A Score Based Approach to Wild Bootstrap Inference

Patrick Kline
Department of Economics
UC Berkeley / NBER
pkline@econ.berkeley.edu

Andres Santos[*]
Department of Economics
UC San Diego
a2santos@ucsd.edu

June, 2011

Abstract

We propose a generalization of the wild bootstrap of Wu (1986) and Liu (1988) based upon perturbing the scores of M-estimators. This "score bootstrap" procedure avoids recomputing the estimator in each bootstrap iteration, making it substantially less costly to compute than the conventional nonparametric bootstrap, particularly in complex nonlinear models. Despite this computational advantage, in the linear model, the score bootstrap studentized test statistic is equivalent to that of the conventional wild bootstrap up to order $O_p(n^{-1})$. We establish the consistency of the procedure for Wald and Lagrange Multiplier type tests and tests of moment restrictions for a wide class of M-estimators under clustering and potential misspecification. In an extensive series of Monte Carlo experiments we find that the performance of the score bootstrap is comparable to competing approaches despite its computational savings.

KEYWORDS: Wild Bootstrap, Robust Inference, Clustered Data.

[*] We thank Justin McCrary, Nicolas Lepage-Saucier, Graham Elliott, Michael Jansson as well as the editor Jason Abrevaya and two anonymous referees for comments that helped greatly improve the paper.

1 Introduction

The bootstrap of Efron (1979) has become a standard tool for conducting inference with economic data. Among the numerous variants of the original bootstrap, the so-called "wild" bootstrap of Wu (1986) and Liu (1988) has been found to yield dramatic improvements in the ability to control the size of Wald tests of OLS regression coefficients in small samples (Mammen (1993), Horowitz (1997, 2001), Cameron, Gelbach, and Miller (2008)).

Originally proposed as an alternative to the residual bootstrap of Freedman (1981), the wild bootstrap has often been interpreted as a procedure that resamples residuals in a manner that captures any heteroscedasticity in the underlying errors. Perhaps for this reason, the applications and extensions of the wild bootstrap have largely been limited to linear models where residuals are straightforward to obtain; see for example Hardle and Mammen (1993) for nonparametric regression, You and Chen (2006) for partially linear regression, Davidson and MacKinnon (2010) for IV regression and Cavaliere and Taylor (2008) for unit root inference.

We propose a new variant of the wild bootstrap (the "score" bootstrap) which perturbs the fitted score contributions of an M-estimator with i.i.d. weights conditional on a fixed Hessian. In the linear model, our score bootstrap procedure is numerically equivalent to the conventional wild bootstrap for unstudentized statistics and higher order equivalent for studentized ones. However, in contrast to the wild bootstrap, our approach is easily adapted to estimators without conventional residuals and avoids recomputing the estimator in each bootstrap iteration. As a result, the score bootstrap possesses an important advantage over existing bootstraps in settings where the model is computationally expensive to estimate or poorly behaved in a subset of the bootstrap draws. Such difficulties often arise even in simple probit or logit models where, for some nonparametric bootstrap draws, the estimator cannot be computed.[1]

The score bootstrap is closely related to several existing bootstrap procedures in the literature. Most notably, it bears a close relationship to the estimating equation bootstrap of Hu and Zidek (1995) who propose resampling score contributions in the linear model conditional on a fixed Hessian. Hu and Kalbfleisch (2000) generalize this approach to nonlinear models by resampling both score and Hessian contributions evaluated at the estimated parameter vector. In the case of score or Lagrange Multiplier tests, our approach can be interpreted as a

[1] Kaido and Santos (2011) employ the score bootstrap to estimate the asymptotic distribution of set estimators, an area where computational considerations are particularly important.

wild bootstrap analogue to their pairs resampling procedure. Also related is the generalized bootstrap of Chatterjee and Bose (2005) which perturbs the objective function of an M-estimator with i.i.d. weights. This approach is closer to the weighted bootstrap (e.g. Barbe and Bertail (1995), Ma and Kosorok (2005)) than the wild bootstrap and, in contrast to the score bootstrap, requires reoptimization of the estimator under each perturbation of the criterion function. Finally, our procedure has an interpretation as a variant of the k-step bootstrap procedure of Davidson and MacKinnon (1999a) which involves taking a finite number of Newton steps towards optimization of an M-estimator in a bootstrap sample. Andrews (2002) showed that this procedure yields an Edgeworth refinement depending on the number of optimization steps taken. Like the conventional nonparametric bootstrap however, the k-step procedure may be difficult to compute if, in some bootstrap samples, the Hessian is poorly behaved or of less than full rank, problems which the score bootstrap avoids.

We provide results establishing the consistency of the score bootstrap for a broad class of test statistics under weak regularity conditions and in the presence of potential misspecification. Our framework is shown to encompass Wald and Lagrange Multiplier (LM) tests as well as tests of moment restrictions. To assess the empirical relevance of these theoretical results, we conduct an extensive series of Monte Carlo experiments comparing the performance of several different bootstrap procedures in settings with clustered data. Our focus on clustered data is motivated by the prevalence of such settings in applied work and a large literature (e.g. Bertrand, Duflo, and Mullainathan (2004), Wooldridge (2003), Donald and Lang (2007), Cameron et al. (2008)) finding that asymptotic cluster robust methods often perform poorly in small samples. We find that variants of our proposed score based bootstrap substantially outperform analytical cluster robust methods. The performance of these procedures is also comparable to that of competing bootstrap methods, despite their large difference in computational cost.

The remainder of the paper is structured as follows: Section 2 reviews the wild bootstrap, while Section 3 introduces the score bootstrap and establishes its higher order equivalence. In Section 4 we develop the consistency of the score bootstrap under weak regularity conditions and illustrate its applicability to a variety of settings. Our simulation study is contained in Section 5, while Section 6 briefly concludes. All proofs are contained in the Appendix.

2 Wild Bootstrap Review

We begin by reviewing the wild bootstrap and the reasons for its consistency in the context of a linear model. A careful examination of the arguments justifying its validity provides us with the intuition necessary for developing the score bootstrap and its extension to M-estimation problems.

While there are multiple approaches to implementing the wild bootstrap, for expository purposes we focus on the original methodology developed in Liu (1988). Suppose $\{Y_i, X_i\}_{i=1}^n$ is an i.i.d. sequence of random variables, with $Y_i \in \mathbf{R}$, $X_i \in \mathbf{R}^m$ and satisfying the linear relationship:

$$Y_i = X_i'\beta_0 + \varepsilon_i \ . \tag{1}$$

Letting $\hat{\beta}$ denote the OLS estimate of β_0 and $e_i \equiv (Y_i - X_i'\hat{\beta})$ the implied residual, the wild bootstrap generates new residuals of the form $\varepsilon_i^* \equiv W_i e_i$ for some randomly generated i.i.d. sequence $\{W_i\}_{i=1}^n$ that is independent of $\{Y_i, X_i\}_{i=1}^n$ and satisfies $E[W_i] = 0$ and $E[W_i^2] = 1$. Common choices of distributions for W_i include the Standard Normal, Rademacher,[2] and the two-point distribution advocated in Mammen (1993).[3] Under these assumptions on $\{W_i\}_{i=1}^n$ it follows that:

$$E[\varepsilon_i^* | \{Y_i, X_i\}_{i=1}^n] = 0 \qquad E[(\varepsilon_i^*)^2 | \{Y_i, X_i\}_{i=1}^n] = e_i^2 \ . \tag{2}$$

Hence ε_i^* is mean independent of $\{Y_i, X_i\}_{i=1}^n$ and, in addition, captures the pattern of heteroscedasticity found in the original sample. This property, originally noted in Wu (1986), enables the wild bootstrap to remain consistent even in the presence of heteroscedasticity or model misspecification.[4]

The wild bootstrap resampling scheme consists of generating dependent variables $\{Y_i^*\}_{i=1}^n$ by

$$Y_i^* \equiv X_i'\hat{\beta} + \varepsilon_i^* \tag{3}$$

and then conducting OLS on the sample $\{Y_i^*, X_i\}_{i=1}^n$ in order to obtain a bootstrap estimate $\hat{\beta}^*$. The distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ *conditional* on $\{Y_i, X_i\}_{i=1}^n$ (but not on $\{W_i\}_{i=1}^n$) is then used as an estimate of the unknown distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$. Since the former distribution can be computed through simulation, the wild bootstrap provides a simple way to obtain critical values for inference.

[2] A Rademacher random variable puts probability one half on the values one and negative one.

[3] Here W_i equals $\frac{1-\sqrt{5}}{2}$ with probability $\frac{\sqrt{5}+1}{2\sqrt{5}}$ and $\frac{\sqrt{5}+1}{2}$ with probability $1 - \frac{\sqrt{5}+1}{2\sqrt{5}}$.

[4] We refer to misspecification in model (1) as $E[\varepsilon_i | X_i] \neq 0$ but $E[\varepsilon_i X_i] = 0$.

We review why the wild bootstrap is consistent by drawing from arguments in Mammen (1993). First, observe that standard OLS algebra and the relationships in (1) and (3) imply that:

$$\sqrt{n}(\hat{\beta}-\beta_0) = H_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\varepsilon_i \qquad \sqrt{n}(\hat{\beta}^*-\hat{\beta}) = H_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\varepsilon_i^* \ , \tag{4}$$

where $H_n \equiv n^{-1}\sum_i X_i X_i'$. When $\hat{\beta}$ is viewed as the maximum likelihood estimator of a normal model, H_n is the Hessian of the likelihood, while $\sum_i X_i\varepsilon_i$ is the gradient (or score) evaluated at the true parameter value β_0. Since both the full sample score contributions $(\{X_i\varepsilon_i\}_{i=1}^n)$ and their bootstrap analogues $(\{X_i\varepsilon_i^*\}_{i=1}^n)$ are properly centered, the expressions in (4) can be expected to converge to a normal limit. Therefore, consistency of the wild bootstrap hinges on whether these limits are the same or, equivalently, whether the asymptotic variances agree. However, since $E[W_i^2]=1$ and $\{W_i\}_{i=1}^n$ is independent of $\{Y_i, X_i\}_{i=1}^n$, we may write:

$$E[(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\varepsilon_i)(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\varepsilon_i)'] = E[X_i X_i'\varepsilon_i^2] \tag{5}$$

$$E[(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\varepsilon_i^*)(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\varepsilon_i^*)'|\{Y_i, X_i\}_{i=1}^n] = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 \ , \tag{6}$$

which implies, by standard arguments, that the second moments indeed agree asymptotically. As a result, $\sqrt{n}(\hat{\beta}-\beta_0)$ and $\sqrt{n}(\hat{\beta}^*-\hat{\beta})$ converge in distribution to the same normal limit and the consistency of the wild bootstrap is immediate.

While the ability of the wild bootstrap to asymptotically match the first two moments of the full sample score provides the basis for establishing its validity, it does not elucidate why it often performs better than a normal approximation. Improvements occur when the bootstrap is able to additionally match higher moments of the score. If, for example, $E[W_i^3]=1$, then the third moments match asymptotically and the wild bootstrap provides a refinement over the normal approximation to a studentized statistic by providing a skewness correction (Liu (1988)). The additional requirement that $E[W_i^3]=1$ is satisfied, for example, by the weights proposed in Mammen (1993), as well as for $W_i=(V_i-2)$ with V_i following a Gamma distribution with mean 2 and variance 1. Alternatively, the Rademacher distribution, which satisfies $E[W_i]=E[W_i^3]=0$ and $E[W_i^2]=E[W_i^4]=1$, is able to match the first four moments for symmetric distributions and can in such cases provide an additional refinement (Liu (1988), Davidson and Flachaire (2008)).

3 The Score Bootstrap

The wild bootstrap resampling scheme is often interpreted as a means of generating a set of bootstrap residuals mimicking the heteroscedastic nature of the true errors. However, the residuals only influence the limiting distribution of the OLS estimator through the score. Thus, we may alternatively view the wild bootstrap as creating a set of bootstrap score contributions ($\{X_i\varepsilon_i^*\}_{i=1}^n$) that mimic the heteroscedastic nature of the true score contributions ($\{X_i\varepsilon_i\}_{i=1}^n$). In this section, we develop the implications of this observation, which provides the basis for our procedure.

The relationship between the wild bootstrap and the score is transparent from the discussion of its consistency in Section 2. Since $\varepsilon_i^* = e_i W_i$, we learn from (4) that the wild bootstrap may be interpreted as a perturbation of the score contributions ($\{X_i(Y_i - X_i'\beta)\}_{i=1}^n$) evaluated at the estimated parameter value ($\hat{\beta}$) that leaves the Hessian ($\frac{1}{n}\sum_i X_i X_i'$) unchanged.[5] More precisely, a numerically equivalent way to implement the wild bootstrap is given by the following algorithm:

STEP 1: Obtain the full sample OLS estimate $\hat{\beta}$ and employ it to generate the fitted score contributions $\{X_i(Y_i - X_i'\hat{\beta})\}_{i=1}^n$. □

STEP 2: Using an i.i.d. sample of random weights $\{W_i\}_{i=1}^n$ independent of $\{Y_i, X_i\}_{i=1}^n$ and satisfying $E[W_i] = 0$ and $E[W_i^2] = 1$, construct a new set of perturbed score contributions $\{X_i(Y_i - X_i'\hat{\beta})W_i\}_{i=1}^n$. □

STEP 3: Multiply the constructed perturbed score by the inverse Hessian to obtain $H_n^{-1} n^{-\frac{1}{2}} \sum_i (Y_i - X_i\hat{\beta}) X_i W_i$ and use its distribution conditional on $\{Y_i, X_i\}_{i=1}^n$ as an estimate of the distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$. □

Unlike the residual based view of the wild bootstrap, the score interpretation is easily generalized to nonlinear models. One may simply perturb the fitted score contributions of such a model while keeping the Hessian unchanged and, provided $E[W_i] = 0$ and $E[W_i^2] = 1$, the first two moments of the perturbed score and true score will match asymptotically. Under the appropriate regularity conditions, this moment equivalence will suffice for establishing the consistency of the proposed bootstrap. For obvious reasons, we term this procedure a "score bootstrap".

