The Validity of Instruments Revisited

Daniel Berkowitz
University of Pittsburgh

Mehmet Caner*
North Carolina State University

Ying Fang
Xiamen University

October 10, 2008

Abstract

Valid instrumental variables must be relevant and exogenous. However, in practice it is difficult to find instruments that are exogenous in that they satisfy the knife-edged orthogonality condition and at the same time are strongly correlated with the endogenous regressors. In this paper we show how a mild violation of the exogeneity assumption affects the limit of the Anderson-Rubin test (1949). This test statistic is frequently used in economics due to the fact that it is robust to identification problems. However, when there is mild violation of exogeneity the test is oversized and with larger samples the problem gets worse. In order to correct this problem, we introduce the fractionally resampled Anderson-Rubin test (FAR) that is derived by modifying the resampling technique introduced by Wu (1990). We show the FAR test does not overreject the null hypothesis when we use half of the sample without replacement as the block size from the original sample.

As a novel scheme, we treat the block size as a random variable and prove that this choice recovers the limit of the Anderson-Rubin (1949) test. We also prove that this optimal choice of block size converges in probability to 1/2. Simulations show that in finite samples the FAR is conservative; thus, we propose a range of block size choices that has very good size and power when there are possible violations of the exogeneity assumption.

Keywords: Berry-Esseen Bound, Finite Sample of Random Variables, Near-Exogeneity

---

*Daniel Berkowitz: Department of Economics, University of Pittsburgh, Pittsburgh, PA 21685, email:dmberk@pitt.edu. Mehmet Caner, Corresponding Author: Department of Economics, 4168 Nelson Hall, Raleigh, NC 27695. email: mcaner@ncsu.edu. Ying Fang: WISE, Xiamen University, China email:yifst1@gmail.com. We thank C.R. Rao and C.F.J. Wu for comments.

1 Introduction

Instrumental variable estimation is one of the most widely used methods in economics. Valid instruments must be relevant and exogenous. Regarding the relevance of instruments, there has recently been a growing interest in the asymptotics of weak instruments. One of the most widely used test statistics in that research line is the Anderson-Rubin (1949) test (for herein denoted the AR test). The AR test statistic can be used when instruments are weak as shown by Stock and Wright (2000). However, recently Berkowitz, Caner and Fang (2008) find that AR test is oversized when there is a minor violation of exogeneity assumption.

Even when researchers are careful in selecting instruments that are relevant and plausibly exogenous, it is still unlikely that an instrument is perfectly exogenous. The exogeneity assumption is a knife-edge condition in which a zero correlation between the instruments and the structural error term must hold exactly. As described by Conley, Hansen and Rossi (2007) and Kray (2008) and others, it is very difficult to perfectly satisfy this orthogonality condition in empirical work.

To correct for this problem, we assume that there can be a mild violation of the exogeneity assumption. This near exogeneity assumption allows for a local to zero correlation between the instruments and the structural error. However, convergence of this correlation to zero is slower than root n rate that is used in Berkowitz, Caner, and Fang (2008). This new imperfect exogeneity assumption enables us to use a powerful resampling method. And, this imperfect exogeneity assumption is realistic because it allows the structural error to get larger as the sample size increases. Furthermore, the case of root n convergence to zero covariance is discussed in the paper. We find that our method works well in both cases. We derive the limit of AR test and we show that the limit depends on the correlation between the instruments and structural error term. Furthermore, in larger samples, using critical values for the AR test based on the perfect exogeneity assumption results in huge size distortions when in reality there is a moderate correlation between the instruments and structural error term.

In this paper we propose a novel resampling technique for the Anderson-Rubin (1949) test. This technique is based on the jackknife histogram estimator in section 2 of Wu (1990). Because we can write the AR test in terms of the sample mean (see equation (4)), we can modify the results in section 2 of Wu (1990). Section 2 in Wu (1990) uses sampling without replacement from the original sample by drawing a fraction of the sample size. This is proportional to the full sample. In this setup, we propose a fractionally resampled version of the AR test. We show that this resampling technique with exactly the half-sample size is valid, but conservative. Then we derive theoretically a block size (which is a random variable) and show that with this choice the resampling technique is consistent.

This random block size converges in probability to 1/2.

Specifically, we find that by drawing half of the sample, the AR test does not overreject the null. We show that subsampling is oversized. We also find that the Kleibergen (2002) test may not be amenable to the resampling technique that we use.

We conduct simulations to check for the size properties and power of this FAR test. We find that a half-sample block size is very conservative; and, we find that resampling without replacement that uses anywhere between one fourth to one third of the sample size gives very good size and power results.

In related work, Kray (2008) and Conley, Hansen and Rossi (2007) both use a Bayesian approach for solving the problem of working with instruments that do not perfectly satisfy the orthogonality condition. They clearly show that even a small violation of the orthogonality condition can lead to entirely different outcomes. When they use the more realistic assumption of a mild violation of exogeneity, they find that the confidence intervals for structural parameters are larger. Conley, Hansen and Rossi (2007) analyze the support of the correlation parameter (i.e., the correlation between the instrument and structural error) and, for each plausible parameter value, they find the confidence interval for the structural parameters and take the union of these intervals. This method provides a conservative solution. Conley, Hansen and Rossi (2007) also use a local to zero approach; here, they assume the correlation parameter comes from a normal distribution and they characterize its asymptotics. In a third approach, Conley et al, attach Bayesian priors to this parameter and derive the posterior distribution. Kray (2008) takes a similar approach to this problem: however, his prior for the correlation parameter is not drawn from a normal distribution. In contrast to these methods that place priors on the correlation between the instrument and structural error term, our method is completely data dependent and we derive confidence intervals using subsamples of the data. In this paper we provide a solution to the problem of drawing inferences when there is a minor violation of exogeneity without attaching distributions to the correlation parameter. We resample the AR test and our critical values are adjusted accordingly.

In Section 2, we describe the problem of making inferences with instruments in violation of exogeneity and we develop a novel way of resampling of the AR test. Section 3 considers subsampling and shows that it will be oversized and will not solve the problem of drawing reliable inferences with instruments violating the exogeneity assumption. Section 3 also contains an analysis of some of the variants of subsampling. Section 4 contains Monte Carlo simulation. Section 5 concludes.

2   Inference with Violation of Exogeneity

We analyze a model that contains a specific violation of the exogeneity assumption. Similar assumptions about the violation of exogeneity have been used by Newey (1985) and Hall and Inoue (2003). Our assumption allows for a local to zero covariance between the instruments and the structural error term and is more flexible than the knife-edged exogeneity assumption used in the instrumental variables estimation literature. The model that we use is:

$$y = Y\theta_0 + u, \tag{1}$$

$$Y = Z\Pi + V, \tag{2}$$

where $cov(u, V) \neq 0, Y : n \times m, Z : n \times k, k \geq m$, for $i = 1, \cdots, n$

$$EZ_i u_i = \frac{C_n}{\sqrt{n}},$$

where C_n is a $k \times 1$ vector, where $C_{nj} \to \infty$, as $n \to \infty$, and $C_{nj}/\sqrt{n} \to 0$, for $j = 1, \cdots, k$. This is the violation of the exogeneity assumption because it allows for a mild correlation between the instruments and the structural error. Simultaneous asymptotics are used, so C_n grows along with the sample size n, but it grows slower than root n as shown in Assumption 1 below. This is different than the near exogeneity used by Berkowitz, Caner and Fang (2008) in which the covariance between the structural error and the instruments are "$C/\sqrt{n}$" where C is a constant vector. The reasons to choose this assumption are that it allows us to benefit from resampling methods and it is realistic in that it allows the covariance between the instruments and structural error to vary with the sample size. Recall that Assumption 1 below is used for obtaining better approximation in finite samples.

Thus, our assumption uses non-contiguous sequence rather than a constant "C" (contiguous sequence). However, we show that our results in Theorem 1 apply to contiguous (constant C) case as well. The only major difference is that when the block size is treated as a random variable (section 2.4, Corollary 1), the optimal block size converges in probability to 1/2 in the non-contiguous case (C_n), and does not converge in probability to 1/2 in the contiguous case (C). We show that in finite samples the size and the power results are similar in the contiguous and non-contiguous cases.

