THE KOREAN ECONOMIC REVIEW
Volume 20, Number 1, Summer 2004

AN ENDOGENEITY-CORRECTED BOOTSTRAP TEST ON INSTRUMENT RELEVANCE IN INSTRUMENTAL VARIABLES ESTIMATION*

JINOOK JEONG**

It is well known that IV estimation produces considerable bias when the instruments are irrelevant. Previous studies have suggested several instrument screening tests to avoid such bias. In this paper, an LR test based on the exact finite sample distribution of R^2 is proposed for a more powerful instrument screening test. It is also analyzed how the degree of endogeneity affects bias in IV estimation. Then, a nonparametric instrument screening test with endogeneity adjustment is suggested. The finite sample performance of the new test is evaluated in a Monte Carlo simulation.

JEL Classification: C12, C14, C15
Keywords: IV Estimation, Instrument Relevance, Bootstrap, Endogeneity

I. INTRODUCTION

Instrumental Variables (IV) estimation has been an important econometric technique for decades. While the IV estimator is consistent and asymptotically normal, its finite sample properties are not flawless. As Nelson and Startz (1990b) and Maddala and Jeong (1992) have shown, especially when the instrument is weakly correlated with the endogenous variable, the IV estimator is considerably biased in small samples.[1] Nelson and Startz (1990a) propose that a

Received for publication: Jan. 16, 2004. *Revision accepted:* May 3, 2004.
* Valuable comments from Hashem Dezhbakhsh, Alastair Hall, G.S. Maddala, Charles Nelson, Joon Y. Park, B. Sam Yoo, the participants in the KEA seminar, and anonymous referees are gratefully acknowledged. This research was financially supported by '95 Research Fund, Ministry of Education, Republic of Korea.
** Department of Economics, Yonsei University, Seoul, Republic of Korea 120-749, Phone: 82-2-2123-2493, Fax: 82-2-313-5331, Email: jinook@yonsei.ac.kr
[1] Staiger and Stock (1997) derive the asymptotic properties of IV estimator using weak (local to zero) instrument. They show the IV estimator is badly biased even in large samples, and suggest to use LIML estimator that is approximately median unbiased.

pretest of instrument relevance be done to avoid erroneous inference due to biased IV estimator. They advise, in the case of one-regressor, one-instrument model, not to use the instrument if R^2 is less than (2/n), where R^2 is the multiple correlation coefficient between the endogenous variable and the instrument, and n is the number of observations.

Bound et al. (1993) extend Nelson-Startz test to the case of multiple instruments. They suggest the quality of IV estimator to be evaluated by the F statistic and R^2 of the 'first stage' regression (regression of the endogenous variable on the multiple instruments). Shea (1997) considers a case of multiple-regressors and multiple-instruments. When there exist multiple endogenous variables, the simple R^2 between each endogenous variable and the instruments can be misleading due to collinearity between endogenous variables. He proposes a partial R^2, which is a multicollinearity-removed R^2 between the endogenous variables and the instruments. Hall et al. (1996) also generalize Nelson-Startz's criterion to multiple-regressor, multiple-instrument case. Naturally, for the place of simple correlation, the canonical correlations between the endogenous variables vector and the instruments vector are used to estimate the relevance of instruments. The Likelihood Ratio (LR) test by Fujikoshi (1974) is used to test if the canonical correlations are all zero. The approach by Hahn and Hausman (2002) is yet different. They develop a test based on the so-called Durbin-Hausman-Wu specification test. The test examines if there exists significant difference between forward 2SLS estimator and reverse 2SLS estimator. Stock, Wright and Yogo (2002) provides with a useful overview of the tests on weak instruments.

In this paper, I will identify two problems in the previous instrument screening tests. First, all the above instrument screening tests use R^2 as the test statistic.[2] While the finite sample distribution of R^2 is well known under normality assumption, no previous tests are based on the distribution of R^2. Except Hall et al. (1996), all the previous tests mentioned above are rather conventional tests based on experience and intuition.[3] The test by Hall et al. (1996) is not a conventional rule. It is, however, based on the *asymptotic* distribution of R^2 (= squared canonical correlation), not the exact finite sample distribution of R^2.[4] Because the exact finite sample distribution of R^2 is available in closed form, it is straightforward to devise an instrument screening test based on the distribution of R^2. I will propose a new LR test based on the exact distribution of R^2 and show that the proposed test is more powerful

[2] Hall et al. (1996) is an exception. However, in the special case of single regressor, the LR test by Hall et al. (1996) is also based on R^2.

[3] This point will be elaborated in section IV.

[4] Although the test by Hall et al. (1996) can be applied to a more general case than a single-regressor case, we consider a single-regressor case only for convenience of comparison.

than the previous tests.

Second, the previous tests are too lenient for some bad instruments. Because the earlier tests examine only the correlation between the instrument and the endogenous regressor without considering other factors affecting the performance of IV estimator, some unqualified instruments would easily pass the screening test. In other words, the pretests accept any instrument that has non-zero correlation with the endogenous variable, although a stronger correlation is often demanded because of tougher estimation environment. One such factor is the degree of endogeneity of the regressor. In this paper, I will identify how the magnitude of endogeneity affects the bias of IV estimator, and I will suggest an adjustment of the LR test I propose. I will also show, through a Monte Carlo simulation, how the new screening test improves the performance of IV estimation.

The puzzling phenomenon demonstrated by Hall et al. (1996) can be partly due to this problem of ignoring endogeneity effect. Hall et al. (1996) show through a Monte Carlo study that, even though their LR test quite successfully detects weak instruments, such pre-estimation screening of instruments may induce even more erroneous statistical inference on IV estimator. They examine the empirical size of the usual t-test on the regression parameter before and after their IV relevance pretest. The result shows the empirical size of t-test becomes even worse when the estimation used only those instruments that passed the pretest: the cure is worse than the disease!

This phenomenon can be explained by two reasons. First, it is basically a choice-based sampling problem, which is inevitable with any two-stage test procedures due to cumulated Type I Error. Because the Type I Error of the pre-estimation test is carried over to the second stage of the estimation, the conditional size of the second stage test (conditional on the first stage test result) becomes less accurate than the unconditional size. Second, as explained earlier, the IV relevance pretests are too lenient so that the unsatisfactory instruments may have deteriorated the performance of IV estimator.

In the next section, I will first derive the exact distribution of the sampling error $\hat{\beta}_{iV}-\beta$ in IV estimation. Then the effect of endogeneity on the sampling error will be shown in section III. Section IV will derive an LR test based on the exact distribution of R^2. An endogeneity adjustment for the proposed LR test and a bootstrap test procedure will be suggested in section V, and a Monte Carlo comparison will be presented in section VI. Section VII will summarize and conclude.

II. EXACT DISTRIBUTION OF THE SAMPLING ERROR OF IV ESTIMATOR

Let us consider the following simple model.

$$y=\beta x+u \tag{1}$$

$$x=\delta z+v \tag{2}$$

where β and δ are scalar and x and z are in deviations from their means. Assume u_i and v_i are serially independent and $[\ u_i\ \ v_i\ ]' \sim N(0,\ \Sigma)$ for all $i=1,2,\ldots,n$, with

$$\Sigma=\begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix} \tag{3}$$

Also assume z is 'exogenous': $E(z'u)=E(z'v)=0$. The IV estimator, $\hat{\beta}_{iV}$ is $\hat{\beta}_{IV=}(z'x)^{-1}(z'y)$, and the sampling error is $(\hat{\beta}_{IV}-\beta)=(z'x)^{-1}(z'u)$. The bias of IV estimator is the expected value of the sampling error.

According to the previous works, it is necessary to pretest the relevance of instrument, because the bias of IV estimator could be far from zero in small sample when the instrument is poor, that is, when δ is close to zero. To evaluate this statement, we need to derive the exact small sample distribution of the sampling error (or the IV estimator).[5] The exact small sample distribution of the IV estimator with multiple-variables and multiple-instruments has been derived by many authors, using infinite series of gamma distributions. Among others, Sawa (1969), Mariano (1973), and Mariano and McDonald (1979) have derived the small sample distribution for exactly identified model. Phillips (1983) has derived a more general distribution of IV estimator for the case of overidentification. Staiger and Stock (1997) have developed a different approach. They have derived an alternative asymptotic approximation of IV estimator's distribution for 'nearly unidentified' case, using local-to-zero asymptotics. Their approximation is very close to the finite sample distribution even when the sample size is as small as 20 and keeps the advantages of asymptotic approximation: no distributional assumption required, computationally tractable, etc.[6]

However, for the above single-regressor-single-instrument model, the exact distribution of the IV estimator, and accordingly of the sampling error, can be derived through much simpler direct method, without employing any infinite series of confluent hyper-geometric function. As shown in Nagar (1959) and Hinkley (1969), the density function of the sampling error (B) is as follows:[7]

[5] Because the sampling error is only (IV estimator - constant), the sampling error has the same distribution as IV estimator after a horizontal shift.

[6] Actually, the test procedure suggested in the current study also has the same advantages because it utilizes bootstrap method for which no distributional assumption or algebraic computation is necessary.