In order to fix ideas, we illustrate in the following example how this intuition may be applied in a nonlinear model.

Example 3.1. Consider a standard probit model in which $\varepsilon_i \sim N(0,1)$ and:

$$Y_i = 1\{X_i'\beta_0 \geq \varepsilon_i\} \ , \tag{7}$$

[5] In contrast, the weighted bootstrap perturbs the score and Hessian (Barbe and Bertail (1995)).

where $1\{\cdot\}$ is the indicator function. Suppose we wish to approximate the distribution of $\sqrt{n}(\hat{\beta}-\beta_0)$, where $\hat{\beta}$ is the maximum likelihood estimate of β_0. The log-likelihood which $\hat{\beta}$ maximizes is then:

$$L_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n}\{Y_i \log(\Phi(X_i'\beta)) + (1-Y_i)\log(1-\Phi(X_i'\beta))\} , \qquad (8)$$

for Φ the cdf of a standard normal random variable. Thus, the score is given by:

$$S_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n} s(Y_i, X_i, \beta) \qquad s(y,x,\beta) \equiv \frac{\phi(x'\beta)(y-\Phi(x'\beta))}{\Phi(x'\beta)(1-\Phi(x'\beta))}x , \qquad (9)$$

where ϕ is the derivative of Φ. The principles derived from the linear model then suggest estimating the distribution of $\sqrt{n}(\hat{\beta}-\beta_0)$ by: (i) Obtaining fitted score contributions $\{s(Y_i,X_i,\hat{\beta})\}_{i=1}^n$; (ii) Perturbing them by random weights $\{W_i\}_{i=1}^n$ to obtain $\{s(Y_i,X_i,\hat{\beta})W_i\}_{i=1}^n$; (iii) Multiplying the perturbed score by the inverse Hessian $([\frac{1}{n}\sum_i \nabla_\beta s(Y_i,X_i,\hat{\beta})]^{-1} n^{-\frac{1}{2}} \sum_i s(Y_i,X_i,\hat{\beta})W_i)$ and employing its distribution conditional on $\{Y_i,X_i\}_{i=1}^n$ to approximate that of $\sqrt{n}(\hat{\beta}-\beta_0)$. □

3.1 Higher Order Equivalence

In the linear model, the wild and score bootstrap statistics for $\sqrt{n}(\hat{\beta}-\beta_0)$ are numerically equivalent. However, in most instances the statistic of interest is studentized, since only in this context is a refinement over an analytical approximation available (Liu (1988), Horowitz (2001)). In accord with the perturbed score interpretation, it is natural to employ the sample variance of the perturbed score contributions when studentizing. For this reason, we define the bootstrap statistics:

$$\begin{aligned} T_n^{*w} &\equiv (H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1})^{-\frac{1}{2}}\sqrt{n}(\hat{\beta}^*-\hat{\beta}) \\ T_n^{*s} &\equiv (H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1})^{-\frac{1}{2}}H_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i\varepsilon_i^* , \end{aligned} \qquad (10)$$

where $\Sigma_n^*(\beta) \equiv \frac{1}{n}\sum_i X_i X_i'(Y_i^* - X_i'\beta)^2$ and T_n^{*w} and T_n^{*s} are the studentized wild and score bootstrap statistics respectively. It is important to note that in the computation of T_n^{*s}, the full sample estimator $\hat{\beta}$ is used in obtaining the standard errors, and hence calculation of $\hat{\beta}^*$ and its implied residuals is unnecessary. As a result, the score bootstrap is computationally simpler to implement than the wild bootstrap which requires obtaining bootstrap residuals.

While for the statistics in (4) the wild and score bootstraps are numerically equivalent, such a relationship fails to hold for the studentized versions. An important concern then is whether the refinement of the wild bootstrap over a normal approximation (Liu (1988)) is lost due to this discrepancy. Somewhat surprisingly, the differences between the wild and score bootstrap are asymptotically negligible even at higher order. Specifically, the wild and score bootstrap statistics are asymptotically equivalent up to a higher order than that of the refinement the wild bootstrap possesses over the normal approximation. As a result, under appropriate regularity conditions, the score bootstrap not only remains consistent despite not recomputing the estimator but can in addition be expected to obtain a refinement over an analytical approximation in precisely the same instances as the wild bootstrap.

In order to establish the higher order equivalence of T_n^{*s} and T_n^{*w}, we impose the following assumption on the data:

Assumption 3.1. *(i)* $\{Y_i, X_i\}_{i=1}^n$ *are i.i.d.* $E[X_i\varepsilon_i] = 0$*, and* $E[X_iX_i']$, $E[X_iX_i'\varepsilon_i^2]$ *are full rank; (ii) The moments* $E[\|X_i\|^4]$, $E[\varepsilon_i^4]$ *and* $E[\|X_i\|^4\varepsilon_i^4]$ *are finite; (iii)* $\{W_i\}_{i=1}^n$ *are i.i.d., independent of* $\{Y_i, X_i\}_{i=1}^n$ *with* $E[W_i] = 0$, $E[W_i^2] = 1$ *and* $E[W_i^4] < \infty$.

Let P^* and E^* denote probability and expectation conditional on $\{Y_i, X_i\}_{i=1}^n$ (but not $\{W_i\}_{i=1}^n$). Under Assumption 3.1 we can then establish the higher order equivalence of T_n^{*w} and T_n^{*s} under P^*.

Lemma 3.1. *Under Assumption 3.1,* $T_n^{*w} = T_n^{*s} + O_{p^*}(n^{-1})$ *almost surely.*

If the conditions for an Edgeworth expansion of the bootstrap statistics T_n^{*w} and T_n^{*s} are satisfied, then Lemma 3.1 implies that they can be expected to disagree only in terms of order n^{-1} or smaller; see Chapter 2.7 in Hall (1992) for such arguments.[6] Therefore, in settings where the wild bootstrap obtains the traditional Edgeworth refinement of order $n^{-\frac{1}{2}}$ over a normal approximation, the score bootstrap should as well. Kline and Santos (2011) show that such a refinement is often available in the linear model even under certain forms of misspecification.

The higher order equivalence of T_n^{*w} and T_n^{*s} is at first glance unexpected since the score bootstrap appears to violate the usual plug-in approach of the standard bootstrap. However, this only introduces a smaller order error due to the

[6] More precisely, Lemma 3.1 is not sufficient for showing the equivalence of the first two terms in the Edgeworth expansions. Such an equivalence can be established if $P(P^*(\|T_n^{*w} - T_n^{*s}\| > (n^{\frac{1}{2}}\log n)^{-1}) > n^{-\frac{1}{2}}) = o(n^{-\frac{1}{2}})$ and the Edgeworth expansion is valid in the bootstrap sample with probability $1 - o(n^{-\frac{1}{2}})$ (Lemma 5 Andrews (2002)).

residuals $\{\varepsilon_i^*\}_{i=1}^n$ being mean independent of $\{X_i\}_{i=1}^n$ under the bootstrap distribution. Importantly, the higher order equivalence would fail to hold if the residuals $\{\varepsilon_i^*\}$ were sampled in a manner under which they were merely uncorrelated with $\{X_i\}_{i=1}^n$ under the bootstrap distribution.

Remark 3.1. The bootstrap estimator $\hat{\beta}^*$ acquired from running OLS in the sample $\{Y_i^*, X_i\}$ may easily be obtained from the score bootstrap procedure by the equality:

$$\hat{\beta}^* = \hat{\beta} + H_n^{-1}\frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i^* \ . \tag{11}$$

Note that the right hand side of equation (11) is a single Newton-Raphson step towards the wild bootstrap estimator $\hat{\beta}^*$ starting from $\hat{\beta}$. Thus, there is a close connection between our approach and the k-step bootstrap procedure studied by Davidson and MacKinnon (1999a) and Andrews (2002). This interpretation, however, does not carry over to nonlinear models where wild bootstrap procedures have yet to be proposed. □

4 Inference

We turn now to establishing the validity of a score bootstrap procedure for estimating the critical values of a large class of tests. Building on our earlier discussion we consider test statistics based upon the fitted parametric scores of M-estimators, using perturbations of those scores to estimate their sampling distribution. Since this approach does not depend upon resampling of residuals, we do not distinguish between dependent and exogenous variables and instead consider a random vector $Z_i \in \mathscr{Z} \subseteq \mathbf{R}^m$ which may contain both.

We study test statistics G_n that are quadratic forms in a vector valued T_n:

$$G_n \equiv T_n' T_n \ . \tag{12}$$

Under the null hypothesis, the underlying statistic T_n is required to be asymptotically pivotal and allow for a linear expansion. More precisely, we require that under the null hypothesis the following hold:

$$T_n = (A_n(\theta_0)\Sigma_n(\theta_0)A_n(\theta_0)')^{-\frac{1}{2}} S_n(\theta_0) + o_p(1) \qquad S_n(\theta) \equiv A_n(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^n s(Z_i, \theta) \ , \tag{13}$$

where $A_n(\theta)$ is a $r \times k$ matrix, $s(z,\theta)$ is a $k \times 1$ vector, $\Sigma_n(\theta)$ is the sample covariance matrix of $\{s(Z_i,\theta)\}_{i=1}^n$ and θ_0 is an unknown parameter vector. Examples are discussed in Sections 4.2-4.4.

In accord with the terminology we employed for the linear model, we refer to $A_n(\theta)$ as the inverse of the Hessian, $\sum_i s(Z_i,\theta)$ as the score of the model, and to $\{s(Z_i,\theta)\}_{i=1}^n$ as the score contributions. Under appropriate regularity conditions, T_n is therefore asymptotically normally distributed with identity covariance matrix and hence G_n is asymptotically Chi-squared distributed with degrees of freedom equal to the dimension of T_n. Though we only consider asymptotically pivotal statistics, our results readily extend to unstudentized ones as well.

The bootstrap statistics employed to estimate the distributions of G_n and T_n are given by $G_n^* \equiv T_n^{*\prime} T_n^*$, with T_n^* defined as:

$$T_n^* \equiv (A_n(\hat\theta)\Sigma_n^*(\hat\theta)A_n(\hat\theta)')^{-\frac{1}{2}} S_n^*(\hat\theta) \qquad S_n^*(\theta) \equiv A_n(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^n s(Z_i,\theta)W_i \quad (14)$$

where $\Sigma_n^*(\theta)$ is the sample covariance matrix of $\{s(Z_i,\theta)W_i\}_{i=1}^n$ and $\hat\theta$ is a consistent estimator for θ_0. As discussed in the previous section, implementation of the score bootstrap only requires calculation of the full sample estimator $\hat\theta$; no additional optimization is needed in each bootstrap iteration.

Remark 4.1. An alternative to perturbing the fitted score contributions by random weights is to instead resample them with replacement. Specifically, for $\tilde s_i \equiv s(Z_i,\hat\theta)$, we may consider drawing from $\{\tilde s_i\}_{i=1}^n$ with replacement and approximating the distribution of G_n by that of $\tilde G_n$, where $\tilde G_n \equiv \tilde T_n' \tilde T_n$ and:

$$\tilde T_n \equiv (A_n(\hat\theta)\tilde\Sigma_n(\hat\theta)A_n(\hat\theta)')^{-\frac{1}{2}} \tilde S_n(\hat\theta) \qquad \tilde S_n(\theta) \equiv A_n(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde s_i$$

for $\tilde\Sigma_n(\hat\theta)$ the sample covariance matrix of $\{\tilde s_i\}_{i=1}^n$. In the linear model, this procedure corresponds to that of Hu and Zidek (1995). For nonlinear problems, Hu and Kalbfleisch (2000) propose a closely related approach that additionally resamples the Hessian. Specifically, when the Hessian takes the form $A_n^{-1}(\hat\theta) = \sum_i a(Z_i,\hat\theta)$, their bootstrap procedure samples with replacement from both $\{\tilde s_i\}_{i=1}^n$ and $\{a(Z_i,\hat\theta)\}_{i=1}^n$ to obtain a new score and inverse Hessian. This may be thought of as an approximation to the traditional nonparametric ("pairs") bootstrap. Like the pairs bootstrap however, their procedure may encounter computational difficulties when the Hessian is poorly behaved in some bootstrap draws, a problem which becomes more likely in small samples when some of the covariates are discrete. □

4.1 Bootstrap Consistency

We establish the consistency of the bootstrap under the following assumptions:

Assumption 4.1. *(i) $\hat{\theta} \xrightarrow{p} \theta_0$ with $\hat{\theta}, \theta_0 \in \Theta \subset \mathbf{R}^p$ and Θ a compact set; (ii) θ_0 satisfies $E[s(Z_i, \theta_0)s(Z_i, \theta_0)'] < \infty$ and $A(\theta_0)E[s(Z_i, \theta_0)s(Z_i, \theta_0)']A(\theta_0)'$ is invertible.*

Assumption 4.2. *(i) Under the null hypothesis T_n satisfies* (13) *and θ_0 is such that $E[s(Z_i, \theta_0)] = 0$; (ii) Under the alternative hypothesis $G_n \xrightarrow{p} \infty$.*

Assumption 4.3. *(i) $\{Z_i\}_{i=1}^n$ is i.i.d.; (ii) $\sup_{\theta \in \Theta} \|A_n(\theta) - A(\theta)\|_F = o_p(1)$ with $A(\theta)$ continuous and $\|\cdot\|_F$ the Frobenius norm.*

Assumption 4.4. *(i) $\{W_i\}_{i=1}^n$ is an i.i.d. sample, independent of $\{Z_i\}_{i=1}^n$ satisfying $E[W_i] = 0$ and $E[W_i^2] = 1$; (ii) For conv(Θ) the convex hull of Θ, $s(z, \theta)$ is continuously differentiable in $\theta \in conv(\Theta)$ and $\sup_{\theta \in conv(\Theta)} \|\nabla s(z, \theta)\|_F \leq F(z)$ for some function $F(z)$ with $E[F^2(Z_i)] < \infty$.*

In Assumption 4.1 we require $\hat{\theta}$ to converge in probability to some parameter vector $\theta_0 \in \Theta$ whose value may depend upon the distribution of Z_i. The compactness of the parameter space Θ is employed to verify the perturbed scores form a Donsker class. This restriction may be relaxed at the expense of a more complicated argument that exploits the consistency of $\hat{\theta}$ for a local analysis. Though in the notation we suppress such dependence, it is important to note that θ_0 may take different values under the null and alternative hypotheses. Assumptions 4.2 and 4.3, in turn enable us to establish the asymptotic behavior of G_n under the null and alternative hypotheses. Assumption 4.4(i) imposes the only requirements on the random weights $\{W_i\}_{i=1}^n$, which are the same conditions imposed for inference on the linear model in previous wild bootstrap studies. Assumption 4.4(ii) allows us to establish that the empirical process induced by functions of the form $ws(z, \theta)$ is asymptotically tight. Differentiability is not necessary for this end, but we opt to impose it due to its ease of verification and wide applicability.[7] We note, however, that estimation of $A(\theta_0)$ may be more challenging in the non-differentiable case as this quantity usually depends on the population Hessian.

