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

December, 2010

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 easier to implement than the conventional 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, Graham Elliott and Michael Jansson for useful comments.

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 scores 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. For example, computational problems often arise in small samples even in simple probit or logit models where, for some bootstrap draws, the estimator cannot be computed.

The score bootstrap is closely related to several existing bootstrap procedures in the literature. Most notably, it bears a close relationship to the k-step bootstrap procedure of Davidson and MacKinnon (1999) which involves taking a finite number of Newton steps towards optimization of an M-estimator in a bootstrap sample. This procedure was subsequently studied by Andrews (2002) and shown to yield an Edgeworth refinement depending on the number of optimization steps taken. Like the pairs bootstrap however, the k-step procedure may be difficult to compute if, in some bootstrap samples, the Hessian is poorly defined or of less than full rank, problems which the score bootstrap avoids. Hu and Zidek (1995) propose resampling sums of scores in the linear model conditional on a fixed Hessian using pairs resampling. This is similar in spirit to our approach

however results are not available for the properties of this bootstrap when applied to nonlinear models. Finally, the generalized bootstrap of Chatterjee and Bose (2005) perturbs the objective function of an M-estimator with i.i.d. weights rather than simply the scores. 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.

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. We find that variants of our proposed score based bootstrap substantially outperform analytical cluster robust methods. In line with our theoretical results, we find negligible differences in the performance of the score and wild bootstraps 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:

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

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 $\epsilon_i^* \equiv W_i e_i$ 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 $E[W_i^2] = 1$. Under these conditions,

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

and 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.[1]

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

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

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.

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:

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

where $H_n \equiv n^{-1} \sum_i X_i X_i'$. Since both the true and bootstrap scores are properly centered, both expressions in (4) can be expected to converge to a normal limit. Therefore, consistency of the wild bootstrap hinges on whether this limit is the same or, equivalently, whether the asymptotic variances agree. However, as $E[W_i^2] = 1$ and $\{W_i\}_{i=1}^n$ is independent of $\{Y_i, X_i\}_{i=1}^n$, we obtain:

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

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

and hence the second moments indeed agree asymptotically by standard arguments. 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 scores 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

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

match higher moments of the statistics. 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)). Alternatively, the Rademacher distribution,[2] 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. An alternative interpretation, however, is that it creates a set of bootstrap scores mimicking the heteroscedastic nature of the true scores. In this section, we develop the implications of this observation, which provides the basis for our proposed procedure.

The relationship between the wild bootstrap and the scores is transparent from the discussion of its consistency in Section 2. Since $\epsilon_i^* = e_i W_i$, we learn from (4) that the wild bootstrap may be interpreted as a perturbation of the scores $(X_i(Y_i - X_i'\beta))$ evaluated at the estimated parameter value $(\hat{\beta})$ that leaves the Hessian $(\frac{1}{n}\sum_i X_i X_i')$ unchanged.[3] More precisely, a numerically equivalent way to implement the wild bootstrap would be to employ the following algorithm:

Step 1: Obtain the OLS estimate $\hat{\beta}$ and generate the fitted scores $\{X_i(Y_i - X_i'\hat{\beta})\}_{i=1}^n$.

Step 2: Using random weights $\{W_i\}_{i=1}^n$ independent of $\{Y_i, X_i\}_{i=1}^n$ and satisfying $E[W_i] = 0$ and $E[W_i^2] = 1$, perturb the original fitted scores to obtain a new set of scores $\{X_i(Y_i - X_i'\hat{\beta})W_i\}_{i=1}^n$.

Step 3: Multiply the perturbed scores by the original Hessian to obtain $H_n^{-1}\frac{1}{\sqrt{n}}\sum_i(Y_i - X_i\hat{\beta})X_i W_i$ 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 more complex nonlinear models. One may simply perturb the fitted scores of such a model while keeping the Hessian unchanged and, provided the perturbations satisfy $E[W_i] = 0$ and $E[W_i^2] = 1$, the first two moments of the perturbed and true scores 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 approach a "score bootstrap."

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

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

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 simply employ the sample variance of the perturbed scores for studentization. For this reason, we define the bootstrap statistics:

$$T^*_{w,n} \equiv (H_n^{-1}\Sigma^*_n(\hat{\beta}^*)H_n^{-1})^{-\frac{1}{2}}\sqrt{n}(\hat{\beta}^* - \hat{\beta}) \qquad T^*_{s,n} \equiv (H_n^{-1}\Sigma^*_n(\hat{\beta})H_n^{-1})^{-\frac{1}{2}}H_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\epsilon_i^* \ , \quad (7)$$

where $\Sigma^*_n(\beta) \equiv \frac{1}{n}\sum_i X_iX_i'(Y_i^* - X_i'\beta)^2$ and $T^*_{w,n}$ and $T^*_{s,n}$ are the studentized wild and score bootstrap statistics respectively. It is important to note that in the computation of $T^*_{s,n}$, the full sample estimator $\hat{\beta}$ is used in obtaining the standard errors, and hence calculation of $\hat{\beta}^*$ and its implied residuals remains 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 this difference is of importance and, in particular, whether the refinement of the wild bootstrap over a normal approximation (Liu (1988)) is lost due to this discrepancy. Somewhat surprisingly, the answer is negative. 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 3.1. *(i)* $\{Y_i, X_i\}_{i=1}^n$ *are i.i.d.* $E[X_i\epsilon_i] = 0$*, and* $E[X_iX_i']$*,* $E[X_iX_i'\epsilon_i^2]$ *are full rank; (ii) The moments* $E[\|X_i\|^4]$*,* $E[\epsilon_i^4]$ *and* $E[\|X_i\|^4\epsilon_i^4]$ *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 $T^*_{w,n}$ and $T^*_{s,n}$.