[7] The derivation and notation for this specific model is in Appendix. I am grateful to Charles

$$f(B)=\frac{\sqrt{\sigma_u^2\sigma_v^2-\sigma_{uv}^2}}{\pi(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)}\exp\left\{-\frac{\delta^2(z'z)\sigma_u^2}{2(\sigma_u^2\sigma_v^2-\sigma_{uv}^2)}\right\}$$
$$+\frac{\delta(z'z)(\sigma_u^2-\sigma_{uv}B)}{\sqrt{2\pi(z'z)(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)^3}}[\Phi(\Delta)-\Phi(-\Delta)]\times$$
$$\exp\left\{\frac{\delta^2(z'z)[(\sigma_u^2-\sigma_{uv}B)^2-\sigma_u^2(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)]}{2(\sigma_u^2\sigma_v^2-\sigma_{uv}^2)(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)}\right\} \tag{4}$$

where Φ is the standard univariate normal distribution function, and

$$\Delta\equiv\frac{\delta(z'z)(\sigma_u^2-\sigma_{uv}B)}{\sqrt{(\sigma_u^2\sigma_v^2-\sigma_{uv}^2)(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)(z'z)}} \tag{5}$$

Ⅲ. INSTRUMENT RELEVANCE AND DEGREE OF ENDOGENEITY

Since the density function is quite complicated even in this simple model, it is extremely difficult to algebraically examine the effect of instrument relevance (and other factors) on the bias, the expected value of the sampling error. Numerical computation, however, will show the response path of the bias to various factors affecting it. In addition to instrument relevance (ρ_{xz}), the factors of interest here are degree of endogeneity (ρ_{xu}), and sample size (n). For numerical computation, the following nuisance parameters were pre-set:

$$\sigma_u=\sigma_v=1, \qquad z'z=10, \qquad var(x)=var(z) \tag{6}$$

With these parameters fixed, four different values of ρ_{xz} and ρ_{xu} were examined: 0, 0.3, 0.6, and 0.9. Note that with the fixed nuisance parameters, $\rho_{xz}=\delta$ and $\rho_{xu}=\delta_{uv}$. The number of observations is fixed at 20.[8]

Figure 1 shows the effect of instrument relevance on the sampling error of IV estimator. It reaffirms the finding of Nelson and Startz (1990b) and Maddala and Jeong (1992). When the instrumental variable becomes less relevant (lower ρ_{xz}), the finite sample bias of IV estimator becomes more severe. This is the basis for pretesting instrument relevance. To avoid biased estimators and erroneous inference, 'bad' instruments should be screened out, and only 'acceptable' instruments should be used. The goal of instrument relevance pretest is to sort out the 'acceptable' instruments.

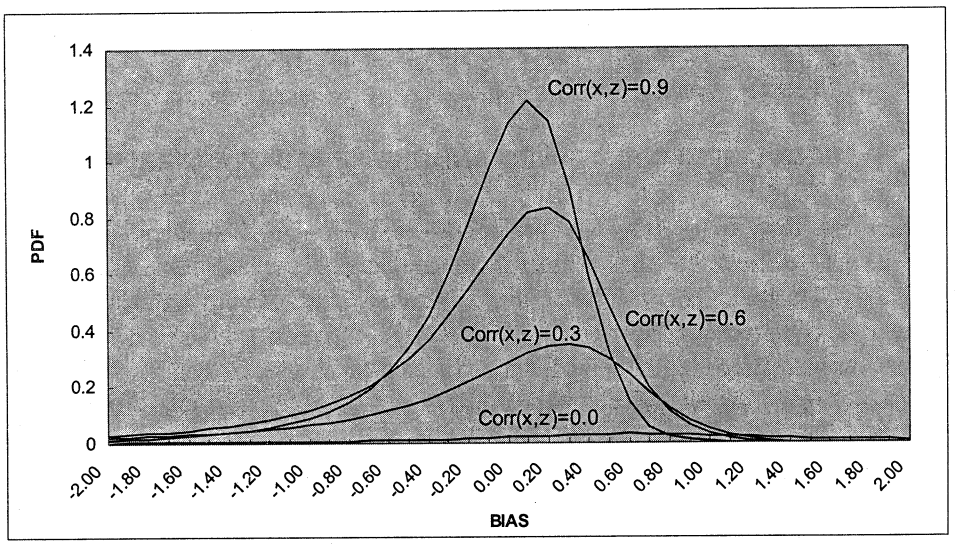
To sort out 'acceptable' instruments, we need to clarify the definition of

Nelson for his valuable comment on the exposition of the distribution.

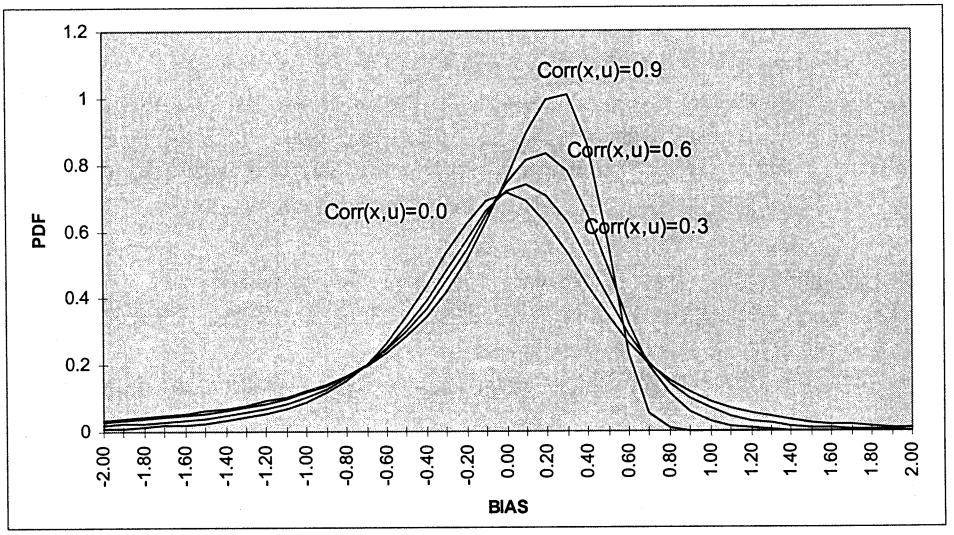
[8] Additionally, the effect of n (sample size) was examined. The bias becomes smaller as n increases, but does not degenerate even with n=100. The result, which is not reported to conserve space, is available from the author upon request.

'acceptable' instruments. First of all, it should be emphasized that poor instrument is not the only factor causing small sample bias of IV estimator. The degree of endogeneity (ρ_{xu}) also affects the magnitude of bias. Figure 2 shows the effect of endogeneity on the bias of IV estimator. It is clear in Figure 2 that, even with a reasonably high correlation between x and z ($\rho_{xz}=0.6$), the magnitude of the bias increases rapidly as the degree of endogeneity increases.

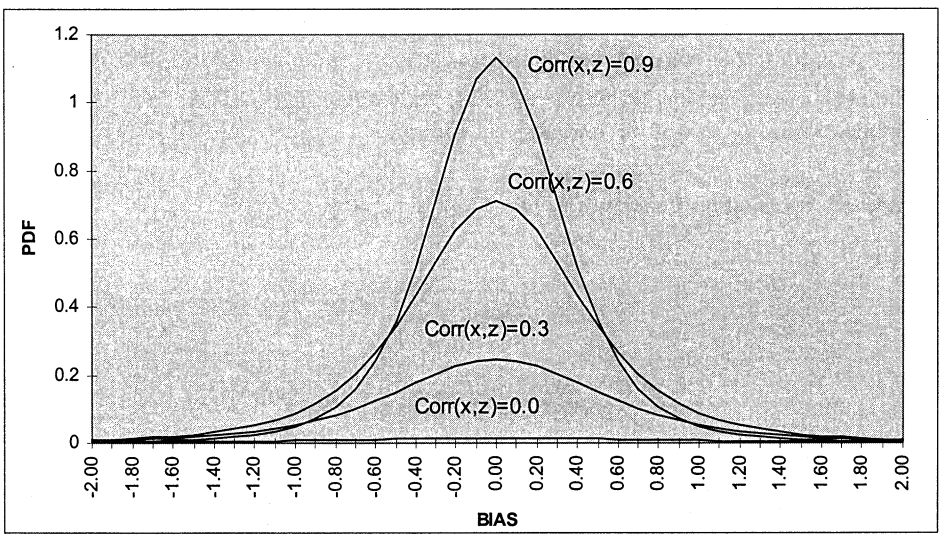
[Figure 1] Effects of Weak Instrument when Corr(x, z)=0.6



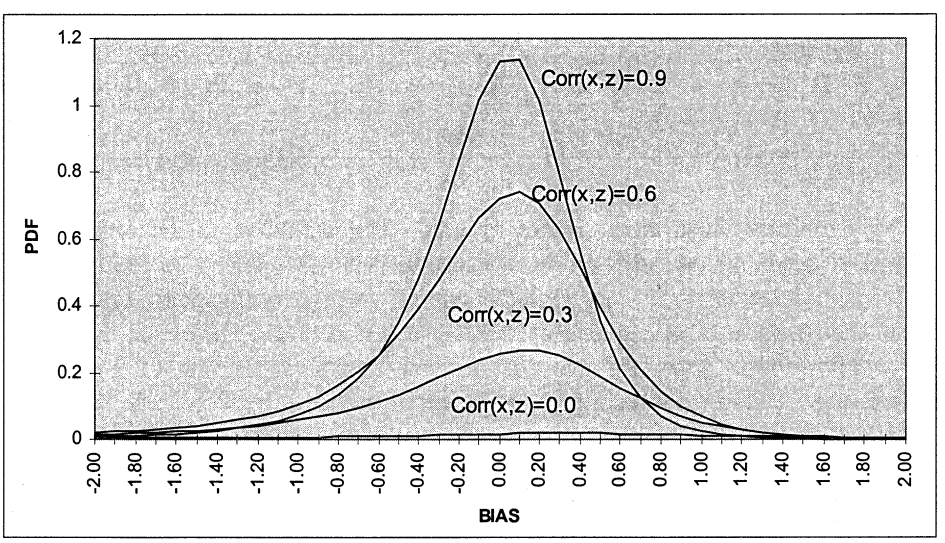
[Figure 2] Effects of Endogeneity when Corrr(x, z)=0.6



[Figure 3] Instrument Relevance when Corrr(x, u)=0.0



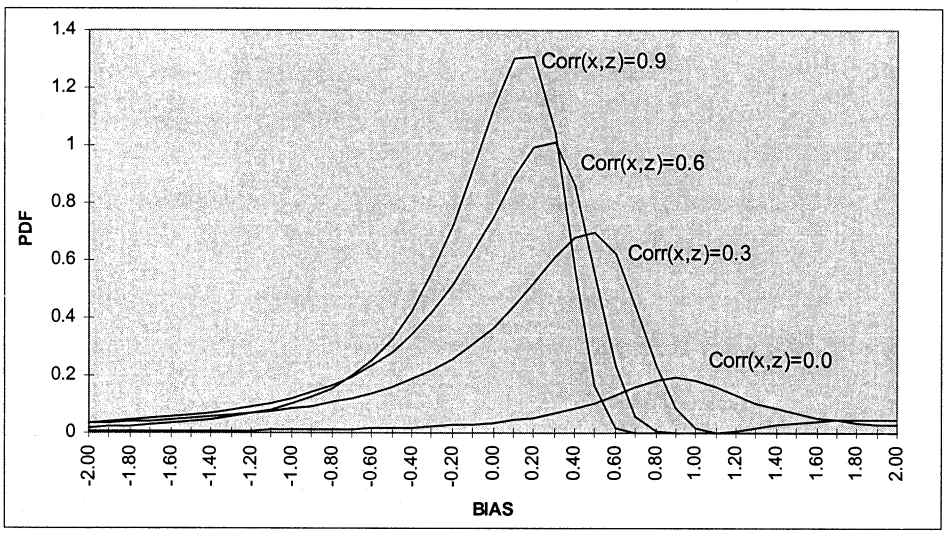
[Figure 4] Instrument Relevance when Corr(x, u)=0.3



A more interesting fact is that the effectiveness of instrument is not independent of the degree of endogeneity. Figures 3-5, along with Figure 1, show the effect of endogeneity on the effectiveness of instrument. Figure 3 is the distribution of sampling error when x is not endogenous at all ($\rho_{xu}=0$). As is seen in the Figure, no matter how 'relevant' the instrument (z) is, the bias is zero.[9] Figure 4 shows the case of weakly endogenous x ($\rho_{xu}=0.3$).

