled on base-case values of $\delta_1 = 0.8$ and $\rho = 0.5$. Since the model is assumed to be stationary, it does not make

sense to generate bootstrap samples using $|\hat{\delta}_1| \geq 1$. We therefore replaced $\hat{\delta}_1$ by -0.99 when $\hat{\delta}_1 < -0.99$ and by 0.99 when $\hat{\delta}_1 > 0.99$.

In the main set of experiments, the results of which we report here, $k = l = 7$, all the β_i are equal to 1, and σ_1 takes on the values 1, 2, 4, and 8. Because there are five variables in each model that are not in the other, and the sample sizes we investigate are small, the asymptotic J test does not work particularly well. As σ_1 increases, $\|\boldsymbol{\theta}\|$ becomes smaller, and the performance of both the asymptotic and bootstrap J tests deteriorates.

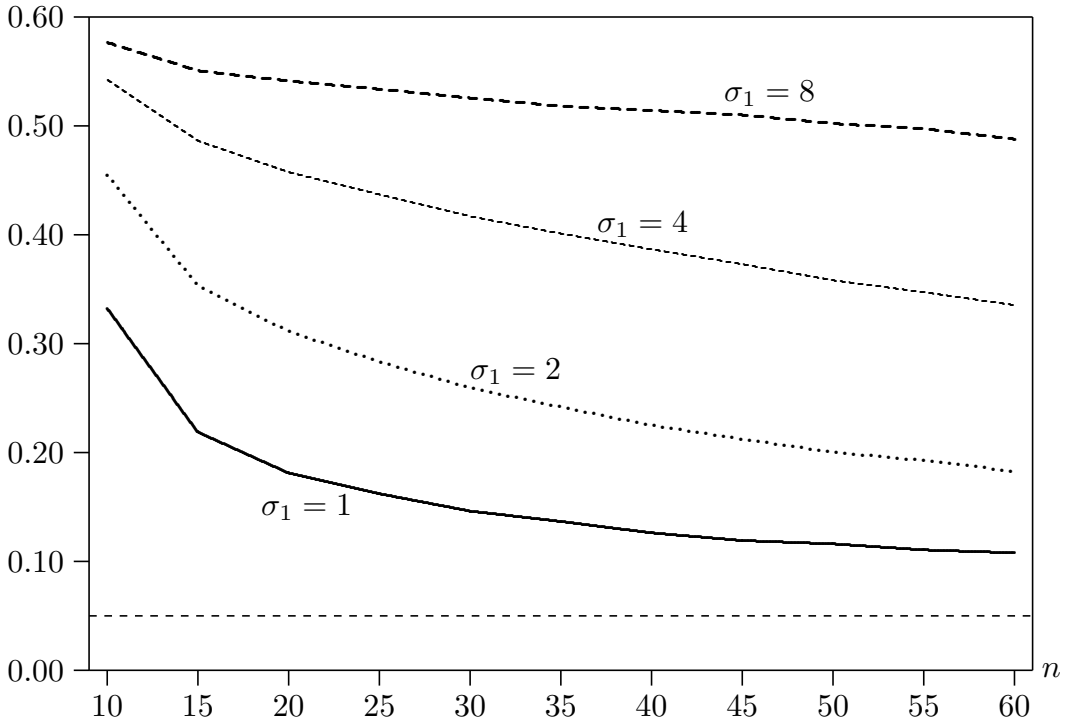

FIG. 1 Rejection frequencies for asymptotic tests

In order to limit experimental randomness, which would make it hard to detect small departures from the nominal level of the tests, each experiment used 100,000 replications. This choice implies that, if the true rejection frequency is .05, the standard error of the estimated rejection frequency will be .00069. The number of bootstrap samples was always 999, which is the smallest number that we would recommend using in practice.

Rejection frequencies for the asymptotic J test (based on the standard normal distribution) at the nominal .05 level for eleven sample sizes ($n = 10, 15, \ldots, 55, 60$) are shown in Figure 1. It is evident that the asymptotic J test overrejects very severely indeed, especially for the larger values of σ_1.

