P- VALUE ADJUSTMENTS FOR ASYMPTOTIC CONTROL OF THE GENERALIZED FAMILYWISE ERROR RATE

by

Christopher J. Bennett

Working Paper No. 09-W05

April 2009

DEPARTMENT OF ECONOMICS
VANDERBILT UNIVERSITY
NASHVILLE, TN 37235

www.vanderbilt.edu/econ

p-Value Adjustments for Asymptotic Control of the Generalized Familywise Error Rate

Christopher J. Bennett
Assistant Professor
Department of Economics
Vanderbilt University, Nashville, TN 37240
E-mail: chris.bennett@vanderbilt.edu

April 8, 2009

Abstract

We introduce a computationally efficient bootstrap procedure for obtaining multiplicity-adjusted p-values in situations where multiple hypotheses are tested simultaneously. This new testing procedure accounts for the mutual dependence of the individual statistics, and is shown under weak conditions to maintain asymptotic control of the generalized familywise error rate. Moreover, the estimated critical values (p-values) obtained via our procedure are less sensitive to the inclusion of true hypotheses and, as a result, our test has greater power to identify false hypotheses even as the collection of hypotheses under test increases in size. Another attractive feature of our test is that it leads naturally to balance among the individual hypotheses under test. This feature is especially attractive in settings where balance is desired but alternative approaches, such as those based on studentization, are difficult or infeasible.

KEYWORDS: bootstrap, familywise error, multiple testing, step-down, balanced testing

JEL CLASSIFICATIONS: C12, C14, C52

1. INTRODUCTION

In this paper we introduce a computationally efficient bootstrap procedure for obtaining multiplicity-adjusted p-values in situations where multiple hypotheses are tested simultaneously. This new testing procedure accounts for the mutual dependence of the individual statistics, delivers a balanced testing procedure where all of the individual tests have approximately equal power, and is shown under weak conditions to maintain asymptotic control of the generalized familywise error rate.

In the classical approach to the multiplicity problem, testing procedures are designed to control the probability of one or more false rejections, otherwise referred to in the literature as the *familywise error rate* and commonly abbreviated as $FWER_P$, where the subscript P reflects the dependence of the FWER on the underlying true probability distribution. Formally,

$$\text{FWER}_P = P\{\text{reject at least one hypothesis } H_i : i \in I_0(P)\}$$

where $I_0(P)$ denotes the set of true null hypotheses under P. If $I_0(P)$ is empty then FWER_P is defined to be zero. We say that a multiple testing procedure satisfying

$$\limsup_T \text{FWER}_P \leq \alpha, \tag{1}$$

where T denotes sample size, maintains asymptotic control of the familywise error rate at level α. This is to be distinguished from control of the FWER when (i) all null hypotheses are true (weak control) or (ii) any configuration of the null hypotheses are true (strong control). Evidently, error control in the sense of (1) is an intermediate case. As argued in Pollard and van der Laan (2003), strong control is "too much and not necessary" since error control under the true data generating distribution is "all that one cares about." This viewpoint, which is shared by the author, also appears to be the prevailing viewpoint in the econometrics literature—see e.g. Romano and Wolf (2005*b*) and Romano, Shaikh and Wolf (2008*b*), and references therein. For the remainder of the paper we therefore focus on

asymptotic error control in the sense of (1). We note however that there has also been recent interest in relaxing FWER control to k-FWER control, where k-FWER is defined as the probability of at most k false rejections.[1] Our proposed testing procedure generalizes in a natural way to k-FWER control and this point is discussed briefly in Section 5, though no additional insight is gained by considering this more general case in the main body of the paper.

The single-step Bonferroni adjustment and the stepdown procedure of Holm (1979) are examples of simple multiplicity adjustments that offer control of the FWER; see e.g. Hochberg and Tamhane (1987). However, these and many other procedures are generally conservative in part because they fail to account for the dependence structure of the individual test statistics (or p-values). White (2000) demonstrated, among other things, how bootstrap methods can be applied to (asymptotically) account for the true dependence structure of the individual test statistics thereby improving upon the power—i.e. the average probabilities of rejecting each false hypothesis—of such tests. White's (2000) influential paper also served as a catalyst for a number of subsequent contributions which have collectively resulted in significant improvements in the power of FWER procedures to detect false hypotheses. Examples of contributions in this area include Hansen (2005), Romano and Wolf (2005a), Romano and Wolf (2005b), and Hsu and Kuan (2008).

In addition to controlling the FWER, it is often desirable to design balanced testing procedures where all of the individual tests have approximately equal power. While studentization of the individual test statistics can sometimes achieve balance and lead to more powerful tests, in many applications the computation of the the standard errors of the individual statistics may be difficult or infeasible. Moreover, studentization will be ineffective in achieving balance if the (limiting) distributions of the studentized statistics are different. This would occur, for instance, whenever the collection of hypotheses being tested includes both one-sided and two-sided tests. Alternatively, even in these situations the marginal dis-

[1] A discussion of various Type I error rates may be found in Dudoit, van der Laan and Pollard (2004).

tributions of the p-values of the individual statistics are, under mild conditions, generally uniform on the boundaries of the null hypotheses and are therefore the natural quantities to use whenever balance is desired.

In this paper we introduce a computationally efficient bootstrap procedure for use in multiple testing problems which is based on the comparison of estimated p-values. Aside from delivering a balanced testing procedure, our multiple comparison test yields appropriately adjusted p-values and is therefore attractive since (i) users are not required to specify a target error rate α in advance; and (ii) reporting p-values is more informative since each p-value provides a measure of strength of evidence against an individual hypothesis.

Our procedure is based on the following simple observation: If $J(\mathbf{x})$ is the joint distribution of the test statistics under the complete null hypothesis, then the exact distribution of the minimum p-value is given by

$$H(p) = \mathbb{E}1\{\min_i [1 - J_{(i)}(X_{(i)})] \leq p\} \tag{2}$$

where $X_{(i)}$ denotes the ith element of a random vector with distribution $J(\mathbf{x})$ and $J_{(i)}(\cdot)$ denotes the marginal (univariate) distribution associated with $X_{(i)}$. If $J(\mathbf{x})$ were known, then (2) may be computed analytically or to an arbitrary degree of accuracy via Monte Carlo simulation. When $J(\mathbf{x})$ is unknown—as is generally the case—we may replace $J(\mathbf{x})$ with a resampling-based estimator to obtain an approximation to (2); and such a strategy will be valid in the sense of (1) under mild conditions whenever resampling methods deliver a consistent estimator of $J(\mathbf{x})$. In this paper we focus on the use bootstrap resampling to estimate (2), though our proposed methodology may easily be generalized to accommodate situations where, say, subsampling or the m out of n bootstrap is required for consistency.