When x is weakly endogenous, ρ_{xz} does not need to be too high to be acceptable. Even an instrument with $\rho_{xz}=0.3$ does not cause a severe bias. Figure 5 shows the case of highly endogenous x ($\rho_{xu}=0.9$). In Figure 5, no instrument is satisfactory in terms of bias with such high endogeneity. Even an instrument with $\rho_{xz}=0.9$ is not very effective.

[Figure 5] Instrument Relevance when Corr(x, u)=0.9



What we learn from these figures is clear. The definition of an 'acceptable' instrument needs to be adjusted in connection with the degree of endogeneity. While any instrument is 'acceptable'[10] in a weakly endogenous model, the instrument needs to be strongly correlated to be acceptable in a highly endogenous model. Thus, the test for instrument screening needs to incorporate the effect of endogeneity into the correlation between x and z. The relevance of instrument must be measured relatively to the magnitude of endogeneity. To date, however, all the previous testing procedures are about ρ_{xz} without any adjustment for the degree of endogeneity. For the existing tests, the instruments are accepted if they have nonzero correlation with x, even when the endogeneity is severe and stronger instrument is necessary, to avoid bias. As a result, existing pretests would over-accept bad instruments when they are not good enough to be 'acceptable.' This is partly responsible for the problem Hall et al. (1996) find. Because their test is, like others, to screen out only the instruments satisfying H_0: $\rho_{xz}=0$,[11] (relatively) bad instruments that are

[9] Naturally, IV estimation becomes more efficient as z is more closely correlated with x.

[10] Again, 'acceptable' means the instrument does not cause serious bias.

correlated with x but not acceptable in connection with endogeneity are passed into the second stage regression. Consequently, due to the bad instruments, the symptoms of biased estimation and false inference become more severe than the nominal cumulated Type I Errors.

Ⅳ. LIKELIHOOD RATIO TEST OF INSTRUMENT RELEVANCE

Thus far, the single-regressor-single-instrument model has been considered for brevity of presentation. As for testing instrument relevance, however, it is of more practical interest to consider a general model. The presentation will be focused on a single-regressor-multiple-instrument case to begin with, but the discussions in this section and the tests in the next section can be easily applied to a multiple-regressor-multiple-instrument model.[12] Note that we will use the estimator using the multiple instruments as the IV estimator.

In a single-regressor-multiple-instrument model, it is common in the literature to use the multiple correlation coefficient (R^2) to measure regressor-instruments relationship. Formally, the new model is

$$y=\beta x+u \tag{7}$$

$$x=\delta_1 z_1+\delta_2 z_2+\ldots+\delta_m z_m+v\equiv z\delta+v \tag{8}$$

In this model, the IV estimator is estimated using two stage least squares (2SLS), that is, $\hat{\beta}_{iv}=(\hat{x}'\hat{x})^{-1}(\hat{x}'y)=(\hat{x}'x)^{-1}(\hat{x}'y)$ where $\hat{x}=z(z'z)^{-1}(z'x)$. Naturally, the R^2 from the regression of equation (8) can be used as a measure of instrument relevance. Let us denote the population multiple correlation coefficient as R_p^2. Nelson and Startz (1990a) suggest a conventional rule of testing H_0: $R_p^2=0$ against H_1: $R_p^2>0$. They argue that the instrument should not be used if $nR^2<2$. Shea (1997) suggests a more stringent test based on $\chi^2(1)$ cut-off point: do not use the instrument (or reject H_0) if $nR^2<3.84$ (5% significance level). Although these two tests are practically usable, a more accurate likelihood ratio test can be derived using the finite sample distribution of R^2.[13]

---

[11] More accurately, the instruments that do not statistically reject H_0: $\rho_{xz}=0$.

[12] In the multiple instruments case, the distribution of bias in IV estimator has a more complicated expression with infinite series of confluent hyper-geometric functions, as shown in Phillips (1983). However, the findings about the effects of various factors on the bias in the earlier section are not qualitatively different from single-instrument case. Numerical plots are available from the author.

[13] Bound et al. (1993) suggest the usual F-test on H_0: $\delta_1=\delta_2=\ldots=\delta_m=0$. Their test is actually identical to this LR test on H_0: $R_p^2=0$.

Assuming the variables have normal distributions, the finite sample distribution of R^2 was originally derived by Fisher (1928).[14] Under H_0: $R_p^2=0$, the distribution takes a much simpler form:

$$f(R^2)=\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\left(\frac{n-m-1}{2}\right)}(R^2)^{\frac{m}{2}-1}(1-R^2)^{\frac{n-m-1}{2}-1} \tag{9}$$

Based on (9), the following results are immediate.[15]

Corollary 1. If the regressors and the regressed variable are normally distributed and if $R_p^2=0$, then

$$\frac{R^2}{1-R^2}\cdot\left(\frac{n-m-1}{m}\right) \sim F(m,\ n-m-1) \tag{10}$$

Corollary 2. Given the sample from a multivariate normal distribution, the likelihood ratio (LR) test at significance level α for H_0: $R_p^2=0$ is given by

$$\frac{R^2}{1-R^2}\cdot\left(\frac{n-m-1}{m}\right) > F_\alpha(m,\ n-m-1) \tag{11}$$

To compare this LR test to the tests of Nelson and Startz (1990a) and Shea (1997), the likelihood ratio test (11) can be rewritten as:

$$R^2>\frac{mF_\alpha(m,n-m-1)}{(n-m-1)+mF_\alpha(m,n-m-1)} \tag{12}$$

Again, Nelson-Startz's rejection rule is $R^2>(2/n)$, and Shea's rejection rule is $R^2>(3.84/n)$. In addition to these, another test should be considered for the comparison. Hall, Rudebusch and Wilcox (1996) (HRW hereafter) suggest a different LR test of instrument relevance using the canonical correlations between the regressors and the instruments. HRW's LR test of H_0: *no correlation between* x *and* z is:

$$-n\log(1-c_m^2)>\chi_\alpha^2(m-k+1) \tag{13}$$

where k is the number of regressors (rank of x), and c_m is the smallest

[14] Simpler forms obtained by other authors like Gurland (1968) are summarized in Johnson and Kotz (1970), chapter 32.

[15] See, for instance, Anderson (1984) pp.138-142.

sample canonical correlation between x and z. Their test is unique in that the correlation between *multiple* regressors ($k>1$) and *multiple* instruments is tested. For the current case of the *single*-regressor-multiple-instruments model, HRW test (13) becomes

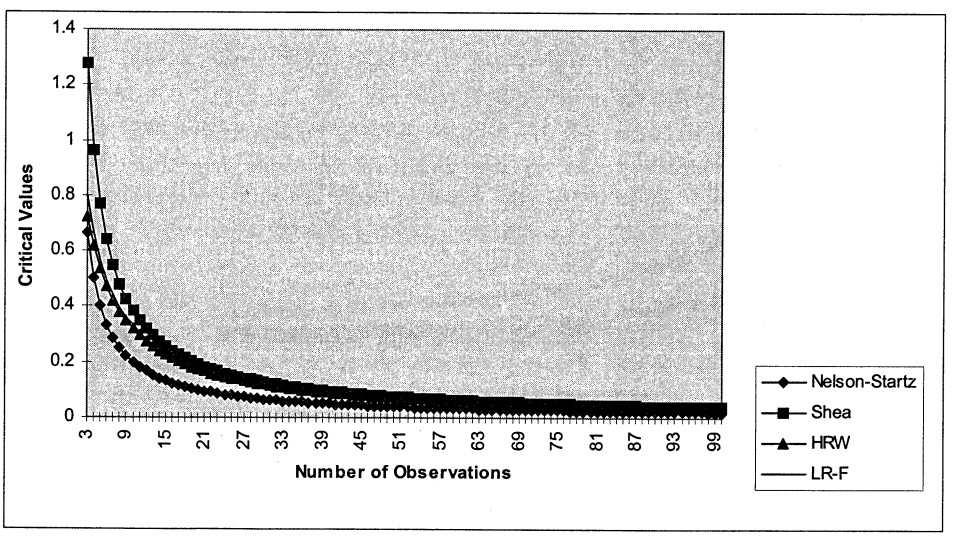
$$-n\log(1-R^2)>\chi_a^2(m) \tag{14}$$

because the canonical correlation is $\sqrt{R^2}$ when $k=1$.[16] For comparison to the other tests, HRW's rejection criterion (14) can be rewritten as:

$$R^2>1-e^{-\frac{\chi_a^2(m)}{n}} \tag{15}$$

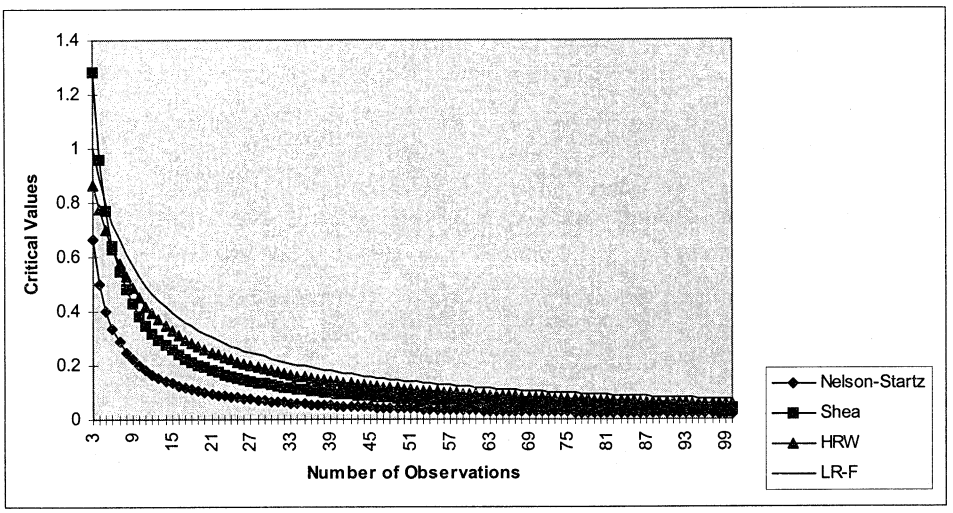
Figure 6 plots the critical values of these four tests at 5% significance level against sample size (n) for the single-regressor-single-instrument model. It is interesting that HRW test is almost identical to the likelihood ratio F-test. While its cut-off point converges to the LR F-test as sample size grows, Shea's test is too strict in small samples. Nelson-Startz test is close to LR in small samples but does not quickly converge to LR.

[Figure 6] Critical Values of IV Screening Tests (1 Regressor, 1 Instrument)

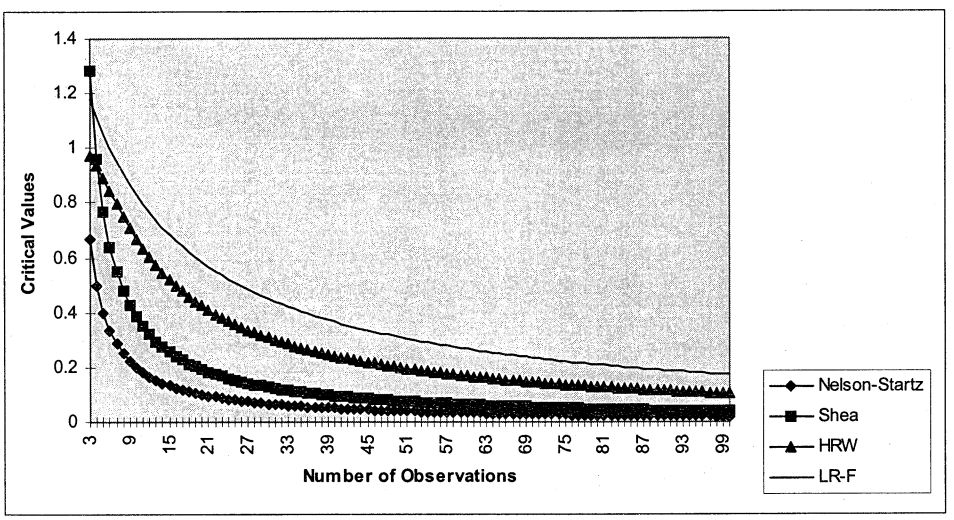


[16] See, for instance, Spanos (1986) pp.313-314.

[Figure 7] Critical Values of IV Screening Tests (1 Regressor, 2 Instruments)



[Figure 8] Critical Values of IV Screening Tests (1 Regressor, 5 Instruments)



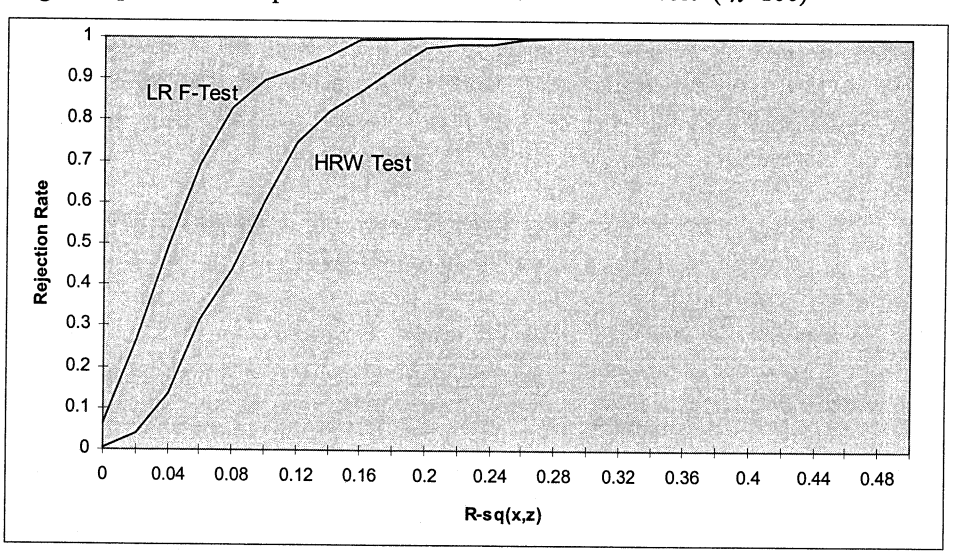
When the model is over-identified (there are more than one instrument), however, the critical values of these tests are quite different. Figure 7 shows, for the case of a single regressor and two instruments, the changes in critical values of the three tests as n increases. It is shown that the rejection region of Nelson-Startz's test is the widest, Shea's test is the second, HRW χ^2-test is the third, and the LR F-test has the smallest rejection region. In other words, Nelson-Startz test is the most lenient test among the four, and the LR F-test is

the most stringent one. Also, the differences do not degenerate as sample size increases.[17] Figure 8 is the one-regressor-five-instruments case. It is clear the discrepancies in critical values become more severe as the number of instruments (m) increases.

Naturally, the LR F-test has better power than HRW test in a *single* regressor case, although the comparison is not actually fair because HRW test is robust to the number of regressors (m) and LR F-test is not. However, it should be noted that the LR F-test is based on finite sample distribution, while HRW χ^2 test is an *asymptotic* test. Thus, the proposed LR F-test is expected to have a better small sample properties than HRW test in any case. Figures 9 and 10 show the simulated power functions of LR F-test and HRW test for a single regressor case.[18] As expected, Figure 9 shows the power of LR F-test is superior to HRW test even in a large sample (n=100). In a small sample (n =20), as Figure 10 shows, the power difference is more evident.

It is not surprising that the LR F-test is most powerful in this setup. The LR F-test, derived from the finite sample distribution of R^2, actually becomes equivalent to the classical F-test for the significance of all the explanatory variables in a regression. It is well known that the F-test is uniformly most powerful for the null hypothesis.

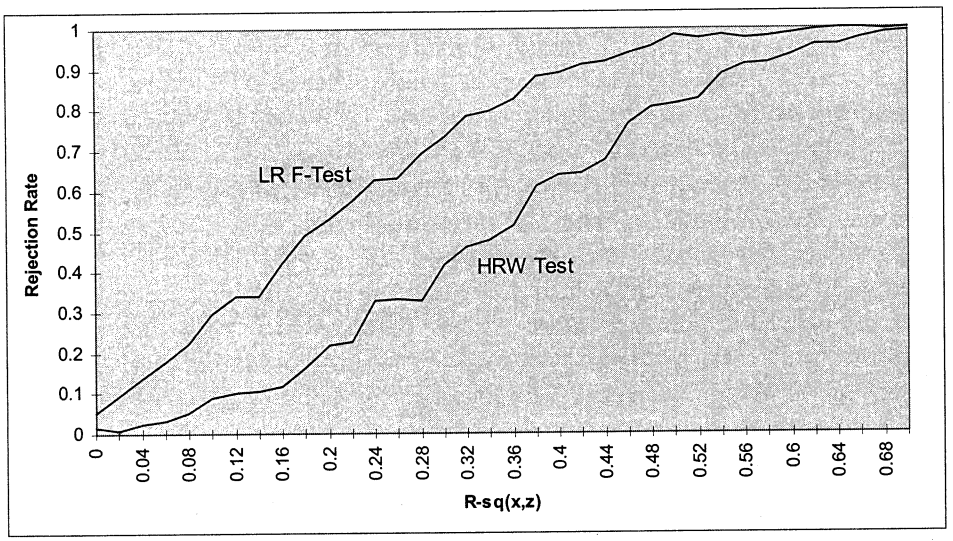
[Figure 9] Power Comparison of HRW Test vs. LR F-Test (n=100)



[17] Even with n=1000, the critical value of LR F-test is twice higher than and Shea's and HRW's, and about four times higher than Nelson-Startz's. More detailed results are available from the author.

[18] The power plots are based on 500 replications.

[Figure 10] Power Comparison of HRW Test vs. LR F-Test (n=20)



For a multiple-regressor-multiple-instrument case ($m>1$), the LR F-test needs to be repeatedly applied for each regressor as Shea (1997) did.[19] While LR F-test does not test the correlation between *multiple* regressors and multiple instruments, it can identify which regressor has weak instruments and which does not. A more important advantage of LR F-test over HRW test is that it can be adjusted for the degree of endogeneity to incorporate the effect of ρ_{xu} on IV estimation.[20] In the next section, I will suggest an adjustment of the LR F-test for endogeneity based on an artificial regression.[21]

V. BOOTSTRAP TEST FOR INSTRUMENT SCREENING WITH ENDOGENEITY ADJUSTMENT

Let us consider the model (7) and (8) again.

$$y=\beta x+u \tag{7}$$

$$x=\delta_1 z_1+\delta_2 z_2+\ldots+\delta_m z_m+v\equiv z\delta+v \tag{8}$$

[19] When severe multicollinearity among multiple regressors is suspected, the correction by Shea (1997) can be applied, too.

[20] In a single-regressor case, HRW test can similarly be adjusted to incorporate the effect of endogeneity. Then, however, HRW test loses its generality to be applied to a multiple-regressor case.