Assumptions 4.1-4.4 are sufficient for establishing the consistency of the proposed score bootstrap procedure under the null hypothesis.

[7] For non-differentiable settings, we need $\mathscr{F} \equiv \{ws(z, \theta) : \theta \in \Theta\}$ to be a Donsker class.

Theorem 4.1. *Let F_n and F_n^* be the cdfs of G_n and of G_n^* conditional on $\{Z_i\}_{i=1}^n$ and Assumptions 4.1, 4.2, 4.3 and 4.4 hold. If the null hypothesis is true, then:*

$$\sup_{c\in\mathbf{R}} |F_n(c) - F_n^*(c)| = o_p(1) \ .$$

Theorem 4.1 justifies the use of quantiles from the distribution of G_n^* conditional on $\{Z_i\}_{i=1}^n$ as critical values for the test statistic G_n. In order to control the size of the test at level α, we may employ:

$$\hat{c}_{1-\alpha} \equiv \inf\{c : P(G_n^* \leq c \,|\{Z_i\}_{i=1}^n) \geq 1-\alpha\} \ . \tag{15}$$

While difficult to compute analytically, $\hat{c}_{1-\alpha}$ may be calculated via simulation. Employing a random number generator, B samples $\{\{W_{i1}\}_{i=1}^n, \ldots, \{W_{iB}\}_{i=1}^n\}$ may be created independently of the data and used to construct B statistics $\{G_{n1}^*, \ldots, G_{nB}^*\}$. Provided B is sufficiently large, the empirical $1-\alpha$ quantile of $\{G_{n1}^*, \ldots, G_{nB}^*\}$ will yield an accurate approximation to $\hat{c}_{1-\alpha}$.

While Theorem 4.1 implies that the critical value $\hat{c}_{1-\alpha}$ in conjunction with the test statistic G_n delivers size control, it does not elucidate the behavior of the test under the alternative hypothesis. As in other bootstrap procedures, the test is consistent due to the bootstrap statistic G_n^* being properly centered even under the alternative. As a result, $\hat{c}_{1-\alpha}$ converges in probability to the $1-\alpha$ quantile of a Chi-squared distribution with r degrees of freedom, while G_n diverges to infinity. Therefore, under the alternative hypothesis, G_n is larger than $\hat{c}_{1-\alpha}$ with probability tending to one and the test rejects asymptotically. We summarize these findings in the following corollary:

Corollary 4.1. *Under Assumptions 4.1, 4.2, 4.3 and 4.4, it follows that under the null hypothesis:*

$$\lim_{n\to\infty} P(G_n \geq \hat{c}_{1-\alpha}) = 1-\alpha \ ,$$

for any $0 < \alpha < 1$. Under the same assumptions, if the alternative hypothesis is instead true, then:

$$\lim_{n\to\infty} P(G_n \geq \hat{c}_{1-\alpha}) = 1 \ .$$

4.2 Parameter Tests

A principal application of the proposed bootstrap is in obtaining critical values for parametric hypothesis tests. We consider a general M-estimation framework in which the parameter of interest θ_M is the unique minimizer of some non-stochastic but unknown function $Q : \Theta \to \mathbf{R}$:

$$\theta_M = \arg\min_{\theta \in \Theta} Q(\theta) \ . \tag{16}$$

We examine the classic problem of conducting inference on a function of θ_M. Specifically, for some known and differentiable mapping $c : \Theta \to \mathbf{R}^l$ with $l \leq p$, the hypothesis we study is:

$$H_0 : c(\theta_M) = 0 \qquad\qquad H_1 : c(\theta_M) \neq 0 \ . \tag{17}$$

Standard tests for this hypothesis include the Wald and Lagrange Multiplier (LM) tests. Intuitively, the Wald test examines whether the value of the function c evaluated at an unrestricted estimator $\hat{\theta}_M$ is statistically different from zero. In contrast, the LM test instead checks whether the first order condition of an estimator $\hat{\theta}_{M,R}$ computed imposing the null hypothesis is statistically different from zero. Therefore, in the nomenclature of Assumption 4.1(i), $\hat{\theta}$ equals $\hat{\theta}_M$ for the Wald test and $\hat{\theta}_{M,R}$ for the LM test. Similarly, if $\theta_{M,R}$ denotes the minimizer of Q over Θ subject to $c(\theta) = 0$, then θ_0 equals θ_M and $\theta_{M,R}$ under the Wald and LM test respectively.

We proceed to illustrate the details of the score bootstrap in this setting for both generalized method of moments (GMM) and maximum likelihood (ML) estimators. We focus on the analytical expressions $A_n(\theta)$ and $s(z,\theta)$ take in those specific settings and provide references for primitive conditions that ensure Assumptions 4.1, 4.2, 4.3 and 4.4 hold.

4.2.1 ML Estimators

For a ML estimator, Q and its sample analogue Q_n are of the general form:

$$Q_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} q(Z_i,\theta) \qquad\qquad Q(\theta) \equiv E[q(Z_i,\theta)] \ , \tag{18}$$

where $q : \mathscr{Z} \times \Theta \to \mathbf{R}$ is the log-likelihood. If q is twice differentiable in θ, then we may define the Hessian $H_n(\theta) \equiv n^{-1}\sum_i \nabla^2 q(Z_i,\theta)$. For notational convenience, it is also helpful to denote the gradient of the function c evaluated at θ by $C(\theta) \equiv \nabla c(\theta)$.

Example 4.1. (Wald) The relevant Wald statistic is the studentized quadratic form of $\sqrt{n}c(\hat{\theta}_M)$, which under both the null and alternative hypothesis satisfies:

$$\sqrt{n}(c(\hat{\theta}_M) - c(\theta_M)) = -C(\theta_M)H_n^{-1}(\theta_M)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla q(Z_i,\theta_M) + o_p(1) \ . \tag{19}$$

Therefore, the Wald statistic fits the formulation in (13) with $A_n(\theta) = -C(\theta)H_n^{-1}(\theta)$ and $s(z,\theta) = \nabla q(z,\theta)$. Under the alternative hypothesis, G_n diverges to infinity since $c(\theta_M) \neq 0$. Refer to Section 3.2 in Newey and McFadden (1994) for a formal justification of these arguments. □

Example 4.2. (LM) The LM statistic is the normalized quadratic form of:

$$C(\hat{\theta}_{M,R})H_n^{-1}(\hat{\theta}_{M,R})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla q(Z_i,\hat{\theta}_{M,R}) \ . \tag{20}$$

Moreover, under conditions stated in Chapter 12.6.2 in Wooldridge (2002),

$$\begin{aligned} &C(\hat{\theta}_{M,R})H_n^{-1}(\hat{\theta}_{M,R})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla q(Z_i,\hat{\theta}_{M,R}) \\ &\qquad = C(\theta_{M,R})H_n^{-1}(\theta_{M,R})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla q(Z_i,\theta_{M,R}) + o_p(1) \ , \end{aligned} \tag{21}$$

under the null hypothesis. Thus, the LM statistic also fits the general formulation in (13) with $A_n(\theta) = C(\theta)H_n^{-1}(\theta)$ and $s(z,\theta) = \nabla q(z,\theta)$. Under the alternative, $G_n \xrightarrow{p} \infty$ provided $\theta_{M,R}$ is not a local minimizer of Q, $C(\theta_{M,R})E[\nabla^2 q(Z_i,\theta_{M,R})]$ is full rank and Assumption 4.1(ii) holds. □

4.2.2 GMM Estimators

In the context of GMM estimation, Q and its sample analogue Q_n are of the form:

$$Q_n(\theta) \equiv [\frac{1}{n}\sum_{i=1}^{n} q(Z_i,\theta)']\Omega_n[\frac{1}{n}\sum_{i=1}^{n} q(Z_i,\theta)] \qquad Q(\theta) \equiv E[q(Z_i,\theta)']\Omega E[q(Z_i,\theta)] \ , \tag{22}$$

where $q : \mathscr{X} \times \Theta \to \mathbf{R}^k$ is a known function and Ω_n, Ω are positive definite matrices such that $\Omega_n \xrightarrow{p} \Omega$. Assuming q is differentiable in θ, let $D_n(\theta) \equiv n^{-1}\sum_i \nabla q(Z_i,\theta)$ and $B_n(\theta) \equiv D_n(\theta)'\Omega_n D_n(\theta)$. As in the discussion of ML estimators, we also denote $C(\theta) \equiv \nabla c(\theta)$.

Example 4.3. (Wald) The Wald statistic for the hypothesis in (17) is given by the studentized quadratic form of $\sqrt{n}c(\hat{\theta}_M)$. In the present context we therefore obtain:

$$\sqrt{n}(c(\hat{\theta}_M) - c(\theta_M)) = -C(\theta_M)B_n^{-1}(\theta_M)D_n(\theta_M)'\Omega_n\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(Z_i,\theta_M) + o_p(1) \tag{23}$$

which implies $A_n(\theta) = -C(\theta)B_n^{-1}(\theta)D_n(\theta)'\Omega_n$ and $s(z,\theta) = q(z,\theta)$ and Assumption 4.2(i) is satisfied provided $E[q(Z_i,\theta_M)] = 0$.[8] Primitive conditions under which Assumptions 4.1-4.4 hold in this context can be found in Section 3.3 of Newey and McFadden (1994). □

Example 4.4. **(LM)** The LM test statistic is the studentized quadratic form of:

$$C(\hat{\theta}_{M,R})B_n^{-1}(\hat{\theta}_{M,R})D_n(\hat{\theta}_{M,R})'\Omega_n \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(Z_i,\hat{\theta}_{M,R}) \ , \tag{24}$$

which under the null hypothesis, as shown in Section 9.1 of Newey and McFadden (1994), is asymptotically equivalent to:

$$C(\theta_{M,R})B_n^{-1}(\theta_{M,R})D_n(\theta_{M,R})'\Omega_n \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(Z_i,\theta_{M,R}) \ . \tag{25}$$

Hence, in this setting $A_n(\theta) = C(\theta)B_n^{-1}(\theta)D_n(\theta)'\Omega_n$ and $s(z,\theta) = q(z,\theta)$. □

4.3 Moment Restrictions

An additional application of the bootstrap procedure we consider is for testing:

$$H_0 : E[m(Z_i,\theta_M)] = 0 \qquad H_1 : E[m(Z_i,\theta_M)] \neq 0 \ , \tag{26}$$

where $m : \mathscr{X} \times \Theta \to \mathbf{R}^l$ is a known function and θ_M is the minimizer of some unknown non-stochastic $Q : \Theta \to \mathbf{R}$. Such restrictions arise, for example, in tests of proper model specification and hypotheses regarding average marginal effects in nonlinear models. As in Section 4.2, the specific nature of the bootstrap statistic is dependent on whether Q is as in (18) (ML) or as in (22) (GMM). For brevity, we focus on the former, though the extension to GMM can be readily derived following manipulations analogous to those in Example 4.3.

The Wald test statistic for the hypothesis in (26) is the quadratic form of the studentized plug-in estimator:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Z_i,\hat{\theta}_M) \ , \tag{27}$$

[8] Notice this is trivially satisfied in a just identified system. The extension to overidentified models in which $E[q(Z_i,\theta_M)] \neq 0$ but $E[\nabla q(Z_i,\theta_M)']\Omega E[q(Z_i,\theta_M)] = 0$ can be accomplished by letting $s(z,\theta)$ depend on n and setting $s_n(z,\theta) = D_n(\theta)'\Omega_n g(z,\theta)$. Though straightforward to establish, we do not pursue such an extension.

where $\hat{\theta}_M$ is in this case the unconstrained minimizer of Q_n on Θ. Hence, in this setting θ_0 equals θ_M and $\hat{\theta}$ equals $\hat{\theta}_M$ in the notation of Assumption 4.1(i). Obtaining an expansion for T_n as in (13) is straightforward provided m and q are once and twice continuously differentiable in θ respectively. Defining the gradient $M_n(\theta) \equiv n^{-1}\sum_i \nabla m(Z_i,\theta)$ and Hessian $H_n(\theta) \equiv n^{-1}\sum_i \nabla^2 q(Z_i,\theta)$, standard arguments imply that under the null hypothesis:

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Z_i,\hat{\theta}_M)\\ &\qquad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Z_i,\theta_M) - M_n(\theta_M)H_n^{-1}(\theta_M)\frac{1}{\sqrt{n}}\sum_{i=1} \nabla q(Z_i,\theta_M) + o_p(1)\ ; \quad (28)\end{aligned}$$

see Newey (1985a) for primitive conditions. Therefore, in this setting:

$$s(z,\theta) = \begin{pmatrix} m(z,\theta) \\ \nabla q(z,\theta)\end{pmatrix} \qquad A_n(\theta) = \left[I \ \vdots\ -M_n(\theta)H_n^{-1}(\theta)\right]. \qquad (29)$$

Moreover, if θ_M is an interior point of Θ, then $E[\nabla q(Z_i,\theta_M)] = 0$ because θ_M minimizes Q. Hence, $G_n \xrightarrow{p} \infty$ under the alternative hypothesis by $E[m(Z_i,\theta_M)] \neq 0$.

