Journal of Econometrics 171 (2012) 54–70

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Higher order properties of the wild bootstrap under misspecification

Patrick Kline [a], Andres Santos [b,*]

[a] Department of Economics, UC Berkeley/NBER, United States

[b] Department of Economics, UC San Diego, United States

ARTICLE INFO

Article history:
Received 31 January 2011
Received in revised form
21 March 2012
Accepted 12 June 2012
Available online 23 June 2012

JEL classification:
C12

Keywords:
Wild bootstrap
Misspecification
Edgeworth expansion

ABSTRACT

We examine the higher order properties of the wild bootstrap (Wu, 1986) in a linear regression model with stochastic regressors. We find that the ability of the wild bootstrap to provide a higher order refinement is contingent upon whether the errors are mean independent of the regressors or merely uncorrelated with them. In the latter case, the wild bootstrap may fail to match some of the terms in an Edgeworth expansion of the full sample test statistic. Nonetheless, we show that the wild bootstrap still has a lower maximal asymptotic risk as an estimator of the true distribution than a normal approximation, in shrinking neighborhoods of properly specified models. To assess the practical implications of this result we conduct a Monte Carlo study contrasting the performance of the wild bootstrap with a normal approximation and the traditional nonparametric bootstrap.

1. Introduction

The wild bootstrap of Wu (1986) and Liu (1988) provides a procedure for conducting inference in the model:

$$Y = X'\beta_0 + \epsilon, \tag{1}$$

where $Y \in \mathbf{R}, X \in \mathbf{R}^{d_x}$ and ϵ may have a heteroscedastic structure of unknown form. This robustness to arbitrary heteroscedasticity provides the wild bootstrap with a distinct advantage over the residual bootstrap of Freedman (1981) which requires homoscedastic errors. Moreover, theoretical results from Mammen (1993) indicate that the wild bootstrap outperforms the nonparametric bootstrap when a large number of regressors are present and the errors obey the mean independence restriction $E[\epsilon|X] = 0$. These properties have led to increasing attention among economists concerned with heteroscedasticity robust inference in small sample environments (Horowitz, 1997, 2001; Cameron et al., 2008; Davidson and Flachaire, 2008), and to a variety of recent extensions beyond the basic linear regression model (Cavaliere and Taylor, 2008; Gonçalves and Meddahi, 2009; Davidson and MacKinnon, 2010; Kline and Santos, in press). To date, however, the higher order properties of the wild bootstrap have only been studied under the assumption of proper model specification, where the errors are mean independent of the regressors. Liu (1988) first established that when this condition holds the wild bootstrap provides a refinement over a normal approximation.

Since the seminal work of White (1980a,b, 1982), economists have sought inference procedures robust to the possibilities of both unmodeled heteroscedasticity and misspecification (see Stock, 2010 for a recent retrospective). In an important contribution, Mammen (1993) established that the wild bootstrap exhibits a form of robustness, showing that it remains consistent in the absence of proper model specification. In this paper, we contribute to the literature by examining whether, in addition to remaining consistent, the wild bootstrap continues to provide a refinement over the normal approximation under misspecification. Concretely, we study the higher order properties of the wild bootstrap when ϵ is uncorrelated with X but not necessarily mean independent of it—a setting commonly encountered in economics where parametric modeling is pervasive. It is precisely in such misspecified environments that heteroscedasticity is likely to arise making the higher order properties of the wild bootstrap of particular interest (White, 1982).

We conduct our analysis in two steps. First, we compute the approximate cumulants (Bhattacharya and Ghosh, 1978) of t-statistics under both the full sample and bootstrap distributions with general assumptions on the wild bootstrap weights. We show that both the first and third approximate cumulants may disagree up to order $O_p(n^{-\frac{1}{2}})$ if higher powers of X are correlated with ϵ—a situation that is ruled out under proper specification. This higher order discordance between the approximate cumulants under the full sample and the bootstrap distribution implies that if valid Edgeworth expansions exist they would only be equivalent

* Correspondence to: 9500 Gilman Drive MS 0508, La Jolla, CA 92093-0508, United States. Tel.: +1 858 534 2407; fax: +1 858 534 7040.

E-mail address: a2santos@ucsd.edu (A. Santos).