[21] Throughout section IV, the random variables are assumed to have a normal distribution. If the normality assumption is violated, the LR F-test becomes invalid. However, the bootstrap test proposed in the next section is robust to distributional assumption.

The problem of conventional instrument relevance tests is that the degree of endogeneity (i.e. correlation of x and u) is ignored when R^2_{xz} is measured from equation (8). To remedy this problem, the following adjustment of R^2_{xz} is considered.

$$\tilde{R}^2_{xz} = R^2_{xz}(1 - R^2_{xu}) \tag{16}$$

where R^2_{xu} is the R^2 from the artificial regression of $\hat{u}$ on x, and $\hat{u}$ is the residual from IV regression of y on x.[22] $\tilde{R}^2_{xz}$ is an endogeneity-penalized measure of R^2 between x and z. It is straightforward to see $\tilde{R}^2_{xz}$ is not higher than R^2_{xz} and becomes lower as the degree of endogeneity (R^2_{xz}) becomes higher.[23]

A problem in testing instrument relevance using $\tilde{R}^2_{xz}$ is that the exact distribution of $\tilde{R}^2_{xz}$ is not tractable. This problem can be overcome by adopting a distribution-free testing procedure. Bootstrap method, originally introduced by Efron (1979, 1982), is one such procedure. There have been introduced numerous 'resampling' methods in the literature, such as cross-validation procedure, recursive residual tests, jackknife methods, Goldfeld-Quant's test for heteroskedasticity, etc. Bootstrap method is a relatively new, computer-oriented resampling method, which utilizes 'random' resampling scheme with replacements. Formally, the basic procedure, when we have a set of observations $\{x_1, x_2, \ldots, x_n\}$ and a test statistic $\hat{\theta}$, is as follows.

1) Draw a 'bootstrap sample' $B_1 = \{x_1^*, x_2^*, \ldots, x_n^*\}$ from the original sample $\{x_1, x_2, \ldots, x_n\}$. Each x_i^* is a random pick from $\{x, x_2, \ldots, x_n\}$ with replacement.
2) Compute $\hat{\theta}_1^B$ using B_1.
3) Repeat steps 1) and 2) m times to obtain $\{\hat{\theta}_1^B, \hat{\theta}_2^B, \ldots, \hat{\theta}_m^B\}$.
4) Approximate the distribution of $\hat{\theta}$ by the bootstrap distribution $\hat{F}$, putting mass $1/n$ at each point $\hat{\theta}_1^B, \hat{\theta}_2^B, \ldots, \hat{\theta}_m^B$.

It has been shown that the bootstrap approximation of the distribution of $\hat{\theta}$ converges to the true one under mild conditions, and its finite sample properties

---

[22] Like $\hat{\beta}_{iv}$, $\hat{u}$ is also affected by the problem of poor instruments. However, $\hat{u}$ is used because 1) there is no better estimator of u, and 2) the correction using $\hat{u}$ turns out enough to correct the problem.

[23] In (16), the endogeneity adjustment factor, R^2_{xu}, has an equal weight to R^2_{xz}. We could consider variations of the weight on endogeneity penalty. We may possibly find some optimal weighting scheme for maximum performance of the suggested test. A generalization to this direction will be investigated in further studies. I am grateful to an anonymous referee for his/her comment on this possibility.

are reasonably satisfactory even in the cases where traditional statistical approach fails. Recent survey on bootstrap methods in econometrics is available in Jeong and Maddala (1994).

When a statistical inference is made using bootstrap method, a bootstrap confidence interval must be constructed. Among numerous alternative methods for constructing bootstrap confidence interval, Efron's 'accelerated bias-corrected percentile (BC_a)' interval is employed in this study. Hall (1988) and Martin (1990), among others, show BC_a method is asymptotically superior to other methods.[24] Efron (1987), Beran (1988), and Diciccio and Tibshirani (1987), among others, show the small sample performance of BC_a method is acceptable.[25] Andrews and Buchinsky (2002) develop a three-step method of choosing the number of bootstrap repetitions for BC_a intervals. Using their method, a researcher can choose B that yields a BC_a interval close to the ideal one with infinite bootstrap repetitions. MacKinnon (2002)'s review of bootstrap inference surveys recent developments in bootstrap confidence intervals.

The BC_a method is to compute the (1-2 α) confidence interval for θ as

$$\theta \in [\hat{G}^{-1}(\Phi(z[\alpha])), \hat{G}^{-1}(\Phi(z[1-\alpha]))] \tag{17}$$

where $\hat{G}$ is the cdf of the bootstrap distribution, Φ is the cdf of the standard normal distribution, and

$$z[i] = z_0 + \frac{(z_0 + z^{(i)})}{1 - a(z_0 + z^{(i)})} \qquad (i = \alpha \text{ or } 1-\alpha) \tag{18}$$

In (18), $z^{(\alpha)}$ is the α-level critical value of the standard normal distribution, z_0 is 'bias constant' for bias correction, and a is 'acceleration constant' for variance stabilization. z_0 and a are computed from the sample as

$$z_0 = \Phi^{-1}(\hat{G}(\hat{\theta})) \tag{19}$$

$$a = \frac{1}{6}\left(\frac{\sum \varepsilon_i^3}{(\sum \varepsilon_i^2)^{3/2}}\right) \tag{20}$$

where ε_i is the finite sample version of the empirical influence function

[24] They show BC_a method, percentile-t method, and Beran's B method are asymptotically better than the other methods.

[25] For a brief review of bootstrap confidence intervals, see Jeong and Maddala (1994), pp.579-582. For more detailed review, see Diciccio and Romano (1988).

$$\varepsilon_i = \lim_{\Delta \to 0} \frac{t((1-\Delta)\hat{F} + \Delta\delta_i) - t(\hat{F})}{\Delta} \tag{21}$$

In (21), $\hat{\theta} = t(\hat{F})$, $\hat{F}$ is the empirical distribution of the original sample, and δ_i is a point mass at x_i. Thus, ε_i is the derivative of the estimate $\hat{\theta}$ with respect to the contamination of point x_i. One-tailed version of the confidence interval (17) is used to derive the critical point of bootstrap test for H_0: $R_p^2 = 0$ against H_1: $R_p^2 > 0$. For more detailed discussion about BC_a method, readers are referred to Efron (1987).

By using $\tilde{R}_{xz}^2$ instead of R_{xz}^2, we can screen out relatively weak instruments in the case of severe endogeneity. For example, consider a case with $R_{xz}^2 = 0.2$ and $R_{xz}^2 = 0.7$. Although the bias of IV estimation would be quite high because of the high endogeneity, the conventional LR tests without endogeneity correction would most likely accept z as a valid set of instruments because R_{xz}^2 is considerably greater than zero. However, if $\tilde{R}_{xz}^2$ is used instead of R_{xz}^2, z would not be accepted because $\tilde{R}_{xz}^2$ is only 0.06. Naturally, by eliminating such an instrument set, bias of IV estimation will be reduced.

VI. MONTE CARLO STUDY

To evaluate the new test, a Monte Carlo study is performed. Artificial data for the simulation were created using the following simple model.

$$y = \beta x + u \tag{22}$$

$$x = \delta z + v \tag{23}$$

$$v = \gamma u + e \tag{24}$$

where the third equation reflects the relationship between x and u. It is straightforward from (22) - (24) to see that γ and δ are determined once R_{xz}^2 and R_{xu}^2 are set. That is,[26]

$$\gamma = \sqrt{\frac{R_{xu}^2 \cdot var(e)}{(1 - R_{xz}^2 - R_{xu}^2) \cdot var(u)}} \tag{25}$$

$$\delta = \sqrt{\frac{R_{xz}^2 \cdot var(e)}{(1 - R_{xz}^2 - R_{xu}^2) \cdot var(z)}} \tag{26}$$

[26] Without loss of generality, we only consider positive γ and positive δ.

Various values of R^2_{xz} and R^2_{xu} are examined for our Monte Carlo study, and y, x, z, u, v, and e are created using (22) - (26). For the random number generation, we assume that $u \sim N(0,1)$, $e \sim N(0,1)$ and z are uniformly distributed with a zero mean and a variance of 10. The true β is set to zero and the sample size (n) is set to 100.

Table 1 shows the rejection rates of the LR F-test and the bootstrap LR test with endogeneity correction for various combinations of R^2_{xz} and R^2_{xu}.[27] We observe the following in Table 1. First, the usual two-tailed t-test of H_0: $\beta=0$ with IV estimator is misleading when the instrument is weak or the regressor is highly endogenous. For example, when $R^2_{xz}=0.00$ and $R^2_{xu}=0.9$, the size of the t-test is as high as 0.382 while the nominal size is 0.05. This phenomenon becomes less serious as R^2_{xz} becomes higher, but the tendency is still present even when R^2_{xz} is as high as 0.25.[28] Second, it is shown in the fourth column, that the empirical size and power of standard LR F-test are satisfactory but do not reflect the effect of endogeneity at all. The rejection rates of standard LR F-test for a particular R^2_{xz} is constant over the whole range of R^2_{xu}, 0.0 through 0.9, in Table 1. Third, as expected, the endogeneity-corrected bootstrap LR test incorporates the effect of endogeneity quite well. For example, let us consider the case with $R^2_{xz}=0.10$. While the LR F-test passes about 90% of the instruments regardless of the endogeneity level, the bootstrap test passes less instruments (0.364 through 0.024) as the endogeneity becomes higher (0.0 through 0.9). More specifically, consider the case of $R^2_{xz}=0.10$ and $R^2_{xu}=0.8$. In this situation, although R^2_{xz} is not zero, the instruments should not easily pass the relevance test because the endogeneity level is too high. While the LR F-test erroneously passes 90.2% of the instruments, the bootstrap test passes only 2.4% of the instruments. This difference is observed in every combination of R^2_{xz} and R^2_{xu} in Table 1. Fourth, it is observed that the bootstrap LR test produces less accurate empirical size and lower power than the LR F-test, when x is not endogenous at all ($R^2_{xu}=0$). This is certainly a weakness of the suggested bootstrap test. Nonparametric tests usually sacrifice power to overcome distributional assumption. The suggested bootstrap test does not seem to be an exception. However, its empirical size is not severely distorted, and its power becomes as high as the parametric LR F-test as R^2_{xz} becomes higher.