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

If the conditions for an Edgeworth expansion of the bootstrap statistics $T^*_{w,n}$ and $T^*_{s,n}$ 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.[4] 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. The higher order equivalence of $T^*_{w,n}$ and $T^*_{s,n}$ 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 residuals $\{\epsilon^*_i\}^n_{i=1}$ being mean independent of $\{X_i\}^n_{i=1}$ under the bootstrap distribution. Importantly, the higher order equivalence would fail to hold if the residuals $\{\epsilon^*_i\}$ were sampled in a manner under which they were merely uncorrelated with $\{X_i\}^n_{i=1}$ 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:

$$\hat{\beta}^* = \hat{\beta} + H_n^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i \epsilon^*_i \ . \tag{8}$$

Note that the right hand side of equation (8) 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 (1999) and Andrews (2002). However, as Lemma 3.1 suggests and our Monte Carlos confirm, in the linear model, computation of $\hat{\beta}^*$ may be avoided while still obtaining a refinement over an analytical approximation. ■

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 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 $Z_i \in \mathcal{Z} \subseteq \mathbf{R}^m$ which may contain both.

We focus on test statistics G_n that are quadratic forms of a vector valued statistic T_n:

$$G_n \equiv T_n' T_n \ . \tag{9}$$

[4]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 $P(P^*(\|T^*_{w,n} - T^*_{s,n}\| > (n^{\frac{1}{2}} \log n)^{-1}) > n^{-\frac{1}{2}}) = o(n^{-\frac{1}{2}})$ and the Edgeworth expansion is valid in the bootstrap sample with probability $1 - o(n^{-\frac{1}{2}})$ (Lemma 5 Andrews (2002)).

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:

$$T_n = (A_n(\theta_0)\Sigma_n(\theta_0)A_n(\theta_0)')^{-\frac{1}{2}}S_n(\theta_0) + o_p(1) \qquad S_n(\theta) \equiv A_n(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i,\theta)\ , \tag{10}$$

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 $s(Z_i,\theta)$ and θ_0 is an unknown parameter vector. 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:

$$G_n^* \equiv T_n^{*\prime}T_n^* \qquad T_n^* \equiv (A_n(\hat{\theta})\Sigma_n^*(\hat{\theta})A_n(\hat{\theta})')^{-\frac{1}{2}}S_n^*(\hat{\theta}) \qquad S_n^*(\theta) \equiv A_n(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i,\theta)W_i \tag{11}$$

where $\Sigma_n^*(\theta)$ is the sample covariance matrix of $s(Z_i,\theta)W_i$ 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 $\hat{\theta}$ and no additional optimization is needed in each bootstrap iteration.

4.1 Bootstrap Consistency

We establish the consistency of the bootstrap under the following set of 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) The limit point* θ_0 *satisfies* $E[s(Z_i,\theta_0)s(Z_i,\theta_0)'] < \infty$ *and the matrix* $A(\theta_0)E[s(Z_i,\theta_0)s(Z_i,\theta_0)']A(\theta_0)'$ *is invertible.*

Assumption 4.2. *(i) Under the null hypothesis* T_n *satisfies* (10) *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.*

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* $conv(\Theta)$ *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. In Assumptions 4.3(ii) and 4.4(ii), $\|\cdot\|_F$ denotes the Frobenius norm. Assumptions 4.2 and 4.3, in turn enable us to establish the asymptotic behavior of G_n under the null and alternative hypotheses; see Lemma 6.3 in the Appendix. 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.[5]

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

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 suppose that Assumptions 4.1, 4.2, 4.3 and 4.4 hold. If the null hypothesis is true, it then follows that:*

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

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

$$\hat{c}_{1-\alpha} \equiv \inf\{c : P(G_n^* \geq c \,|\{Z_i\}_{i=1}^n) \geq 1-\alpha\} \ . \tag{12}$$

While difficult to compute analytically, $\hat{c}_{1-\alpha}$ may easily 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:

[5]For non-differentiable settings, the relevant condition is that $\mathcal{F} \equiv \{ws(z,\theta) : \theta \in \Theta\}$ be a Donsker class.

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 hypothesis tests on parametric models. 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 $Q : \Theta \to \mathbf{R}$:

$$\theta_M = \arg\min_{\theta\in\Theta} Q(\theta) \ . \tag{13}$$

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:

$$H_0 : c(\theta_M) = 0 \qquad H_1 : c(\theta_M) \neq 0 \ . \tag{14}$$

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 an ML estimator, the criterion function Q and its sample analogue Q_n are of the form:

$$Q_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} q(Z_i, \theta) \qquad Q(\theta) \equiv E[q(Z_i, \theta)] \ , \tag{15}$$

where $q : \mathcal{Z} \times \Theta \to \mathbf{R}$ 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 the asymptotic expansion:

$$\sqrt{n}(c(\hat{\theta}_M) - c(\theta_M)) = -C(\theta_M)H_n^{-1}(\theta_M)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla q(Z_i, \theta_M) + o_p(1) \ . \tag{16}$$

Therefore, the Wald statistic fits the formulation in (10) with $A_n(\theta) = -C(\theta)H_n^{-1}(\theta)$ and $s(Z_i, \theta) = \nabla q(Z_i, \theta)$. 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) In this setting, the LM statistic is the normalized quadratic form of:

$$C(\hat{\theta}_{M,R})H_n^{-1}(\hat{\theta}_{M,R})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla q(Z_i, \hat{\theta}_{M,R}) \ . \tag{17}$$

Moreover, under conditions stated in Chapter 12.6.2 in Wooldridge (2002), we additionally have:

$$C(\hat{\theta}_{M,R})H_n^{-1}(\hat{\theta}_{M,R})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla q(Z_i, \hat{\theta}_{M,R}) = C(\theta_{M,R})H_n^{-1}(\theta_{M,R})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla q(Z_i, \theta_{M,R}) + o_p(1) \ , \tag{18}$$

under the null hypothesis. Thus, the LM statistic also fits the general formulation in (10) with $A_n(\theta) = C(\theta)H_n^{-1}(\theta)$ and score $s(z, \theta) = \nabla q(z, \theta)$. 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, the criterion function Q and its sample analogue Q_n are:

$$Q_n(\theta) \equiv [\frac{1}{n}\sum_{i=1}^{n} q(Z_i, \theta)']\Omega_n[\frac{1}{n}\sum_{i=1}^{n} q(Z_i, \theta)] \qquad Q(\theta) \equiv E[q(Z_i, \theta)']\Omega E[q(Z_i, \theta)] \ , \tag{19}$$

where $q : \mathcal{Z} \times \Theta \to \mathbf{R}^k$ 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 (14) is given by the studentized quadratic form of $\sqrt{n}c(\hat{\theta}_M)$. In the present context we therefore obtain an expansion of the form:

$$\sqrt{n}(c(\hat{\theta}_M) - c(\theta_M)) = -C(\theta_M)B_n^{-1}(\theta_M)D_n(\theta_M)'\Omega_n\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(Z_i, \theta_M) + o_p(1) \ , \tag{20}$$