Notably only two bootstrap samples are required of our procedure, and of these the second bootstrap sample is constructed by random sampling from the first bootstrap sample and hence without re-computing test statistics. It is also remarkable that the same bootstrap

samples generated for a single-step test can be used to obtain asymptotically valid critical values (adjusted p-values) for the related stepdown procedures. This is of considerable practical importance since it implies that the computational burden of our testing procedure does not increase dramatically when we move from single-step to stepdown procedures. In contrast, even in the case of single-step tests existing procedures generally require an iterated or double bootstrap to approximate the joint distribution of the p-values, whereby, denoting by $B1$ and $B2$ the number of replications in the first and second bootstrap procedures, $1+B1+B\times B2$ statistics must be calculated; see for example Godfrey (2005) or the discussion in MacKinnon (2007). As opposed to a p-value approach Romano and Wolf (2008) have recently introduced a procedure based on inverting balanced simultaneous confidence regions as originally proposed in Beran (1988). However, their procedure delivers critical values for a pre-specified target error rate α as opposed to p-values. Additionally, their proposed methodology, unlike ours, becomes computationally burdensome when FWER control is relaxed in favor of k-FWER control.

In the discussion of our testing procedure we consider the case of testing multiple *one-sided* hypotheses. One-sided tests are composite and thus introduce subtleties which are not present when only two-sided hypotheses are considered. In particular, in the case of one-sided tests the lower quantiles of the null distribution of the MinP statistic are predominantly determined by those test statistics associated with true hypotheses for which the parameters are at or near the boundary of equality. In their recently proposed tests of parameters defined by many moment inequalities, for example, Andrews and Soares (2007) and Andrews and Jia (2008) exploit sample information to help identify binding or "near" binding parameters and use only these in computing critical values.[2] Since the quantiles of the null distribution of the MinP statistic are non-decreasing in the number of hypotheses under test, excluding those hypotheses associated with parameters that appear to be "deep" within the null will

[2]In fact the basic intuition underlying these tests can be traced back to Andrews (2000) and has been exploited by several authors in various contexts including Hansen (2005), and Hsu and Kuan (2008), among others.

generally lead to more powerful tests while maintaining asymptotic validity. We show how to incorporate these recent developments in our resampling procedure. The result is a more powerful test which is also less sensitive to the inclusion of "irrelevant" hypotheses when compared to a MinP test based on the canonical fully recentered bootstrap. The generalization to two-sided hypotheses and mixtures of one- and two-sided hypotheses is shown to be straightforward.

The plan for the rest of the paper is as follows. In the next section we discuss the basic formulation of the multiple testing problem. We then introduce a single-step multiple testing procedure based on a comparison of nonstudentized test statistics, and use these results to develop our single-step p-value comparison test. We subsequently consider extensions to stepdown procedures as well as to k-FWER control, and then briefly illustrate the performance of our tests via Monte Carlo simulation. Proofs are collected in the appendix.

2. MULTIPLE TESTING: FORMULATION AND EXAMPLES

The class of testing problems under consideration is the same as that considered in Romano and Wolf (2005*b*). In order to conveniently draw from their examples and facilitate comparison to their results we have adopted their notation where applicable.

Let $X_1, \ldots, X_T$ be a data set generated from some probability distribution P. Our interest centers on simultaneous testing of

$$H_s : \theta_s(P) \leq 0 \text{ vs. } H_s' : \theta_s(P) > 0, \quad s \in \{1, \ldots, S\}. \tag{3}$$

We denote by $w_{T,(s)} = w_{T,(s)}(X_1, \ldots, X_T)$ a statistic for testing H_s and assume, without loss of generality, that a large value of $w_{T,(s)}$ constitutes evidence against the individual null hypothesis H_s. The following is but one example from the general class of testing problems under consideration.

Example 1 (Model Selection)**.** *For model s let θ_s denote a performance measure relative*

*to a given benchmark model, where $\theta_s > 0$ is indicative of model s being "superior" to the benchmark. Let $w_{t,s}$ denote a statistic for testing the hypothesis $H_s : \theta_s \leq 0$ (for specific performance measures and their corresponding test statistics see White (2000) and Romano and Wolf (2005*b*)). Interest in this context centers on simultaneously testing*

$$H_s : \theta_s \leq 0 \text{ vs. } H_s' : \theta_s > 0, \quad s \in \{1, \ldots, S\}$$

Failure to account for multiplicity in this context will generally result in "too many" models being found superior to the benchmark.

As mentioned at the outset our focus in this paper is on resampling-based multiple testing procedures, or more specifically bootstrap-based tests. Accordingly, let $J_T(P)$ denote the sampling distribution under the true probability mechanism P of the scaled and centered $S \times 1$ statistic $\sqrt{T}(W_T - \theta)$ and denote by $J_T(\hat{P}_T)$ the sampling distribution under $\hat{P}_T$ of the bootstrap counterpart $\sqrt{T}(W_T^* - \theta^*)$. The appropriate bootstrap procedure for constructing $\sqrt{T}(W_T^* - \theta^*)$ will naturally depend on whether the underlying data are i.i.d. or temporally dependent. In the first case, Efron's (1979) bootstrap procedure may be appropriate whereas in the latter case a bootstrap procedure for dependent data would be required; see Lahiri (2003) for further details. In any event, the testing procedures discussed in this paper depend critically on the validity of the bootstrap approximation to the asymptotic distribution. Assumption 1 below is a high level assumption concerning the validity of the bootstrap approximation and is sufficient for establishing all of our main results.

Assumption 1. *Let P denote the true probability mechanism and let $\hat{P}_T$ denote an estimate of P based on the data X_T. Assume that $J_T(P)$ converges in distribution to a limit distribution $J(P)$, which is continuous and strictly increasing. Further, assume that $J_T(\hat{P}_T)$ consistently estimates this limit distribution: $\rho(J_T(\hat{P}_T), J(P)) \to 0$ in probability for any metric ρ metrizing weak convergence.*

Remark 1. *Assumption 1, which is a slightly modified version of assumption 3.1 of Romano*

*and Wolf (2005*b*), is rather common in the bootstrap literature. For discussion and examples pertaining to cases for which Assumption 1 is satisfied see Politis, Romano and Wolf (1999) p. 9 and Shao and Tu (1995) p. 80. See also Romano and Wolf (2005*b*) pp. 1249-1251 for specific examples.*

3. TESTING PROCEDURES

We begin our discussion with a single-step bootstrap test involving basic (nonstudentized) statistics, which we then use as a basis for developing our p-value comparison test. The generalization to stepwise procedures along the lines of Hsu and Kuan (2008) is natural and therefore discussed only briefly in Section 4.

3.1 A Partially Recentered Bootstrap Approach

The decision rules of single-step procedures for testing H_s at the nominal level α can invariably be written as

$$\text{Reject } H_s \text{ if } 0 \notin [\sqrt{T}w_{T,s} - c(\alpha), \infty), \tag{4}$$

where $c(\alpha)$ denotes an appropriately chosen critical value. In this section we describe a bootstrap procedure for estimating $c(\alpha)$ that ensures asymptotic control of the FWER at level α. For the purpose of describing this testing procedure, let b_T denote a positive sequence satisfying $b_T \to 0$ and $\sqrt{T}b_T \to \infty$ as T approaches infinity . Additionally, let $J_T^{\max}(\hat{P}_T)$ denote the sampling distribution under $\hat{P}_T$ of

$$M_T = \max_{s\in\{1,\ldots,S\}} \left\{ \sqrt{T}(w_{T,s}^* - \theta_{T,s}^*) + \sqrt{T}w_{T,s}\mathbb{1}\{w_{T,s} < -b_T\} \right\}. \tag{5}$$

Our proposal for estimating $c(\alpha)$ in (4) is to select the value $\hat{c}^{PC}(\alpha)$ defined as

$$\hat{c}^{PC}(\alpha) = \inf\{x : J_T^{\max}(\hat{P}_T)(x) \geq 1 - \alpha\}. \tag{6}$$