0304-4076/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2012.06.001

up to order $O_p(n^{-\frac{1}{2}})$ (Hall, 1992). As a result, despite remaining consistent under misspecification, the wild bootstrap may fail to provide a higher order refinement over a normal approximation.

We complement this result by formally establishing the existence of valid one term Edgeworth expansions when the distribution of the wild bootstrap weights is additionally assumed to be strongly nonlattice (Bhattacharya and Rao, 1976). In accord with Liu (1988) we note that one-sided wild bootstrap tests obtain a refinement to order $O_p(n^{-1})$ under proper specification. However, this result is undermined by certain forms of misspecification under which only some, but not all, of the second order terms in the full sample Edgeworth expansion are matched by their bootstrap counterparts. Despite this discordance, we establish that the wild bootstrap still possesses a lower asymptotic risk as an estimator of the true distribution of studentized test statistics than a normal approximation in shrinking neighborhoods of properly specified models. Heuristically, these results suggest that the wild bootstrap should outperform a normal approximation provided misspecification is not "too severe". To assess the practical implications of this result, we conclude by conducting a Monte Carlo study contrasting the performance of the wild bootstrap with that of a normal approximation and the traditional nonparametric bootstrap in the presence of misspecification.

The rest of the paper is organized as follows. Section 2 contains our theoretical results while Section 3 examines the implications of our analysis in a simulation study. We briefly conclude in Section 4 and relegate the main proofs to the Appendix. Some auxiliary results are collected in a supplementary Appendix available in our respective websites.

2. Theoretical results

While numerous variants of the wild bootstrap exist, we study the original version proposed by Wu (1986) and Liu (1988). Succinctly, given a sample $\{Y_i, X_i\}_{i=1}^n$ and $\hat{\beta}$ the OLS estimator from such sample, the wild bootstrap generates new errors and dependent variables:

$$Y_i^* \equiv X_i'\hat{\beta} + \epsilon_i^* \quad \epsilon_i^* \equiv (Y_i - X_i'\hat{\beta})W_i, \tag{2}$$

where $\{W_i\}_{i=1}^n$ is an i.i.d. sample independent of the original data $\{Y_i, X_i\}_{i=1}^n$. A bootstrap estimator $\hat{\beta}^*$ can then be computed from the sample $\{Y_i^*, X_i\}_{i=1}^n$ and the distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ conditional on $\{Y_i, X_i\}_{i=1}^n$ (but not $\{W_i\}_{i=1}^n$) is used to approximate that of $\sqrt{n}(\hat{\beta} - \beta_0)$. While it may not be possible to compute the bootstrap distribution analytically, it is straightforward to simulate it.

We focus our analysis on inference on linear contrasts of β_0, which includes both individual coefficients and predicted values as special cases. In particular, for an arbitrary $c \in \mathbf{R}^{d_x}$ we examine:

$$T_n \equiv \frac{\sqrt{n}}{\hat{\sigma}} c'(\hat{\beta} - \beta_0) \qquad \hat{\sigma}^2 \equiv c'H_n^{-1}\Sigma_n(\hat{\beta})H_n^{-1}c, \tag{3}$$

where the $d_x \times d_x$ matrices H_n and $\Sigma_n(\beta)$ are defined by

$$H_n \equiv \frac{1}{n}\sum_{i=1}^n X_iX_i' \qquad \Sigma_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^n X_iX_i'(Y_i - X_i'\beta)^2. \tag{4}$$

The bootstrap statistic T_n^* is then the analogue to T_n but computed on $\{Y_i^*, X_i\}_{i=1}^n$ instead. Namely,

$$T_n^* \equiv \frac{\sqrt{n}}{\hat{\sigma}^*} c'(\hat{\beta}^* - \hat{\beta}) \qquad (\hat{\sigma}^*)^2 \equiv c'H_n^{-1}\Sigma_n^*(\hat{\beta}^*)H_n^{-1}c, \tag{5}$$

where H_n is as in (4), and $\Sigma_n^*(\beta) \equiv \frac{1}{n}\sum_i X_iX_i'(Y_i^* - X_i'\beta)^2$.

As argued in Mammen (1993), under mild assumptions on the wild bootstrap weights $\{W_i\}_{i=1}^n$, the distribution of T_n^* conditional on $\{Y_i, X_i\}_{i=1}^n$ (but not $\{W_i\}_{i=1}^n$) provides a consistent estimator for the distribution of T_n. Consequently, tests based upon a comparison of the statistic T_n to the quantiles of the wild bootstrap distribution of T_n^* can provide size control asymptotically. In what follows, we explore whether such a procedure is additionally able to provide a refinement over the standard normal approximation.