The second question is whether the pretest of instrument relevance improves

---

[27] Note that the sum of R^2_{xz} and R^2_{xu} cannot exceed 1 from (25) and (26). The restriction is necessary to keep $R^2_{zu}=0$. Intuitively, if x and z, and at the same time x and u are closely correlated, it is impossible for z and u to be independent.

28 Note that, when R^2_{xz} is 0.25, the correlation coefficient ($\rho_{xz}=\sqrt{R^2_{xz}}$) between x and z is as high as 0.50.

the accuracy of statistical inference on the parameter of interest, β. Let us consider again the t-test of H_0: $\beta=0$ against H_1: $\beta\neq0$. It has been already shown in Table 1 that the t-test becomes seriously misleading, if irrelevant instruments are used with no screening pretest. The alternative is a two-step test: first, perform LR F-test or the bootstrap test on the instruments to check if the instruments are relevant, and then perform the t-test on β only using the relevant instruments. Thus, we have three options: (1) direct t-test with no instrument relevance test, (2) two-step t-test with LR-F pretest for instrument relevance, (3) two-step t-test with the endogeneity-corrected bootstrap pretest for instrument relevance. Table 2 compares the accuracy of the three alternative procedures.

As was explained above, column (1) of Table 2 reaffirms that the direct t-test with no instrument screening procedure is inaccurate when the instrument is not good enough. The empirical size of the direct t-test is not satisfactory until R^2_{xz} becomes as high as 0.30.[29] It is noticed that the t-test tends to over-reject the null hypothesis as the endogeneity of x increases. Column (2) of Table 2 presents the rejection rate of the two-step t-test with LR-F screening test. The rejection rate in each case is 'unconditional' rate, defined as

$$\text{Rej. rate=P[reject } H_0\text{]}=\frac{\text{number of second stage rejections}}{\text{number of cases for first stage screening test}} \tag{27}$$

By defining the rejection rate as in (27), we can directly compare the rejection rate of two-step test to the rejection rate of direct t-test. Note the denominator of (27) is the same as the denominator of the direct t-test rejection rate.[30]

Column (2) of Table 2 shows that the two-step t-test with LR-F screening test eventually produces almost identical rejection pattern to the direct t-test. Except the cases of $R^2_{xz}=0.00$, the rejection rates in column (1) and column (2) are almost the same. Column (2) shows the same tendency of over-rejection as column (1) in every combination of R^2_{xz} and R^2_{xu}, when endogeneity of x is considerable. What this implies is that the LR-F screening test does not improve the accuracy of statistical inference except when $R^2_{xz}=0.00$. It is impressive in column (2), however, that the LR-F screening test corrects the severe bias with high endogeneity of x when $R^2_{xz}=0.00$.

Column (3) in Table 2 presents the unconditional rejection rate of the two-step t-test with the endogeneity-corrected bootstrap screening test. As shown in Table 2, the null hypothesis tends to be under-rejected with the bootstrap

[29] $R^2_{xz}=0.30$ mplies that $\rho_{xz}\approx0.55$.

[30] As explained in Introduction, Hall et al. (1994) compared 'conditional' rejection rates of two-step test to 'unconditional' rejection rates of direct test. While such comparison has its own implication, one should note the difference between the current comparison and their comparison.

pretest. However, the empirical sizes are generally closer to the nominal rate (5%) than the other procedures, (1) and (2). Especially when the endogeneity of x is high, unlike the other two competitors, the test does not show the tendency of over-rejection.

To summarize, although the results are mixed in some cases, it is shown that the endogeneity-corrected bootstrap pretest is generally helpful for more accurate statistical inference with IV estimation. On the contrary, the LR-F pretest in most cases does not improve the accuracy of inference with IV estimation.[31]

[31] In addition to the power comparison, the bias reduction patterns after alternative screening tests could interest readers. Unfortunately, however, there existed no significant difference in bias reduction between the two screening tests, LR F-test and endogeneity-corrected bootstrap LR test. Though there were a few exceptions, the average biases of IV estimator after the screening tests were almost identical. The results are available from the author upon request.

[Table 1] Rejection Rate of LR F-Test and Endogeneity Adjusted Bootstrap Test

R^2_{xz}	R^2_{xu}	Rejection Rate t-test of H_0: $\beta=0$	Rejection Rate Standard LR F-test	Rejection Rate Endo. Adjusted Bootstrap LR test
0.00	0.0	.000	.042	.034
0.00	0.1	.004	.052	.040
0.00	0.2	.010	.046	.026
0.00	0.3	.026	.050	.024
0.00	0.4	.038	.056	.054
0.00	0.5	.088	.064	.024
0.00	0.6	.114	.064	.038
0.00	0.7	.182	.074	.030
0.00	0.8	.270	.048	.042
0.00	0.9	.382	.046	.034
0.03	0.0	.000	.404	.042
0.03	0.1	.026	.438	.034
0.03	0.2	.030	.378	.046
0.03	0.3	.040	.392	.044
0.03	0.4	.066	.426	.036
0.03	0.5	.088	.462	.042
0.03	0.6	.084	.372	.022
0.03	0.7	.094	.444	.016
0.03	0.8	.114	.402	.020
0.07	0.0	.008	.790	.208
0.07	0.1	.034	.774	.178
0.07	0.2	.026	.736	.148
0.07	0.3	.064	.772	.132
0.07	0.4	.052	.786	.086
0.07	0.5	.052	.752	.062
0.07	0.6	.064	.706	.030
0.07	0.7	.098	.770	.028
0.07	0.8	.090	.748	.012
0.10	0.0	.018	.908	.364
0.10	0.1	.028	.914	.298
0.10	0.2	.036	.908	.274
0.10	0.3	.058	.910	.208
0.10	0.4	.044	.926	.172
0.10	0.5	.068	.912	.122
0.10	0.6	.074	.908	.062
0.10	0.7	.060	.934	.040
0.10	0.8	.110	.902	.024

(continued)

[Table 1] Rejection Rate of LR F-Test and Endogeneity Adjusted Bootstrap Test
(continued)

R^2_{xz}	R^2_{xu}	Rejection Rate t-test of H_0: $\beta=0$	Rejection Rate Standard LR F-test	Rejection Rate Endo. Adjusted Bootstrap LR test
0.15	0.0	.030	.980	.658
0.15	0.1	.036	.984	.570
0.15	0.2	.032	.986	.506
0.15	0.3	.056	.990	.404
0.15	0.4	.072	.978	.358
0.15	0.5	.062	.982	.234
0.15	0.6	.074	.988	.144
0.15	0.7	.074	.982	.094
0.15	0.8	.084	.986	.068
0.20	0.0	.044	.998	.882
0.20	0.1	.032	1.000	.804
0.20	0.2	.056	1.000	.696
0.20	0.3	.068	1.000	.578
0.20	0.4	.044	.998	.488
0.20	0.5	.056	1.000	.364
0.20	0.6	.060	.996	.292
0.20	0.7	.076	.998	.210
0.25	0.0	.036	1.000	.974
0.25	0.1	.046	1.000	.942
0.25	0.2	.046	1.000	.868
0.25	0.3	.042	1.000	.778
0.25	0.4	.048	1.000	.686
0.25	0.5	.052	.996	.554
0.25	0.6	.052	1.000	.462
0.25	0.7	.078	1.000	.396
0.30	0.0	.050	1.000	.990
0.30	0.1	.052	1.000	.980
0.30	0.2	.048	1.000	.956
0.30	0.3	.064	1.000	.894
0.30	0.4	.056	1.000	.826
0.30	0.5	.048	1.000	.756
0.30	0.6	.056	1.000	.684

* The results are based on 500 simulations.
* Bootstrap test is based on 500 bootstrap resamples from the original (y_i, x_i, z_i).

[Table 2] Unconditional Rejection Rate of H_0: $\beta = 0$ with and without Instrument Screening Tests

R^2_{xz}	R^2_{xu}	(1) Rejection Rate with no screening test	(2) Rejection Rate after LR F-test	(3) Rejection Rate after Bootstrap test
0.00	0.0	.000	.000	.000
0.00	0.1	.004	.004	.000
0.00	0.2	.010	.006	.000
0.00	0.3	.026	.012	.000
0.00	0.4	.038	.022	.000
0.00	0.5	.088	.044	.000
0.00	0.6	.114	.058	.002
0.00	0.7	.182	.068	.002
0.00	0.8	.270	.048	.002
0.00	0.9	.382	.046	.002
0.03	0.0	.000	.000	.000
0.03	0.1	.026	.026	.004
0.03	0.2	.030	.030	.010
0.03	0.3	.040	.036	.008
0.03	0.4	.066	.062	.018
0.03	0.5	.088	.084	.026
0.03	0.6	.084	.072	.012
0.03	0.7	.094	.082	.006
0.03	0.8	.114	.114	.006
0.07	0.0	.008	.008	.000
0.07	0.1	.034	.034	.018
0.07	0.2	.026	.024	.012
0.07	0.3	.064	.064	.034
0.07	0.4	.052	.052	.030
0.07	0.5	.052	.052	.036
0.07	0.6	.064	.064	.024
0.07	0.7	.098	.098	.028
0.07	0.8	.090	.090	.012
0.10	0.0	.018	.018	.012
0.10	0.1	.028	.028	.014
0.10	0.2	.036	.036	.030
0.10	0.3	.058	.058	.040
0.10	0.4	.044	.044	.030
0.10	0.5	.068	.068	.056
0.10	0.6	.074	.074	.034
0.10	0.7	.060	.060	.028
0.10	0.8	.110	.110	.024

(continued)