Another point for clarification is that we need $C_n/\sqrt{n}$ rather than $d/n^\kappa, 0 < \kappa < 1/2$ where d is a constant that we use for expositional purposes and for deriving Lemma 1. The proof of Lemma 1 is an extension of the contiguous case in Berkowitz, Caner, Fang (2008) to the non- contiguous case. Formulating Assumption 1 in our way makes clarifies this extension.

3

Note that there are no exogenous control variables in the system. This can be projected out easily. In order to simplify the notation, we do not include exogenous control variables in the equations above.

Our goal is to test $H_0 : \theta = \theta_0$. We also assume $EZ_iV_i' = 0$, for $i = 1, \cdots n$.

Now we start describing the Anderson-Rubin (1949) test. This is as follows:

$$AR(\theta_0) = [(y - Y\theta_0)'Z'/n^{1/2}]\hat{\Omega}^{-1}[Z'(y - Y\theta_0)/n^{1/2}], \tag{3}$$

where $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} Z_iZ_i'u_i^2$.

We can rewrite the Anderson-Rubin test (AR (θ_0)) as

$$AR(\theta_0) = \bar{S}_n'(var\bar{S}_n)^{-1}\bar{S}_n = n\bar{S}_n'\hat{\Omega}^{-1}\bar{S}_n, \tag{4}$$

where $\bar{S}_n = \frac{\sum_{i=1}^{n} Z_iu_i}{n} = \frac{Z'(y-Y\theta_0)}{n}$, and $var\bar{S}_n = \hat{\Omega}/n$. We can also demean Z_i in the variance formula and this does not change the asymptotics.

2.1 Assumptions

In this section we introduce our assumptions and discuss them.

Assumptions

Assumption 1. *For $i = 1\cdots, n$,*

(i).

$$EZ_iu_i = \frac{C_n}{\sqrt{n}}, \text{ for } i = 1, \cdots, n,$$

where C_n is $k \times 1$ vector and $C_{nj}/\sqrt{n} \to 0$, as $n \to \infty$, and $C_{nj} \to \infty$ as $n \to \infty$ for each cell $j = 1, \cdots, k$ in the vector C_n.

(ii).

$$\Omega = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} EZ_iZ_i'u_i^2,$$

where Ω is nonsingular and finite. Note that various types of data is discussed after the proof of Theorem 1 and here.

(iii).

$$cov(u_i, V_i) \neq 0.$$

(iv).

$$EZ_iV_i' = 0.$$

Assumption 2.

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\|Z_iu_i\|^3 < \infty.$$

4

Assumption 1 allows for a small covariance between the instruments and the structural error: it is how we operationalize minor violations of exogeneity. Assumption 1 is discussed above in this section. The situation when $C_n = O(\sqrt{n})$ is discussed in the remarks after Theorem 1.

Assumption 2 is needed for the strong law of large numbers approximation for obtaining the Berry-Esseen bounds. This assumption is discussed in Zhao, Wang, Wu (2004), as Remark 3 after Corollary 1 in their paper. This is a sufficient condition for the Berry-Esseen bound for the independent case. The triangular array case can also be obtained by Theorem 1 in Zhao, Wang, Wu (2004), and Assumption 2 is again sufficient.

2.2 Full Sample Result

In this subsection we derive the limiting distribution of the full sample Anderson-Rubin (1949) test under our violation of exogeneity condition in Assumption 1. Let q_α^1 be the $1 - \alpha$ quantile of a noncentral chi-square distribution with k degrees of freedom and with the noncentrality parameter as $C_n' \Omega^{-1} C_n / 2$, $(\chi_{k, C_n' \Omega^{-1} C_n / 2}^2)$.

Lemma 1. *Under Assumptions 1 and 2, we have*

$$P(AR(\theta_0) \geq q_\alpha^1) \to \alpha. \tag{5}$$

In this noncentrality parameter as indicated in Assumption 1, $C_n \to \infty$ but $C_n = o(n^{1/2})$. This shows that if we use the standard χ_k^2 critical values when there is a violation of the exogeneity assumption, then increasing the sample size makes the size of the AR-test worse. This is also what we observe in the simulations in Table 1 for the setups discussed in Section 4. This extends the limit result for contiguous sequences in Berkowitz, Caner, Fang (2008) which is a noncentral chi-square distribution, and the noncentrality parameter depends on constants rather than a slowly diverging sequence.

2.3 Resampling Technique

In this section we first describe the resampling technique that we use. We take a subset of size "b" (block size) from n observations. We resample from data "x" where $x = (x_1, \cdots x_n)$. The blocks in this resampling from "x" are "x_b" with size "b", and equal probability of $\binom{n}{b}^{-1}$. This is done via simple random sampling without replacement from the population. The size of the blocks plays a crucial role in achieving our result. Denote this resampling technique by "*". Notation such as P_*, E_* refer to calculations under $*$.

For our setup, we are interested in resampling from the following quantity: $Z'u = \sum_{i=1}^n Z_i u_i$, where $u = y - Y\theta_0$. Denoting the sample average by $\bar{S}_n = n^{-1} \sum_{i=1}^n Z_i u_i$,

$\bar{S}_b$ is the mean of the simple random sample of size "b" drawn without replacement from "n" observations (the mean of the b observations that are drawn out of the sample of n observations). Thus, for block size $b = fn, f \in [1/2, f_u], f_u < 1$ [1]. Note that we are not using directly the estimators in equations (1.2)(1.3) of Wu (1990). Instead we benefit from section 2, equations (2.2)(2.3) of Wu (1990). We also extend his case to independent random variables and the extension from iid to triangular arrays is simple and is discussed after Theorem 1. In terms of block size choice, our approach is also different than Wu (1990) where he considers all fractions between 0 and 1. In our problem, this results in overrejections of the null in large samples. We discuss our choice of block size after Theorem 1. Note that fraction of the sample "f" is the choice of the researcher. We now describe the fractionally resampled Anderson-Rubin test (FAR (θ_0)).

$$FAR(\theta_0) = \bar{S}_b'(var_* \bar{S}_b)^{-1}\bar{S}_b = \frac{b\bar{S}_b'\hat{\Omega}^{-1}\bar{S}_b}{(1-f)}, \tag{6}$$

where immediately after equation (2.2) on p.1440 of Wu (1990) or Theorem 2.2. of Cochran (1977) shows that $var_* \bar{S}_b = \frac{1-f}{b}\hat{\Omega}$. Observe that the right-hand side in (6) is slightly different than the right hand side in (4). This is due to the property of $var_* \bar{S}_b = \left[\frac{\hat{\Omega}}{b}\right](1-f)$. This will play an important role in the derivation of our main result.

We show that with $f = 1/2$, the technique is conservative. In our method, we basically resample from $n/2$ observations to obtain the score, and the estimate of the variance term is not resampled. We will describe why $n/2$ is chosen in the proofs rigorously. This is also intuitively described in the remarks after Theorem 1. The half-sampled FAR (θ_0) is

$$FAR_o(\theta_0) = n\bar{S}_{b_o}'\hat{\Omega}^{-1}\bar{S}_{b_o}, \tag{7}$$

where $b_o = n/2$. The algorithm for obtaining the empirical distribution function of these tests is described below after Theorem 1. The following Theorem is one of the main results of the paper. Theorem 1 (ii) clearly shows that the half-sampled FAR (θ_0) does not over-reject the null. Note that first we consider the general test in (6) in Theorem 1i, then the half-resampled test in (7) is given as its subcase in Theorem 1ii.

Theorem 1.*Given (5), under Assumptions 1 and 2,*

(i). For $f \in [1/2, f_u], f_u < 1$, and the test in (6), define

$$J_b(t) = P_*(\frac{b\bar{S}_b'\hat{\Omega}^{-1}\bar{S}_b}{(1-f)} \leq t),$$

then

$$\sup_t |J_b(t) - \phi_{\chi_f^2}(t)| \to 0 \quad a.s.,$$

---

[1] We can set $b = [fn]$, where $b/n \to f$ and [.] is the integer part of the number fn. But to save from notation and to keep up with section 2 of Wu (1990) framework we do not use that.

where $\phi_{\chi_f^2}(t)$ is the noncentral chi-square distribution with k degrees of freedom and with the noncentrality parameter $\frac{f}{1-f}\frac{(C_n+L)'(\Omega)^{-1}(C_n+L)}{2}$, and L is a multivariate normal distribution $N(0,\Omega)$.