Remark 4.2. Similar manipulations may be employed to show the score bootstrap can be applied to Wald tests in two stage parametric estimation problems. Unlike the nonparametric bootstrap, however, such a procedure would require an analytical derivation of the gradient of the second stage influence function with respect to the first stage parameters. □

4.3.1 ML Specification Tests

A prominent application of hypotheses as in (26) is in model specification testing. In particular, this setting encompasses moment based specification tests ("m-tests") for maximum likelihood models, as considered in White (1982, 1994), Newey (1985b) and Tauchen (1985).[9] Computations are simplified for ML models by the generalized information matrix equality, which implies:

$$\begin{aligned} E[\nabla^2 q(Z_i,\theta_M)] &= -E[\nabla q(Z_i,\theta_M)\nabla q(Z_i,\theta_M)'] \\ E[\nabla m(Z_i,\theta_M)] &= -E[m(Z_i,\theta_M)\nabla q(Z_i,\theta_M)']\ . \qquad (30)\end{aligned}$$

[9] For a bootstrap construction for the Information Matrix Equality test see Horowitz (1994).

For example, as noted in Chesher (1984) and Newey (1985b), computation of the Wald test statistic for the null hypothesis in (26) can be performed through the auxiliary regression:

$$1 = m(Z_i, \hat{\theta}_M)'\delta + \nabla q(Z_i, \hat{\theta}_M)'\gamma + \varepsilon_i \ . \tag{31}$$

Equation (31) is often termed the Outer Product of Gradient (OPG) regression form for moment tests; see Chapter 15.2 in Davidson and MacKinnon (2004) or Chapter 8.2.2 in Cameron and Trivedi (2005). If R^2 is the uncentered R-squared of the regression in (31), then under the generalized information matrix equality result in (30) the Wald test statistic is asymptotically equivalent to:

$$G_n = nR^2 \ . \tag{32}$$

The calculation of the score bootstrap simplifies in an analogous fashion. Under a uniform law of large numbers, $A_n(\hat{\theta}_M)$ as defined in (29) satisfies:

$$\begin{aligned} & A_n(\hat{\theta}_M) \\ & = \Big[I \vdots -\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\theta}_M)\nabla q(Z_i, \hat{\theta}_M)' \Big[\frac{1}{n}\sum_{i=1}^{n} \nabla q(Z_i, \hat{\theta}_M)\nabla q(Z_i, \hat{\theta}_M)'\Big]^{-1} \Big] + o_p(1) \end{aligned}$$

under the null hypothesis. As a result, the score bootstrap has a simple interpretation in terms of the multivariate regression of the moments $m(Z_i, \hat{\theta}_M)$ on $\nabla q(Z_i, \hat{\theta}_M)$:

$$\begin{array}{rl} m^{(1)}(Z_i, \hat{\theta}_M) & = \nabla q(Z_i, \hat{\theta}_M)'\beta_1 + \varepsilon_{1,i} \\ \vdots & = \quad \vdots \\ m^{(l)}(Z_i, \hat{\theta}_M) & = \nabla q(Z_i, \hat{\theta}_M)'\beta_l + \varepsilon_{l,i} \end{array} , \tag{33}$$

where $m^{(j)}(Z_i, \hat{\theta}_M)$ is the j^{th} component of $m(Z_i, \hat{\theta}_M)$. In particular, letting $e_{j,i} \equiv m^{(j)}(Z_i, \hat{\theta}_M) - \nabla q(Z_i, \hat{\theta}_M)'\hat{\beta}_j$ be the fitted residual of the j^{th} regression and denoting $e_i = (e_{1,i}, \ldots, e_{l,i})'$, we obtain that:

$$S_n^*(\hat{\theta}_M) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i W_i \ . \tag{34}$$

Hence, G_n^* is simply the Wald test statistic for the null hypothesis that $E[e_i W_i] = 0$.

In summary, if the generalized information matrix equality holds, then in testing (26) we may employ the following simple algorithm:

STEP 1: Run the regression in (31) and compute the uncentered R-squared to obtain the test statistic G_n as in (32). □

STEP 2: Regress $\{m(Z_i,\hat{\theta}_M)\}_{i=1}^n$ on $\{\nabla q(Z_i,\hat{\theta}_M)\}_{i=1}^n$ to generate $\{e_i\}_{i=1}^n$. □

STEP 3: Using random weights $\{W_i\}_{i=1}^n$ independent of $\{Y_i,X_i\}_{i=1}^n$ with $E[W_i]=0$ and $E[W_i^2]=1$, perturb the original residual vectors $\{e_i\}_{i=1}^n$ to obtain a new set of residual vectors $\{e_iW_i\}_{i=1}^n$. □

STEP 4: Let G_n^* be the Wald test statistic for the null that $E[e_iW_i]=0$ calculated using $\{e_iW_i\}_{i=1}^n$. To control size at level α, reject if G_n is larger than the $1-\alpha$ quantile of G_n^* conditional on $\{Z_i\}_{i=1}^n$. □

4.4 Clustered Data

Theorem 4.1 and Corollary 4.1 may be applied to clustered data provided clusters are i.i.d. with the same number of observations. Extensions to settings where the clusters are unbalanced or there is heteroskedasticity across them are feasible, essentially requiring an extension of Theorem 4.1 to independent but not identically distributed observations.

Let Z_{ic} denote observation number i in cluster c, J be the total number of observations per cluster, n be the total number of clusters and $Z_c=\{Z_{1c},\ldots,Z_{Jc}\}$. Following (13), we consider test statistics of the form $\tilde{G}_n\equiv\tilde{T}_n'\tilde{T}_n$, where $\tilde{T}_n$ satisfies:

$$\begin{aligned}\tilde{S}_n(\theta) &\equiv A_n(\theta)\frac{1}{\sqrt{n}}\sum_{c=1}^{n}\frac{1}{\sqrt{J}}\sum_{i=1}^{J}\tilde{s}(Z_{ic},\theta)\\ \tilde{T}_n &= (A_n(\theta_0)\tilde{\Sigma}_n(\theta_0)A_n(\theta_0)')^{-\frac{1}{2}}\tilde{S}_n(\theta_0)+o_p(1)\ ,\end{aligned} \tag{35}$$

where $A_n(\theta)$ is again a $r\times m$ matrix, $\tilde{s}(z,\theta)$ maps each (Z_{ic},θ) into a $m\times 1$ vector, and $\tilde{\Sigma}_n(\theta)$ is a robust covariance matrix that allows for arbitrary correlation within cluster. The Wald and LM test statistics, as well as the moment restriction tests previously discussed all extend to this setting when observations are allowed to be dependent within clusters.

The applicability of Theorem 4.1 and Corollary 4.1 to the present context is immediate once we define $s(z,\theta)$, mapping each (Z_c,θ) into a $m\times 1$ vector, by:

$$s(Z_c,\theta)=\frac{1}{\sqrt{J}}\sum_{i=1}^{J}\tilde{s}(Z_{ic},\theta)\ . \tag{36}$$

The statistics $\tilde{T}_n$ and $\tilde{S}_n(\theta)$ are then special cases of T_n and $S_n(\theta)$ as considered in (13) but with Z_c in place of Z_i. Hence, equations (13) and (36) indicate that the relevant bootstrap statistic should perturb the data at the level of the cluster rather than the individual observation. We thus define $\tilde{G}_n^* \equiv \tilde{T}_n^{*\prime}\tilde{T}_n^*$, for:

$$\tilde{T}_n^* \equiv (A_n(\hat{\theta})\tilde{\Sigma}_n^*(\hat{\theta})A_n(\hat{\theta})')^{-\frac{1}{2}}\tilde{S}_n^*(\hat{\theta}) \qquad \tilde{S}_n^*(\theta) \equiv A_n(\theta)\frac{1}{\sqrt{n}}\sum_{c=1}^{n}\frac{W_c}{\sqrt{J}}\sum_{i=1}^{J}\tilde{s}(Z_{ic},\theta)$$

where $\tilde{\Sigma}_n^*(\theta)$ is a robust bootstrap covariance matrix for $s(Z_{ic},\theta)W_c$.

Given these definitions, it is readily apparent that $\tilde{G}_n^*$, $\tilde{T}_n^*$ and $\tilde{S}_n^*(\theta)$ are themselves special cases of the bootstrap statistics G_n^*, T_n^* and $S_n^*(\theta)$. The consistency of the proposed score bootstrap then follows immediately provided the clusters are i.i.d., the number of clusters tends to infinity and $s(z,\theta)$ as defined in (36) satisfies Assumption 4.1(ii), 4.2(i) and 4.4(ii).

Corollary 4.2. *Under Assumptions 4.1, 4.2, 4.3 and 4.4, it follows that under the null hypothesis:*

$$\lim_{n\to\infty} P(\tilde{G}_n \geq \hat{c}_{1-\alpha}) = 1-\alpha \ ,$$

for any $0 < \alpha < 1$*. Under the same assumptions, if the alternative hypothesis is instead true, then:*

$$\lim_{n\to\infty} P(\tilde{G}_n \geq \hat{c}_{1-\alpha}) = 1 \ .$$

5 Simulation Evidence

To assess the small sample behavior of the score bootstrap we conduct a series of Monte Carlo experiments examining the performance of bootstrap Wald and LM tests of hypotheses regarding the parameters of a linear model estimated by OLS and a nonlinear probit model estimated by maximum likelihood. We also examine the performance of a test for the presence of intra cluster correlation in the probit model. Because small sample issues often arise in settings with dependent data, we work with hierarchical data generating processes (DGPs) exhibiting dependence of micro-units i within independent clusters c. We consider balanced panels with 20 observations per cluster and sampling designs ranging from 5 to 200 clusters.[10]

In order to allow a comparison of the wild and score bootstraps with the traditional nonparametric block bootstrap, we consider a setting with continuous regressors so that the block bootstrap distribution may be computed in small samples. It is important to note, however, that in many studies the regressor of interest

[10] In unreported results we found our results to be insensitive to variation in the number of observations per cluster.

will have discrete or binary support, in which case the statistic of interest will be undefined in bootstrap samples where only one value of the regressor is sampled. Moreover, even in bootstrap draws where the regressor of interest does exhibit variation, the Hessian may not be full rank. In such settings the traditional resampling based bootstrap will not be viable and the case for consideration of the wild and score bootstraps will be much stronger.

5.1 Designs

As pointed out by Chesher (1995), symmetric Monte Carlo designs are likely to yield an overly optimistic assessment of the ability of testing procedures to control size. For this reason we study the performance of our proposed bootstrap procedures under a variety of different designs meant to reflect realistic features of microeconomic datasets. Throughout, the linear model we examine is given by:

$$Y_{ic} = X_{ic} + D_c + \eta_c + \varepsilon_{ic} \ , \tag{37}$$

where the regressors (X_{ic}, D_c) and cluster level error (η_c) are generated by:

$$X_{ic} = X_c + \xi_{ic} \qquad\qquad \eta_c = (1 + D_c + X_c)\upsilon_c \ . \tag{38}$$

The regressor of interest is D_c, which varies only at the cluster level. Note that the cluster level random effect η_c exhibits heteroscedasticity with respect to D_c and X_{ic}.

Design I: (baseline) We let $(X_c, D_c, \xi_{ic}, \varepsilon_{ic})$ be normally distributed with identity covariance matrix, and υ_c independent of other variables with a t-distribution with six degrees of freedom. □

Design II: (skewed regressor) Design I is modified to generate D_c according to a mixture between a $N(0,1)$ with probability 0.9 and a $N(2,9)$ with probability 0.1 as in Horowitz (1997). This yields a regressor with occasional "outliers" and substantial skew and kurtosis in its marginal distribution. □

Design III: (misspecification) The model estimated is still (37), but the DGP is modified to:

$$Y_{ic} = X_{ic} + D_c + .1D_c^2 + \eta_c + \varepsilon_{ic} \ , \tag{39}$$

and other features remain as in Design I. Hence, the quadratic term in the regressor of interest is ignored in estimation which yields a skewed reduced form regression error. Note that $E[D_c^3] = E[X_{ic}D_c^2] = 0$ which ensures the population regression coefficient on D_c is still one. □

To study the performance of the score bootstrap in a nonlinear model we consider probit estimation of the following data generating process:

$$Y_{ic} = 1\{X_{ic} + D_c + \eta_c + \varepsilon_{ic} \geq 0\} \qquad X_{ic} = X_c + \xi_{ic} \; . \tag{40}$$

This is essentially a latent variable representation of the model in (37) without heteroscedasticity in the group error η_c. We consider the following two designs:

Design IV: (baseline probit) In model (40), we let $(X_c, D_c, \xi_{ic}) \sim N(0, I_3/16)$ and $(\eta_c, \varepsilon_{ic}) \sim N(0, I_2/2)$.[11] □

Design V: (skew probit) We modify Design IV by generating D_c according to a mixture distribution as in Design II, so that the distribution of the regressor of interest is heavily skewed. □

Finally, we illustrate the methods of Section 4.3 by testing for the presence of intra cluster correlation in the probit model. Specifically, we test the hypothesis:

$$H_0 : E[\sum_{i=1}^{20} \nu_{ic} \sum_{j \neq i} \nu_{jc}] = 0 \qquad H_1 : E[\sum_{i=1}^{20} \nu_{ic} \sum_{j \neq i} \nu_{jc}] \neq 0 \; , \tag{41}$$

where $\nu_{ic} = [Y_{ic} - p_{ic}]/\sqrt{p_{ic}(1 - p_{ic})}$ is a generalized residual and $p_{ic} = \Phi(X_{ic} + D_c)$ is the conditional probability that Y_{ic} equals one given D_c and X_{ic}.[12] Note that under the probit model $E[\nu_{ic}] = 0$ and $E[\nu_{ic}^2] = 1$. A test of H_0 examines whether within cluster dependence is present in the data, the finding of which might suggest the presence of an unmodeled cluster level random effect. In order to ensure the null hypothesis is true, we employ designs IV and V but set $\eta_c = 0$ almost surely and change the variance of ε_{ic} to equal one.