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 $E[q(Z_i,\theta_M)] = 0$.[6] 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) In this setting, the LM test statistic is the studentized quadratic form of:

$$C(\hat{\theta}_{M,R})B_n^{-1}(\hat{\theta}_{M,R})D_n(\hat{\theta}_{M,R})'\Omega_n\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(Z_i,\hat{\theta}_{M,R}) \ , \tag{21}$$

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

$$C(\theta_{M,R})B_n^{-1}(\theta_{M,R})D_n(\theta_{M,R})'\Omega_n\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(Z_i,\theta_{M,R}) \tag{22}$$

under the null hypothesis. Hence, $A_n(\theta) = C(\theta)B_n^{-1}(\theta)D_n(\theta)'\Omega_n$ and $s(z,\theta) = q(z,\theta)$. ■

4.3 Moment Restrictions

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

$$H_0 : E[m(Z_i,\theta_M)] = 0 \qquad H_1 : E[m(Z_i,\theta_M)] \neq 0 \ , \tag{23}$$

where $m : \mathcal{Z} \times \Theta \to \mathbf{R}^l$ 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 (15) (ML) or as in (19) (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 (23) is based on the studentized plug-in estimator:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Z_i,\hat{\theta}_M) \ , \tag{24}$$

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 (10) is straightforward provided m and q are once and twice continuously differentiable in θ respectively.

[6]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.

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:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Z_i, \hat{\theta}_M) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Z_i, \theta_M) - M_n(\theta_M) H_n^{-1}(\theta_M) \frac{1}{\sqrt{n}} \sum_{i=1} \nabla q(Z_i, \theta_M) + o_p(1) \ ; \qquad (25)$$

see Newey (1985a) for primitive conditions for (25). Thus, in this setting $s(z, \theta)$ and $A_n(\theta)$ are:

$$s(z, \theta) = \begin{pmatrix} m(z, \theta) \\ \nabla q(z, \theta) \end{pmatrix} \qquad A_n(\theta) = \left[I \ \vdots \ - M_n(\theta) H_n^{-1}(\theta) \right] \ . \qquad (26)$$

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

4.3.1 ML Specification Tests

A prominent application of hypotheses as in (23) 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).[7] Computations are simplified for ML models by the generalized information matrix equality, which implies:

$$E[\nabla^2 q(Z_i, \theta_M)] = -E[\nabla q(Z_i, \theta_M) \nabla q(Z_i, \theta_M)'] \qquad E[\nabla m(Z_i, \theta_M)] = -E[m(Z_i, \theta_M) \nabla q(Z_i, \theta_M)']$$

For example, as noted in Chesher (1984) and Newey (1985b), computation of the Wald test statistic for the null hypothesis in (23) can be performed through the auxiliary regression:

$$1 = m(Z_i, \hat{\theta}_M)' \delta + \nabla q(Z_i, \hat{\theta}_M)' \gamma + \epsilon_i \ . \qquad (27)$$

If R^2 is the uncentered R-squared of the regression in (27), then under the generalized information matrix equality result in (27) the Wald test statistic is asymptotically equivalent to:[8]

$$G_n = nR^2 \ . \qquad (28)$$

The calculation of the score bootstrap simplifies in an analogous fashion. Under a uniform law of large numbers, we obtain that $A_n(\hat{\theta}_M)$ as defined in (25) satisfies,

$$A_n(\hat{\theta}_M) = \left[I \ \vdots \ - \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \hat{\theta}_M) \nabla q(Z_i, \hat{\theta}_M)' \times \left[\frac{1}{n} \sum_{i=1}^{n} \nabla q(Z_i, \hat{\theta}_M) \nabla q(Z_i, \hat{\theta}_M)' \right]^{-1} \right] + o_p(1) \ , \ (29)$$

[7] A bootstrap construction for the Information Matrix Equality test was also developed in Horowitz (1994).

[8] See also Chapter 8.2.2 in Cameron and Trivedi (2005) for a summary of these results.

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 the score $\nabla q(Z_i, \hat{\theta}_M)$:

$$\begin{aligned} m^{(1)}(Z_i, \hat{\theta}_M) &= \nabla q(Z_i, \hat{\theta}_M)'\beta_1 + \epsilon_{1,i} \\ \vdots &= \quad \vdots \quad , \\ m^{(l)}(Z_i, \hat{\theta}_M) &= \nabla q(Z_i, \hat{\theta}_M)'\beta_l + \epsilon_{l,i} \end{aligned} \tag{30}$$

where $m^{(j)}(Z_i, \hat{\theta}_M)$ is the j^{th} component of $m(Z_i, \hat{\theta}_M)$. 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 $e_i = (e_{1,i}, \ldots, e_{l,i})'$, we obtain that:

$$S_n^*(\hat{\theta}_M) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i W_i \ . \tag{31}$$

Therefore, G_n^* is simply the Wald test for the null hypothesis that the mean of $e_i W_i$ equals zero.

In summary, if the generalized information matrix equality holds, then in testing (23) we may employ the following simple algorithm:

STEP 1: Run the regression in (27) and compute the uncentered R-squared to obtain G_n as in (28).

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

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 $\{e_i W_i\}_{i=1}^n$.