(ii). Set $b, b_o = n/2, f = 1/2$, and define for the test in (7)

$$J_H(t) = P_*(n\bar{S}'_{b_o}\hat{\Omega}^{-1}\bar{S}_{b_o} \leq t),$$

then

$$\sup_t |J_H(t) - \phi_{\chi_{1/2}^2}(t)| \to 0 \quad a.s.,$$

where $\phi_{\chi_{1/2}^2}(t)$ is the noncentral chi-square distribution with k degrees of freedom and with the noncentrality parameter $(C_n + L)'(\Omega)^{-1}(C_n + L)/2$.

Remarks.

1. With the case of $f_o = 1/2$, we see that the noncentrality parameter in Theorem 1ii is closer to the noncentrality parameter in the full sample distribution of $AR(\theta_0)$ in Lemma 1. To expand on that point, we use the critical values on the right-hand side of the distribution in FAR (θ_0) for inference. This is where $L > 0$. So

$$\frac{1}{2}(C_n + L)'\Omega^{-1}(C_n + L) > \frac{1}{2}C_n'\Omega^{-1}C_n.$$

Since $\frac{1}{2}C_n'\Omega^{-1}C_n$ is the noncentrality parameter in Lemma 1, this shows that resampling technique is valid for $f_o = 1/2$, but conservative. The other choices of f such that $f > 1/2$, proves to be more conservative by analyzing the noncentrality parameters since in that case $f/(1-f) > 1$. If we had chosen $f < 1/2$, then it is not clear that we have

$$\frac{f}{1-f}\frac{(C_n + L)'\Omega^{-1}(C_n + L)}{2} > \frac{1}{2}\frac{C_n'\Omega^{-1}C_n}{2}.$$

The above inequality may be the other way around and the test may be oversized.

2. The proof uses Assumption 2 which is for independent data. The triangular arrays can be easily done, and this is discussed and shown after the Assumption 2.

3. An important question is what if in Assumption 1, we have $EZ_iu_i = \frac{C}{n^{1/2}}$, where C is a constant? In that case by following the proofs of Lemma 1, and Theorem 1 that all the results, Lemma 1, Theorem 1 will hold with C replacing C_n everywhere.

4. An important issue is that theoretically we can get a better result than Theorem 1, if we treat the fraction (f) as a random variable and choose the block size such that we recover the limit in Lemma 1 with our resampling technique. Corollary 1 in section 2.4 proves that point.

2.4 The Block Size as a Random Variable

The overriding issue is that whether we can set up the fraction of the sample size as a random variable and then recover the limit in Lemma 1, even though this is infeasible in exact samples, but feasible in large sample setup (i.e. half of the sample). We can show below that there is such a choice f_n^* that $f_n^* \xrightarrow{p} 1/2$, and

$$\frac{f_n^*}{1 - f_n^*}(C_n + L)'\Omega^{-1}(C_n + L) = C_n'\Omega^{-1}C_n. \tag{8}$$

We try to find the optimal f_n^* such that we satisfy (8) in the limit. To that end see that, by simple algebra,

$$f_n^* = \frac{C_n'\Omega^{-1}C_n}{2C_n'\Omega^{-1}C_n + C_n'\Omega^{-1}L + L'\Omega^{-1}C_n + L'\Omega^{-1}L}.$$

We can further arrange that to have

$$f_n^* = \frac{1}{2 + \left[\frac{C_n'\Omega^{-1}L + L'\Omega^{-1}C_n + L'\Omega^{-1}L}{C_n'\Omega^{-1}C_n}\right]}. \tag{9}$$

Clearly $f_n^* \xrightarrow{p} 1/2$. Also it is easy to verify that f_n^* choice in (9) satisfies (8).

To be rigorous (an infeasible estimate of f_n^*), we propose

$$\hat{f}_n^* = \frac{1}{2 + \left[\frac{C_n'\hat{\Omega}^{-1}\hat{L} + \hat{L}'\hat{\Omega}^{-1}C_n + \hat{L}'\hat{\Omega}^{-1}\hat{L}}{C_n'\Omega^{-1}C_n}\right]}, \tag{10}$$

where $\hat{L} = n^{-1/2}\sum_{i=1}^n (Z_i u_i - EZ_i u_i)$, and $\hat{\Omega} = n^{-1}\sum_{i=1}^n Z_i Z_i' u_i^2$. This is asymptotically equivalent to f_n^* (still infeasible in small samples), and is needed to derive the proof of Corollary 1.

The main point of the paper is the following Corollary. It also shows why half the sample size makes theoretical sense and shows we can recover the limit in Lemma 1 with $\hat{f}_n^* \xrightarrow{p} 1/2$.

Corollary 1. *Given (5), Assumption 1 with $\max_i E\|Z_i u_i\|^3 < \infty$, with $\hat{b}_n^* = \hat{f}_n^* n$, (where $\hat{f}_n^* \xrightarrow{p} 1/2$), and $\hat{f}_n$ is defined in (10),*

$$J_*(t) = P_* \left(\frac{\hat{b}_n^* \bar{S}_{\hat{b}_n^*}' \hat{\Omega}^{-1} \bar{S}_{\hat{b}_n^*}}{1 - \hat{f}_n^*} \leq t \right),$$

then

$$\sup_t |J_*(t) - \phi_*(t)| \to 0, \quad a.s.,$$

where $\phi_*(t)$ is the noncentral χ^2 distribution with k degrees of freedom with the noncentrality parameter $C_n'\Omega^{-1}C_n/2$.

Note that $\bar{S}_{\hat{b}_n^*}$ is the sample average of $Z_i u_i$ using $\hat{b}_n^*$ observations out of n without replacement.

Remarks.

1. One issue is how important is the choice of specific f_n^* (Note that we talk about f_n^* rather than $\hat{f}_n^*$ which is asymptotically equivalent. This is done since in the limit we have f_n^*). Assume now instead we select $\tilde{f}_n \xrightarrow{p} 1/3$, and

$$\tilde{f}_n = \frac{1}{3 + \frac{C_n'\Omega^{-1}L + L'\Omega^{-1}C_n + L'\Omega^{-1}L}{C_n'\Omega^{-1}C_n}}.$$

Then by simple algebra

$$\frac{\tilde{f}_n}{1 - \tilde{f}_n} = \frac{C_n'\Omega^{-1}C_n}{C_n'\Omega^{-1}C_n + +(C_n + L)'\Omega^{-1}(C_n + L)},$$

which means that

$$\frac{\tilde{f}_n}{1 - \tilde{f}_n}(C_n + L)'\Omega^{-1}(C_n + L) < C_n'\Omega^{-1}C_n.$$

So with $\tilde{f}_n$ the FAR (θ_0) is oversized.

2. Another issue is with large C_n we need large block size near $1/2$ in practice. Related to that point, even though in large samples theoretically choosing $1/2$ makes sense, this is not obvious from a finite sample standpoint. One possible way to choose the block size in finite samples is to take the expectation of f_n^* (analysis is basically the same with sample version $\hat{f}_n^*$, hence it is skipped). First,

$$Ef_n^* = \frac{1}{2 + \frac{k}{C_n'\Omega^{-1}C_n}},$$

since $E[C_n'\Omega^{-1}L + L'\Omega^{-1}C_n] = 0$, via $L \equiv N(0, \Omega)$, and $E(L'\Omega^{-1}L) = E\chi_k^2 = k$. Next assume diagonal Ω^{-1} (i.e. conditional homoskedasticity) and no correlation among instruments to have

$$C_n'\Omega^{-1}C_n \leq kn(corr(Z_i, u_i))^2,$$

via Assumption 1, where correlation between Z_i, u_i is denoted as $corr(Z_i u_i)$.

$$Ef_n^* \leq \frac{1}{2 + \frac{1}{n[corr(Z_i, u_i)]^2}}.$$

With $k = 1$, $n = 100$, trying a reasonable correlation of 0.1 between Z_i, u_i, the right hand side of the above equation is $1/3$. With correlation as 0.3, the right hand side is 0.473, which is chosen with 47 observations. With larger correlations the block size is almost 50 observations out of 100. In simulations we see that the block size choice of $1/3$ does well in terms of size and power. A recipe may be use the above formula given your n, by trying reasonable correlations and checking whether the results are robust to differing block sizes. Note that this is an upper bound, smaller values may be used but lead to higher size than the block size that corresponds to the bound.