2.1. Assumptions

In model (1), the regression can be made to include a constant by setting one of the components of the vector X to equal one almost surely. Because such a setting will require special care in our notation, we let $X = (1, \tilde{X}')'$ if X contains a constant and set $\tilde{X} = X$ otherwise. Throughout, for a matrix A, we also let $\|\cdot\|_F$ denote the Frobenius norm $\|A\|_F^2 \equiv \text{trace}\{A'A\}$. Given this notation, we introduce the following assumptions on the data generating process.

Assumption 2.1. (i) $\{Y_i, X_i\}_{i=1}^n$ is i.i.d., satisfying (1) with $E[X\epsilon] = 0$; (ii) $E[\|XX'\|_F^{\nu}] < \infty$ and $E[\|XX'\epsilon^2\|_F^{\nu}] < \infty$ for some $\nu \geq 9$; (iii) $H_0 \equiv E[XX']$ and $\Sigma_0 \equiv E[XX'\epsilon^2]$ are full rank; (iv) for $Z \equiv (\tilde{X}', X'\epsilon, \text{vech}(\tilde{X}\tilde{X}')', \text{vech}(XX'\epsilon^2)')'$, ξ_Z its characteristic function, $\limsup_{\|t\|\to\infty} |\xi_Z(t)| < 1$.[1]

Assumption 2.2. (i) $\{W_i\}_{i=1}^n$ is i.i.d., independent of $\{Y_i, X_i\}_{i=1}^n$ with $E[W] = 0$, $E[W^2] = 1$ and $E[|W|^{\omega}] < \infty$, $\omega \geq 9$; (ii) for $U \equiv (W, W^2)'$, ξ_U its characteristic function, $\limsup_{\|t\|\to\infty} |\xi_U(t)| < 1$.

Assumption 2.1(i) allows for misspecification of the conditional mean function by requiring $E[X\epsilon] = 0$ rather than $E[\epsilon|X] = 0$. In Assumption 2.1(ii), we demand the existence of certain higher order moments of (Y, X) so that the approximate cumulants of T_n are finite. The requirements on the weights $\{W_i\}_{i=1}^n$ in Assumption 2.2(i) are standard in the wild bootstrap literature and satisfied by all commonly used choices of wild bootstrap weights.

Assumptions 2.1(i)–(iii) and 2.2(i) suffice for showing that the approximate cumulants of T_n and of T_n^* under the bootstrap distribution may disagree up to order $O_p(n^{-\frac{1}{2}})$ under misspecification. In order to additionally establish the existence of Edgeworth expansions, however, we also impose Assumptions 2.1(iv) and 2.2(ii). These requirements, also known as Cramer's condition, are standard in the Edgeworth expansion literature (Bhattacharya and Rao, 1976). Unfortunately, this requirement rules out two frequently used wild bootstrap weights: Rademacher random variables and a weighting scheme originally proposed in Mammen (1993). Thus, while our results on approximate cumulants are applicable to these choices of weights, our results on Edgeworth expansions are not.

2.2. Approximate cumulants

In what follows, for notational simplicity, we denote expectations, probability and law statements conditional on $\{Y_i, X_i\}_{i=1}^n$ (but not $\{W_i\}_{i=1}^n$) by E^*, P^* and L^* respectively. Additionally, we define the following parameters which play a fundamental role in our higher order analysis:

$$\sigma^2 \equiv c'H_0^{-1}\Sigma_0H_0^{-1}c \qquad \kappa \equiv E[(c'H_0^{-1}X)^3\epsilon^3]$$
$$\gamma_0 \equiv E[(c'H_0^{-1}X)^2X\epsilon] \qquad \gamma_1 \equiv E[(c'H_0^{-1}X)(X'H_0^{-1}X)\epsilon]. \tag{6}$$

Finally, we let Φ denote the distribution of a standard normal random variable and ϕ its density.

We begin our analysis by obtaining an asymptotic expansion for T_n and T_n^*.

[1] For a symmetric matrix A, vech (A) denotes a column vector composed of its unique elements.