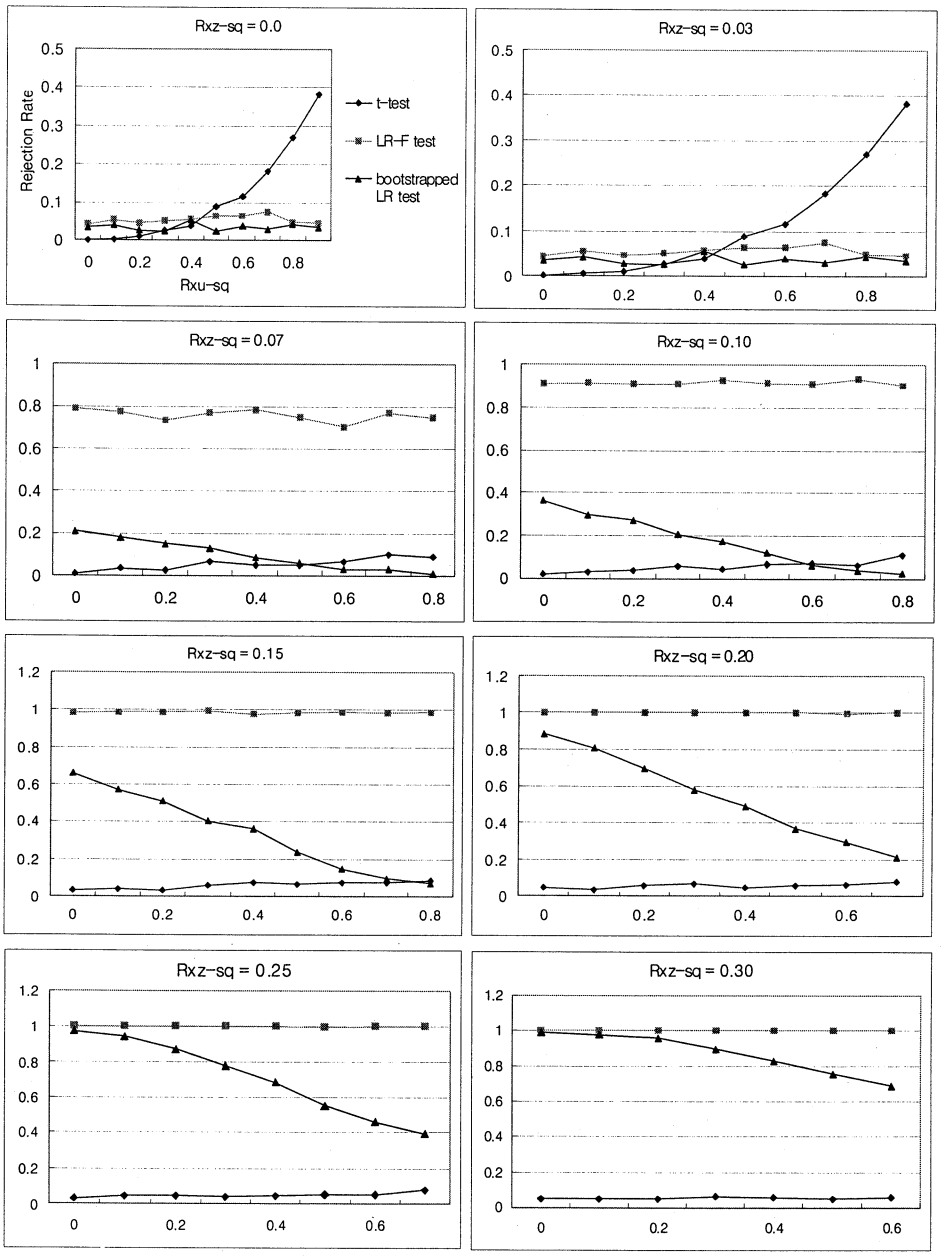
[Table 2] Unconditional Rejection Rate of H_0: $\beta = 0$ with and without Instrument Screening Tests

(continued)

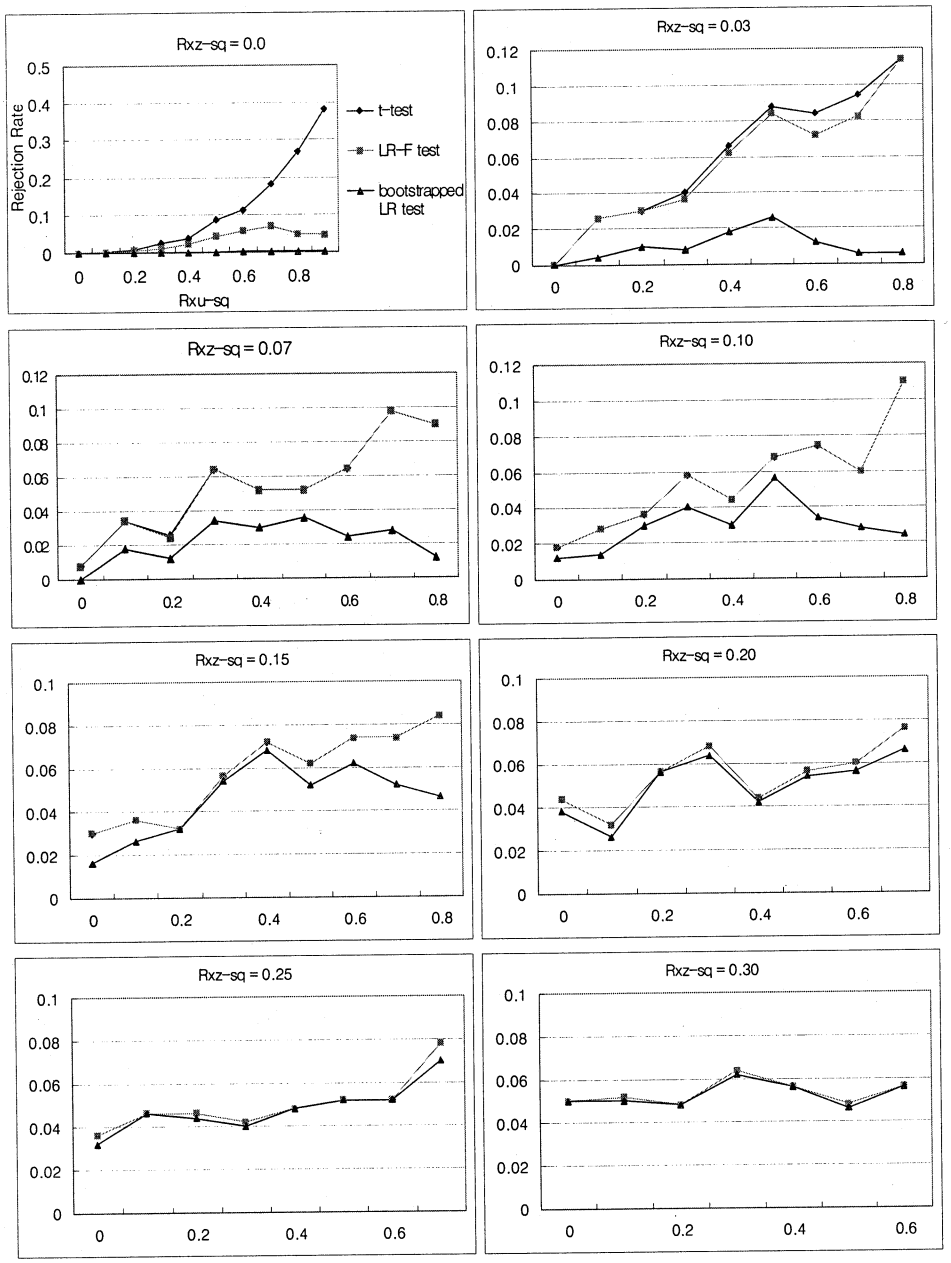
R^2_{xz}	R^2_{xu}	(1) Rejection Rate with no screening test	(2) Rejection Rate after LR F-test	(3) Rejection Rate after Bootstrap test
0.15	0.0	.030	.030	.016
0.15	0.1	.036	.036	.026
0.15	0.2	.032	.032	.032
0.15	0.3	.056	.056	.054
0.15	0.4	.072	.072	.068
0.15	0.5	.062	.062	.052
0.15	0.6	.074	.074	.062
0.15	0.7	.074	.074	.052
0.15	0.8	.084	.084	.046
0.20	0.0	.044	.044	.032
0.20	0.1	.032	.032	.026
0.20	0.2	.056	.056	.056
0.20	0.3	.068	.068	.064
0.20	0.4	.044	.044	.042
0.20	0.5	.056	.056	.054
0.20	0.6	.060	.060	.056
0.20	0.7	.076	.076	.066
0.25	0.0	.036	.036	.032
0.25	0.1	.046	.046	.046
0.25	0.2	.046	.046	.044
0.25	0.3	.042	.042	.040
0.25	0.4	.048	.048	.048
0.25	0.5	.052	.052	.052
0.25	0.6	.052	.052	.052
0.25	0.7	.078	.078	.070
0.30	0.0	.050	.050	.050
0.30	0.1	.052	.052	.050
0.30	0.2	.048	.048	.048
0.30	0.3	.064	.064	.062
0.30	0.4	.056	.056	.056
0.30	0.5	.048	.048	.046
0.30	0.6	.056	.056	.056

* The results are based on 500 simulations.
* Bootstrap test is based on 500 bootstrap resamples from the original (y_i, x_i, z_i).

[Figure 11] Graphical Presentation of Table 1 (Rejection Rate of LR F-Test and Endogeneity Adjusted Bootstrap Test)



[Figure 12] Graphical Presentation of Table 2 (Unconditional Rejection Rate of H_0: $\beta = 0$ with and without Instrument Screening Tests)



Ⅶ. SUMMARY AND CONCLUSION

In this paper, two problems in the previous instrument screening tests are considered. First, an LR test based on the finite sample distribution of R^2 is proposed. It is shown that the proposed test is more powerful than the previous tests at least in small samples. Second, while the degree of endogeneity is an obvious contributor to bias in Ⅳ estimation, no previous tests incorporate the magnitude of endogeneity into instrument screening procedure. An adjustment for endogeneity for the LR F-test is suggested in this paper, and the performance of the tests are presented in a Monte Carlo simulation.