9

3. An important fact to note here is that the larger the n is the larger the bound on the expectation of the block size.

4. In the case of a random variable choice of block size, replacing C_n with C makes a difference. Even though we can recover the limit in Lemma 1 by the resampling technique, we do not have the fraction converging in probability to $1/2$. To see these points, express

$$\hat{f} = \frac{C'\hat{\Omega}^{-1}C}{2C'\hat{\Omega}^{-1}C + C'\hat{\Omega}^{-1}\hat{L} + \hat{L}'\hat{\Omega}^{-1}C + \hat{L}'\hat{\Omega}^{-1}\hat{L}}.$$

So basically we replace C_n's in $\hat{f}_n^*$ with C. It is clearly not the case now that $\hat{f} \xrightarrow{p} 1/2$. But we have still

$$\frac{\hat{f}}{1 - \hat{f}}(C + \hat{L})'\hat{\Omega}^{-1}(C + \hat{L}) = C'\hat{\Omega}^{-1}C.$$

Also we have

$$E\hat{f} \leq \frac{1}{2 + \frac{1}{ncorr(Z_iu_i)^2}},$$

as before. So even though still there is an optimal block size (fraction) choice, it does not converge in probability to $1/2$, but in small samples we can still benefit from the same formula for $E\hat{f}$.

2.5 Power Issues of $AR(\theta_0), FAR(\theta_0)$

The analysis of power will be conducted with fixed f choice to show the results without too much notation and without losing any generality. Here we briefly show that the $AR(\theta_0)$ is consistent against fixed alternatives when there is strong identification and imperfect exogeneity. So in equation (2), assume Π has full column rank, but Assumption 1 is still holding. If the true parameter value is θ_1 (for power exercise here), and $\theta_1 \neq \theta_0, \theta_1 - \theta_0 = c$, where c is a nonzero constant, we test $H_0 : \theta = \theta_0$.

In (3) in $AR(\theta_0)$ we have the following term

$$\frac{Z'(y - Y\theta_0)}{\sqrt{n}} = \left(\frac{\sum_{i=1}^n Z_iY_i'}{\sqrt{n}}\right)(\theta_1 - \theta_0) + \frac{\sum_{i=1}^n Z_iu_i}{\sqrt{n}}, \tag{11}$$

where

$$Y_i = \Pi Z_i + V_i.$$

Then in (11)

$$\left(\frac{\sum_{i=1}^n Z_iY_i'}{\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^n Z_iZ_i'}{\sqrt{n}}\right)\Pi + \left(\frac{\sum_{i=1}^n Z_iV_i'}{\sqrt{n}}\right). \tag{12}$$

Clearly if Π is full rank, then

$$\left(\frac{\sum_{i=1}^n Z_iZ_i'}{\sqrt{n}}\right)\Pi \xrightarrow{p} \infty. \tag{13}$$

Then use (13) into (11)(12) with $\theta_1 - \theta_0 = c \neq 0$ to have $AR(\theta_0) \xrightarrow{p} \infty$.

10

The larger issue is whether the resampling technique that we suggest will have power, and whether this is consistent against fixed alternatives. The key term in the analysis is $\bar{S}_b$ in equation (6). Note that

$$\bar{S}_b = \frac{\sum_{i=1}^{b}(Z_i(y_i - Y_i\theta_0)')_*}{\sqrt{b}},$$

where $(.)_*$ represents resampling b observations out of n without replacement. As in (11) we can express that

$$
\begin{aligned}
\bar{S}_b &= \frac{\sum_{i=1}^{b}(Z_iY_i')_*}{\sqrt{b}}(\theta_1 - \theta_0) + \frac{\sum_{i=1}^{b}(Z_iu_i)_*}{\sqrt{b}} \\
&= \frac{\sum_{i=1}^{b}(Z_iZ_i')_*}{\sqrt{b}}\Pi(\theta_1 - \theta_0) + \frac{\sum_{i=1}^{b}(Z_iV_i')_*}{\sqrt{b}}(\theta_1 - \theta_0) + \frac{\sum_{i=1}^{b}(Z_iu_i)_*}{\sqrt{b}}.
\end{aligned}
\tag{14}
$$

Note that the first term on the right hand side can be rewritten as

$$\sqrt{b}\left(\frac{\sum_{i=1}^{b}(Z_iZ_i')_*}{b}\right).$$

That term converges in probability to infinity at rate $\sqrt{b} = \sqrt{n}\sqrt{f}$ rate. The test will not be consistent if we use the resampled critical values, even though the test diverges at rate n. However, in finite sample we may have power with smaller "f". So simulation exercises become crucial for small sample choice of block size. We should note that since the main problem is size, a reasonable power loss is acceptable.

An important fact to note is that the power will be less with large C_n (C_n being near $\sqrt{n}$). We observe this also in our simulation exercise below.

Note that power results also go through with constant C replacing C_n.

2.6   The Algorithm

Next we write the algorithm to test the null of $H_0 : \theta = \theta_0$ by using the critical values obtained from empirical distribution function of the half-sampled $FAR(\theta_0)$ in (7).

Step 1: First calculate the terms $\hat{\Omega}$ from the full sample of Z_i, u_i as described at the beginning of this section.

Step 2: Denote $y_{b_o}, Y_{b_o}, Z_{b_o}$ as draws of block size $b_o = n/2$ from full sample y, Y, Z without replacement, respectively. Note that $y_{b_o} : n/2 \times 1, Y_{b_o} : n/2 \times m, Z_{b_o} : n/2 \times k$. Form

$$\bar{S}_{b_o} = [Z_{b_o}'(y_{b_o} - Y_{b_o}\theta_0)]/(n/2)$$

.

Step 3. Form $FAR_o(\theta_0)$ in (7) by using steps 1-2.

Step 4. Repeat steps 2-3, J times. (J may be 1000, or 5000) Then sort J values of $FAR_o(\theta_0)$ to form the empirical distribution function.

Step 5. For a 5% test find the 95 percentile of the empirical distibution function in step 4.

Step 6. Reject the null of $H_0 : \theta = \theta_0$, if the full sample $AR(\theta_0)$ as described in equation (3) is larger than the 95th percentile in step 5.

Note that in the above algorithm, $\hat{\Omega}$ is calculated from the full sample. Only the numerator of the test statistic, the score, has to be resampled. The main technical reason for that is shown in the proof of Theorem 1. Basically for any block size b, p.1440 of Wu (1990) or Theorem 2.2 of Cochran (1977) shows that $var_* \bar{S}_b = \frac{1-f}{b} \hat{\Omega}$ in our case.

3 Comparison With Subsampling and Variants

In this section we compare the resampling technique employed here in Theorem 1 with subsampling. Note that subsampling the AR (θ_0) test will not work because it will be oversized. A simple counterexample can be seen from our results since in that case $\sqrt{f} = \sqrt{b/n} = o(C_n^{-1})$, so $b/n \to 0$, as $b \to \infty, n \to \infty$, meaning $f \to 0$. That means the noncentrality parameter in Theorem 1 becomes with $C_n = o(n^{1/2})$:

$$\frac{f}{1-f} \frac{C_n' \Omega C_n}{2} \to 0.$$

Hence, the subsampled $AR(\theta_0)$ test statistics will converge to the standard χ^2 limit. Clearly this is oversized. The subsampled limit is stochastically less than the one in (5).

The work of Andrews and Guggenberger (2007a) is important for this section because it analyzes cases in which the subsampling approach works and cases in which it fails. Their assumptions do not cover the resampling used here. Andrews and Guggenberger (2007a) specifically assume $b_{n,s}/n \to 0$ as $b_{n,s} \to \infty, n \to \infty$, where $b_{n,s}$ represents the block size in subsampling. In our case $b_n/n = f, f \in [1/2, f_u], f_u < 1$, where b_n denotes the block size in the resampling method we use. This is also different from Wu (1990) where he covers all $0 < f < 1$. In other words, subsampling is only concerned with small blocks and the technique that we use takes at least half of the sample as the block size.

Apart from the treatment of the block size issue, another very important difference between subsampling and the fractional resampling is the variance terms. In order to illustrate this, take any sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and resample with the technique here (sample $b_n = fn$ observations out of n without replacement) and denote the sample mean averaged over b observations as $\bar{X}_b$. Then the variance of $\bar{X}_b$ under simple random sampling without replacement is $(1 - f)\hat{\sigma}^2/b_n$, where $\hat{\sigma}^2$ is the standard sample variance of each X_i (p.1440, after equation (2.2), Wu, 1990). This is entirely different from the

12

subsampling. In subsampling the counterpart is $\hat{\sigma}^2/b_{n,s}$. This point will be analyzed in detail with an example below.