5.2 Results

Tables 1 and 2 provide empirical rejection rates from 10,000 Monte Carlo repetitions of Wald and LM tests of the null that the population least squares coefficient on D_c in (37) is one. All tests have a nominal size of 5% and are studentized using a recentered variance estimator.[13] Bootstrap tests are computed via bootstrap p-values based on 199 repetitions, and reject the null hypothesis whenever the p-

[11] Though the DGP contains a cluster level random effect, the marginal model for the outcome given covariates is a probit ensuring conventional maximum likelihood estimation is consistent.

[12] See McCall (1994) and Card and Hyslop (2005) for further examples of the use of generalized residual correlations as specification diagnostics.

[13] We make a finite sample degrees of freedom correction of $n/(n-1)$ to all variance estimators.

Table 1: Empirical Rejection Rates - Linear Model, Designs I&II

	Design I: Normal Regressor				
	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
			Wald		
Analytical	0.381	0.216	0.130	0.085	0.059
Pairs	0.013	0.046	0.041	0.039	0.045
Wild Rademacher	0.178	0.098	0.066	0.057	0.051
Wild Mammen	0.175	0.123	0.090	0.069	0.054
Wild2 Rademacher	0.057	0.059	0.055	0.052	0.051
Wild2 Mammen	0.051	0.024	0.027	0.036	0.048
Score Rademacher	0.267	0.184	0.127	0.092	0.063
Score Mammen	0.374	0.248	0.165	0.104	0.063
Score2 Rademacher	0.042	0.058	0.053	0.055	0.051
Score2 Mammen	0.043	0.036	0.059	0.063	0.053
			LM		
Analytical	0.157	0.090	0.065	0.056	0.052
Pairs	0.020	0.041	0.039	0.039	0.047
Score Rademacher	0.105	0.101	0.079	0.065	0.053
Score Mammen	0.080	0.031	0.035	0.044	0.050
	Design II: Mixture Regressor				
	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
			Wald		
Analytical	0.443	0.322	0.255	0.168	0.087
Pairs	0.018	0.082	0.084	0.068	0.039
Wild Rademacher	0.229	0.191	0.143	0.086	0.053
Wild Mammen	0.229	0.185	0.149	0.107	0.063
Wild2 Rademacher	0.077	0.083	0.070	0.056	0.050
Wild2 Mammen	0.066	0.035	0.041	0.036	0.032
Score Rademacher	0.325	0.301	0.268	0.203	0.107
Score Mammen	0.437	0.363	0.310	0.222	0.114
Score2 Rademacher	0.047	0.082	0.077	0.063	0.052
Score2 Mammen	0.046	0.045	0.067	0.068	0.054
			LM		
Analytical	0.143	0.088	0.054	0.037	0.039
Pairs	0.021	0.039	0.045	0.044	0.035
Score Rademacher	0.121	0.145	0.110	0.074	0.054
Score Mammen	0.096	0.053	0.049	0.040	0.034

value is less than 0.05. Stata code for our Monte Carlo experiments is available online.[14]

We consider implementations of the score bootstrap using both Rademacher weights and the skew correcting weights suggested by Mammen (1993). For comparison with the various score bootstraps we also compute the empirical rejection rates of Wald and LM tests based upon analytical clustered standard errors, the original wild bootstrap of Liu (1988), and the pairs-based block bootstrap. Following the results of Davidson and MacKinnon (1999b) on the value of "imposing the null hypothesis" on bootstrap tests, we include in our exercise a variant of the wild bootstrap studied in Cameron et al. (2008), which perturbs the restricted score contributions obtained from estimates constraining the coefficient on D_c to equal one, a procedure we term "Wild2". We also examine the performance of an analogous score bootstrap, which we term "Score2," that works with an estimator's restricted score but employs an unrestricted variance estimate. Details of the various procedures are described in the Implementation Appendix.

The standard clustered Wald test severely overrejects in samples with few clusters, with performance further degrading when the regressor of interest is generated according to a mixture distribution. Mild misspecification of the sort captured by Design III has little effect on the rejection rates of any of the procedures. A conventional pairs bootstrap of the Wald test yields dramatic improvements in size control though its performance degrades somewhat when the regressor of interest exhibits outliers. Wild bootstrapping the Wald test yields improvements over analytical methods but underperforms relative to pairs regardless of whether Mammen or Rademacher weights are used. As suggested by our theoretical results, the score bootstrap yields improvements over analytical methods but is somewhat outperformed by the wild bootstrap particularly in the skewed regressor design. Our variants of the score and wild bootstrap Wald tests that work with restricted residuals perform much better than their unrestricted counterparts. Both Wald2 and Score2 yield performance on par with Pairs even under the relatively difficult skewed regressor design.

In contrast to the Wald tests, the LM tests appear to perform well across a range of sample sizes, regardless of the distribution of the regressors. It is only in samples with very few clusters that the analytical LM test yields significant

[14] URL: http://www.bepress.com/jem/

[15] We were unable to compute the Wald and LM statistics in the majority of pairs draws with 5 clusters.

Table 2: Empirical Rejection Rates - Linear Model, Design III

	Design III: Misspecification, Normal Regressor				
	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 200$
	Wald				
Analytical	0.383	0.219	0.132	0.087	0.058
Pairs	0.013	0.046	0.042	0.040	0.046
Wild Rademacher	0.179	0.100	0.067	0.056	0.051
Wild Mammen	0.177	0.125	0.093	0.071	0.053
Wild2 Rademacher	0.062	0.060	0.056	0.052	0.050
Wild2 Mammen	0.050	0.027	0.026	0.035	0.046
Score Rademacher	0.269	0.189	0.130	0.093	0.062
Score Mammen	0.375	0.253	0.165	0.105	0.062
Score2 Rademacher	0.044	0.059	0.053	0.054	0.051
Score2 Mammen	0.043	0.037	0.058	0.061	0.052
	LM				
Analytical	0.157	0.092	0.067	0.055	0.050
Pairs	0.020	0.041	0.040	0.038	0.048
Score Rademacher	0.098	0.104	0.080	0.064	0.054
Score Mammen	0.078	0.032	0.035	0.043	0.049

overrejection. Score bootstrapping the LM statistic with Mammen weights largely removes these distortions as does application of the nonparametric pairs bootstrap.

Table 3 examines the performance of Wald and LM tests in the probit model. Here both Wald and LM tests tend to overreject when asymptotic critical values are used. Use of the pairs bootstrap corrects for this overrejection though in small samples we were sometimes unable to compute the bootstrap distribution.[16] Score bootstrapping the Wald test yields improvements over analytical clustered standard errors but substantial overrejection remains in small samples. Use of the restricted score variant of the test yields smaller improvements than were found with OLS. Score bootstrapping the LM test with Mammen weights, on the other hand, yields size control roughly on par with the pairs bootstrap.

Table 4 examines the performance of tests for intra cluster correlation of the

[16] We discarded simulations for which we were unable to compute an estimate in some bootstrap draws.

Table 3: Empirical Rejection Rates - Probit, Designs IV&V

	Design IV: Normal Regressor				
	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
			Wald		
Analytical	0.326	0.167	0.104	0.070	0.051
Pairs[15]	n.a.	0.055	0.060	0.055	0.048
Score Rademacher	0.186	0.128	0.090	0.065	0.049
Score Mammen	0.306	0.177	0.108	0.070	0.049
Score2 Rademacher	0.140	0.090	0.069	0.059	0.050
Score2 Mammen	0.159	0.096	0.070	0.057	0.050
			LM		
Analytical	0.171	0.106	0.080	0.062	0.050
Pairs	n.a.	0.083	0.082	0.064	0.053
Score Rademacher	0.081	0.079	0.066	0.057	0.049
Score Mammen	0.061	0.023	0.038	0.050	0.048
	Design V: Mixture Regressor				
	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
			Wald		
Analytical	0.308	0.170	0.100	0.065	0.058
Pairs	n.a.	0.052	0.060	0.052	0.050
Score Rademacher	0.167	0.125	0.085	0.060	0.056
Score Mammen	0.279	0.176	0.105	0.065	0.057
Score2 Rademacher	0.140	0.113	0.092	0.071	0.058
Score2 Mammen	0.160	0.121	0.092	0.070	0.058
			LM		
Analytical	0.158	0.110	0.080	0.060	0.055
Pairs	n.a.	0.076	0.081	0.066	0.053
Score Rademacher	0.075	0.082	0.069	0.057	0.054
Score Mammen	0.058	0.025	0.041	0.053	0.053

generalized residuals in the probit model, as in (41).[17] Because the information matrix equality holds under both DGPs we use the outer product version of the test

[17] A description of the implementation of this test can be found in the Implementation Appendix

Table 4: Empirical Rejection Rates - Probit, m-Test

	Normal Regressor				
	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 200$
Analytical	0.767	0.448	0.257	0.136	0.076
Score Rademacher	0.441	0.394	0.233	0.130	0.073
Score Mammen	0.370	0.237	0.223	0.128	0.069
	Mixture Regressor				
	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 200$
Analytical	0.772	0.439	0.253	0.142	0.072
Score Rademacher	0.437	0.386	0.229	0.134	0.072
Score Mammen	0.353	0.234	0.218	0.134	0.068

described in 4.3.1 generalized to allow for clustering. We see that the analytical m-test procedure overrejects substantially in small samples. Both score bootstraps partially correct this problem, though they significantly overreject as well. With 200 clusters, the analytical and bootstrap approaches appear to work equally well.

Finally, to illustrate the dramatic computational advantages of the score bootstrap relative to the wild bootstrap and pairs resampling, Table 5 presents the time elapsed in conducting bootstrap Wald tests using 9,999 bootstrap repetitions of the Score LM, Score2, Wild2, and pairs bootstrap procedures on a simulated dataset with twenty clusters. These computations were performed in Stata/SE 11.1 on a single core of a 2.3 Ghz Quad Core AMD Opteron Processor running Linux.

Table 5: COMPUTATIONAL TIME (IN SECONDS)

	Score LM	Score2	Wild2	Pairs
OLS	13	15	116	278
Probit	18	21	n.a.	30,718

For OLS, the score bootstrap yields nearly an order of magnitude improvement in computational time over the Wild bootstrap and more than a twenty fold improvement over pairs. For the probit model the results are even more striking. The score bootstrap is more than 1,000 times faster than nonparametric pairs resampling.

6 Conclusion

The score bootstrap provides a substantial computational advantage over the wild and pairs bootstraps and may easily be applied to estimators that lack conventional residuals. Our Monte Carlo experiments suggest these computational advantages come at little cost in terms of performance. Particularly when applied to LM test statistics, the score bootstrap tends to yield substantial improvements over traditional asymptotic testing procedures in small sample environments and exhibits performance comparable to more computationally expensive bootstrap procedures.

7 Appendix

7.1 Implementation Details

We provide here implementation details for the various bootstrap procedures discussed in Section 5. We restrict our discussion to the linear model and to the test of hypothesis (41), as the generalization to the probit model is straightforward but notationally intensive. Throughout, the linear model we consider is given by:

$$Y_{ic} = X_{ic}'\beta_0 + \varepsilon_{ic} \ , \tag{42}$$

where $\beta_0 \in \mathbf{R}^m$ and (Y_{ic}, X_{ic}) denotes observation i in cluster c with J the total number of observations per cluster and n the total number of clusters. We examine:

$$H_0 : R\beta_0 = r \qquad\qquad H_1 : R\beta_0 \neq r \ , \tag{43}$$

where R is a $d \times m$ matrix and r a $d \times 1$ column vector. In our Monte Carlo, R is a vector that selects the coefficient corresponding to D_c (in (37)) and $r = 1$.

7.1.1 Wald Tests

A number of bootstrap procedures we consider provide approximations to the distribution of the test statistic:

$$T_n \equiv nJ(R\hat{\beta}_u - r)'(RH_n^{-1}\Sigma_n(\hat{\beta}_u)H_n^{-1}R')^{-1}(R\hat{\beta}_u - r) \ , \tag{44}$$

where $\hat{\beta}_u$ is the unrestricted OLS estimator, $H_n \equiv \frac{1}{nJ}\sum_{c=1}^n \sum_{i=1}^J X_{ic}X_{ic}'$ and

$$\Sigma_n(\beta) \equiv \frac{1}{(n-1)} \sum_{c=1}^n \Big(\frac{1}{J}\sum_{i=1}^J \{X_{ic}e_{ic}(\beta) - \overline{Xe(\beta)}\}\Big)\Big(\frac{1}{J}\sum_{i=1}^J \{X_{ic}e_{ic}(\beta) - \overline{Xe(\beta)}\}\Big)' , \tag{45}$$

where $e_{ic}(\beta) \equiv (Y_{ic} - X_{ic}'\beta)$ and $\overline{Xe(\beta)} \equiv \frac{1}{nJ}\sum_{c=1}^n \sum_{i=1}^J X_{ic}e_{ic}(\beta)$. Tests employing analytical critical values reject at level α if T_n is larger than the $1-\alpha$ quantile of a $\chi^2(d)$ random variable. Note that in Section 5, $d = 1$ and hence the analytical critical value is approximately 3.841.