Notice that the recentering in (5) differs from the canonical fully-recentered bootstrap;

namely the bootstrap distribution corresponding to the indices $s \in \{1, \ldots, S\}$ with $w_{T,s} \leq -b_T$ are shifted downwards by the amount $T^{1/2} w_{T,s}$. Asymptotically, this adjustment reduces the set of indices over which the maximum of the random vector M ($M_T \to^d M$) is computed to only those indices for which $\theta_s \geq 0$, and hence results in smaller critical values and greater power than tests based on the canonical bootstrap recentering scheme. Upon inspection the procedure can be viewed as a simple hybrid of the procedures proposed in White (2000) and Hansen (2005). Specifically, the testing procedure above may be obtained by dropping the outer maximum on the statistic employed by Hansen (2005)—which leads to White's (2000) test statistic—while retaining Hansen's (2005) proposed resampling scheme. An appropriate sequence $\{b_T\}$ may be obtained by appealing to the Law of the Iterated Logarithm. This choice is discussed in some detail in Hansen (2003) and Hansen (2005).

The asymptotic properties of the proposed test are summarized below in Theorem 1. In the statement of the theorem we denote by $I_0(P)$ the set of indices corresponding to the true hypotheses.

Theorem 1. *Suppose Assumption 1 holds and $b_T = o(1)$ with $\sqrt{T} b_T \to \infty$. Then the following statements are true.*

i. If $\theta_s > 0$, then

$$Prob_P\{Reject\ H_s\} \to 1 \ as\ T \to \infty$$

ii. The procedure provides asymptotic control the familywise error rate, i.e. at the nominal level α

$$\lim_{T \to \infty} FWER_P \leq \alpha$$

iii. The limiting probability in (ii) is equal to α if and only if $I_0(P) = \{1, \ldots, S\}$ and $\theta_s = 0$ for as least one $s \in I_0(P)$.

Theorem 1 is the analogue of Theorem 3.1 of Romano and Wolf (2005*b*) . Parts (i) and (ii) of the theorem show that the test is consistent and provides asymptotic control of the FWER.

In part (iii) it is shown that the partially recentered bootstrap test is correctly sized for any configuration of the parameters where all of the null hypotheses are satisfied and at least one parameter is on the boundary. Intimately connected to property (iii) is the improved ability of the partially recentered bootstrap test to detect false hypotheses. Formally, letting $\hat{c}^{FC}(\alpha)$ denote the α-level critical value obtained via the canonical bootstrap we have:

Theorem 2. *Suppose Assumption 1 holds and $b_T = o(1)$ with $\sqrt{T}b_T \to \infty$. Then,*

$$\lim_{T\to\infty} Prob_P\{T^{1/2}w_{T,s} > \hat{c}^{PC}(\alpha)\} \geq \lim_{T\to\infty} Prob_P\{T^{1/2}w_{T,s} > \hat{c}^{FC}(\alpha)\}$$

with strict inequality holding whenever $I_0(P)$ is nonempty and $\theta_i < 0$ for some $i \in I_0(P)$.

3.2 An approach based on the comparison of p-values

In this section we present a single-step testing procedure based on the comparison of p-values. There are several attractive features of this new test including: (i) in contrast to other resampling-based p-value tests, our procedure provides asymptotic control of the FWER without imposing any assumptions beyond Assumption 1; (ii) the procedure is computationally efficient requiring only two separate bootstrap samples as opposed to a full double bootstrap; and (iii) the test is invariant to monotonic transformations of the component statistics making studentization of the component statistics unnecessary.

In order to describe the testing procedure we introduce the following notation. First, denote by $J_T^{PC}(\hat{P}_T)$ the joint sampling distribution under $\hat{P}_T$ of

$$w_{T,s}^{PC} = \sqrt{T}(w_{T,s}^* - \theta_{T,s}^*) + \sqrt{T}w_{T,s}\mathbb{1}\{w_{T,s} < -b_T\}. \tag{7}$$

Additionally, let $J_{T,(i)}(\hat{P})$ denote the ith marginal distribution of $J_T(\hat{P})$. It is relatively straightforward that the p-values of the component statistics $w_{T,s}$ may be consistently esti-

mated from the marginal bootstrap distributions as

$$p_{T,s} = 1 - J_{T,(s)}(\hat{P})(T^{1/2} w_{T,s}).$$

Since a small p-value corresponds to evidence against a null hypothesis, the appropriate decision rule in this case is

$$\text{Reject } H_s \text{ if } p_{T,s} < p_T(\alpha),$$

where $p_T(\alpha)$ is a data-dependent critical that is to be estimated in a manner that ensures, at least asymptotically, control of the FWER. Our proposed strategy in this paper is to use as an estimator

$$p_T(\alpha) = \inf\{x : J_T^{\min}(\hat{P}_T)(x) \geq 1 - \alpha\}, \tag{8}$$

where $J_T^{\min}(\hat{P}_T)$ is the sampling distribution under $\hat{P}$ of

$$\min_{s \in \{1,\ldots,S\}} \left[1 - J_{T,(s)}(\hat{P})(W_{(s)})\right], \tag{9}$$

and $W_{(s)}$ is the sth element of the random vector $W \sim J_T^{PC}(\hat{P}_T)$. Note that the corresponding adjusted p-values are easily obtained from $J_T^{\min}(\hat{P}_T)(p_{T,s})$ for $s \in \{1, \ldots, S\}$.

It is noteworthy that the bootstrap procedure used to estimate the sampling distribution of the minimum p-value involves only resampling with replacement from the tabulated distribution $J_T^{PC}(\hat{P}_T)$. In other words, the second stage bootstrap procedure is nothing other than Efron's (1979) i.i.d. bootstrap applied to $J_T^{PC}(\hat{P}_T)$.

To gain some intuition for the mechanics of this procedure first consider the case where all of the hypotheses are on the boundary, i.e. $\theta_s = 0$ for every s. In this case $J_T^{PC}(\hat{P}_T)$ and $J_T(\hat{P})$ both converge to $J(P)$, and consequently $\left[1 - J_{T,(s)}(\hat{P})(W_s)\right]$ converges to a uniform random variable for every $s \in \{1, \ldots, S\}$. Thus, asymptotically, the minimum is over an $S \times 1$ vector random variable with uniform (univariate) marginals, as should be expected when all of the $\theta_s = 0$. It is also worth noting here that the dependence structure among the

uniform random variables is captured implicitly by the nature of the bootstrap resampling. When $\theta_s < 0$ the sth marginal distribution $J^{PC}_{T,(s)}(\hat{P}_T)$ converges in probability to a degenerate distribution at $-\infty$. It follows that

$$\left[1 - J_{T,(s)}(\hat{P})(W_s)\right] \to 1$$

in probability as $T \to \infty$, and the index set over which the minimum is computed is effectively reduced. Since $p_T(\alpha)$ is increasing in the number of indices for which $\theta_s < 0$, the ability to detect false hypotheses is therefore not as adversely affected by the inclusion of true hypotheses that are strictly in the null as would otherwise be the case if the canonical fully recentered bootstrap scheme is employed.