Even though the assumptions employed in Andrews and Guggenberger (2007a) do not apply to our setup, whenever the resampled distribution of any test statistic is stochastically greater than or equal to the original distribution (the target distribution that we want to replicate), then the nominal level $\alpha \in (0,1)$ of the resampled test has the asymptotical level α. This is true since the critical value of the resampled test is greater than or equal to critical values from the original test limit. For this point see equation (9.14) and Comment 2 after Theorem 2 in Andrews and Guggenberger (2007a). This is clearly true in our case. By comparing the limits in Theorem 1 and (5), it is obvious that the limit of fractionally resampled $AR(\theta_0)$ test is stochastically greater than or equal to (5). At $f = 1/2$, the limit is stochastically equal to (5).

Now we illustrate the difference between subsampling and fractional resampling in a simple example used in section 2 of Andrews and Guggenberger (2007a) analyzing the issue of a simple boundary problem. The true parameter θ_0 is nonnegative. Assume that X_i is iid with N (0,1), for $i = 1, \cdots, n$. The Maximum Likelihood Estimator (MLE) of θ_0 is $\hat{\theta}_n = max\{\bar{X}_n, 0\}$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. The distribution of $\hat{\theta}_n$ is

$$\hat{\theta}_n \sim max\{Z_n, 0\}, \quad Z_n \sim N(\theta_0, \frac{1}{n}).$$

Then we subsample ($b_{n,s}/n \to 0$, as $b_{n,s} \to \infty, n \to \infty$), which $b_{n,s} = o(n)$, the subsampled estimator is $\hat{\theta}_{b_{n,s},j} = max\{\bar{X}_{b_{n,s},j}, 0\}$ where $\bar{X}_{b_{n,s},j} = \frac{1}{b_{n,s}}\sum_{i=j}^{j+b_{n,s}-1} X_i$ and the distribution is

$$\hat{\theta}_{b_{n,s},j} \sim max\{Z_{b_{n,s}}, 0\}, \quad Z_{b_{n,s}} \sim N(\theta_0, \frac{1}{b_{n,s}}).$$

It is clear that the distribution of $\hat{\theta}_{b_n,j}$ does not replicate the distribution of $\hat{\theta}_n$. This is thoroughly discussed in Andrews and Guggenberger (2007a). The main reason is that the subsample estimator is closer to the boundary of parameter space than the full sample estimator. To see this $var\hat{\theta}_{b_{n,s},j}$ when $Z_n > 0$ is $1/b_{n,s}$, and $var\hat{\theta}_n$ is $1/n$ when $Z_n > 0$. Since $b_{n,s} = o(n)$, $1/b_{n,s}$ is larger than $1/n$ and hence more variable near the boundary.

In the fractional resampling $b_n = fn, f \in [1/2, f_u], f_u < 1$, so $b_n = O(n)$. When we use this technique in the case of the former example (with $Z_{b_n} > 0$), the variance of the fractionally resampled estimator is:

$$var\tilde{\theta}_{b_n} = \frac{1-f}{b_n}.$$

To understand this, we use section 2 of Wu (1990). The variance of the resampled mean is $\frac{(1-f)varX_i}{b_n}$, where $varX_i = 1$ in this example. Then note that compared to the original variance and the subsampled ones

$$var\hat{\theta}_{b_{n,s},j} > var\hat{\theta} \geq var\tilde{\theta}_{b_n},$$

since

$$\frac{1}{b_{n,s}} > \frac{1}{n} \geq \frac{(1-f)}{b_n} = \left(\frac{1-f}{f}\right)\frac{1}{n},$$

and $f \in [1/2, f_u], f_u < 1$. At $f = 1/2$ we have the optimal choice and capture the variance. This shows that variability in this technique is less than or equal to the subsampling technique.

There are also other techniques suggested by Andrews and Guggenberger (2007b) besides subsampling for recovering the limits of tests under nonstandard situations. One test is the hybrid subsample method. In this test the researcher takes the maximum of the subsampling and the limit under the "stochastically largest" critical value of the distribution. The critical values using that approach will be very large and will waste power in our case. The second approach suggested by Andrews and Guggenberger (2007b) to overcome the problems of subsampling is size-corrected subsampling. This adjusts the critical values by adding constants depending on the problem. This may help in our case, but when the system is over-identified, this requires a grid search over R^k that is computationally very intensive. Our method does not require a grid search, and is not computationally demanding. Andrews and Guggenberger (2007b) show that these two methods give mixed results in small-samples in terms of size and power of the tests in discontinuous limit cases.

4   Simulation

This section shows the small sample properties of the tests that are proposed in equations (6) and (7). We consider several block sizes for (6). We use the setup in section 2, namely

$$y_i = Y_i \theta_0 + u_i,$$

$$Y_i = Z_i \Pi + V_i,$$

for $i = 1, \cdots, n$. The sample size is n and varies between 100 and 200. We consider the case of one instrument and one endogenous regressor, so $k = 1$, $m = 1$ (exact identification). A case with overidentification is also considered, but not reported here because the results are very similar. Π can take the values of 1 (strong identification), and 0.1 (weak identification). The iid data (Z_i, u_i, V_i) are generated from a joint normal distribution $N(0, \Omega)$ where

$$\Omega = \begin{bmatrix} 1 & cov(Z_i, u_i) & 0 \\ cov(Z_i, u_i) & 1 & 0.5 \\ 0.5 & 0 & 1 \end{bmatrix}.$$

So $var Z_i = var u_i = var V_i = 1$, $cov(Z_i, V_i) = 0$, $cov(u_i, V_i) = 0.5$. For the size exercise $\theta_0 = 0$, we test $H_0 : \theta = 0$. For the power $\theta_0 = -1, -0.8, -0.5, -0.2, 0.2, 0.5, 0.8, 1$.

14

We consider three setups for the $cov(Z_i, u_i)$ term. The first setup is consistent with the near exogeneity assumption in Berkowitz, Caner and Fang (2008):

$$cov(Z_i, u_i) = \frac{C}{n^{1/2}}, \tag{15}$$

and C takes the values of 2, 3, and 5. As C becomes larger, endogeneity becomes more problematic. And, the researcher picks a terrible instrument when $C = 5$.

In the second setup we have:

$$cov(Z_i, u_i) = D, \tag{16}$$

where D is a constant and takes values 0.1, 0.2, 0.3. With this setup we expect large size distortions to emerge as the sample becomes larger because the drift D is multiplied with the square root of the block size in the score in the test statistic. We also used negative values for the covariance term but the results do not change and hence are not reported.

In third setup, we return to Assumption 1 of this paper:

$$cov(Z_i, u_i) = \frac{an^{1/3}}{n^{1/2}}, \tag{17}$$

where $C_n = an^{1/3}$, and a takes the values of 0.25, 0.5, and 1. At $n = 100$, these correspond to covariance (correlation, since variances are normalized at 1) of 0.12, 0.23, and 0.46 respectively for $a = 0.25, 0.5, 1$.

For both getting the distribution of the resampled FAR (θ_0) in the algorithm in section 2.3, and for calculating the rejection rates of the null (i.e. comparing the full sample AR (θ_0) with the 90% critical value calculated from resampled $FAR(\theta_0)$), 1000 iterations are used. We try the block sizes $b = 25, 32, 37, 50$ for $n = 100$ and for $n = 200$, we set $b = 50, 65, 75, 100$. For the power exercise, $n = 100$ is used only with $\Pi = 1$. When $\Pi = 0.1$, the power is very low due to weak identification.

When we use half the sample size our results are very conservative in every simulation that is conducted (0 size with 0 power) and they are not reported. This result is related to the behavior of the higher order terms as well as the power problems associated with our technique. Clearly we obtain the asymptotic result in Theorem 1 when we use half the sample. In the simulation that we have done with $n = 200$, and C larger than 5, we approach the perfect size with a block size near one half. This simulation is conducted to check if we approach the asymptotics (it is not reported). Also we consider $a = 2$ and $n = 200$ and we see that we approach the asymptotics and got excellent size with $f = 0.41$. But as mentioned above, it is clear that the optimal f_n can be thought as $\hat{f}_n^* \to 1/2$.