Wald - Pairs Cluster Bootstrap

Let $X_c = \{X_{ic}\}_{i=1}^J$ and $Y_c = \{Y_{ic}\}_{i=1}^J$. The pairs cluster bootstrap draws with replacement n observations from $\{Y_c, X_c\}_{c=1}^n$ – here an "observation" is an entire cluster. For $\{\tilde{Y}_c, \tilde{X}_c\}_{c=1}^n$ the bootstrap sample, we then compute the unrestricted OLS estimator $\tilde{\beta}_u$, matrix $\tilde{H}_n \equiv \frac{1}{nJ}\sum_{c=1}^n \sum_{i=1}^J \tilde{X}_{ic}\tilde{X}_{ic}'$ and:

$$\tilde{T}_n \equiv nJ(R\tilde{\beta}_u - R\hat{\beta}_u)'(R\tilde{H}_n^{-1}\tilde{\Sigma}_n(\tilde{\beta}_u)\tilde{H}_n^{-1}R')^{-1}(R\tilde{\beta}_u - R\hat{\beta}_u) , \tag{46}$$

where $\tilde{\Sigma}_n(\beta)$ is computed as $\Sigma_n(\beta)$ with $\{\tilde{Y}_c, \tilde{X}_c\}_{c=1}^n$ in place of $\{Y_c, X_c\}_{c=1}^n$. The pairs cluster bootstrap test then rejects the null hypothesis at level α if:

$$P(\tilde{T}_n \geq T_n \,|\{Y_c, X_c\}_{c=1}^n) < \alpha . \tag{47}$$

In practice, the probability in (47) can be accurately approximated by simulation of $\tilde{T}_n$. That is we: (i) Draw B samples $\{\tilde{Y}_c, \tilde{X}_c\}_{c=1}^n$ and obtain for sample b a statistic $\tilde{T}_{n,b}$; (ii) Reject the null hypothesis if the proportion of bootstrap draws $\{\tilde{T}_{n,b}\}_{b=1}^B$ that is larger than T_n is less than α.

Wald - Score Cluster Bootstrap

The score bootstraps we study are constructed employing the statistics:

$$S_n^*(\beta) \equiv \frac{1}{nJ}\sum_{c=1}^n W_c \sum_{i=1}^J X_{ic}e_{ic}(\beta) \tag{48}$$

$$\Sigma_n^{*s}(\beta) \equiv \frac{1}{(n-1)}\sum_{c=1}^n (\frac{W_c}{J}\sum_{i=1}^J \{X_{ic}e_{ic}(\beta) - \overline{Xe(\beta)}^*\})(\frac{W_c}{J}\sum_{i=1}^J \{X_{ic}e_{ic}(\beta) - \overline{Xe(\beta)}^*\})' \tag{49}$$

where $\overline{Xe(\beta)}^* \equiv \frac{1}{nJ}\sum_{c=1}^n W_c(\sum_{i=1}^J X_{ic}e_{ic}(\beta))$, and as in Section 4.4 $\{W_c\}_{c=1}^n$ is a sample of random weights satisfying $E[W_c] = 0$ and $E[W_c^2] = 1$. The score bootstrap statistics we consider are then given by:

$$T^{*s}_{n,1} \equiv (RH_n^{-1}S_n^*(\hat{\beta}_u))'(RH_n^{-1}\Sigma_n^{*s}(\hat{\beta}_u)H_n^{-1}R')^{-1}(RH_n^{-1}S_n^*(\hat{\beta}_u)) \qquad (50)$$

$$T^{*s}_{n,2} \equiv (RH_n^{-1}S_n^*(\hat{\beta}_r))'(RH_n^{-1}\Sigma_n^{*s}(\hat{\beta}_u)H_n^{-1}R')^{-1}(RH_n^{-1}S_n^*(\hat{\beta}_r)) \qquad (51)$$

for $\hat{\beta}_r$ the OLS estimate of (42) restricted to satisfy $R\hat{\beta}_r = r$. Here, $T^{*s}_{n,1}$ corresponds to "Score Wald" in Section 5 and $T^{*s}_{n,2}$ to "Score2 Wald". Both procedures reject if:

$$P(T^{*s}_{n,i} \geq T_n \, |\{Y_c, X_c\}_{c=1}^n) < \alpha \qquad (52)$$

for $i \in \{1,2\}$ and given level α. As is the case for the pairs cluster bootstrap, the probability in (52) need not be computed analytically but may be approximated through simulation. To this end: (i) Draw B samples of $\{W_c\}_{c=1}^n$ and employ each b sample to compute $T^{*s}_{n,i,b}$ for either $i \in \{1,2\}$; (ii) Reject the null hypothesis if the proportion of draws $\{T^{*s}_{n,i,b}\}_{b=1}^B$ that is larger than T_n is smaller than α.

Wald - Wild Cluster Bootstrap

The wild bootstrap procedures require the generation of new dependent variables:

$$Y^*_{ic}(\beta) \equiv X'_{ic}\beta + e_{ic}(\beta)W_c \, , \qquad (53)$$

where $\{W_c\}_{c=1}^n$ is independent of $\{Y_c, X_c\}_{c=1}^n$ and satisfy $E[W_c] = 0$ and $E[W_c^2] = 1$. For $Y_c^*(\beta) \equiv \{Y^*_{ic}(\beta)\}_{i=1}^J$, let $\hat{\beta}^*_2$ and $\hat{\beta}^*_1$ denote the OLS estimator on the bootstrap samples $\{Y_c^*(\hat{\beta}_r), X_c\}_{c=1}^n$ and $\{Y_c^*(\hat{\beta}_u), X_c\}_{c=1}^n$ respectively – i.e. $\hat{\beta}^*_2$ differs from $\hat{\beta}^*_1$ in that for the former the null hypothesis is imposed in the bootstrap distribution. Similarly, let $\Sigma^{*w}_{n,1}(\beta)$ and $\Sigma^{*w}_{n,2}(\beta)$ denote the analogues of $\Sigma_n(\beta)$ (in (45)) but respectively employing $\{Y_c^*(\hat{\beta}_u), X_c\}_{c=1}^n$ and $\{Y_c^*(\hat{\beta}_r), X_c\}_{c=1}^n$ in place of $\{Y_c, X_c\}_{c=1}^n$. The two wild bootstrap statistics we consider are then:

$$T^{*w}_{n,1} \equiv (RH_n^{-1}S_n^*(\hat{\beta}_u))'(RH_n^{-1}\Sigma^{*w}_{n,1}(\hat{\beta}^*_1)H_n^{-1}R')^{-1}(RH_n^{-1}S_n^*(\hat{\beta}_u)) \qquad (54)$$

$$T^{*w}_{n,2} \equiv (RH_n^{-1}S_n^*(\hat{\beta}_r))'(RH_n^{-1}\Sigma^{*w}_{n,2}(\hat{\beta}^*_2)H_n^{-1}R')^{-1}(RH_n^{-1}S_n^*(\hat{\beta}_r)) \, , \qquad (55)$$

where $T^{*w}_{n,1}$ corresponds to "Wild" in Section 5 and $T^{*w}_{n,2}$ to "Wild2". Both wild bootstrap procedures then reject the null hypothesis at level α whenever:

$$P(T^{*w}_{n,i} \geq T_n \, |\{Y_c, X_c\}_{c=1}^n) < \alpha \qquad (56)$$

for $i \in \{1,2\}$. As in (52), we may: (i) Draw B samples of $\{W_c\}_{c=1}^n$ and compute $T^{*w}_{n,i,b}$ for either $i \in \{1,2\}$; (ii) Reject the null hypothesis if the proportion of computed bootstrap statistics $\{T^{*w}_{n,i,b}\}_{b=1}^B$ that is larger than T_n is smaller than α.

7.1.2 Lagrange Multiplier (LM) Tests

To test the hypothesis in (43) employing a Lagrange Multiplier test, let

$$S_n(\beta) \equiv \frac{1}{nJ}\sum_{c=1}^{n}\sum_{i=1}^{J} X_{ic}e_{ic}(\beta)\ , \tag{57}$$

which is the full sample analogue of $S_n^*(\beta)$ as in (48). The LM test statistic is then:

$$L_n \equiv nJ(RH_n^{-1}S_n(\hat{\beta}_r))'(RH_n^{-1}\Sigma_n(\hat{\beta}_r)H_n^{-1}R')^{-1}(RH_n^{-1}S_n(\hat{\beta}_r))\ , \tag{58}$$

which converges in distribution to a $\chi^2(d)$ under the null hypothesis. The analytical LM test rejects at level α when L_n is larger than the $1-\alpha$ quantile of $\chi^2(d)$.

LM - Pairs Cluster Bootstrap

As in the Wald pairs cluster bootstrap, let $\{\tilde{Y}_c, \tilde{X}_c\}_{c=1}^n$ denote a bootstrap sample drawn with replacement from $\{Y_c, X_c\}_{c=1}^n$. For $\tilde{\beta}_r$ the OLS estimator on $\{\tilde{Y}_c, \tilde{X}_c\}_{c=1}^n$ restricted to satisfy $R\tilde{\beta}_r = r$, $\tilde{H}_n \equiv \frac{1}{nJ}\sum_{c=1}^n\sum_{i=1}^J \tilde{X}_{ic}\tilde{X}_{ic}'$ and $\tilde{\Sigma}_n(\beta)$ and $\tilde{S}_n(\beta)$ the analogues of $\Sigma_n(\beta)$ (in (45)) and $S_n(\beta)$ (in (57)) but employing $\{\tilde{Y}_c, \tilde{X}_c\}_{c=1}^n$ in place of $\{Y_c, X_c\}_{c=1}^n$, define the bootstrap LM statistic:

$$\tilde{L}_n \equiv nJ\tilde{\Gamma}_n'(R\tilde{H}_n^{-1}\tilde{\Sigma}_n(\tilde{\beta}_r)\tilde{H}_n^{-1}R')^{-1}\tilde{\Gamma}_n\ , \tag{59}$$

where $\tilde{\Gamma}_n \equiv R(\tilde{H}_n^{-1}\tilde{S}_n(\tilde{\beta}_r) - H_n^{-1}S_n(\hat{\beta}_r))$. The pairs cluster bootstrap for the Lagrange multiplier test then rejects at level α if:

$$P(\tilde{L}_n \geq L_n \,|\{Y_c, X_c\}_{c=1}^n) < \alpha\ . \tag{60}$$

In practice, we: (i) Draw B samples with replacement $\{\tilde{Y}_c, \tilde{X}_c\}_{c=1}^n$ and for each sample b compute the bootstrap statistic $\tilde{L}_{n,b}$; (ii) Reject the null hypothesis if the proportion of bootstrap statistics $\{\tilde{L}_{n,b}\}_{b=1}^B$ larger than L_n is smaller than α.

LM - Score Cluster Bootstrap

Given a sample $\{W_c\}_{c=1}^n$ of weights, independent of $\{Y_c, X_c\}_{c=1}^n$ satisfying $E[W_c] = 0$ and $E[W_c^2] = 1$, and for $S_n^*(\beta)$ as in (48) and $\Sigma_n^{*s}(\beta)$ as in (49), define:

$$L_n^* \equiv nJ(RH_n^{-1}S_n^*(\hat{\beta}_r))'(RH_n^{-1}\Sigma_n^{*s}(\hat{\beta}_r)H_n^{-1}R')^{-1}(RH_n^{-1}S_n^*(\hat{\beta}_r))\ . \tag{61}$$

The score bootstrap Lagrange Multiplier procedure then rejects at level α whenever:

$$P(L_n^* \geq L_n \,|\{Y_c, X_c\}_{c=1}^n) < \alpha\ . \tag{62}$$

We approximate this decision by: (i) Drawing B samples $\{W_c\}_{c=1}^n$ and employing each sample b to obtain a score bootstrap LM statistics $L_{n,b}^*$; (ii) Rejecting the null hypothesis if the proportion of $\{L_{n,b}^*\}_{b=1}^B$ larger than L_n is smaller than α.