The larger is σ_1, the less rapidly does the performance of the test improve as the sample size increases. The figure suggests that, for the two largest values of σ_1, the sample size would have to be very large indeed for the asymptotic J test to perform well.

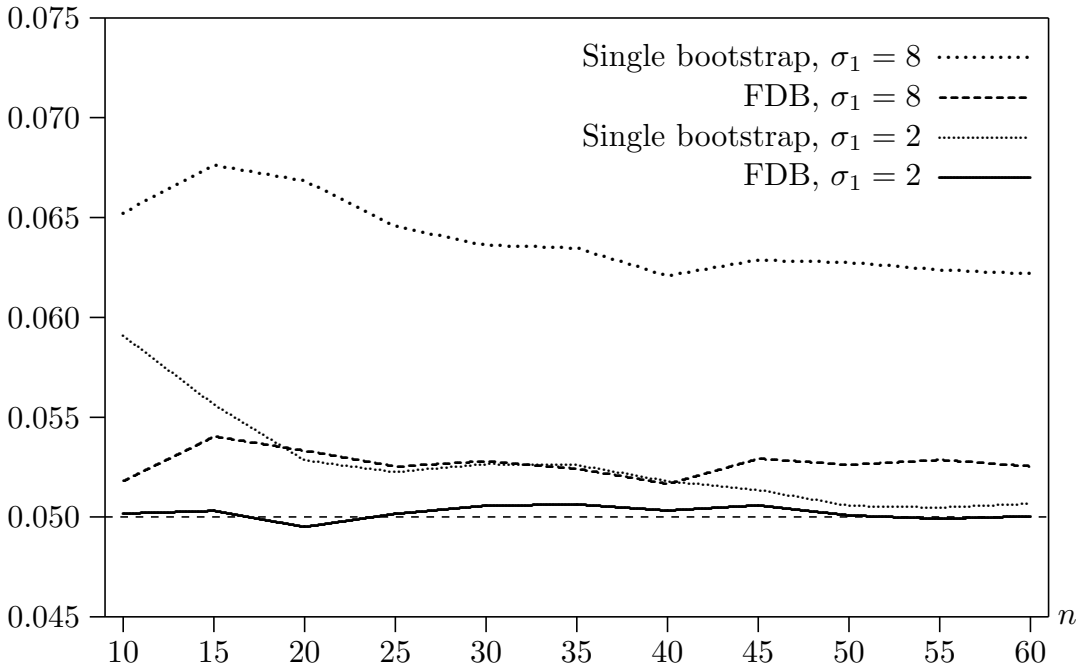

FIG. 2 Rejection frequencies for bootstrap tests

When $\sigma_1 = 1$, the ordinary (or single) bootstrap test worked almost perfectly, and the FDB test worked even better. These results are therefore not reported. Results for $\sigma_1 = 2$ and $\sigma_1 = 8$, which are much more interesting, are shown in Figure 2. When $\sigma_1 = 2$, the single bootstrap test overrejects quite noticeably for small sample sizes, but its performance improves rapidly as n increases. In contrast, the FDB test performs about as well as any test could be expected to perform, given some experimental error, for all sample sizes. When $\sigma_1 = 8$, the single bootstrap test overrejects for all sample sizes. The FDB test also overrejects, but very much less severely. The performance of both tests, mirroring that of the asymptotic test, improves very slowly as n increases beyond about 30.

We remarked in the previous section that it is very important to rescale the residuals prior to resampling from them. The consequences of not doing so are shown in Figure 3, which deals with the same cases, and uses the same random numbers, as Figure 2. We see that the single bootstrap performs dramatically worse, especially for smaller sample sizes, when the