However, it is not possible to replicate the second order term in our test statistics via resampling, as can be seen in p.1450 of Wu (1990). This is an undesirable characteristic. This is one of the reasons we get very conservative results at half-sample. However, with block sizes smaller than $1/2$ we can do well in finite samples as shown below.

15

Table 1 reports the size of the full sample regular $AR(\theta_0)$ test in (3). This is compared with asymptotic critical values for χ_1^2 distribution at 10% level. We report the rejection rates of the true null in Table 1. We see that both in setups 1 and 2 the actual size is very large. In setup 1, at $C = 2$, the size is 66% with $n = 200$. This shows there is a major size distortion problem if we use $AR(\theta_0)$ for tests when there is a violation of exogeneity. This can also be seen for t-tests in Berkowitz, Caner and Fang (2008). The size calculations are done for $\Pi = 1$. Simulations for the case when $\Pi = 0.1$ are also done, but the results are not reported because they are very similar to the case of $\Pi = 1$. Another point to make is the size gets worse with large sample size in setup 2. This is an important warning to applied researchers who believe that increasing the sample size can correct for size distortions! In fact, as is very clear from setup 3, a larger sample size can also increase size distortions. With $a = 1$, and $n = 100$, the correlation is 0.46, the instrument is poorly selected and there is a huge size distortion.

Tables 2-4 and 5-7 report the actual size when we use FAR (θ_0) in (6) with different block sizes when there is strong identification and weak identification respectively. Compared to Table 1, Tables 2-4 show that at each block size the size is reduced. An important question is whether we can achieve the perfect size of 10%? For setup 1 in Table 2, at $n = 100$ and $b = 25$, this block size results in a 12.1% size at $C = 2$, and when $b = 32$ for $C = 3$ the size is 6.5%. For setup 2 in Table 3, when cov $Z_i, u_i = 0.2$, and $b = 25$ the size is 12% size for $n = 100$. In setup 3 in Table 4, the size is 18.4% when $a = 0.50$, with $b = 25$,. In Table 4, when n increases to 200, with block size of 50 at the same covariance level of 0.2 the size is 43.0%. When $a = 0.5$ (corresponding to 0.23 correlation) we have sizes of 1.7% and 7.7% with $n = 100, 200$ respectively. Generally we see that block sizes above 1/3 of the sample size are very conservative when there is some low/mild correlation between the instrument and the structural error. We see from the simulations that at low/mild correlation levels the block size choice between 1/4 to 1/3 of the sample size gives good results. In an unrealistic case of $C = 5$ we see that block size of 37 at $n = 100$ gives 9.6% size at 10% level. If we are concerned with reducing size distortions, a good approach is to use a fraction of 1/3 for minor/moderate violations of the exogeneity assumption.

Even though it is not reported in Table 4, we have also simulated $b = 80$ for $a = 1$ with $n = 200$. The size is 2.9%. So there is a sharp decline of size from $b = 75$ to $b = 80$. We should note that we also try the very unrealistic case of $a = 2$ (this corresponds to 0.83 correlation between the instrument and the structural error) with $n = 200$, with $b = 85$ and we get a very good size that is close to 10%.

Next we analyze the power properties of the FAR (θ_0). Tables 8-10 consider the power of FAR (θ_0) when $b = 25, 32, 37$ with $n = 100, \Pi = 1$. We see that $b = 25$ has the best power, and also $b = 32$ has reasonably good power. If we are concerned with by size and

16

Table 1: Size at 10%, $AR(\theta_0)$ test, $\Pi = 1$

Sample Size	Setup 1			Setup 2			Setup 3		
	$C = 2$	$C = 3$	$C = 5$	$D = 0.1$	$D = 0.2$	$D = 0.3$	$a = 0.25$	$a = 0.5$	$a = 1$
100	65.0	93.0	100.0	26.0	64.0	93.0	32.8	77.9	93.0
200	66.0	93.0	100.0	43.0	88.0	99.0	42.2	91.7	100.0

Note: Setup 1 is explained in (15), Setup 2 is explained in (16). "D" represents the covariance between the instrument and the structural error. Setup 3 and constant "a" is explained in (17).

Table 2: Setup 1, Size at 10%, $FAR(\theta_0)$, $\Pi = 1$

	$n = 100$			$n = 200$		
Block size	$b = 25$	$b = 32$	$b = 37$	$b = 50$	$b = 65$	$b = 75$
$C = 2$	12.1	0.1	0.0	12.6	1.0	0.0
$C = 3$	42.0	6.5	0.1	46.1	8.8	0.1
$C = 5$	96.0	66.7	9.6	98.0	74.0	12.7

Note: This is the test statistic in (6) and setup 1 is (15).

power, then a good strategy is first to resample with 1/4 of the sample size and then try the 1/3 block size; and, if the results are the same, report them them with the 1/3 block size. With larger sample sizes, if the results are in conflict, a block of 1/3 should used; if the sample is small then a 1/4 block size is preferred.

We also try the minimum volatility method used in subsampling to obtain better finite sample results, but the block $b = n/3$ has better size properties.

5   Conclusion

This paper shows that it is possible to conduct inference using instrumental variables when there is a mild violation of the strict exogeneity assumption. We use a resampling technique

Table 3: Setup 2, Size at 10%, $FAR(\theta_0)$, $\Pi = 1$

	$n = 100$			$n = 200$		
Block size	$b = 25$	$b = 32$	$b = 37$	$b = 50$	$b = 65$	$b = 75$
$D = 0.1$	2.0	0.0	0.0	5.0	0.0	0.0
$D = 0.2$	12.0	0.9	0.0	11.9	5.4	0.2
$D = 0.3$	42.0	9.0	0.1	88.0	8.8	2.3

Note: This is the test statistic in (6) and setup 2 is (16). D represents the covariance between the instrument and the structural error.

Table 4: Setup 3, Size at 10%, $FAR(\theta_0)$, $\Pi = 1$

	$n = 100$			$n = 200$		
Block size	$b = 25$	$b = 32$	$b = 37$	$b = 50$	$b = 65$	$b = 75$
$a = 0.25$	2.6	0.0	0.0	4.0	0.2	0.0
$a = 0.50$	18.4	1.7	0.0	43.0	7.7	0.0
$a = 1$	93.6	51.9	4.9	99.4	92.4	35.1

Note: This is the test statistic in (6) and setup 3 is (17). "a" is in (17) as $C_n = an^{1/3}$.

Table 5: Setup 1, Size at 10%, $FAR(\theta_0)$, $\Pi = 0.1$

	$n = 100$			$n = 200$		
Block size	$b = 25$	$b = 32$	$b = 37$	$b = 50$	$b = 65$	$b = 75$
$C = 2$	10.8	1.3	0.0	17.0	0.1	0.0
$C = 3$	39.1	8.1	0.1	43.4	8.9	0.1
$C = 5$	96.8	70.4	10.3	97.6	72.6	12.7

Note: This is the test statistic in (6) and setup 1 is (15).

Table 6: Setup 2, Size at 10%, $FAR(\theta_0)$, $\Pi = 0.1$

	$n = 100$			$n = 200$		
Block size	$b = 25$	$b = 32$	$b = 37$	$b = 50$	$b = 65$	$b = 75$
$D = 0.1$	0.8	0.2	0.0	0.2	0.2	0.0
$D = 0.2$	12.2	1.1	0.0	38.7	6.0	0.1
$D = 0.3$	44.4	5.4	0.2	87.3	46.3	3.4

Note: This is the test statistic in (6) and setup 2 is (16). D represents the covariance between the instrument and the structural error.

Table 7: Setup 3, Size at 10%, $FAR(\theta_0)$, $\Pi = 0.1$

	$n = 100$			$n = 200$		
Block size	$b = 25$	$b = 32$	$b = 37$	$b = 50$	$b = 65$	$b = 75$
$a = 0.25$	2.7	0.1	0.0	5.3	0.1	0.0
$a = 0.50$	20.7	2.1	0.0	41.0	6.9	0.1
$a = 1$	91.3	56.5	5.6	99.7	93.4	36.6

Note: This is the test statistic in (6) and setup 3 is (17). "a" is in (17) as $C_n = an^{1/3}$.