STEP 4: Let G_n^* be the Wald test statistic for the null that $E[e_i W_i] = 0$ calculated using $\{e_i W_i\}_{i=1}^n$. 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 all clusters have the same number of observations. An extension to unbalanced clusters is 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 (10), we consider test statistics of the general form $\tilde{G}_n \equiv \tilde{T}_n'\tilde{T}_n$, where $\tilde{T}_n$ satisfies:

$$\tilde{T}_n = (A_n(\theta_0)\tilde{\Sigma}_n(\theta_0)A_n(\theta_0)')^{-\frac{1}{2}}\tilde{S}_n(\theta_0) + o_p(1) \qquad \tilde{S}_n(\theta) \equiv A_n(\theta)\frac{1}{\sqrt{n}}\sum_{c=1}^{n}\frac{1}{\sqrt{J}}\sum_{i=1}^{J}\tilde{s}(Z_{ic}, \theta) \ , \tag{32}$$

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 notice that we may define $s(z,\theta)$, mapping each (Z_c,θ) into a $m \times 1$ vector, to be given by:

$$s(Z_c,\theta) = \frac{1}{\sqrt{J}} \sum_{i=1}^{J} \tilde{s}(Z_{ic},\theta) \ . \tag{33}$$

The statistics $\tilde{T}_n$ and $\tilde{S}_n(\theta)$ are then special cases of T_n and $S_n(\theta)$ as considered in (10) but with Z_c in place of Z_i. Hence, equations (10) and (33) indicate that the relevant bootstrap statistic should perturb the data at the cluster rather than at the individual observation level. We thus define:

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

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 (33) 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 performance 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 proper specification 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 independent clusters.[9]

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

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 will have discrete support or be binary, in which case the statistic of interest will be undefined in many block bootstrap samples. 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 the different bootstrap procedures in conducting inference on a linear model under four different designs meant to reflect realistic features of microeconomic datasets. Throughout, the linear model we examine is given by:

$$Y_{ic} = X_{ic} + D_c + \eta_c + \epsilon_{ic} \ , \tag{34}$$

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

$$X_{ic} = X_c + \xi_{ic} \qquad D_c = X_c\omega_c \qquad \eta_c = (1 + D_c)\upsilon_c \ . \tag{35}$$

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. The designs are:

Design I: (baseline) We let $(X_c, \xi_{ic}, \omega_c, \epsilon_{ic})$ 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 ω_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 (34), but the DGP is modified to:

$$Y_{ic} = X_{ic} + D_c + .1D_c^2 + \eta_c + \epsilon_{ic} \ , \tag{36}$$

and other features remain as in Design I. Hence, the quadratic term in the regressor of interest is ignored in estimation which yields an asymmetric 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.

Design IV: (skew and misspecification) Design III is modified so that ω_c is as in Design II.

Our baseline design for the linear model exhibits fat tails in the random cluster effect but no skew in the score. Design II introduces skewness into the experiment by modifying the regressor (and hence the reduced form error) to contain outliers. Finally, in Designs III and IV, we explore the effects of misspecification.

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

$$Y_{ic} = 1\{X_{ic} + D_c + \eta_c + \epsilon_{ic} \geq 0\} \qquad X_{ic} = X_c + \xi_{ic} \qquad D_c = X_c\omega_c \ . \tag{37}$$

This is essentially a latent variable representation of the model in (34) without heteroscedasticity in the group error η_c. We consider the following two designs for our probit analysis:

Design V: (baseline probit) In (37), we let $(X_{ic}, \xi_{ic}, \omega_c) \sim N(0, I_3)$ and $(\eta_c, \epsilon_{ic}) \sim N(0, I_2/2)$.[10]

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

Finally, we illustrate the methods of Section 4.3 by testing the following moment restrictions implied by the probit model:

$$E[e_{ic}D_c^2] = E[e_{ic}D_c^3] = E[e_{ic}X_{ic}^2] = E[e_{ic}X_{ic}^3] = E[e_{ic}X_{ic}D_c] = 0 \tag{38}$$

where $e_{ic} = [Y_{ic} - p_{ic}]\phi(X_{ic} + D_c)/[p_{ic}(1 - p_{ic})]$ 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}.[11] A test of these five moment conditions examines the probit model for unmodeled nonlinearities in the response function.

5.2 Results

Table 1 provides empirical false 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 (34) is one. Stata code for all our Monte Carlo experiments is available online.[12] All tests have a nominal size of 5% and are studentized using a recentered variance matrix estimator.[13] We consider implementations of

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

[11] Note that the ML probit scores are of the form $e_{ic}[1, X_{ic}, D_c]$.

[12] URL: `http://www.econ.berkeley.edu/~pkline/papers/score.zip`

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

Table 1: Empirical Rejection Rates, OLS (Properly Specified)

	Normal Regressor					Mixture Regressor				
Wald Tests	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
Analytical	0.442	0.328	0.240	0.153	0.083	0.467	0.398	0.317	0.240	0.140
Pairs	0.020	0.088	0.079	0.054	0.040	0.030	0.134	0.110	0.080	0.052
Wild Rademacher	0.243	0.185	0.128	0.078	0.052	0.273	0.250	0.193	0.127	0.075
Wild Mammen	0.252	0.187	0.146	0.105	0.060	0.282	0.240	0.187	0.138	0.094
Score Rademacher	0.263	0.194	0.142	0.091	0.048	0.270	0.223	0.188	0.142	0.091
Score Mammen	0.288	0.220	0.162	0.104	0.053	0.292	0.245	0.206	0.156	0.096
	Normal Regressor					Mixture Regressor				
LM Tests	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
Analytical	0.001	0.023	0.030	0.037	0.038	0.001	0.021	0.024	0.026	0.030
Pairs	0.051	0.061	0.051	0.043	0.039	0.054	0.057	0.047	0.045	0.035
Wild Rademacher	0.103	0.065	0.039	0.039	0.046	0.112	0.067	0.036	0.031	0.038
Wild Mammen	0.165	0.103	0.068	0.057	0.051	0.177	0.102	0.065	0.049	0.046
Score Rademacher	0.105	0.077	0.062	0.053	0.048	0.121	0.088	0.060	0.052	0.052
Score Mammen	0.084	0.034	0.026	0.026	0.033	0.097	0.036	0.024	0.017	0.026

the score bootstrap using both Rademacher weights and the skew correcting weights suggested by Mammen (1993).[14] 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 conventional wild bootstrap, and the pairs-based block bootstrap. All bootstrap tests were computed using 200 bootstrap repetitions.

The standard clustered Wald test severely over-rejects in samples with few clusters, with performance further degrading when the regressors are generated according to a mixture distribution. 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 under performs relative to pairs regardless of whether Mammen or Rademacher weights are used. As suggested by our theoretical results, the score bootstrap yields results roughly in line with those of the corresponding Wild bootstrap.

In contrast to the Wald tests, the clustered LM tests appear to perform well across a range

[14] Rademacher weights impose $E[W_i^4] = 1$ while Mammen's weights impose $E[W_i^3] = 1$.