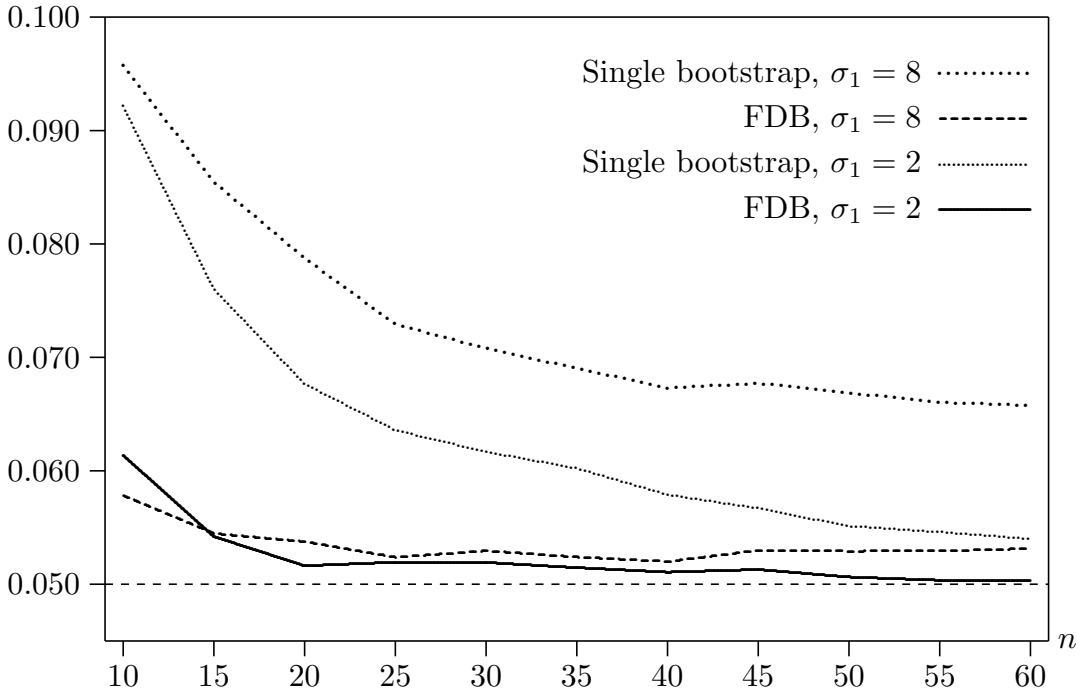

FIG. 3 Bootstrap rejection frequencies, ordinary residuals

residuals are not rescaled. In contrast, the FDB procedure performs only a little worse for the smaller sample sizes, and its performance is almost unchanged for the larger ones. Thus it appears that the correction implicit in the FDB procedure can compensate for flaws in the underlying method of bootstrapping.

In the experiments discussed so far, both the regressors and the error terms were symmetrically distributed. It seems likely that the bootstrap will perform less well when this is not the case; see Hall (1992). To investigate this possibility, we conducted a second set of experiments in which the components of $\boldsymbol{X}_t$, except the first, were $\chi^2(2)$ variates, recentered to have mean zero and rescaled to have variance unity. In these experiments, the second through $(k-1)^{\text{th}}$ components of $\boldsymbol{Z}_t$ were linear combinations of the corresponding components of $\boldsymbol{X}_t$ and of independent recentered and rescaled $\chi^2(2)$ variates. These components of $\boldsymbol{Z}_t$ were constructed in such a way that they had mean zero, variance unity, and correlation 0.5 with the corresponding components of $\boldsymbol{X}_t$. As before, the final components of $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$ were lagged dependent variables, the coefficients on which were 0.8.

Figure 4 shows the results of three sets of experiments in which the regressors were as just described and $\sigma_1 = 4$. In the first set of experiments, described as "highly skewed errors" in the figure, the error terms were $\chi^2(2)$