18

Table 8: Setup 1, Power , $FAR(\theta_0)$

	$b = 25, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$C = 2$	99.2	98.8	55.2	0.2	57.2	93.7	98.0	99.1
$C = 3$	99.9	96.2	21.0	2.7	84.3	98.4	99.6	100.0
$C = 5$	88.8	66.0	0.2	65.4	99.4	99.9	100.0	100.0
	$b = 32, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$C = 2$	95.4	83.6	13.2	0.0	16.8	58.3	81.1	87.7
$C = 3$	89.6	67.0	2.5	0.1	42.5	77.9	88.1	92.3
$C = 5$	43.6	21.3	0.0	20.3	87.7	95.4	97.2	96.8
	$b = 37, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$C = 2$	43.9	21.3	0.6	0.0	0.5	8.2	15.3	23.7
$C = 3$	30.0	8.5	2.5	0.0	1.9	14.6	24.5	33.1
$C = 5$	0.3	0.7	0.0	0.6	25.1	44.6	51.0	52.7

Note: This is the test statistic in (6) and setup 1 is (15).

Table 9: Setup 2, Power , $FAR(\theta_0)$

	$b = 25, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$D = 0.1$	100.0	99.6	85.4	1.5	26.6	83.3	96.4	99.1
$D = 0.2$	99.9	99.2	55.6	0.2	56.3	94.5	98.5	99.5
$D = 0.3$	99.4	97.3	20.3	1.8	83.8	98.1	99.8	100.0
	$b = 32, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$D = 0.1$	97.3	93.0	40.3	0.3	4.3	33.7	66.1	80.8
$D = 0.2$	95.2	86.5	12.1	0.0	14.7	56.3	80.0	86.9
$D = 0.3$	90.9	66.6	2.1	0.2	40.5	76.8	88.9	94.3
	$b = 37, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$D = 0.1$	56.2	38.1	2.6	0.0	0.1	3.2	9.5	16.3
$D = 0.2$	41.7	21.5	0.3	0.0	0.5	5.8	18.3	25.8
$D = 0.3$	29.3	7.6	0.0	0.0	2.0	12.7	28.9	34.9

Note: This is the test statistic in (6) and setup 2 is (16).

Table 10: Setup 3, Power , $FAR(\theta_0)$

	$b = 25, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$a = 0.25$	100.0	99.8	81.3	1.2	32.1	85.8	97.8	98.6
$a = 0.50$	99.9	99.0	44.1	0.1	67.2	95.5	99.2	99.8
$a = 1.00$	96.4	78.2	0.4	48.3	98.5	99.7	99.7	99.9
	$b = 32, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$a = 0.25$	97.7	91.5	34.5	0.0	4.0	37.1	68.6	82.3
$a = 0.50$	94.3	79.5	7.5	0.0	20.7	64.7	84.6	88.0
$a = 1.00$	73.0	28.3	0.0	9.7	82.4	93.2	96.3	97.4
	$b = 37, n = 100$							
$\theta_0 =$	-1	-0.8	-0.5	-0.2	0.2	0.5	0.8	1
$a = 0.25$	52.0	32.1	2.2	0.0	0.0	4.4	9.0	16.7
$a = 0.50$	38.2	16.9	0.0	0.0	0.5	9.0	21.8	27.7
$a = 1.00$	13.1	1.4	0.0	0.3	16.7	36.7	47.7	51.8

Note: This is the test statistic in (6) and setup 3 is (17). "a" is in (17) as $C_n = an^{1/3}$

which draws half the sample from the all of the sample without replacement. The fractionally resampled Anderson-Rubin test $(FAR(\theta_0))$ does not overreject the null in large samples and this result is robust when the instruments are weak. Instruments that perfectly satisfy the knife-edge orthogonality assumption are few and far between. If researchers carefully pick instruments and use our method with a block between one-quarter and one-third, they can draw reliable inferences using instrumental variable methods.

APPENDIX

Proof of Lemma 1. First see that

$$\sqrt{n}\bar{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i u_i = [n^{-1/2} \sum_{i=1}^{n} (Z_i u_i - E Z_i u_i)] + \sqrt{n} E Z_i u_i.$$

Then via Central Limit Theorem

$$[n^{-1/2} \sum_{i=1}^{n} (Z_i u_i - E Z_i u_i)] \xrightarrow{d} N(0, \Omega),$$

and by Assumption 1

$$\sqrt{n} E Z_i u_i = C_n.$$

Then with $C_n = o(\sqrt{n})$, and using these in $AR(\theta_0)$ we get the desired result. To see this in detail, set the concentration parameter as $\lambda_n = C_n' \Omega^{-1} C_n / 2$. Then by the noncentral chi-square distribution (with mean $k + \lambda_n$, and variance $2(k + 2\lambda_n)$)

$$\frac{AR(\theta_0) - (k + \lambda_n)}{\sqrt{2(k + 2\lambda_n)}} \xrightarrow{d} N(0, 1).$$

Then by the standard result for the non central chi-square distributions, as $C_n \to \infty$, the $(1 - \alpha)$ th quantile q_α^1 of χ^2_{k,λ_n} we have

$$\frac{q_\alpha^1 - (k + \lambda_n)}{\sqrt{2(k + 2\lambda_n)}}$$

converges to the $1 - \alpha$ th quantile of $N(0, 1)$. The points above can also be seen by noncentral chi-square distribution properties when the concentration parameter converges to infinity as in p.51-52 of Evans, Hastings, Peacock (1993). A similar argument is used (when only the degrees of freedom converge to infinity and concentration parameter is constant) in proof of Theorem 4 in Newey and Windmeijer (2007).

Next we have

$$P(AR(\theta_0) \geq q_\alpha^1) = P\left(\frac{AR(\theta_0) - (k + 2\lambda_n)}{\sqrt{2(k + 2\lambda_n)}} \geq \frac{q_\alpha^1 - (k + \lambda_n)}{\sqrt{2(k + 2\lambda_n)}} \right) \to \alpha.$$

In a sense the Anderson-Rubin limit converging to a limit with a parameter that depends on sample size (C_n) is also observed in many weak moments context of Theorem 4 in Newey and Windmeijer (2007) without the violation of exogeneity. In their case the proof proceeds in the same way, but they do not have a noncentrality parameter, however in their case $k \to \infty$ (the degrees of freedom converge to infinity albeit slower than root n). So it is still the case, the mean and the variance of the distribution converges to infinity when the sample size goes to infinity.**Q.E.D.**

Proof of Theorem 1.

(i). The first part of the proof extends Wu (1990) (equations (18)-(22)) to independent case. Triangular array case is discussed after the proof of Theorem 1 here. From page 1440 in Wu (1990), specifically benefiting from equations (2.1)(2.2) of Wu (1990):

$$P_* \left(\frac{\bar{S}_b - E_* \bar{S}_b}{(var_* \bar{S}_b)^{1/2}} \leq t \right) = P_* \left(\frac{\sqrt{b}(\bar{S}_b - \bar{S}_n)}{[(1-f)v\hat{a}rS]^{1/2}} \leq t \right), \tag{18}$$

where $f = b/n$, and by the argument immediately after equation (2.2) of Wu (1990) (or Theorem 2.2 of Cochran (1977))

$$var_* \bar{S}_b = \frac{1-f}{b} v\hat{a}rS. \tag{19}$$

See that in our case

$$v\hat{a}rS = \hat{\Omega}. \tag{20}$$

Define as in Wu (1990)

$$J(t) = P_* \left[\frac{\sqrt{b}(\bar{S}_b - \bar{S}_n)}{[(1-f)v\hat{a}rS]^{1/2}} \leq t \right]. \tag{21}$$

Then under Ω being finite and nonsingular, via Assumption 2, Corollary 1 of Zhao, Wang, Wu (2004) shows that

$$\sup_t |J(t) - \phi(t)| \to 0, \quad a.s. \tag{22}$$

where $\phi(t)$ is the standard normal distribution.