Table 2: Empirical Rejection Rates, OLS (Misspecified)

	Normal Regressor					Mixture Regressor				
Wald Tests	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
Analytical	0.448	0.333	0.248	0.162	0.086	0.473	0.411	0.331	0.262	0.157
Pairs	0.022	0.091	0.084	0.059	0.042	0.033	0.135	0.110	0.073	0.042
Wild Rademacher	0.249	0.192	0.135	0.078	0.051	0.278	0.257	0.198	0.135	0.068
Wild Mammen	0.254	0.195	0.150	0.105	0.060	0.286	0.247	0.191	0.146	0.090
Score Rademacher	0.253	0.184	0.135	0.087	0.045	0.259	0.214	0.185	0.144	0.101
Score Mammen	0.277	0.210	0.154	0.099	0.047	0.281	0.234	0.200	0.156	0.102
	Normal Regressor					Mixture Regressor				
LM Tests	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
Analytical	0.001	0.022	0.032	0.037	0.040	0.001	0.021	0.026	0.024	0.026
Pairs	0.051	0.062	0.053	0.045	0.040	0.055	0.062	0.049	0.039	0.031
Wild Rademacher	0.106	0.062	0.042	0.041	0.044	0.116	0.062	0.033	0.032	0.037
Wild Mammen	0.167	0.103	0.072	0.059	0.048	0.181	0.099	0.063	0.052	0.042
Score Rademacher	0.110	0.082	0.061	0.057	0.051	0.118	0.090	0.061	0.055	0.051
Score Mammen	0.094	0.033	0.026	0.027	0.034	0.094	0.038	0.025	0.021	0.024

of sample sizes and regardless of the distribution of the regressors. While the analytical LM test yields mild underrejection with few clusters, its Wild bootstrapped analogue actually yields slight over-rejection.[15] The score bootstrapped LM tests perform as well as or better than the wild bootstrapped LM tests under both regressor designs. They also perform comparably to the pairs bootstrapped Wald tests. However the pairs bootstrapped LM tests yield the best performance of the group, with coverage rates closest to nominal levels across a range of sample sizes.

Table 2 examines the performance of Wald and LM tests when the model is misspecified. Again the performance of the analytical clustered Wald test appears to be very poor in small samples or when the regressor of interest exhibits outliers. Correcting the critical values of the Wald test with the pairs bootstrap yields much improved though still sometimes unsatisfactory performance. As before, the Wild bootstrap improves on the performance of analytical Wald tests but still overrejects substantially. Score bootstrapping the Wald statistic yields results mimicking those of the Wild

[15]We note that the wild bootstrapped LM test is similar to the Wild bootstrap procedure of Cameron, Gelbach, and Miller (2008) who impose the null $\hat{\beta} = \beta_0$ when generating the bootstrap distribution of outcomes as in (3). In results not shown we found the results from this procedure (which is akin to comparing the bootstrap critical values of an LM statistic to a full sample Wald) to be quite similar to those found in wild or score bootstrap LM tests.

Table 3: Empirical Rejection Rates, Probit (Properly Specified)

	Normal Regressor					Mixture Regressor				
Wald Tests	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
Analytical	0.319	0.192	0.132	0.094	0.063	0.320	0.210	0.137	0.092	0.063
Pairs	0.032	0.065	0.073	0.064	0.053	0.037	0.062	0.075	0.064	0.052
Score Rademacher	0.272	0.153	0.097	0.064	0.037	0.283	0.174	0.102	0.062	0.039
Score Mammen	0.303	0.179	0.110	0.067	0.038	0.316	0.204	0.115	0.063	0.038
	Normal Regressor					Mixture Regressor				
LM Tests	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$	$n=5$	$n=10$	$n=20$	$n=50$	$n=200$
Analytical	0.004	0.070	0.079	0.080	0.060	0.004	0.080	0.089	0.079	0.062
Pairs[16]	n.a.	0.140	0.125	0.096	0.065	n.a.	0.144	0.138	0.091	0.059
Score Rademacher	0.101	0.124	0.099	0.085	0.060	0.111	0.142	0.108	0.085	0.064
Score Mammen	0.079	0.054	0.077	0.081	0.063	0.089	0.063	0.085	0.082	0.061

bootstrap.

With misspecification and an asymmetric regressor, the clustered LM test underrejects substantially in small samples. Wild or score bootstrapping the LM test leads to slight overrejection in small samples. The score bootstrap with Rademacher weights seems to perform particularly well.

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.[17] Score bootstrapping the Wald test yields improvements over analytical clustered standard errors but substantial overrejection remains in small samples. Score bootstrapping the LM test with Mammen weights, on the other hand, yields size control roughly on par with the pairs bootstrap.

Finally, Table 4 examines the performance of tests for proper specification of the probit model via the restrictions in (38). Because the information matrix equality holds under both DGPs we use the outer product version of the test described in 4.3.1 generalized to allow for clustering and a recentered variance matrix estimator. We see that the analytical m-test procedure overrejects substantially in small samples and continues to exhibit poor control over size even with 200 clusters. Surprisingly the score bootstrapped versions of the test work well in small samples, but appear to

[16] We were unable to compute the LM statistic in the majority of pairs draws with 5 clusters.

[17] We discarded bootstrap draws for which we were unable to compute a maximum likelihood estimate.

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

	Normal Regressor					Mixture Regressor				
Wald Tests	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 200$	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 200$
Analytical	0.766	0.593	0.341	0.190	0.130	0.776	0.584	0.348	0.208	0.131
Score Rademacher	0.067	0.135	0.174	0.159	0.138	0.068	0.129	0.180	0.177	0.142
Score Mammen	0.049	0.034	0.092	0.134	0.141	0.046	0.032	0.093	0.147	0.143

degrade slightly as the number of clusters increase. With 200 clusters, the analytical and bootstrap approaches appear to work equally well.

6 Conclusion

Score bootstrap tests provide a computational advantage over conventional wild and pairs bootstraps and may easily be applied to estimators that lack conventional residuals. Both our theoretical and Monte Carlo results suggest the wild bootstrap possesses no inferential advantage over the score bootstrap despite the potentially significant increase in computational cost. Moreover, the score bootstrap provides improvements over analytical techniques in a variety of testing environments of interest where pairs bootstraps might encounter computational difficulties.