7.1.3 Intra-Cluster Correlation Test (Probit)

The score bootstrap examined in Table 3 follows the discussion of Section 4.3.1.
STEP 1 Obtain a probit estimate of model (40) and let $\hat{\theta}_M = (\hat{\theta}_0, \hat{\theta}_X, \hat{\theta}_D)$ where $\hat{\theta}_X$ is the estimated coefficient for X_{ic} and $\hat{\theta}_D$ for D_c and $\hat{\theta}_0$ the intercept. As in Section 4.4 also let $Z_c \equiv \{Y_{ic}, X_{ic}, D_c\}_{i=1}^{20}$.
STEP 2: Employing the probit estimate, construct the fitted prediction $\hat{p}_{ic} \equiv \Phi(\hat{\theta}_0 + X_{ic}\hat{\theta}_X + D_c\hat{\theta}_D)$ and let $\hat{v}_{ic} = (Y_{ic} - \hat{p}_{ic})/\sqrt{\hat{p}_{ic}(1-\hat{p}_{ic})}$.
STEP 3: Following the notation of Section 4.3.1, we may then define the statistics:

$$m(Z_c, \hat{\theta}_M) \equiv \sum_{i=1}^{20} \hat{v}_{ic} \sum_{j \neq i} \hat{v}_{jc} \quad \nabla q(Z_c, \hat{\theta}_M) \equiv \sum_{i=1}^{20} \frac{(Y_{ic} - \hat{p}_{ic})\phi(\hat{\theta}_0 + X_{ic}\hat{\theta}_X + D_c\hat{\theta}_D)}{\hat{p}_{ic}(1-\hat{p}_{ic})} X_{ic} \tag{63}$$

and obtain the cluster level quantities $\{m(Z_c, \hat{\theta}_M)\}_{c=1}^n$ and $\{\nabla q(Z_c, \hat{\theta}_M)\}_{c=1}^n$.
STEP 4: As in Section 4.3.1, (i) Regress $\{m(Z_c, \hat{\theta}_M)\}_{c=1}^n$ on $\{\nabla q(Z_c, \hat{\theta}_M)\}_{c=1}^n$ and obtain residuals $\{e_c\}_{c=1}^n$; (ii) Perturb the cluster level residuals to obtain $\{W_c e_c\}_{c=1}^n$; (iii) For M_n and M_n^* the Wald test statistic for $E[e_c] = 0$ and $E[e_c W_c] = 0$ respectively, the bootstrap test then rejects the null hypothesis in (41) if:

$$P(M_n \geq M_n^* \mid \{Z_c\}_{c=1}^n) < \alpha \ . \tag{64}$$

7.2 Proofs of Main Results

PROOF OF LEMMA 3.1: First note that by Markov's inequality and $E[W_i^2] = 1$:

$$P^*(\|\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \varepsilon_i^*\| > C) \leq \frac{1}{nC^2} E^*[(\sum_{i=1}^n X_i \varepsilon_i^*)'(\sum_{i=1}^n X_i \varepsilon_i^*)] = \frac{1}{nC^2} \sum_{i=1}^n X_i' X_i e_i^2 \ . \tag{65}$$

Let $H \equiv E[X_i X_i']$ and $\Sigma \equiv E[X_i X_i' \varepsilon_i^2]$. Hence, since $n^{-1}\sum_i X_i' X_i e_i^2 \overset{a.s.}{\to} \Sigma$ and $H_n \overset{a.s.}{\to} H$,

$$\|\sqrt{n}(\hat{\beta}^* - \hat{\beta})\| \leq \|H_n^{-1}\|_F \times \|\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \varepsilon_i^*\| = O_{p^*}(1) \qquad a.s. \ , \tag{66}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Also, for $\|\cdot\|_o$ the operator norm:

$$\|(H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1})^{-1} - (H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1})^{-1}\|_o \leq \|(H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1})^{-1}\|_o \\ \times \|H_n^{-1}(\Sigma_n^*(\hat{\beta}) - \Sigma_n^*(\hat{\beta}^*))H_n^{-1}\|_o \times \|(H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1})^{-1}\|_o \, . \quad (67)$$

Let $X_i^{(k)}$ denote the k^{th} element of the vector X_i. Arguing as in (65), it is straightforward to show $n^{-\frac{1}{2}}\sum_i X_i^{(k)}X_i^{(l)}X_i^{(s)}\varepsilon_i^* = O_{p^*}(1)$ almost surely for any indices k,l,s. Therefore, since $\|\cdot\|_o \leq \|\cdot\|_F$ and $E[\|X_i\|^4] < \infty$, we conclude from (66) that:

$$\|\Sigma_n^*(\hat{\beta}) - \Sigma_n^*(\hat{\beta}^*)\|_o \leq \|\frac{1}{n}\sum_{i=1}^n X_iX_i'\{(Y_i^* - X_i'\hat{\beta})^2 - (Y_i^* - X_i'\hat{\beta}^*)^2\}\|_F \\ = \|\frac{1}{n}\sum_{i=1}^n X_iX_i'\{2\varepsilon_i^*(X_i'(\hat{\beta} - \hat{\beta}^*)) + (X_i'(\hat{\beta} - \hat{\beta}^*))^2\}\|_F = O_{p^*}(n^{-1}) \qquad a.s. \quad (68)$$

Moreover, since $E[(\varepsilon_i^*)^k] = E[W_i^k]e_i^k$, we also obtain from $E[\|X_i\|^4\varepsilon_i^4] < \infty$ that:

$$E^*[\|\frac{1}{n}\sum_{i=1}^n X_iX_i'\{(\varepsilon_i^*)^2 - e_i^2\}\|_F^2] = \sum_{l=1}^m\sum_{s=1}^m E^*[(\frac{1}{n}\sum_{i=1}^n X_i^{(l)}X_i^{(s)}\{(\varepsilon_i^*)^2 - e_i^2\})^2] \\ = \frac{1}{n}\sum_{l=1}^m\sum_{s=1}^m \frac{1}{n}\sum_{i=1}^n (X_i^{(l)}X_i^{(s)})^2\{(E[W_i^4] - 1)e_i^4\} = o_{a.s.}(1) \, . \quad (69)$$

Hence, since $n^{-1}\sum_i X_iX_i'e_i^2 \overset{a.s.}{\to} \Sigma$ and $H_n^{-1} \overset{a.s.}{\to} H^{-1}$, results (68) and (69) imply:

$$\|H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1} - H^{-1}\Sigma H^{-1}\|_F = o_{p^*}(1) \\ \|H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1} - H^{-1}\Sigma H^{-1}]\|_F = o_{p^*}(1) \quad (70)$$

almost surely. Next, for any normal matrix A, let $\xi(A)$ denote its smallest eigenvalue. Since Corollary III.2.6 in Bhatia (1997) implies $|\xi(A) - \xi(B)| \leq \|A - B\|_F$, it then follows from result (70) that:

$$\xi(H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1}) = \xi(H^{-1}\Sigma H^{-1}) + o_{p^*}(1) \\ \xi(H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1}) = \xi(H^{-1}\Sigma H^{-1}) + o_{p^*}(1) \quad (71)$$

almost surely. However, since for any normal matrix A, $\|A^{-1}\|_o = 1/\xi(A)$, result (71) and Assumption 3.1(i) imply $\|H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1}\|_o = O_{p^*}(1)$ and in addition $\|H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1}\|_o = O_{p^*}(1)$ almost surely. Hence,

$$\|(H_n^{-1}\hat{\Sigma}_n^*(\hat{\beta}^*)H_n^{-1})^{-1} - (H_n^{-1}\hat{\Sigma}_n^*(\hat{\beta})H_n^{-1})^{-1}\|_o = O_{p^*}(n^{-1}) \qquad a.s. \quad (72)$$

as a result of (67) and (68). In turn, (72) yields that $\|(H_n^{-1}\hat{\Sigma}_n^*(\hat{\beta}^*)H_n^{-1})^{-\frac{1}{2}} - (H_n^{-1}\hat{\Sigma}_n^*(\hat{\beta})H_n^{-1})^{-\frac{1}{2}}\|_o = O_{p^*}(n^{-1})$ almost surely, and the claim of the Lemma then follows by result (66). □

Lemma 7.1. *Let $\{W_i\}_{i=1}^n$ be i.i.d. independent of $\{Z_i\}_{i=1}^n$ satisfying $E[W_i^2] = 1$. If Assumptions 4.1, 4.3(i) and 4.4(ii) hold, then $\mathscr{F} = \{s(z,\theta)s(z,\theta)'w^2 : \theta \in \Theta\}$ is Glivenko-Cantelli.*

PROOF: By Assumption 4.4(ii), $s(z,\theta)w$ is continuous in $\theta \in \Theta$, and hence so is $s(z,\theta)s(z,\theta)'w^2$. Let $s^{(l)}(z,\theta)$ be the l^{th} component of the vector $s(z,\theta)$. By the mean value theorem and Assumption 4.4(ii):

$$|s^{(l)}(z,\theta)| \leq |s^{(l)}(z,\theta) - s^{(l)}(z,\theta_0)| + |s^{(l)}(z,\theta_0)| \leq F(z)\|\theta - \theta_0\| + |s^{(l)}(z,\theta_0)| \ . \tag{73}$$

Hence, for $D = \text{diam}(\Theta)$ it then follows that $|s^{(l)}(z,\theta)s^{(j)}(z,\theta)w^2| \leq w^2(F(z)D + |s^{(l)}(z,\theta_0)|)(F(z)D + |s^{(j)}(z,\theta_0)|)$, which is integrable for all $1 \leq l \leq j \leq k$ by Assumption 4.1(i)-(ii) and 4.4(ii). We conclude that $\mathscr{F}$ has an integrable envelope, and the Lemma follows by Example 19.8 in van der Vaart (1999). □

Lemma 7.2. *If Assumptions 4.1(i), 4.3(i) and 4.4(i)-(ii) hold, then the class $\mathscr{F} \equiv \{ws(z,\theta) : \theta \in \Theta\}$ is Donsker.*

PROOF: Let $\|\cdot\|_o$ and $\|\cdot\|_F$ denote the operator and Frobenious norms respectively. Using $\|\cdot\|_o \leq \|\cdot\|_F$, Assumption 4.4(ii) and the mean value theorem, we obtain that for some $\bar{\theta}$ a convex combination of θ_1 and θ_2:

$$\begin{aligned} \|ws(z,\theta_1) - ws(z,\theta_2)\| &= |w| \times \|\nabla s(z,\bar{\theta})(\theta_1 - \theta_2)\| \\ &\leq |w| \times \|\nabla s(z,\bar{\theta})\|_o \times \|\theta_1 - \theta_2\| \leq |w| \times F(z) \times \|\theta_1 - \theta_2\| \ . \end{aligned} \tag{74}$$

Hence, the class $\mathscr{F}$ is Lipschitz in $\theta \in \Theta$. For $s^{(l)}(z,\theta)$ the l^{th} component of the vector $s(z,\theta)$, let $\mathscr{F}^l \equiv \{ws^{(l)}(z,\theta) : \theta \in \Theta\}$ and note that Theorem 2.7.11 in van der Vaart and Wellner (1996) implies:

$$N_{[\,]}(2\varepsilon\|\tilde{F}\|_{L^2}, \mathscr{F}^l, \|\cdot\|_{L^2}) \leq N(\varepsilon, \Theta, \|\cdot\|) \ , \tag{75}$$

where $\tilde{F}(w,z) = |w|F(z)$. Let $D = \text{diam}(\Theta)$ and $M^2 = E[\tilde{F}^2(W_i,Z_i)]$ and notice that Assumptions 4.4(i)-(ii) imply $M < \infty$. Since by (74), the diameter of $\mathscr{F}^l$ under $\|\cdot\|_{L^2}$ is less than or equal to MD, we then obtain:

$$\begin{aligned} \int_0^\infty \sqrt{\log N_{[\,]}(\varepsilon, \mathscr{F}^l, \|\cdot\|_{L^2})}d\varepsilon &= 2M\int_0^{\frac{D}{2}} \sqrt{\log N_{[\,]}(2Mu, \mathscr{F}^l, \|\cdot\|_{L^2})}du \\ &\leq 2M\int_0^{\frac{D}{2}} \sqrt{\log N(u, \Theta, \|\cdot\|)}du \leq 2M\int_0^{\frac{D}{2}} \sqrt{p\log(D/u)}du < \infty \ , \end{aligned} \tag{76}$$

where the equality follows by the change of variables $u = \varepsilon/2M$, the first inequality from (75) and the second by $N(u, \Theta, \|\cdot\|) \leq (\text{diam}(\Theta)/u)^p$. Since $E[\tilde{F}^2(Z_i, W_i)] < \infty$, (76) and Theorem 2.5.6 in van der Vaart and Wellner (1996) implies $\mathscr{F}^l$ is Donsker for $1 \leq l \leq k$, and the claim of the Lemma follows. □

Lemma 7.3. *Suppose Assumptions 4.1, 4.2, 4.3 and 4.4(ii) hold. If the null hypothesis is true, it then follows that* $G_n \xrightarrow{L} \chi^2(r)$*. On the other hand, if the alternative hypothesis is true, then* $G_n \xrightarrow{p} \infty$*.*

PROOF: We first study the limiting behavior of G_n under the null hypothesis. For this purpose, notice that Assumption 4.3(ii) implies that $A_n(\theta_0) = A(\theta_0) + o_p(1)$, while Lemma 7.1 applied to $W_i = 1$ with probability one yields $\Sigma_n(\theta_0) = \Sigma(\theta_0) + o_p(1)$ for $\Sigma(\theta_0) = E[s(Z_i, \theta_0)s(Z_i, \theta_0)']$. Therefore, we conclude:

$$A_n(\theta_0)\Sigma_n(\theta_0)A_n(\theta_0)' = A(\theta_0)\Sigma(\theta_0)A(\theta_0)' + o_p(1) \; . \qquad (77)$$

It follows that $A_n(\theta_0)\Sigma_n(\theta_0)A_n(\theta_0)'$ is then invertible with probability tending to one by Assumption 4.1(ii). Therefore, by Assumptions 4.2(i), 4.1(ii), 4.3(i) and the central limit theorem we conclude that:

$$T_n = (A_n(\theta_0)\Sigma_n(\theta_0)A_n(\theta_0)')^{-1}A_n(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i, \theta_0) + o_p(1) \xrightarrow{L} N(0, I) \; . \quad (78)$$