Appendix: Derivation of Exact Distribution of $(\hat{\beta}_{iv}-\beta)$

Given that $\hat{\beta}_{iv}-\beta=(z'x)^{-1}(z'u)$, let us define $W_2\equiv(z'u)$ and $W_2\equiv(z'x)$. Then the goal is to find the distribution of (W_1/W_2). Recall that $[W_1, W_2]$ has a bivariate normal distribution with mean (μ_1, μ_2) and variance Ω where

$$\mu_1=0 \qquad \mu_2=\delta(z'z)$$
$$\Omega\equiv\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}=\begin{bmatrix}\sigma_u^2 z'z & \sigma_{uv}z'z\\ \sigma_{uv}z'z & \sigma_v^2 z'z\end{bmatrix} \tag{A1}$$

Since $\hat{\beta}_{iv}-\beta\equiv B=W_1/W_2$,

$$\begin{aligned}F(B)&=P\left(\frac{W_1}{W_2}\le B\right)\\&=P(W_1-W_2B\le 0 \mid W_2>0)P(W_2>0)+P(W_1-W_2B\ge 0 \mid W_2<0)P(W_2<0)\\&=P(W_1-W_2B\le 0 \text{ and } W_2>0)+P(W_1-W_2B\ge 0 \text{ and } W_2<0)\end{aligned} \tag{A2}$$

Thus,

$$\begin{aligned}F(B)=&\Psi\left(\frac{\mu_1-\mu_2B}{\sqrt{\sigma_1^2-2\rho\sigma_1\sigma_2B+B^2\sigma_2^2}},-\frac{\mu_2}{\sigma_2},\frac{\sigma_2B-\rho\sigma_1}{\sqrt{\sigma_1^2-2\rho\sigma_1\sigma_2B+B^2\sigma_2^2}}\right)\\&+\Psi\left(\frac{\mu_2B-\mu_1}{\sqrt{\sigma_1^2-2\rho\sigma_1\sigma_2B+B^2\sigma_2^2}},\frac{\mu_2}{\sigma_2},\frac{\sigma_2B-\rho\sigma_1}{\sqrt{\sigma_1^2-2\rho\sigma_1\sigma_2B+B^2\sigma_2^2}}\right)\end{aligned} \tag{A3}$$

where Ψ is the standard bivariate normal distribution function:

$$\Psi(a,b;c)=\frac{1}{2\pi\sqrt{(1-c^2)}}\int_a^\infty\int_b^\infty\exp\left\{-\frac{x^2-2cxy+y^2}{2(1-c^2)}\right\}dxdy \tag{A4}$$

A differentiation of $F(B)$ gives the density function of B:

$$\begin{aligned}f(B)=&\frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{\pi(\sigma_1^2-2\rho\sigma_1\sigma_2B+\sigma_2^2B^2)}\exp\left\{-\frac{\mu_1^2\sigma_2^2-2\rho\sigma_1\sigma_2\mu_1\mu_2+\mu_2^2\sigma_1^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right\}\\&+\frac{\mu_1\sigma_2^2B-\rho\sigma_1\sigma_2(\mu_1+\mu_2B)+\mu_2\sigma_1^2}{\sqrt{2\pi(\sigma_1^2-2\rho\sigma_1\sigma_2B+\sigma_2^2B^2)^3}}[\Phi(\Delta)-\Phi(-\Delta)]\times\\&\exp\left\{\frac{(\mu_1\sigma_2^2B-\rho\sigma_1\sigma_2(\mu_1+\mu_2B)+\mu_2\sigma_1^2)^2}{2(1-\rho^2)\sigma_1^2\sigma_2^2(\sigma_1^2-2\rho\sigma_1\sigma_2B+\sigma_2^2B^2)}\right\}\times\\&\exp\left\{-\frac{(\mu_1^2\sigma_2^2-2\rho\sigma_1\sigma_2\mu_1\mu_2+\mu_2^2\sigma_1^2)(\sigma_1^2-2\rho\sigma_1\sigma_2B+\sigma_2^2B^2)}{2(1-\rho^2)\sigma_1^2\sigma_2^2(\sigma_1^2-2\rho\sigma_1\sigma_2B+\sigma_2^2B^2)}\right\}\end{aligned} \tag{A5}$$

where Φ is the standard univariate normal distribution function, and

$$\Delta \equiv \frac{\mu_1\sigma_2^2 B - \rho\sigma_1\sigma_2(\mu_1+\mu_2 B)+\mu_2\sigma_1^2}{\sigma_1\sigma_2\sqrt{(1-\rho^2)(\sigma_1^2-2\rho\sigma_1\sigma_2 B+\sigma_2^2 B^2)}} \tag{A6}$$

Substituting (A1),

$$\begin{aligned} f(B) = & \frac{\sqrt{\sigma_u^2\sigma_v^2-\sigma_{uv}^2}}{\pi(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)} \exp\left\{-\frac{\delta^2(z'z)\sigma_u^2}{2(\sigma_u^2\sigma_v^2-\sigma_{uv}^2)}\right\} \\ & + \frac{\delta(z'z)(\sigma_u^2-\sigma_{uv}B)}{\sqrt{2\pi(z'z)(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)^3}}[\Phi(\Delta)-\Phi(-\Delta)]\times \\ & \exp\left\{\frac{\delta^2(z'z)\left[(\sigma_u^2-\sigma_{uv}B)^2-\sigma_u^2(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)\right]}{2(\sigma_u^2\sigma_v^2-\sigma_{uv}^2)(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)}\right\} \end{aligned} \tag{A7}$$

where

$$\Delta \equiv \frac{\delta(z'z)(\sigma_u^2-\sigma_{uv}B)}{\sqrt{(\sigma_u^2\sigma_v^2-\sigma_{uv}^2)(\sigma_u^2-2\sigma_{uv}B+\sigma_v^2B^2)(z'z)}} \tag{A8}$$

REFERENCE

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley and Sons, New York.

Andrews, D.W.K. and M. Buchinsky (2002), "On the Number of Bootstrap Repetitions for BCa Confidence Intervals," *Econometric Theory*, 18, pp.962-984.

Beran, R. (1988), "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements," *Journal of American Statistical Association*, 83, pp.687-697.

Bound, J., D.A. Jaeger, and R. Baker (1993), "The Cure Can Be Worse Than The Disease: A Cautionary Tale Regarding Instrumental Variables," NBER Technical Paper, No. 137.

Diciccio, T.J. and J.P. Romano (1988), "A Review of Bootstrap Confidence Intervals," *Journal of Royal Statistical Society: Series B*, 50, pp.338-354.

Diciccio, T. and R. Tibshirani (1987), "Bootstrap Confidence Intervals and Bootstrap Approximations," *Journal of the American Statistical Association*, 82, pp.163-170.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, pp.1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38, Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1987), "Better Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 82, pp.171-200.

Fisher, R.A. (1928), "The General Sampling Distribution of the Multiple Correlation Coefficient," *Proceedings of the Royal Society of London, Series A*, 121, pp.654-673.

Fujikoshi, Y. (1974), "The Likelihood Ratio Tests Of the Dimensionality of Regression Coefficients," *Journal of Multivariate Analysis*, 4, pp.327-340.

Gurland, J. (1968), "A Relatively Simple Form of the Distribution of the Multiple Correlation Coefficient," *Journal of the Royal Statistical Society, Series B*, 30, pp.276-283.

Hahn, J. and J. Hausman (2002), "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, 70, pp.163-189.

Hall, P. (1988), "Theoretical Comparison of Bootstrap Confidence Intervals," *The Annals of Statistics*, 16, pp.927-953.

Hall, A.R., G.D. Rudebusch and D.W. Wilcox (1996), "Judging Instrument Relevance in Instrumental Variables Estimation," *International Economic Review*, 37, pp.283-289.

Hinkley, D.V. (1969), "On the Ratio of Two Correlated Normal Random Variables," *Biometrika*, 56, pp.635-639.

Jeong, J. and S. Chung (2001), "Bootstrap Tests for Autocorrelation," *Computational Statistics and Data Analysis*, 38, pp.49-69.

Jeong, J. and K.W. Lee (1999), "Bootstrapped White's Test for Heteroskedasticity in Regression Models," *Economics Letters*, 63, pp.261-267.

Jeong, J. and G.S. Maddala (1994), "A Perspective on Application of Bootstrap Methods in Econometrics," *Handbook of Statistics Vol. 11: Econometrics*, (ed) by G.S. Maddala, C.R. Rao and H.D. Vinod, North-Holland, pp.573-610.

Johnson, N.L. and S. Kotz (1970), *Distributions in Statistics: Continuous Univariate Distributions-2*, Houghton Mifflin Company, Boston.

MacKinnon, J.G. (2002), "Bootstrap Inference in Econometrics," *Canadian Journal of Economics*, 35, pp.615-645.

Maddala, G.S. and J. Jeong (1992), "On the Exact Small Sample Distribution of the Instrumental Variable Estimator," *Econometrica*, 60, pp.181-183.

Mariano, R.S. (1973), "Approximations to the Distribution Functions of the Ordinary Least Squares and Two-Stage Least Squares Estimators in the Case of Two Included Endogenous Variables," *Econometrica*, 41, pp.67-77.

Mariano, R.S. and J. McDonald (1979), "A Note on the Distribution of LIML and 2SLS Coefficient Estimators in the Exactly Identified Case," *Journal of the American Statistical Association*, 74, pp.847-848.

Martin, M.A. (1990), "Bootstrap Iteration for Coverage Correction in Confidence Intervals," *Journal of the American Statistical Association*, 85, pp.1105-1118.

Nagar, A.L. (1959), "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 60, pp.575-595.

Nelson, C.R. and R. Startz (1990a), "The Distribution of the Instrumental Variable Estimator and Its t-Ratio When the Instrument Is a Poor One," *Journal of Business*, 63, pp.125-140.

Nelson, C.R. and R. Startz (1990b), "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator," *Econometrica*, 58, pp.967-976.

Phillips, P.C.B. (1983), "Exact Small Sample Theory in the Simultaneous Equations Model," *Handbook of Econometrics*, Vol. 1, Chapter 8, North Holland.

Sawa, T. (1969), "The Exact Finite Sample Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators," *Journal of the American Statistical Association*, 64, pp.923-936.

Shea, J. (1997), "Instrument Relevance in Multivariate Linear Models: A Simple Measure," *Review of Economics and Statistics*, 79, pp.348-352.

Spanos, A. (1986), *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge.

Staiger, D. and J.H. Stock (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, pp.557-586.

Stock, J.H., J.H. Wright, and M. Yogo (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20, pp.518-529.