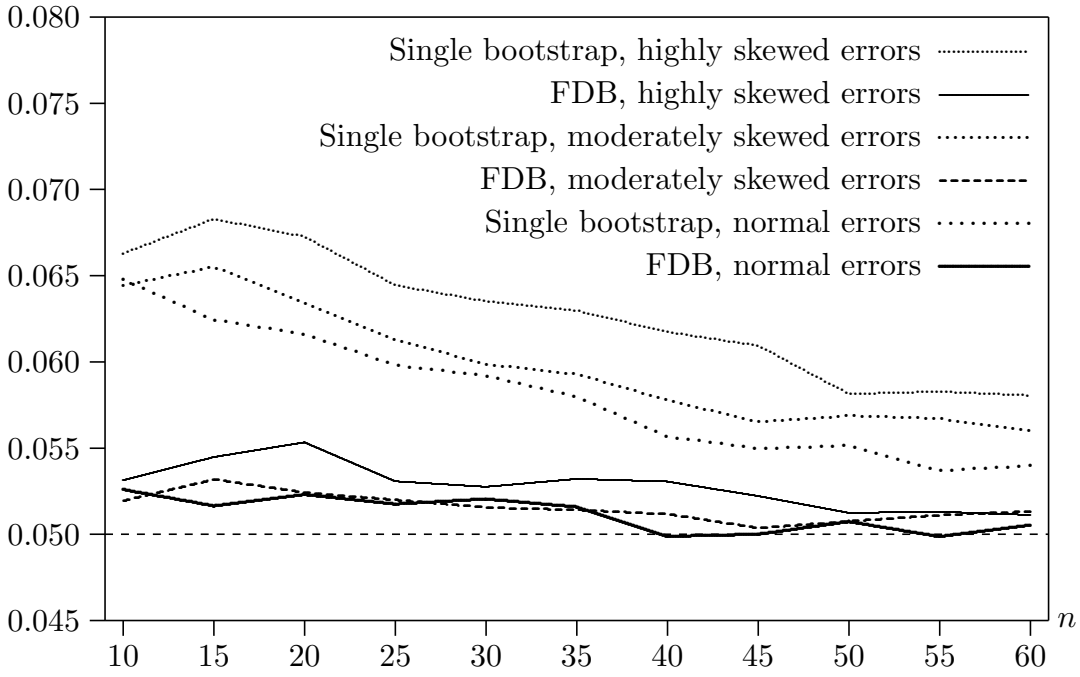

FIG. 4 Bootstrap rejection frequencies, skewed regressors

variates, recentered to have mean zero and rescaled to have variance unity. In the second set, described as "moderately skewed errors," they were $\chi^2(8)$ variates, similarly recentered and rescaled. In the third set, they were normally distributed. It is evident that the performance of both the single bootstrap and the FDB deteriorates somewhat as the error terms become more skewed. However, the FDB continues to perform remarkably well. The worst performance by the FDB, in the "highly skewed" case, is much better than the best performance by the single bootstrap, in the normal case.

In one final set of experiments, we made the experimental design even more extreme, making the situation very unfavorable for the finite-sample performance of the J test. We used the same skewed regressors as in the experiments just discussed, and the error terms were recentered and rescaled $\chi^2(2)$ variates. We set $k = 8$ and $l = 9$ (the largest values that allow calculation of the J test for $n = 10$), and we set $\sigma_1 = 16$. In these experiments, the values of $\|\boldsymbol{\theta}\|^2$ ranged from 0.0038 (for $n = 10$) to 0.2624 (for $n = 60$), and the rejection frequencies for the asymptotic tests at the .05 level ranged from 0.81 to 0.73. Results for the single bootstrap and FDB tests, with both rescaled and ordinary residuals, are shown in Figure 5. The FDB procedures do not perform perfectly, but they always work very much better than the corresponding single bootstrap procedures. Curiously, except when $n = 10$,