Not surprisingly, given the intimate connection between the two tests, the MinP testing procedure has the same basic asymptotic properties as the underlying partially recentered bootstrap test. This is the essence of Theorem 3 below.

Theorem 3. *Suppose Assumption 1 holds and $b_T = o(1)$ with $\sqrt{T}b_T \to \infty$. Then statements (i), (ii), and (iii) of Theorem 1 apply to the p-value testing procedure.*

By letting $p^{FC}_T(\alpha)$ denote the p-value obtained by replacing $W_{(s)}$ in (9) with $\tilde{W}_{(s)}$, where $\tilde{W}_{(s)}$ is the sth element of $\tilde{W} \sim J_T(\hat{P})$, we also obtain the p-value analogue of theorem 2:

Theorem 4. *Suppose Assumption 1 holds and $b_T = o(1)$ with $\sqrt{T}b_T \to \infty$. Then,*

$$\lim_{T\to\infty} Prob_P\{p_{T,s} < p^{PC}_T(\alpha)\} \geq \lim_{T\to\infty} Prob_P\{p_{T,s} < p^{FC}_T(\alpha)\}$$

with strict inequality holding whenever $I_0(P)$ is nonempty and $\theta_i < 0$ for some $i \in I_0(P)$.

As mentioned previously, one of the principle advantages of the p-value approach is that balance of power among the individual tests is obtained without having to studentize the individual statistics—that is, if studentization is feasible—and balance continues to hold

even when one- and two-sided tests are included in the family of hypotheses. This basic property of the p-value test is the content of Theorem 5. For the statement of the theorem we denote by $I_0(P)$ and $I_1(P)$ the set of indices $i \in \{1, \ldots, S\}$ for which $\theta_i(P) \leq 0$ and $\theta_i(P) \geq 0$.

Theorem 5. *Suppose Assumption 1 holds and $b_T = o(1)$ with $\sqrt{T}b_T \to \infty$. Then, for all $i, j \in I_0(P) \bigcap I_1(P)$, the p-value test at nominal level α satisfies*

$$\lim_{T\to\infty} Prob_P\left(Reject\ H_i\right) = \lim_{T\to\infty} Prob_P\left(Reject\ H_j\right)$$

whenever $I_1(P) \backslash I_0(P) = \emptyset$.

Remark 2. *Under the same conditions as stated in the theorem, it may also be shown that in a stepdown test any two hypotheses on the boundary of their respective nulls have identical rejection probabilities equal to $1 - \alpha$..*

4. EXTENSIONS TO STEPDOWN TESTING PROCEDURES

Denote by $w_T^{(1)}, \ldots, w_T^{(S)}$ the order statistics defined by sorting $w_{1,T}, \ldots, w_{S,T}$ in increasing order. Additionally, denote by $H^{(1)}, \ldots, H^{(S)}$ the corresponding null hypotheses associated with each of the order statistics. A stepdown testing procedure is defined as a sequential testing procedure that rejects $H^{(S)}$ if

$$0 \notin [\sqrt{T}w_T^{(1)} - c_1(\alpha), \infty)$$

and rejects $H^{(j)}$ for $1 \leq j < S$ only if both $H^{(j+1)}$ is rejected and

$$0 \notin [\sqrt{T}w_T^{(j)} - c_j(\alpha), \infty),$$

where $c_j(\alpha)$ is a suitably chosen critical value with the property that $c_i \leq c_j$ for $i < j$. Notice that single-step procedures are subsumed by this definition by simply taking $c_i = c_j$ for all

$i \neq j$.

Given $w_T^{(1)}, \ldots, w_T^{(S)}$, define the sequence $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \mathcal{H}_S$ with $\mathcal{H}_S = \{H^{(1)}, \ldots, H^{(S)}\}$ and $\mathcal{H}_j = \mathcal{H}_S \backslash \{H^{(S)}, \ldots, H^{(S-j+1)}\}$ for $1 \leq j < S$. Additionally, let $R_{\mathcal{H}}$ denote a resampling-based procedure for estimating the α-level critical value in a test of the family of hypotheses in $\mathcal{H}$. We now state a general result on the asymptotic control of FWER via stepdown procedures.

Theorem 6. *Suppose the α-level critical value $c_j(\alpha)$ obtained from $R_{\mathcal{H}_j}$ is such that a single-step test based on $c_j(\alpha)$ satisfies*

i. $\limsup_T FWER_P^{\mathcal{H}_j} \leq \alpha$, *and*

ii. $\limsup_T FWER_P^{\mathcal{H}_j} = \alpha$ *whenever all of the hypotheses in $\mathcal{H}_j$ are binding under P.*

Then, the sequence$\{c_j(\alpha), 1 \leq j \leq S\}$ is weakly monotonically increasing, and a stepdown procedure based on $\{c_j(\alpha), 1 \leq j \leq S\}$ satisfies

$$\limsup_T FWER_P \leq \alpha$$

Remark 3. *Theorem 6, which is essentially a reformulation of results obtained by Romano and Wolf (2005*a*), formalizes the notion that a stepdown test is simply a sequence of single-stage tests, and that FWER control of a single-stage bootstrap test implies that control is maintained when the same bootstrap procedure is applied to estimate the critical values in a stepdown procedure.*

Theorem 6, which forms the basis for the generalization of a single-step procedure to a stepdown approach, also applies to the minimum p-value testing procedure. To see this, let $r_T^{(1)}, \ldots, r_T^{(S)}$ denote the order statistics of $1 - p_{T,1}, \ldots, 1 - p_{T,S}$ and define

$$c_{T,j}(\alpha) = 1 - \inf\{x : J_{T,j}^{\min}(\hat{P}_T)(x) \geq 1 - \alpha\} \tag{10}$$

where $J_{T,j}^{\min}(\hat{P}_T)$ is the sampling distribution under $\hat{P}$ of

$$\min_{\{j:1-p_{T,j}\geq r_T^{(j)}\}} \left[1 - J_{T,(j)}(\hat{P})(W_s)\right]$$

and W_s denotes the sth element of the random vector $W \sim J_T^{PC}(\hat{P}_T)$. A stepdown procedure based on the minimum p-value approach rejects $H^{(S)}$ if

$$0 \notin [r_T^{(S)} - c_1(\alpha), \infty),$$

and rejects $H^{(j)}$ for $1 \leq j < S$ only if both $H^{(j+1)}$ is rejected and

$$0 \notin [r_T^{(j)} - c_j(\alpha), \infty).$$

From the results in Theorem 3 it is straightforward to check that conditions (i) and (ii) of Theorem 6 are satisfied. The asymptotic control of the familywise error rate is then obtained as a corollary. We simply state this result and other basic asymptotic properties of the step-down procedure more formally as Theorem 7 below.

Theorem 7. *Suppose Assumption 1 holds and $b_T = o(1)$ with $\sqrt{T}b_T \to \infty$. Then the step-down procedure based on the minimum p-value approach as defined above satisfies*

i. If $\theta_s > 0$, then

$$Prob_P\{Reject\ H_s\} \to 1\ as\ T \to \infty$$

ii. The procedure provides asymptotic control of the familywise error rate, i.e. at the nominal level α

$$\limsup_{T\to\infty} FWER_P \leq \alpha$$

iii. The limiting probability in (ii) is equal to α if and only if $I_0(P)$ is non-empty and $\theta_s = 0$ for as least one $s \in I_0(P)$.