By the continuous mapping theorem and (78), $G_n \xrightarrow{L} \chi^2(r)$, establishing the first claim. The second claim of the Lemma was assumed in Assumption 4.2(ii). □
PROOF OF THEOREM 4.1: Let $\Sigma(\theta) = E[s(Z_i, \theta)s(Z_i, \theta)']$. As argued in (73), $s(z, \theta)s(z, \theta)'$ has an integrable envelope, which together with Assumption 4.4(ii) and the dominated convergence theorem imply $\Sigma(\theta)$ is continuous in θ. Therefore, by Lemma 7.1 and Assumption 4.1(i), we obtain $\Sigma_n^*(\hat{\theta}) = \Sigma(\theta_0) + o_p(1)$. In addition, $A_n(\hat{\theta}) = A(\theta_0) + o_p(1)$ by Assumption 4.3(ii) and hence Assumption 4.1(ii), and $\sup_{\theta \in \Theta} \|n^{-\frac{1}{2}} \sum_i s(Z_i, \theta)W_i\| = O_p(1)$ by Lemma 7.2 imply:

$$\begin{aligned}
&(A_n(\hat{\theta})\Sigma_n^*(\hat{\theta})A_n(\hat{\theta})')^{-\frac{1}{2}}A_n(\hat{\theta})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i, \hat{\theta})W_i \\
&\qquad = (A(\theta_0)\Sigma(\theta_0)A(\theta_0)')^{-\frac{1}{2}}A(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i, \hat{\theta})W_i + o_p(1) \\
&\qquad = (A(\theta_0)\Sigma(\theta_0)A(\theta_0)')^{-\frac{1}{2}}A(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i, \theta_0)W_i + o_p(1) \; , \quad (79)
\end{aligned}$$

where the second equality follows by Assumption 4.1(i) and Lemma 7.2. Let BL_c be the set of Lipschitz real valued functions whose Lipschitz constant and level are less than c. For two random variables Y, V:

$$\|Y - V\|_{BL_1} \equiv \sup_{f \in BL_1} |E[f(Y)] - E[f(V)]| \ , \tag{80}$$

metrizes weak convergence between the distributions of Y and V; see for example Theorem 1.12.4 in van der Vaart and Wellner (1996). Next, define:

$$\bar{T}_n^* \equiv (A(\theta_0)\Sigma(\theta_0)A(\theta_0)')^{-\frac{1}{2}} A(\theta_0) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(Z_i, \theta_0) W_i \ . \tag{81}$$

Using that all $f \in BL_1$ are bounded in level and Lipschitz constant by one,

$$\sup_{f \in BL_1} |E[f(\bar{T}_n^*) - f(T_n^*)|\{Z_i\}_{i=1}^n]| \\ \leq \eta P(|\bar{T}_n^* - T_n^*| \leq \eta |\{Z_i\}_{i=1}^n) + 2P(|\bar{T}_n^* - T_n^*| > \eta |\{Z_i\}_{i=1}^n) \ , \tag{82}$$

for any $\eta > 0$. However, by the law of iterated expectations and (79), we have that $P(|\bar{T}_n^* - T_n^*| > \eta |\{Z_i\}_{i=1}^n)$ converges to zero in mean, and hence in probability. As a result, since η is arbitrary, result (82) in fact implies:

$$\sup_{f \in BL_1} |E[f(\bar{T}_n^*)|\{Z_i\}_{i=1}^n] - E[f(T_n^*)|\{Z_i\}_{i=1}^n]| = o_p(1) \ . \tag{83}$$

Let $T_\infty^* \sim N(0, I)$. Since $\|\cdot\|_{BL_1}$ metrizes weak convergence, Assumptions 4.3(i), 4.4(i) and Lemma 2.9.5 in van der Vaart and Wellner (1996) imply:

$$\sup_{f \in BL_1} |E[f(\bar{T}_n^*)|\{Z_i\}_{i=1}^n] - E[f(T_\infty^*)]| = o_p(1) \ . \tag{84}$$

For any $M > 0$, define $g_M : \mathbf{R}^r \to \mathbf{R}$ by $g_M(a) = \min\{a'a, M\}$ and notice that for any $a, b \in \mathbf{R}^r$ we have $|g_M(a) - g_M(b)| \leq 2\sqrt{M}\|a - b\|$ and $g_M(a) \leq M$ so that for $M \geq 4$ we have $g_M \in BL_M$. As a result, for any $f \in BL_1$, $f \circ g_M \in BL_M$ and $M^{-1} f \circ g_M \in BL_1$, which by (83) and (84) implies:

$$\sup_{f \in BL_1} |E[f(g_M(T_n^*))|\{Z_i\}_{i=1}^n] - E[f(g_M(T_\infty^*))]| \\ \leq M \sup_{f \in BL_1} |E[f(T_n^*)|\{Z_i\}_{i=1}^n] - E[f(T_\infty^*)]| = o_p(1) \ . \tag{85}$$

Moreover, since $G_n^* = T_n^{*\prime} T_n^*$ and every $f \in BL_1$ is bounded by one,

$$\sup_{f \in BL_1} |E[f(G_n^*) - f(g_M(T_n^*)) | \{Z_i\}_{i=1}^n]| \le 2P(T_n^{*\prime} T_n^* > M | \{Z_i\}_{i=1}^n) \ . \tag{86}$$

By (79) and the continuous mapping theorem, $T_n^{*\prime} T_n^* \xrightarrow{L} \chi^2(r)$ unconditionally and hence is asymptotically tight. For an arbitrary $\eta > 0$ it then follows by Markov's inequality that for M sufficiently large:

$$\limsup_{n \to \infty} P(2P(T_n^{*\prime} T_n^* > M | \{Z_i\}_{i=1}^n) > \eta) \le \limsup_{n \to \infty} \frac{2}{\eta} P(T_n^{*\prime} T_n^* > M) < \eta \ . \tag{87}$$

Similarly, let $G_\infty^* \sim \chi^2(r)$ and notice that for M appropriately large we obtain:

$$\sup_{f \in BL_1} |E[f(G_\infty^*) - f(g_M(T_\infty^*))]| \le 2P(T_\infty^{*\prime} T_\infty^* > M) < \eta \ . \tag{88}$$

Since η is arbitrary, (85), (86), (87) and (88) in turn allow us to conclude that:

$$\sup_{f \in BL_1} |E[f(G_n^*) | \{Z_i\}_{i=1}^n] - E[f(G_\infty^*)]| = o_p(1) \ , \tag{89}$$

which establishes the weak convergence of the distribution of G_n^* conditional on $\{Z_i\}_{i=1}^n$ to that of G_∞^* in probability. Letting F be the cdf of G_∞^*, we obtain by the Portmanteau theorem, G_∞^* having a continuous distribution, result (89) and Lemma 7.3 that for any $c \in \mathbf{R}$, $F_n^*(c) = F(c) + o_p(1)$ and $F_n(c) = F(c) + o(1)$. Moreover, convergence is uniform in $c \in \mathbf{R}$ by Lemma 2.11 in van der Vaart (1999). □

Lemma 7.4. *Let $F_n : \mathbf{R} \to [0,1]$, $F : \mathbf{R} \to [0,1]$ be monotonic with probability one, and satisfying $\sup_{c \in \mathbf{R}} |F_n(c) - F(c)| = o_p(1)$. Define:*

$$c_\alpha \equiv \inf\{c : F(c) \ge \alpha\} \qquad c_{n,\alpha} \equiv \inf\{c : F_n(c) \ge \alpha\} \ .$$

If F is strictly increasing at c_α, it then follows that $c_{n,\alpha} = c_\alpha + o_p(1)$.

PROOF: Fix $\varepsilon > 0$. Since by hypothesis F is strictly increasing at c_α, we obtain:

$$F(c_\alpha - \varepsilon) < \alpha < F(c_\alpha + \varepsilon) \ . \tag{90}$$

Moreover, since $F_n(c_\alpha + \varepsilon) > \alpha$ implies that $c_{n,\alpha} \le c_\alpha + \varepsilon$ and similarly $F_n(c_\alpha - \varepsilon) < \alpha$ implies that $c_{n,\alpha} > c_\alpha - \varepsilon$, we conclude

$$\lim_{n \to \infty} P(|c_\alpha - c_{n,\alpha}| \le \varepsilon) \ge \lim_{n \to \infty} P(F_n(c_\alpha - \varepsilon) < \alpha < F_n(c_\alpha + \varepsilon)) = 1 \ , \tag{91}$$

where the final equality follows from (90) and $\sup_c |F_n(c) - F(c)| = o_p(1)$. □

PROOF OF COROLLARY 4.1: Let F denote the cdf of a $\chi^2(r)$ random variable and $c_{1-\alpha}$ be its $1-\alpha$ quantile. As argued following (89), $\sup_c |F_n^*(c) - F(c)| = o_p(1)$, and hence Lemma 7.4 implies $\hat{c}_{1-\alpha} = c_{1-\alpha} + o_p(1)$ provided $0 < \alpha < 1$. The first claim then follows by Lemma 7.3 and the continuous mapping theorem.

For the second claim of the Corollary, observe that the bootstrap statistic $S_n^*(\hat{\theta})$ remains properly centered. In fact, (89) was established without appealing to Assumption 4.2(i). Therefore, $\hat{c}_{1-\alpha} = c_{1-\alpha} + o_p(1)$ under the alternative hypothesis as well. However, under the alternative hypothesis $G_n \xrightarrow{p} \infty$ by Lemma 7.3 and therefore the second claim of the Corollary follows. □

PROOF OF COROLLARY 4.2: This is a special case of Corollary 4.1. □

References

Andrews, D. W. K. (2002): "Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators," *Econometrica*, 70, 119–162.

Barbe, P. and P. Bertail (1995): *The Weighted Bootstrap*, New York: Springer-Verlag.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences in differences estimates?" *Quarterly Journal of Economics*, 119, 249–275.

Bhatia, R. (1997): *Matrix Analysis*, New York: Springer.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90, 414–427.

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics - Methods and Applications*, New York: Cambridge University Press.

Card, D. and D. R. Hyslop (2005): "Estimating the effects of a time-limited earnings subsidy for welfare-leavers," *Econometrica*, 73, 1723–1770.

Cavaliere, G. and A. M. R. Taylor (2008): "Bootstrap unit root tests for time series with nonstationary volatility," *Econometric Theory*, 24, 43–71.

Chatterjee, S. and A. Bose (2005): “Generalized bootstrap for estimating equations,” *Annals of Statistics*, 33, 414–436.

Chesher, A. (1984): “Testing for neglected heterogeneity,” *Econometrica*, 52, 865–872.

Chesher, A. (1995): “A mirror image invariance for m-estimators,” *Econometrica*, 63, 207–211.

Davidson, R. and E. Flachaire (2008): “The wild bootstrap, tamed at last,” *Journal of Econometrics*, 146, 162–169.

Davidson, R. and J. G. MacKinnon (1999a): “Bootstrap testing in nonlinear models,” *International Economic Review*, 40, 487–508.

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

Davidson, R. and J. G. MacKinnon (2004): *Econometric Theory and Methods*, New York, Oxford: Oxford University Press.

Davidson, R. and J. G. MacKinnon (2010): “Wild bootstrap tests for iv regression,” *Journal of Business and Economic Statistics*, 28, 128–144.

Donald, S. G. and K. Lang (2007): “Inference with differences-in-differences and other panel data,” *Review of Economics and Statistics*, 89, 221233.

Efron, B. (1979): “Bootstrap methods: Another look at the jacknife,” *The Annals of Statistics*, 7, 1–26.

Freedman, D. A. (1981): “Bootstrapping regression models,” *The Annals of Statistics*, 9, 1218–1228.

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

Hardle, W. and E. Mammen (1993): “Comparing nonparametric versus parametric regression fits,” *The Annals of Statistics*, 21, 1926–1947.

Horowitz, J. L. (1994): “Bootstrap-based critical values for the information matrix test,” *Journal of Econometrics*, 61, 395–411.

Horowitz, J. L. (1997): “Bootstrap methods in econometrics: Theory and numerical performance,” in D. M. Kreps and K. F. Wallis, eds., *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress*, volume 3, Cambridge University Press.

Horowitz, J. L. (2001): “The bootstrap,” in J. J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, volume 5, Elsevier, chapter 52.

Hu, F. and J. D. Kalbfleisch (2000): “The estimating function bootstrap,” *The Canadian Journal of Statistics*, 28, 449–481.

Hu, F. and J. V. Zidek (1995): “A bootstrap based on the estimating equations of the linear model,” *Biometrika*, 82, 263–275.

Kaido, H. and A. Santos (2011): “Asymptotically efficient estimation of models defined by convex moment inequalities,” Working paper, University of California - San Diego.

Kline, P. and A. Santos (2011): “Higher order properties of the wild bootstrap under misspecification,” Working paper, University of California - Berkeley.

Liu, R. Y. (1988): “Bootstrap procedures under some non-i.i.d. models,” *The Annals of Statistics*, 16, 1696–1708.

Ma, S. and M. R. Kosorok (2005): “Robust semiparametric m-estimation and the weighted bootstrap,” *Journal of Multivariate Analysis*, 96, 190–217.

Mammen, E. (1993): “Bootstrap and wild bootstrap for high dimensional linear models,” *The Annals of Statistics*, 21, 255–285.

McCall, B. P. (1994): “Specification diagnostics for duration models: A martingale approach,” *Journal of Econometrics*, 60, 293–312.

Newey, W. K. (1985a): “Generalized method of moments specifcation testing,” *Journal of Econometrics*, 29, 229–256.

Newey, W. K. (1985b): “Maximum likelihood specification testing and conditional moment tests,” *Econometrica*, 53, 1047–1070.

Newey, W. K. and D. L. McFadden (1994): “Large sample estimation and hypothesis testing,” in R. F. Engle and D. L. McFadden, eds., *Handbook of Econometrics*, volume IV, Elsevier Science B.V., 2113–2245.

Tauchen, G. (1985): “Diagnostic testing and evaluation of maximum likelihood models,” *Journal of Econometrics*, 30, 415–443.

van der Vaart, A. (1999): *Asymptotic Statistics*, New York: Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: with Applications to Statistics*, New York: Springer.

White, H. (1982): “Maximum likelihood estimation of misspecified models,” *Econometrica*, 50, 1–25.

White, H. (1994): *Estimation, Inference, and Specification Analysis*, New York: Cambridge University Press.

Wooldridge, J. (2003): “Jacknife, bootstrap, and other resampling methods in regression analysis,” *American Economic Review*, 93, 133–138.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge: The MIT Press.

Wu, C. F. J. (1986): “Jacknife, bootstrap, and other resampling methods in regression analysis,” *Annals of Statistics*, 14, 1261–1295.

You, J. and G. Chen (2006): “Wild bootstrap estimation in partially linear models with heteroscedasticity,” *Statistics and Probability Letters*, 76, 340–348.