Like Moreira, Porter, and Suarez (2009) we find that bootstrapping Lagrange Multiplier type tests yields improved small sample size control in a number of difficult testing environments of substantial applied interest. Economists have typically shied away from bootstrap LM tests perhaps due to the difficulty of constructing confidence intervals by test inversion. The score bootstrap methods developed here substantially reduce the cost of such an exercise and may enable researchers to conduct inference in a wider range of small sample environments than previously contemplated.

Appendix

Proof of Lemma 3.1: First notice that by Markov's inequality, $E[W_i^2] = 1$ and the i.i.d. assumption

$$P^*(\|\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\epsilon_i^*\| > C) \leq \frac{1}{nC^2}E^*[(\sum_{i=1}^n X_i\epsilon_i^*)'(\sum_{i=1}^n X_i\epsilon_i^*)] = \frac{1}{nC^2}\sum_{i=1}^n X_i'X_ie_i^2 \ . \tag{39}$$

Let $H \equiv E[X_iX_i']$ and $\Sigma \equiv E[X_iX_i'\epsilon_i^2]$. Since $n^{-1}\sum_i X_i'X_ie_i^2 \xrightarrow{a.s.} \Sigma$ and $H_n \xrightarrow{a.s.} H$, we obtain from (39),

$$\|\sqrt{n}(\hat{\beta}^* - \hat{\beta})\| \leq \|H_n^{-1}\|_F \times \|\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\epsilon_i^*\| = O_{p^*}(1) \qquad a.s. \ , \tag{40}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Next observe that for $\|\cdot\|_o$ the operator norm, we have:

$$\begin{aligned}\|(H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1})^{-1} - (H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1})^{-1}\|_o \\ \leq \|(H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1})^{-1}\|_o \times \|H_n^{-1}(\Sigma_n^*(\hat{\beta}) - \Sigma_n^*(\hat{\beta}^*))H_n^{-1}\|_o \times \|(H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1})^{-1}\|_o \ .\end{aligned} \tag{41}$$

Let $X_i^{(k)}$ denote the k^{th} element of the vector X_i. Arguing as in (39), it is straightforward to show that $n^{-\frac{1}{2}}\sum_i X_i^{(k)}X_i^{(l)}X_i^{(s)}\epsilon_i^* = O_{p^*}(1)$ almost surely for any indices k, l, s. Therefore, since $\|\cdot\|_o \leq \|\cdot\|_F$ we conclude from (40) and direct calculation that we must have:

$$\begin{aligned}\|\Sigma_n^*(\hat{\beta}) - \Sigma_n^*(\hat{\beta}^*)\|_o &\leq \|\frac{1}{n}\sum_{i=1}^n X_iX_i'\{(Y_i^* - X_i'\hat{\beta})^2 - (Y_i^* - X_i'\hat{\beta}^*)^2\}\|_F \\ &= \|\frac{1}{n}\sum_{i=1}^n X_iX_i'\{2\epsilon_i^*(X_i'(\hat{\beta} - \hat{\beta}^*)) + (X_i'(\hat{\beta} - \hat{\beta}^*))^2\}\|_F = O_{p^*}(n^{-1}) \qquad a.s. \ .\end{aligned} \tag{42}$$

Moreover, since $E[(\epsilon_i^*)^k] = E[W_i^k]e_i^k$, we can also obtain from the i.i.d. assumption that:

$$\begin{aligned}E^*[\|\frac{1}{n}\sum_{i=1}^n X_iX_i'\{(\epsilon_i^*)^2 - e_i^2\}\|_F^2] &= \sum_{l=1}^m\sum_{s=1}^m E^*[(\frac{1}{n}\sum_{i=1}^n X_i^{(l)}X_i^{(s)}\{(\epsilon_i^*)^2 - e_i^2\})^2] \\ &= \frac{1}{n}\sum_{l=1}^m\sum_{s=1}^m\frac{1}{n}\sum_{i=1}^n(X_i^{(l)}X_i^{(s)})^2\{(E[W_i^4] - 1)e_i^4\} = o_{a.s.}(1) \ .\end{aligned} \tag{43}$$

Therefore, since $n^{-1}\sum_i X_iX_i'e_i^2 \xrightarrow{a.s.} \Sigma$ and $H_n^{-1} \xrightarrow{a.s.} H^{-1}$, results (42) and (43) establish:

$$\|H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1} - H^{-1}\Sigma H^{-1}\|_F = o_{p^*}(1) \qquad \|H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1} - H^{-1}\Sigma H^{-1}]\|_F = o_{p^*}(1) \tag{44}$$

almost surely. Next, for any normal matrix A, let $\xi(A)$ denote its smallest eigenvalue. Since Corollary III.2.6 in Bhatia (1997) implies that $|\xi(A) - \xi(B)| \leq \|A - B\|_F$, it then follows from (44) that:

$$\xi(H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1}) = \xi(H^{-1}\Sigma H^{-1}) + o_{p^*}(1) \qquad \xi(H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1}) = \xi(H^{-1}\Sigma H^{-1}) + o_{p^*}(1) \tag{45}$$

almost surely. However, since for any normal matrix A, we have $\|A^{-1}\|_o = \xi(A)$, result (45) and Assumption 3.1(i) imply $\|H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1}\|_o = O_{p^*}(1)$ and $\|H_n^{-1}\Sigma_n^*(\hat{\beta})H_n^{-1}\|_o = O_{p^*}(1)$ almost surely. The claim of the Lemma then follows by combining results (40), (41) and (42). ■

Lemma 6.1. *Let $\{W_i\}_{i=1}^n$ be an i.i.d. sample 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 the class $\mathcal{F} = \{s(z,\theta)s'(z,\theta)w^2 : \theta \in \Theta\}$ 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):

$$|s^{(l)}(z,\theta)| \leq |s^{(l)}(z,\theta) - s^{(l)}(z,\theta_0)| + |s^{(l)}(z,\theta_0)| \leq F(z)\|\theta - \theta_0\| + |s^{(l)}(z,\theta_0)| \ . \tag{46}$$

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

Lemma 6.2. *Under Assumptions 4.1(i), 4.3(i) and 4.4(i)-(ii), $\mathcal{F} \equiv \{ws(z,\theta) : \theta \in \Theta\}$ is Donsker.*