Remark 4. *The ability of the stepdown procedures to reject more false hypotheses than their single-step counterparts follows immediately from the monotonicity of the sequence* $\{c_j(\alpha), 1 \leq j \leq S\}$ *together with the fact that a single-step procedure simply takes* $c_S(\alpha)$ *as the critical value for testing all of the individual hypotheses. Moreover, the weak dominance of the partial recentering schemes over their full centering counterparts, i.e. Theorems 2 and 4, have natural extensions to the stepdown testing framework.*

5. ASYMPTOTIC CONTROL OF THE GENERALIZED FAMILYWISE ERROR RATE

In this section we briefly describe the *single* modification of the MinP testing procedure that is necessary if one wishes to maintain asymptotic control of the k-FWER defined as

$$k\text{-FWER}_P = Prob_P\{\text{at most } k \text{ true } H_i \text{ are rejected}\}. \tag{11}$$

Recalling that $J_T^{PC}(\hat{P}_T)$ denotes the sampling distribution under $\hat{P}_T$ of

$$W_T^{PC} = \sqrt{T}(W_T^* - \theta_T^*) + \sqrt{T} w_{T,s} \mathbb{1}\{w_{T,s} < -b_T\},$$

and that $J_{T,(i)}(\hat{P})$ denotes the ith marginal distribution of $J_T(\hat{P})$, we define $J_T^k(\hat{P}_T)$ to be the sampling distribution under $\hat{P}$ of the kth order statistic in the collection

$$\{P_{(1)}, P_{(2)}, \ldots, P_{(S)}\},$$

where $P_{(s)} = 1 - J_{T,(s)}(\hat{P})(W_{(s)})$ and $W_{(s)}$ is the sth element of the random vector $W \sim J_T^{PC}(\hat{P}_T)$.

In a single-step procedure, the estimated critical value

$$p_T^k(\alpha) = \inf\{x : J_T^k(\hat{P}_T)(x) \geq 1 - \alpha\},$$

together with the decision rule

$$\text{Reject } H_s \text{ if } p_{T,s} < p_T^k(\alpha),$$

can be shown—with little modification to the proofs which are provided in the Appendix for the case $k = 1$— to maintain asymptotic control of the k-FWER. Additionally, the adjusted p-values corresponding to a test of $H_{(s)}$ may be reported by evaluating $J_T^k(\hat{P}_T)(x)$ at $p_{T,s}$.

As for the associated stepdown procedure, asymptotic k-FWER$_P$ control is maintained if the individual single-step tests that are employed in the Section 4 are replaced by the analogous single-step k-FWER tests proposed above. Again, this result is a straightforward extension of the case when $k = 1$ and is thus stated without proof.

6. SIMULATION EXPERIMENT

In this section we conduct a simple simulation experiment to illustrate the size and power properties of the multiplicity-adjusted p-value test. In our simulations we employ a slight variation of the experimental design of Hansen (2003). For various choices of $\theta \in \mathbb{R}^S$, pseudo random numbers satisfying $\bar{X}_T \sim N_S(\theta, T^{-1}\Sigma)$ are generated and used to test the hypotheses

$$H_s : \theta_s \leq 0 \text{ against } H_s' : \theta_s > 0 \text{ for } s \in \{1, \ldots, S\}. \tag{12}$$

Prior to each draw of $\bar{X}_T$, a random covariance matrix Σ is generated using the "clusterGeneration" package in R which is based on the algorithm of Joe (2006). This allows us to examine the performance of the tests across a wide array of covariance structures (this algorithm is also employed in a simulation study of Romano, Shaikh and Wolf (2008*a*)). Additionally, in the partial recentering of the bootstrap we set $b_T = \sqrt{2 \log \log T}$, a choice which is motivated by the Law of the Iterated Logarithm.

To complete our description of our experimental design, let ρ_b and ρ_a denote the fraction of binding inequalities and the fraction of inequalities in the alternative, respectively. For

designs in the null, $\rho_a = 0$ and $(1 - \rho_b)$ of the inequalities θ_s are set equal to $-1/10$. Alternatively, for designs in the alternative, ρ_a of the inequalities θ_i are set equal to $+1/100$ and the remaining fraction $(1 - \rho_a)$ are set equal to $-1/10$. Similar to the parameter choices in Hansen (2003) we consider all combinations $S = 10, 100$; $\rho_a = \rho_b = 0.1, 0.2, 0.5, 0.8, 0.9$ or 1.0; and $T = 100, 200, 500$, or 1000. In every case the nominal size is set equal to 5%, and the number of first- and second-stage bootstrap replications are set equal to 2,999 and 3,999.

Tables 1 and 2 contain empirical FWER's based on 2,000 Monte Carlo simulations for various single-step tests based on the fully recentered and partially recentered bootstrap schemes, which we label as "White" and "Hansen", respectively. The studentized and p-value versions are labeled with the modifier's "St." and "p".

In all situations in which the fully recentered bootstrap is employed, the corresponding tests are generally correctly sized only in the case where $\rho_b = 1$. Otherwise, the empirical FWER's tend to drop off quite rapidly as ρ_b falls to 0.1. On the other hand, those tests based on Hansen's partial recentering scheme generally maintain an empirical FWER close to the nominal size of 5% over the entire range of ρ_b, albeit the p-value test appears to be slightly undersized when $S = 100$ and ρ_b takes on a value close to one. Overall, however, the results are as expected from the theory.

[Table 1 about here.]

[Table 2 about here.]

Tables 3 and 4 contain the average proportion of correctly rejected hypotheses, again based on 2,000 Monte Carlo simulations. Aside from the cases when $S = 10$ and $\rho_a = 1$, and when $S = 100$ and $\rho_a = 1$ or $\rho = 0.9$, where the p-value test based on Hansen's recentering is marginally outperformed, the p-value test generally performs as well as any other competing procedure. The most striking feature that emerges from these simulations is the performance of the studentized test based on White's full recentering bootstrap. In either the case of

$S = 10$ or $S = 100$, the power of the test to reject false hypotheses drops off dramatically as p_a falls to 0.1. Interestingly, this loss of power does not occur to the same extent for the nonstudentized or p-value tests based on White's procedure. This observation suggests that caution should be exercised whenever S is large and one is considering implementing White's procedure combined with studentization.

[Table 3 about here.]

[Table 4 about here.]

7. SUMMARY AND CONCLUDING REMARKS

In this paper we have proposed a computationally efficient procedure for obtaining multiplicity-adjusted p-values. Our approach is shown to maintain asymptotic control of the FWER under weak conditions, and to weakly dominate multiplicity adjustments based on the canonical bootstrap.

Throughout the paper we have concentrated exclusively on multiplicity adjustments via bootstrap procedures. Yet subsampling is known to yield consistent estimates of the sampling distribution in certain situations where the bootstrap fails. In such situations it may be of interest to consider the subsampling analogues of the bootstrap procedures proposed herein. Very few modifications are generally necessary to transform a bootstrap tests to its' subsampling counterpart and, under suitable regularity conditions, it is typically straightforward to show that the basic asymptotic properties of the tests will continue to hold. However interesting, for the sake of expositional clarity we have chosen not to pursue the details here.