Then note that

$$\sqrt{b} E_* \bar{S}_b = \frac{\sqrt{b}}{\sqrt{n}} \left(n^{-1/2} \sum_{i=1}^n Z_i u_i - E Z_i u_i \right) + \frac{\sqrt{b}}{\sqrt{n}} (\sqrt{n} E Z_i u_i). \tag{23}$$

See that by the definition $\sqrt{b/n} = \sqrt{f}$,

$$\frac{\sqrt{b}}{\sqrt{n}} \left(n^{-1/2} \sum_{i=1}^n Z_i u_i - E Z_i u_i \right) \xrightarrow{d} N(0, f\Omega) \equiv \sqrt{f} L, \tag{24}$$

where L is $N(0, \Omega)$. Then by Assumption 1,

$$\frac{\sqrt{b}}{\sqrt{n}} (\sqrt{n} E Z_i u_i) = \sqrt{b} E Z_i u_i = \frac{\sqrt{b}}{\sqrt{n}} C_n = \sqrt{f} C_n. \tag{25}$$

The test is defined as by (18)(19)

$$FAR(\theta_0) = (\sqrt{b}\bar{S}_b)'((1-f)v\hat{a}rS)^{-1}(\sqrt{b}\bar{S}_b).$$

Rewrite $FAR(\theta_0)$ as

$$FAR(\theta_0) = [\sqrt{b}(\bar{S}_b - E_* \bar{S}_b + E_* \bar{S}_b)]'((1-f)v\hat{a}rS)^{-1}[\sqrt{b}(\bar{S}_b - E_* \bar{S}_b + E_* \bar{S}_b)]. \tag{26}$$

22

The using the above expression, by (22)-(25), and the discussion at the end of the proof of Lemma 1, with $J_H(t) = P_*(FAR(\theta_0) \leq t)$,

$$\sup_t |J_H(t) - \phi_{\chi_f^2}(t)| \to 0, \quad a.s.,$$

where $\phi_{\chi_f^2}(t)$ is the noncentral chi-square distribution with noncentrality parameter

$$\frac{f}{1-f} \frac{(C_n + L)'(\Omega)^{-1}(C_n + L)}{2}.$$

(ii).

Note that with choice of $f_o = 1/2$, we have the noncentrality parameter as $(C_n + L)'(\Omega)^{-1}(C_n + L)/2$.

Q.E.D.

Remark. The only difference between the proof of iid case in Wu (1990) and the one here is the Berry-Esseen bounds. The iid case in Wu (1990) is satisfied under finite second moments as well. The extension to triangular arrays can be done using Theorem 1 of Zhao, Wu and Wang (2004). The Berry-Esseen bounds are for a sample sum from a finite set of independent random variables in Zhao, Wu, Wang (2004).

Proof of Corollary 1. The proof is very similar to the proof of Theorem 1.

First we consider the drift term in the proof of Theorem 1. From the test statistic in (26) , we see that the noncentrality parameter comes from the expression below

$$\sqrt{\hat{b}_n^*}[E_*\bar{S}_{\hat{b}_n^*}]'[(1 - \hat{f}_n^*)\hat{\Omega}]^{-1}\sqrt{\hat{b}_n^*}[E_*\bar{S}_{\hat{b}_n^*}]. \tag{27}$$

Then rewrite

$$\sqrt{\hat{b}_n^*}[E_*\bar{S}_{\hat{b}_n^*}] = \frac{\sqrt{\hat{b}_n^*}}{\sqrt{n}}(n^{-1/2}\sum_{i=1}^n[Z_iu_i - EZ_iu_i]) + \frac{\sqrt{\hat{b}_n^*}}{\sqrt{n}}[\sqrt{n}EZ_iu_i]$$

$$= \sqrt{\hat{f}_n^*}[C_n + \hat{L}], \tag{28}$$

via $\hat{f}_n^*$ and $\hat{L} = n^{-1/2}\sum_{i=1}^n[Z_iu_i - EZ_iu_i]$, and Assumption 1. Note the $\hat{f}_n^*$ definition

$$\hat{f}_n^* = \frac{C_n'\hat{\Omega}^{-1}C_n}{2C_n'\hat{\Omega}^{-1}C_n + C_n'\hat{\Omega}^{-1}\hat{L} + \hat{L}'\hat{\Omega}^{-1}C_n + \hat{L}'\hat{\Omega}^{-1}\hat{L}},$$

then we can use that to have

$$\frac{\hat{f}_n^*}{1 - \hat{f}_n^*} = \frac{C_n'\hat{\Omega}^{-1}C_n}{(C_n + \hat{L})'\hat{\Omega}^{-1}(C_n + \hat{L})}.$$

Now we substitute the above expression into (27) taking into account the right-hand side term in (28)

$$\frac{\hat{f}_n^*}{1 - \hat{f}_n^*}(C_n + \hat{L})'\hat{\Omega}^{-1}(C_n + \hat{L}) = C_n'\hat{\Omega}^{-1}C_n.$$

23

We concentrate on the terms of the numerator and show that all the terms approximate the term in Theorem 1ii. Namely

$$\sqrt{\hat{b}_n^*}[\bar{S}_{\hat{b}_n^*} - \bar{S}_{n/2} + \bar{S}_{n/2} - \bar{S}_n] = \sqrt{\hat{b}_n^*}[\bar{S}_{\hat{b}_n^*} - \bar{S}_{n/2}] + (\sqrt{\hat{b}_n^*} - \sqrt{n/2})[\bar{S}_{n/2} - \bar{S}_n] + \sqrt{n/2}[\bar{S}_{n/2} - \bar{S}_n]. \quad (29)$$

Note that the third term on the right-hand side of the equation divided by $1/2v\hat{a}rS$ above converges to standard normal distribution by $(21)(22)$. In the second term we have by $\hat{b}_n^*$ definition $\sqrt{\hat{b}_n^*} - n/2 \xrightarrow{p} 0$, and by the third term's analysis the second term converges in probability to zero as well.

Now consider the first term

$$\sqrt{\hat{b}_n^*}[\bar{S}_{\hat{b}_n^*} - \bar{S}_{n/2}] \leq \frac{\sqrt{\hat{b}_n^*}max_i\|Z_i u_i\||\hat{b}_n^* - n/2|}{min(\hat{b}_n^*, n/2)} = o_p(1),$$

with $\hat{b}_n^* \xrightarrow{p} n/2, max_i\|Z_i u_i\| = o_p(n^{1/2})$ (in which $max_i E\|Z_i u_i\|^3 < \infty$ is a primitive condition). So basically the left hand side term in (29) is approximated by the third term in (29).

Then note that in the denominator of the resampled test statistic

$$(1 - \hat{f}_n^*) \xrightarrow{p} 1/2.$$

So we do not consider the denominator.

Q.E.D.

References

Anderson, T.W., H. Rubin (1949). "Estimators of the parameters of a single equation in a complete set of stochastic equations," *The Annals of Mathematical Statistics,* 21, 570-582.

Andrews, D.W.K. and P. Guggenberger (2007a). "The Limit of Finite-Sample Size and a Problem with Subsampling," unpublished manuscript, Cowles Foundation, Yale University.

Andrews, D.W.K. and P. Guggenberger (2007b). "Hybrid and Size-Corrected Subsample Methods," unpublished manuscript, Cowles Foundation, Yale University.

Berkowitz, D., M. Caner, Y. Fang (2008). "Are nearly exogenous instruments reliable?" *Economics Letters,* Article in press.

Cochran, W.G. (1977). Sampling Techniques, 3rd Ed. John Wiley.

Conley, T., C. Hansen, P.E. Rossi (2007). "Plausibly Exogenous", unpublished manuscript. Graduate School of Business, University of Chicago.

Evans, M., N. Hastings, B. Peacock (1993). Statistical Distributions, 2nd ed. Wiley Interscience.

Hall, A.R., A. Inoue (2003). "The large sample behaviour of the generalized method of moments estimator in misspecified models," *Journal of Econometrics,* 114, 361-394.

Kleibergen, F. (2002). "Pivotal Statistics for Testing Structural Parameters in Instrumental Variable Regression," *Econometrica,* 70, 1781-1803.

Kray, A. (2008). "Instrumental Variables Regressions with Honestly Uncertain Exclusion Restrictions," unpublished manuscript. World Bank.

Newey, W.K. (1985). "Generalized Method of Moments Specification Testing," *Journal of Econometrics,* 29, 229-256.

Newey, W. K. and F. Windmeijer (2007). "GMM with Many Weak Moment Conditions," Working Paper, Department of Economics, MIT.

Stock, J. and J. Wright (2000). "GMM With Weak Identification," *Econometrica,* 68, 1055-1096.

Wu, C.F.J (1990)."On The Asymptotic Properties of the Jackknife Histogram," *Annals of Statistics,* 18, 1438-1452.

Zhao, L.C., C.Q. Wu, and Q. Wang (2004). "Berry-Esseen Bound for a Sample Sum from a Finite Set of Independent Random Variables," *Journal of Theoretical Probability,* 17, 557-571.