PROOF: Let $\|\cdot\|_o$ and $\|\cdot\|_F$ denote the operator and Frobenious norms. 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:

$$\begin{aligned}&\|ws(z,\theta_1) - ws(z,\theta_2)\| \\ &\quad = |w| \times \|\nabla s(z,\bar{\theta})(\theta_1 - \theta_2)\| \leq |w| \times \|\nabla s(z,\bar{\theta})\|_o \times \|\theta_1 - \theta_2\| \leq |w| \times F(z) \times \|\theta_1 - \theta_2\| \ .\end{aligned} \tag{47}$$

Hence, the class $\mathcal{F}$ is Lipschitz in $\theta \in \Theta$, and by Theorem 2.7.11 in van der Vaart and Wellner (1996):

$$N_{[\,]}(2\epsilon\|\tilde{F}\|_{L^2}, \mathcal{F}, \|\cdot\|_{L^2}) \leq N(\epsilon, \Theta, \|\cdot\|) \ , \tag{48}$$

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 (47), the diameter of $\mathcal{F}$ under $\|\cdot\|_{L^2}$ is less than or equal to MD:

$$\begin{aligned}\int_0^\infty \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{L^2})}d\epsilon &\leq \int_0^{MD} \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{L^2})}d\epsilon = 2M\int_0^{\frac{D}{2}} \sqrt{\log N_{[\,]}(2Mu, \mathcal{F}, \|\cdot\|_{L^2})}du \\ &\leq 2M\int_0^{\frac{D}{2}} \sqrt{\log N(u, \Theta, \|\cdot\|)}du \leq 2M\int_0^{\frac{D}{2}} \sqrt{p\log(D/u)}du < \infty\end{aligned} \tag{49}$$

where in the first equality we made a change of variables $u = \epsilon/2M$, the second inequality follows from (48) and the third by $N(u, \Theta, \|\cdot\|) \leq (\text{diam}(\Theta)/u)^p$. The claim of the Lemma then follows from (49), $E[\tilde{F}^2(Z_i, W_i)] < \infty$ and Theorem 2.5.6 in van der Vaart and Wellner (1996). ■

Lemma 6.3. *Suppose Assumptions 4.1, 4.2, 4.3 and 4.4(ii) hold. If the null hypothesis is true, it then follows that $G_n \xrightarrow{L} \mathcal{X}_r^2$. 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 $A_n(\theta_0) = A(\theta_0) + o_p(1)$. Lemma 6.1 applied to $W_i = 1$ with probability one imply $\Sigma_n(\theta_0) = \Sigma(\theta_0) + o_p(1)$ for $\Sigma(\theta_0) = E[s(Z_i,\theta_0)s(Z_i,\theta_0)']$. Hence, by Assumption 4.3(ii):

$$A_n(\theta_0)\Sigma_n(\theta_0)A_n(\theta_0)' = A(\theta_0)\Sigma(\theta_0)A(\theta_0)' + o_p(1) \ . \tag{50}$$

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:

$$T_n = (A_n(\theta_0)\Sigma_n(\theta_0)A_n(\theta_0)')^{-1}A_n(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i,\theta_0) + o_p(1) \xrightarrow{L} N(0,I) \ . \tag{51}$$

Hence, by the continuous mapping theorem and (51), $G_n \xrightarrow{L} \mathcal{X}_r$, which establishes the first claim of the Lemma. 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 (46), the matrix $s(z,\theta)s(z,\theta)'$ has an integrable envelope. Hence, since $s(z,\theta)s(z,\theta)'$ is continuous in θ for all z by Assumption 4.4(ii), the dominated convergence theorem implies $\Sigma(\theta)$ is continuous in θ. Therefore, by Lemma 6.1 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 6.2 imply:

$$\begin{aligned}(A_n(\hat{\theta})\Sigma_n^*(\hat{\theta})A_n(\hat{\theta})')^{-\frac{1}{2}}A_n(\hat{\theta})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i,\hat{\theta})W_i& \\ = (A(\theta_0)\Sigma(\theta_0)A(\theta_0)')^{-\frac{1}{2}}A(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i,\hat{\theta})W_i + o_p(1)& \\ = (A(\theta_0)\Sigma(\theta_0)A(\theta_0)')^{-\frac{1}{2}}A(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i,\theta_0)W_i + o_p(1) \ ,& \end{aligned} \tag{52}$$

where the second equality follows by Assumption 4.1(i) and Lemma 6.2. 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:

$$\|Y - V\|_{BL_1} \equiv \sup_{f\in BL_1} |E[f(Y)] - E[f(V)]| \ , \tag{53}$$

metrizes weak convergence, see for example Theorem 1.12.4 in van der Vaart and Wellner (1996). Define:

$$\bar{T}_n^* \equiv (A(\theta_0)\Sigma(\theta_0)A(\theta_0)')^{-\frac{1}{2}}A(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i,\theta_0)W_i \ . \tag{54}$$

Using that all $f \in BL_1$ are bounded in level and Lipschitz constant by one, we obtain for any $\eta > 0$:

$$\sup_{f\in BL_1} |E[f(\bar{T}_n^*) - f(T_n^*)|\{Z_i\}_{i=1}^n]| \leq \eta P(|\bar{T}_n^* - T_n^*| \leq \eta|\{Z_i\}_{i=1}^n) + 2P(|\bar{T}_n^* - T_n^*| > \eta|\{Z_i\}_{i=1}^n) \ . \tag{55}$$

However, by the law of iterated expectations and (52), 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 (55) in fact implies:

$$\sup_{f\in BL_1} |E[f(\bar{T}_n^*)|\{Z_i\}_{i=1}^n] - E[f(T_n^*)|\{Z_i\}_{i=1}^n]| = o_p(1) \ . \tag{56}$$

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

$$\sup_{f\in BL_1} |E[f(\bar{T}_n^*)|\{Z_i\}_{i=1}^n] - E[f(T_\infty^*)]| = o_p(1) \ . \tag{57}$$