In a related literature there is growing interest on relaxing control of the FWER and instead focusing on the control of generalized error rates such as the false discovery proportion (Romano et al. 2008*b*). Although not explored here, we are interested in how the p-value approach can be adapted and applied to control these generalized error rates. Further, the innovations in this paper have immediate extensions to joint tests of equality and or

inequality restrictions. For example, the p-value comparison approach advocated herein can be used to develop goodness-of-fit tests that distribute power uniformly over the support of the distributions being compared. In light of the findings of this paper we feel that extensions along these lines clearly merit further investigation.

APPENDIX A. PROOFS

Let $I_0(P)$ and $I_1(P)$ denote the set of indices $s \in \{1, \ldots, S\}$ for which $\theta_s(P) \leq 0$ and $\theta_s(P) \geq 0$, respectively . The following lemma establishes weak convergence of the maximum of the partially centered bootstrap statistic.

Lemma 1. *Suppose Assumption 1 holds and $b_T = o(1)$ with $\sqrt{T}b_T \to \infty$. Then,*

$$J_T^{\max}(\hat{P}_T) \Rightarrow \max_{s \in I_1(P)} J(P) \tag{A.1}$$

in probability.

Proof of Lemma 1. Recall that $J_T^{\max}(\hat{P}_T)$ denotes the sampling distribution under $\hat{P}_T$ of

$$M_T = \max_{s \in \{1,\ldots,S\}} \left\{ \sqrt{T}(w_{T,s}^* - \theta_{T,s}^*) + \sqrt{T} w_{T,s} \mathbb{1}\{w_{T,s} < -b_T\} \right\}. \tag{A.2}$$

By assumption 1,

$$\sqrt{T}(w_{T,s}^* - \theta_{T,s}^*) \Rightarrow J(P) \tag{A.3}$$

in probability. Since $T^{1/2} w_{T,s} = O_p(1)$ only if $\theta_s(P) = 0$ and $b_T = o(1)$ with $\sqrt{T}b_T \to \infty$ it is also the case that

$$\sqrt{T} w_{T,s} \mathbb{1}\{w_{T,s} < -b_T\} \to^p \begin{cases} 0 & \theta_s(P) \geq 0 \\ -\infty & \theta_s(P) < 0 \end{cases} \tag{A.4}$$

Combining the results in (A.3) and (A.4), the proof is completed upon applying both Slutzky's theorem and the continuous mapping theorem. □

Proof of Theorem 1.

(i) From lemma 1 we have $M_T = O_p(1)$ whereas $T^{1/2} w_{T,s}$ diverges for any $\theta_s(P) > 0$.

(ii) Suppose $I_0(P)$ and $I_1(P)$ are non-empty. We then have

$$\begin{aligned}
\lim_{T\to\infty} FWER_P &= \lim_{T\to\infty} Prob_P\left[\max_{s\in I_0(P)}\{T^{1/2}w_{T,s}\} > \hat{c}^{PC}(\alpha)\right] \\
&= \lim_{T\to\infty} Prob_P\left[\max_{s\in I_0(P)\bigcap I_1(P)}\{T^{1/2}w_{T,s}\} > \hat{c}^{PC}(\alpha)\right] \\
&\leq \lim_{T\to\infty} Prob_P\left[\max_{s\in I_1(P)}\{T^{1/2}(w_{T,s}-\theta_s(P))\} > \hat{c}^{PC}(\alpha)\right] \\
&= \alpha
\end{aligned} \tag{A.5}$$

where the last line follows from standard arguments (Beran 1984, p. 17).

(iii) If all of the inequalities are binding then $I_0(P) = I_1(P)$ and it follows from Lemma 1 that the asymptotic familywise error rate is equal to the nominal size of the test. If, on the other hand, $I_1(P)\backslash I_0(P) \neq \emptyset$ the inequality in the third line of (A.5) is strict. Similarly, if $I_1(P) = \emptyset$ then $\lim_{T\to\infty} FWER_P = 0$. □

Proof of Theorem 2. The proof follows Lemma 1 together with the fact that $c^{FC}(\alpha) \geq c^{PC}(\alpha)$. □

Lemma 2. *Define the functions* $H_T, H : \mathbb{R}^S \to \mathbb{R}^S$ *as*

$$H_T(x, \hat{P}) = \left(J_{T,(1)}(x_{(1)}, \hat{P}), \ldots, J_{T,(S)}(x_{(S)}, \hat{P})\right) \tag{A.6}$$

and

$$H(x, P) = \left(J_{(1)}(x_{(1)}, P), \ldots, J_{(S)}(x_{(S)}, P)\right). \tag{A.7}$$

Then, if assumption 1 holds,

$$H_T(T^{1/2}W_T) \Rightarrow H(W + \lim_{T\to\infty} T^{1/2}\theta),$$

where $W \sim J(\cdot, P)$.

Proof of Lemma 2. Write

$$H_T(T^{1/2}W_T) = H(T^{1/2}(W_T - \theta) + T^{1/2}\theta) + H_T(W_T) - H(W_T) \tag{A.8}$$

Since

$$\sup_{x\in\mathbb{R}^S} \|H_T(x) - H(x)\| \leq C \sum_{i=1}^{S} \sup_{x\in\mathbb{R}} |J_{T,(i)}(x, \hat{P}_T) - J_{(i)}(x, P)| \tag{A.9}$$

for some constant C, and $\rho_\infty \left(J_{T,(s)} \left(\cdot, \hat{P}_T \right), J_{(s)} \left(\cdot, P \right) \right) \to^p 0$ (follows from pointwise convergence together with the continuity of $J_{(s)}(\cdot, P)$ and Pólya's theorem (Serfling 1981, p. 18)) it follows that

$$\sup_{x\in\mathbb{R}^S} \|H_T(x) - H(x)\| \to^p 0. \tag{A.10}$$

The desired result then follows from (A.8) via Slutzky's theorem together with the continuous mapping theorem. □

Proof of Theorem 3.

(i) From lemma 2, the consistency of the bootstrap distribution, and the continuity of the min function, we have

$$\min_{s\in\{1,\ldots,S\}} \left[1 - J_{T,(s)}(T^{1/2}w_{T,(s)})\right] \Rightarrow \min_{s\in\{1,\ldots,S\}} \left[1 - J_{(s)} \left(W_{(s)} + \lim_{T\to\infty} T^{1/2}\theta_{(s)} \right)\right] \tag{A.11}$$

and

$$\min_{s\in\{1,\ldots,S\}} \left[1 - J_{T,(s)}(\tilde{W}_{(s)})\right] \Rightarrow \min_{s\in\{1,\ldots,S\}} \left[1 - J_{(s)} \left(W_{(s)} + \lim_{T\to\infty} T^{1/2}\mathbb{1}\{\theta_{(s)} < 0\}\theta_{(s)} \right)\right] \tag{A.12}$$

in probability, where $\tilde{W} \sim J_T^{PC}(\hat{P}_T)$ and $W \sim J(P)$. From (A.11) it is clear that

$$\min_{s\in\{1,\ldots,S\}} \left[1 - J_{(s)} \left(W_{(s)} + \lim_{T\to\infty} T^{1/2}\theta_{(s)} \right)\right] \to^p 0 \tag{A.13}$$

if, for any $s \in \{1, \ldots, S\}$, $\theta_{(s)} > 0$. On the other hand, $\text{plim}_{T\to\infty} p_T(\alpha)$ defined in (10) is never less than the α quantile of

$$\min_{s\in\{1,\ldots,S\}} \left[1 - J_{(s)}(W_{(s)})\right] =^d \min_{s\in\{1,\ldots,S\}} U_{(s)} \tag{A.14}$$

where U is a $S \times 1$ random vector with uniform marginals. The results in (A.13) and (A.14) together are sufficient for consistency.