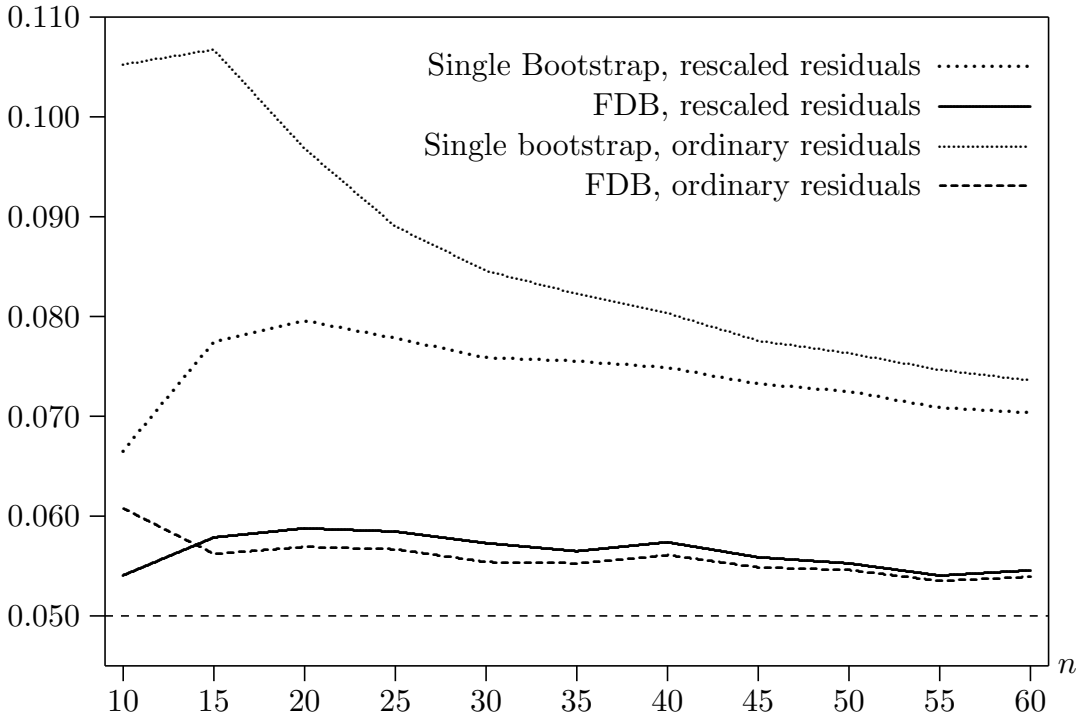

FIG. 5 Bootstrap rejection frequencies, extreme case

the FDB procedure that uses ordinary residuals always performs slightly better than the FDB procedure that uses rescaled residuals.

5. CONCLUSION

In this paper, we have proposed a simple bootstrap procedure for the J test of nonnested linear regression models that works extraordinarily well, even in extreme cases where the usual single bootstrap procedure overrejects quite noticeably. Our FDB procedure may be used in place of the usual one, or it may be used in addition to it in cases where the usual bootstrap P value is near the level of the test. Our simulation experiments deliberately focused on extreme cases in which the asymptotic test often rejects more than half the time and the single bootstrap test does not work very well. In practice, we would expect the single bootstrap to work extremely well, and our FDB procedure to work nearly perfectly, in virtually every case that an econometrician would be likely to encounter.

ACKNOWLEDGEMENTS

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to two anonymous referees for comments on an earlier version.

REFERENCES

Beran, R. (1988). "Prepivoting test statistics: a bootstrap view of asymptotic refinements," *J. Amer. Statist. Ass.*, 83, 687–697.

Davidson, R. and J. G. MacKinnon (1981). "Several tests for model specification in the presence of alternative hypotheses," *Econometrica*, 49, 781–793.

Davidson, R. and J. G. MacKinnon (1999). "The size distortion of bootstrap tests," *Econometric Theory*, 15, 361–376

Davidson, R. and J. G. MacKinnon (2000). "Bootstrap tests: How many bootstraps?," *Econometric Rev.*, 19, 55–68.

Davidson, R. and J. G. MacKinnon (2001). "Improving the reliability of bootstrap tests," Queen's University Institute for Economic Research Discussion Paper No. 995, revised.

Davidson, R. and J. G. MacKinnon (2002). "Bootstrap J tests of nonnested linear regression models," *J. Econometrics*, 109, 167–193.

Fan, Y., and Q. Li (1995). "Bootstrapping J-type tests for non-nested regression models," *Econ. Letters*, 48, 107–112.

Godfrey, L. G. (1998). "Tests of non-nested regression models: Some results on small sample behaviour and the bootstrap," *J. Econometrics*, 84, 59–74.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, New York, Springer-Verlag.

McAleer, M. (1995). "The significance of testing empirical non-nested models," *J. Econometrics*, 67, 149–171.