For any $M > 0$, define the map $g_M : \mathbf{R}^r \to \mathbf{R}$ to be given 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 implies:

$$\sup_{f \in BL_1} |E[f(g_M(T_n^*))|\{Z_i\}_{i=1}^n] - E[f(g_M(T_\infty^*))]| \leq M \sup_{f \in BL_1} |E[f(T_n^*)|\{Z_i\}_{i=1}^n] - E[f(T_\infty^*)]| = o_p(1) \ , \quad (58)$$

where the final result follows by (56) and (57). Since $G_n^* = T_n^{*\prime} T_n^*$ and every $f \in BL_1$ is bounded by one,

$$\sup_{f \in BL_1} |E[f(G_n^*) - f(g_M(T_n^*))|\{Z_i\}_{i=1}^n]| \leq 2P(T_n^{*\prime} T_n^* > M|\{Z_i\}_{i=1}^n) \ . \quad (59)$$

By (52) and the continuous mapping theorem, $T_n^{*\prime} T_n^* \xrightarrow{L} \mathcal{X}_r^2$ unconditionally and hence is asymptotically tight. For an arbitrary $\eta > 0$ it then follows by Markov's inequality that for M sufficiently large:

$$\limsup_{n \to \infty} P(2P(T_n^{*\prime} T_n^* > M|\{Z_i\}_{i=1}^n) > \eta) \leq \limsup_{n \to \infty} \frac{2}{\eta} P(T_n^{*\prime} T_n^* > M) < \eta \ . \quad (60)$$

Similarly, let $G_\infty^* \sim \mathcal{X}_r^2$ and notice that by selecting M appropriately large we may also obtain:

$$\sup_{f \in BL_1} |E[f(G_\infty^*) - f(g_M(T_\infty^*))]| \leq 2P(T_\infty^{*\prime} T_\infty^* > M) < \eta \ . \quad (61)$$

Since η is arbitrary, results (58), (59), (60) and (61) in turn allow us to conclude that:

$$\sup_{f \in BL_1} |E[f(G_n^*)|\{Z_i\}_{i=1}^n] - E[f(G_\infty^*)]| = o_p(1) \ , \quad (62)$$

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 (62) and Lemma 6.3 that for any $c \in \mathbf{R}$, $F_n^*(c) = F(c) + o_p(1)$ and $F_n(c) = F(c) + o(1)$. The Theorem follows since convergence is uniform in $c \in \mathbf{R}$ by Lemma 2.11 in van der Vaart (1999). ■

Lemma 6.4. *Let $F_n : \mathbf{R} \to [0,1]$, $F : \mathbf{R} \to [0,1]$ be monotonic, $\sup_{c \in \mathbf{R}} |F_n(c) - F(c)| = o_p(1)$ and define:*

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

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

PROOF: Fix $\epsilon > 0$. Since by hypothesis F is strictly increasing at c_α it follows by definition of c_α:

$$F(c_\alpha - \epsilon) < \alpha < F(c_\alpha + \epsilon) \ . \quad (63)$$

Moreover, since $F_n(c_\alpha + \epsilon) > \alpha$ implies that $c_{n,\alpha} \leq c_\alpha + \epsilon$ and $F_n(c_\alpha - \epsilon) < \alpha$ implies that $c_{n,\alpha} > c_\alpha - \epsilon$,

$$\lim_{n \to \infty} P(|c_\alpha - c_{n,\alpha}| \leq \epsilon) \geq \lim_{n \to \infty} P(F_n(c_\alpha - \epsilon) < \alpha < F_n(c_\alpha + \epsilon)) = 1 \quad (64)$$

where the final equality follows from (63) and $\sup_c |F_n(c) - F(c)| = o_p(1)$ by hypothesis. ■

PROOF OF COROLLARY 4.1: Let F denote the cdf of a $\mathcal{X}^2_r$ random variable and $c_{1-\alpha}$ be its $1-\alpha$ quantile. As argued following (62), $\sup_c |F^*_n(c) - F(c)| = o_p(1)$, and hence by Lemma 6.4 it follows that $\hat{c}_{1-\alpha} = c_{1-\alpha} + o_p(1)$ provided $0 < \alpha < 1$. The first claim of the Corollary then follows by Lemma 6.3 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, (62) 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 6.3 and therefore the second claim of the Corollary follows. ■

PROOF OF COROLLARY 4.2: Given the definitions, 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(1), 119–162.

BARBE, P., AND P. BERTAIL (1995): *The Weighted Bootstrap*. Springer- Verlag, New York.

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

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*. Cambridge University Press, New York.

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(4), 865–872.

——— (1995): "A Mirror Image Invariance for M-estimators," *Econometrica*, 63(1), 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 (1999): "Bootstrap Testing in Nonlinear Models," *International Economic Review*, 40, 487–508.

——— (2010): "Wild Bootstrap Tests for IV Regression," *Journal of Business and Economic Statistics*, 28, 128–144.

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

Freedman, D. A. (1981): "Bootstrapping Regression Models," *The Annals of Statistics*, 9(6), 1218–1228.

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

Hardle, W., and E. Mammen (1993): "Comparing Nonparametric Versus Parametric Regression Fits," *The Annals of Statistics*, 21(4), 1926–1947.

Horowitz, J. L. (1994): "Bootstrap-based Critical Values for the Information Matrix Test," *Journal of Econometrics*, 61, 395–411.

——— (1997): "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress*, ed. by D. M. Kreps, and K. F. Wallis, vol. 3. Cambridge University Press.

——— (2001): "The Bootstrap," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 5, chap. 52. Elsevier.

Hu, F., and J. V. Zidek (1995): "A Bootstrap Based on the Estimating Equations of the Linear Model," *Biometrika*, 82(2), 263–275.

Liu, R. Y. (1988): "Bootstrap Procedures under some Non-I.I.D. Models," *The Annals of Statistics*, 16(4), 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(1), 255–285.

Moreira, M. J., J. R. Porter, and G. A. Suarez (2009): "Bootstrap Validity for the Score Test when Instruments are Weak," *Journal of Econometrics*, 149, 52–64.

Newey, W. K. (1985a): "Generalized Method of Moments Specifcation Testing," *Journal of Econometrics*, 29, 229–256.

——— (1985b): "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, 53(5), 1047–1070.

Newey, W. K., and D. L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, pp. 2113–2245. Elsevier Science B.V.

Tauchen, G. (1985): "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics*, 30, 415–443.

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

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

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

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

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

Wu, C. F. J. (1986): "Jacknife, Bootstrap, and other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14(4), 1261–1295.

You, J., and G. Chen (2006): "Wild Bootstrap Estimation in Partially Linear Models with Heteroscedasticity," *Statistics and Probability Letters*, 76, 340–348.