(ii) Suppose $I_0(P)$ and $I_1(P)$ are non-empty. We then have

$$\begin{aligned}
\lim_{T\to\infty} FWER_P &= \lim_{T\to\infty} Prob_P \left[\min_{s\in I_0(P)} \{p_{T,s}\} < p_T(\alpha)\right] \\
&= \lim_{T\to\infty} Prob_P \left[\min_{s\in I_0(P)\bigcap I_1(P)} \{p_{T,s}\} < p_T(\alpha)\right] \\
&\leq \lim_{T\to\infty} Prob_P \left[\min_{s\in I_1(P)} \left\{1 - J_{T,(s)}\left(T^{1/2}(w_{T,s} - \theta_s(P))\right)\right\} < p_T(\alpha)\right] \\
&= \alpha
\end{aligned} \tag{A.15}$$

where, again, the last line follows from standard arguments (Beran 1984, p. 17).

(iii) If all of the inequalities are binding then $I_0(P) = I_1(P)$ and it follows from Lemma 2 and the continuity of the min function that the asymptotic familywise error rate is equal to the nominal size of the test. If, on the other hand, $I_1(P)\backslash I_0(P) \neq \emptyset$ the inequality in the third line of (A.15) is strict. Similarly, if $I_1(P) = \emptyset$ then $\lim_{T\to\infty} FWER_P = 0$. □

Proof of Theorem 4 . The proof follows from the proof of Theorem 3 together with the fact that $p_T^{PC}(\alpha) \geq p_T^{FC}(\alpha)$. □

Proof of Theorem 5. From (A.12) and the conditions of the theorem, we have for all

$i, j \in I_0(P) \cap I_1(P)$

$$\begin{aligned}\lim_{T\to\infty} Prob_P(\text{Reject } H_i) &= \mathbb{P}\left(U_{(i)} < \min_{s\in I_0(P)\cap I_1(P)} U_{(s)}\right) \\ &= \mathbb{P}\left(U_{(j)} < \min_{s\in I_0(P)\cap I_1(P)} U_{(s)}\right) \\ &= \lim_{T\to\infty} Prob_P(\text{Reject } H_j)\end{aligned} \tag{A.16}$$

□

Proof of Theorem 6. That the sequence$\{c_j(\alpha), 1 \leq j \leq S\}$ is weakly monotonically increasing is immediate from condition (ii) of the theorem. Asymptotic control of the FWER can be shown by replicating the arguments in the proof of Theorem 3.1 on page 1273 of Romano and Wolf (2005*b*). □

REFERENCES

Andrews, D. W. (2000), "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68, 399.

Andrews, D. W., and Jia, P. (2008), Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure,, Cowles Foundation Discussion Papers 1676, Cowles Foundation, Yale University.

Andrews, D. W., and Soares, G. (2007), Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,, Cowles Foundation Discussion Papers 1631, Cowles Foundation, Yale University.

Beran, R. (1984), "Bootstrap Methods in Statistics," *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 86(1), 14–30.

Beran, R. (1988), "Balanced Simultaneous Confidence Sets," *Journal of the American Statistical Association*, 83(403), 679.

Dudoit, S., van der Laan, M., and Pollard, K. (2004), Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates,, U.C. Berkeley Division of Biostatistics Working Paper Series 1137, Berkeley Electronic Press.

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

Godfrey, L. G. (2005), "Controlling the overall significance level of a battery of least squares diagnostic tests," *Oxford Bulletin of Economics and Statistics*, 67, 267.

Hansen, P. (2003), Asymptotic Tests of Composite Hypotheses,, Working Papers 2003-09, Brown University, Department of Economics.

Hansen, P. R. (2005), "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380.

Hochberg, Y., and Tamhane, A. C. (1987), *Multiple Comparison Procedures* John Wiley & Sons.

Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6(7), 65–70.

Hsu, Po-Hsuan, H. Y.-C., and Kuan, C.-M. (2008), Testing the Predictive Ability of Technical Analysis Using a New Stepwise Test Without Data Snooping Bias,. Unpublished Working Paper Available at SSRN: http://ssrn.com/abstract=1087044.

Joe, H. (2006), "Generating random correlation matrices based on partial correlations," *J. Multivar. Anal.*, 97(10), 2177–2189.

Lahiri, S. N. (2003), *Resampling Methods for Dependent Data* Springer.

MacKinnon, J. G. (2007), Bootstrap Hypothesis Testing,, Working Papers 1127, Queen's University, Department of Economics.

Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling (Springer Series in Statistics)*, 1 edn Springer.

Pollard, K. S., and van der Laan, M. J. (2003), Resampling-based Multiple Testing: Asymptotic Control of Type I Error and Applications to Gene Expression Data,, U.C. Berkeley Division of Biostatistics Working Paper Series 121, Berkeley Electronic Press.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008*a*), Control of the False Discovery Rate under Dependence using the Bootstrap and Subsampling,, IEW - Working Papers iewwp337, Institute for Empirical Research in Economics - IEW.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008*b*), "Formalized Data Snooping Based On Generalized Error Rates," *Econometric Theory*, 24(2), 404–447.

Romano, J. P., and Wolf, M. (2005*a*), "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100(469), 94–108.

Romano, J. P., and Wolf, M. (2005*b*), "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73(4), 1237–1282.

Romano, J. P., and Wolf, M. (2008), Balanced Control of Generalized Error Rates,, IEW - Working Papers iewwp379, Institute for Empirical Research in Economics - IEW.

Serfling, R. J. (1981), *Approximation Theorems of Mathematical Statistics (Wiley Series in Probability and Statistics)* Wiley-Interscience.

Shao, J., and Tu, D. (1995), *The Jackknife and Bootstrap (Springer Series in Statistics)* Springer.

White, H. (2000), "A Reality Check for Data Snooping," *Econometrica*, 68(5), 1097.

List of Tables

Table 1: Familywise Error Rates ($S = 10$)

ρ_b	T	White	Hansen	St.White	St.Hansen	pWhite	pHansen
	100	0.048	0.050	0.034	0.037	0.049	0.054
1.0	200	0.055	0.056	0.054	0.056	0.059	0.062
	500	0.043	0.044	0.049	0.051	0.052	0.054
	1000	0.048	0.049	0.049	0.049	0.046	0.049
	100	0.043	0.048	0.031	0.037	0.044	0.058
0.9	200	0.050	0.058	0.048	0.054	0.052	0.062
	500	0.040	0.045	0.036	0.050	0.046	0.051
	1000	0.044	0.046	0.025	0.049	0.042	0.050
	100	0.038	0.047	0.030	0.038	0.039	0.058
0.8	200	0.046	0.057	0.042	0.054	0.047	0.063
	500	0.035	0.046	0.027	0.048	0.040	0.050
	1000	0.040	0.046	0.014	0.051	0.039	0.052
	100	0.027	0.047	0.020	0.040	0.027	0.059
0.5	200	0.030	0.052	0.022	0.049	0.027	0.057
	500	0.0025	0.051	0.012	0.056	0.030	0.055
	1000	0.025	0.044	0.001	0.052	0.024	0.054
	100	0.009	0.030	0.007	0.031	0.009	0.049
0.2	200	0.010	0.048	0.006	0.048	0.009	0.051
	500	0.011	0.058	0.004	0.055	0.014	0.056
	1000	0.014	0.050	0.001	0.049	0.012	0.049
	100	0.005	0.026	0.004	0.029	0.005	0.045
0.1	200	0.005	0.051	0.005	0.048	0.006	0.051
	500	0.005	0.049	0.002	0.047	0.007	0.049
	1000	0.007	0.045	0.000	0.044	0.005	0.043

Of the $S = 10$ hypotheses under test, $\rho_b S$ are binding, and the remaining $(1 - \rho_b)S$ are strictly in the null. Table reports the familywise error rate at the nominal %5 level as estimated from 2,000 Monte Carlo simulations

Table 2: Familywise Error Rates ($S = 100$)

ρ_b	T	White	Hansen	St.White	St.Hansen	pWhite	pHansen
	100	0.053	0.054	0.059	0.060	0.065	0.067
1.0	200	0.047	0.048	0.043	0.044	0.041	0.044
	500	0.053	0.055	0.055	0.056	0.038	0.041
	1000	0.051	0.051	0.053	0.055	0.035	0.039
	100	0.049	0.053	0.054	0.058	0.058	0.066
0.9	200	0.043	0.051	0.037	0.048	0.039	0.044
	500	0.040	0.051	0.014	0.054	0.032	0.047
	1000	0.047	0.050	0.004	0.057	0.032	0.037
	100	0.045	0.055	0.049	0.059	0.054	0.069
0.8	200	0.034	0.046	0.031	0.043	0.032	0.043
	500	0.044	0.050	0.028	0.055	0.036	0.046
	1000	0.043	0.050	0.001	0.054	0.028	0.045
	100	0.029	0.050	0.037	0.055	0.035	0.063
0.5	200	0.021	0.046	0.014	0.050	0.020	0.050
	500	0.023	0.049	0.001	0.038	0.017	0.038
	1000	0.026	0.047	0.000	0.052	0.018	0.046
	100	0.013	0.044	0.019	0.048	0.015	0.065
0.2	200	0.007	0.056	0.003	0.047	0.006	0.053
	500	0.009	0.050	0.000	0.051	0.007	0.052
	1000	0.010	0.048	0.000	0.051	0.007	0.047
	100	0.004	0.027	0.008	0.035	0.006	0.050
0.1	200	0.005	0.043	0.004	0.048	0.003	0.052
	500	0.005	0.054	0.000	0.056	0.003	0.059
	1000	0.005	0.049	0.000	0.044	0.003	0.046

Of the $S = 100$ hypotheses under test, $\rho_b S$ are binding, and the remaining $(1 - \rho_b)S$ are strictly in the null. Table reports the familywise error rate at the nominal %5 level as estimated from 2,000 Monte Carlo simulations

Table 3: Average Proportion of False Hypotheses Rejected ($S = 10$)

ρ_a	T	White	Hansen	St.White	St.Hansen	pWhite	pHansen
	100	0.013	0.013	0.017	0.017	0.022	0.022
1.0	200	0.032	0.032	0.062	0.063	0.070	0.070
	500	0.256	0.254	0.420	0.420	0.426	0.426
	1000	0.931	0.931	0.915	0.915	0.916	0.917
	100	0.013	0.014	0.016	0.018	0.021	0.023
0.9	200	0.032	0.034	0.062	0.067	0.071	0.075
	500	0.254	0.271	0.396	0.432	0.425	0.437
	1000	0.930	0.936	0.879	0.919	0.916	0.921
	100	0.013	0.015	0.016	0.019	0.021	0.025
0.8	200	0.032	0.038	0.060	0.073	0.071	0.080
	500	0.254	0.293	0.375	0.444	0.426	0.450
	1000	0.930	0.941	0.841	0.923	0.915	0.926
	100	0.013	0.0212	0.017	0.029	0.022	0.036
0.5	200	0.032	0.056	0.053	0.092	0.066	0.101
	500	0.252	0.383	0.317	0.494	0.419	0.501
	1000	0.932	0.959	0.754	0.942	0.914	0.945
	100	0.013	0.044	0.017	0.057	0.022	0.070
0.2	200	0.035	0.137	0.053	0.170	0.069	0.182
	500	0.254	0.591	0.283	0.618	0.428	0.623
	1000	0.929	0.975	0.689	0.970	0.912	0.971
	100	0.014	0.070	0.019	0.096	0.025	0.108
0.1	200	0.041	0.264	0.054	0.253	0.073	0.266
	500	0.253	0.709	0.259	0.705	0.418	0.702
	1000	0.926	0.983	0.666	0.983	0.913	0.982

Of the $S = 10$ hypotheses under test $\rho_a S$ are strictly in the null, and the remaining $(1 - \rho_a)S$ are in the alternative. The table reports the average proportion of false hypotheses that are rejected at the nominal %5 level as estimated from 2,000 Monte Carlo simulations

Table 4: Average Proportion of False Hypotheses Rejected ($S = 100$)

ρ_a	T	White	Hansen	St.White	St.Hansen	pWhite	pHansen
	100	0.002	0.002	0.002	0.002	0.003	0.003
1.0	200	0.004	0.004	0.013	0.013	0.012	0.012
	500	0.054	0.054	0.221	0.221	0.191	0.192
	1000	0.720	0.720	0.783	0.783	0.735	0.737
	100	0.002	0.002	0.003	0.003	0.003	0.004
0.9	200	0.004	0.005	0.012	0.012	0.012	0.013
	500	0.054	0.058	0.188	0.228	0.191	0.202
	1000	0.719	0.734	0.627	0.790	0.735	0.747
	100	0.001	0.002	0.002	0.001	0.003	0.004
0.8	200	0.004	0.0049	0.012	0.015	0.012	0.016
	500	0.054	0.063	0.162	0.236	0.191	0.214
	1000	0.719	0.746	0.534	0.797	0.735	0.765
	100	0.002	0.003	0.003	0.004	0.003	0.006
0.5	200	0.004	0.008	0.010	0.020	0.011	0.021
	500	0.054	0.087	0.116	0.268	0.190	0.261
	1000	0.720	0.806	0.400	0.826	0.736	0.815
	100	0.002	0.005	0.003	0.009	0.003	0.012
0.2	200	0.004	0.017	0.009	0.038	0.011	0.042
	500	0.053	0.157	0.094	0.346	0.190	0.349
	1000	0.717	0.887	0.342	0.878	0.735	0.877
	100	0.002	0.007	0.003	0.014	0.003	0.018
0.1	200	0.003	0.031	0.008	0.060	0.011	0.067
	500	0.051	0.249	0.092	0.416	0.192	0.423
	1000	0.711	0.928	0.326	0.912	0.735	0.913

Of the $S = 100$ hypotheses under test $\rho_a S$ are strictly in the null, and the remaining $(1 - \rho_a)S$ are in the alternative. The table reports the average proportion of false hypotheses that are rejected at the nominal %5 level as estimated from 2,000 Monte Carlo simulations