Inference Based on the Wild Bootstrap

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

email: **jgm@econ.queensu.ca**

Ottawa

September 14, 2012

The Wild Bootstrap

Consider the linear regression model

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + u_i, \quad \mathrm{E}(u_i^2) = \sigma_i^2, \quad i = 1, \ldots, n. \tag{1}$$

One natural way to bootstrap this model is to use the **residual bootstrap**. We condition on $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}$, and the empirical distribution of the residuals (perhaps transformed). Thus the bootstrap DGP is

$$y_i^* = \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + u_i^*, \quad u_i^* \sim \mathrm{EDF}(\hat{u}_i). \tag{2}$$

Strong assumptions! This assumes that $\mathrm{E}(y_i \,|\, \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ and that the error terms are IID, which implies homoskedasticity.

At the opposite extreme is the **pairs bootstrap**, which draws bootstrap samples from the joint EDF of $[y_i, \;\; \boldsymbol{X}_i]$. This assumes that there exists such a joint EDF, but it makes no assumptions about the properties of the u_i or about the functional form of $\mathrm{E}(y_i \,|\, \boldsymbol{X}_i)$.

The wild bootstrap is in some ways intermediate between the residual and pairs bootstraps. It assumes that $\mathrm{E}(y_i \,|\, \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, but it allows for heteroskedasticity by conditioning on the (possibly transformed) residuals.

If no restrictions are imposed, wild bootstrap DGP is

$$y_i^* = \boldsymbol{X}_i \hat{\boldsymbol{\beta}} + f(\hat{u}_i) v_i^*, \tag{3}$$

where $f(\hat{u}_i)$ is a transformation of the i^{th} residual $\hat{u}_i$, and v_i^* has mean 0. Thus $\mathrm{E}\big(f(\hat{u}_i)v_i^*\big) = 0$ even if $\mathrm{E}\big(f(\hat{u}_i)\big) \neq 0$. Common choices:

$$\begin{aligned}
\text{w1:} \quad & f(\hat{u}_i) = \sqrt{n/(n-k)}\,\hat{u}_i, \\
\text{w2:} \quad & f(\hat{u}_i) = \frac{\hat{u}_i}{(1-h_i)^{1/2}}, \\
\text{w3:} \quad & f(\hat{u}_i) = \frac{\hat{u}_i}{1-h_i}.
\end{aligned}$$

Here h_i is the i^{th} diagonal element of the “hat matrix” $\boldsymbol{P_X} \equiv \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. The w1, w2, and w3 transformations are analogous to the ones used in the HC1, HC2, and HC3 covariance matrix estimators.

We would like functions of the bootstrap error terms $f(\hat{u}_i)v_i^*$, such as $n^{-1/2}\boldsymbol{X}^\top \boldsymbol{u}^*$, to have properties similar to those of the same functions of the actual error terms.

Ideally, the bootstrap error terms would have the same moments as the transformed residuals. For that to be the case, we need

$$\mathrm{E}(v_i^*) = 0,\ \ \mathrm{E}(v_i^{*2}) = 1,\ \ \mathrm{E}(v_i^{*3}) = 1,\ \ \mathrm{E}(v_i^{*4}) = 1. \tag{4}$$

But this is impossible!

Consider the outer product of the vector $[1\ \ v\ \ v^2]^\top$ with itself for a random variable v with expectation zero:

$$\mathrm{E}\begin{bmatrix} 1 & v & v^2 \\ v & v^2 & v^3 \\ v^2 & v^3 & v^4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \sigma^2 \\ 0 & \sigma^2 & \mu_3 \\ \sigma^2 & \mu_3 & \mu_4 \end{bmatrix}. \tag{5}$$

Determinant must be nonnegative since the matrix is positive semidefinite:

$$\sigma^2\mu_4 - \mu_3^2 - \sigma^6 \geq 0. \tag{6}$$

But $1 * 1 - 1^2 - 1^6 = -1$. If $\sigma^2 = 1$ and $\mu_3 = 1$, then $\mu_4 \geq 2$. So there exists no distribution for the v_i^* that satisfies (4).

This means that there is no “ideal” distribution for the v_i^*. We either need to relax the requirement that $\mu_3 = 1$ or allow $\mu_4 \geq 2$.

Most common choice for v_i^* is **Mammen's two-point distribution**:

$$v_i^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}). \end{cases} \tag{7}$$

It was suggested in Mammen (1993). In this case,

$$\mathrm{E}(v_i^*) = 0, \;\; \mathrm{E}(v_i^{*2}) = 1, \;\; \mathrm{E}(v_i^{*3}) = 1, \;\; \mathrm{E}(v_i^{*4}) = 2. \tag{8}$$

Thus (6) is satisfied as an equality. No distribution that has the correct third moment can have a fourth moment smaller than 2.

Mammen must have obtained his distribution by solving the equations

$$\begin{aligned} p_1 v_1 + (1-p_1)v_2 &= 0, \\ p_1 v_1^2 + (1-p_1)v_2^2 &= 1, \\ p_1 v_1^3 + (1-p_1)v_2^3 &= 1. \end{aligned} \tag{9}$$

The result is $p_1 = (\sqrt{5}+1)/(2\sqrt{5})$, $v_1 = -(\sqrt{5}-1)/2$, and $v_2 = (\sqrt{5}+1)/2$, which leads to (7).

Besides getting the fourth moment wrong, Mammen's distribution involves two very different probabilities (0.72361 and 0.27639). Thus, about 72% of the time, the sign of the bootstrap error term for observation i will be the opposite of the sign of the residual.

Davidson and Flachaire (2008) proposed the **Rademacher distribution**:

$$v_i^* = \begin{cases} -1 & \text{with probability } \frac{1}{2}, \\ 1 & \text{with probability } \frac{1}{2}, \end{cases} \tag{10}$$

for which

$$\mathrm{E}(v_i^*) = 0, \;\; \mathrm{E}(v_i^{*2}) = 1, \;\; \mathrm{E}(v_i^{*3}) = 0, \;\; \mathrm{E}(v_i^{*4}) = 1. \tag{11}$$

This has the desired fourth moment, and each bootstrap error is positive with probability one-half, which is appealing. But it imposes symmetry.

If the error terms really are symmetric, it is clearly good to impose symmetry. Even if they are not, getting μ_4 right may well be more important than getting μ_3 wrong. D&F provide evidence, and see below.

Using the Rademacher distribution means conditioning on $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}$, and the absolute values of the (transformed) residuals.

Alternatives to Two-Point Distributions

Two-point distributions seem unnatural, as each observation can only have two bootstrap error terms associated with it. In the usual case, this means that there are only 2^n possible bootstrap samples.

Since the standard normal distribution has mean 0 and variance 1, it may seem natural to use it for v^*. But $\mu_3 = 0$ and $\mu_4 = 3$. So its fourth moment is worse than for Mammen, and it has the same, sometimes undesirable, symmetry property as Rademacher.

Mammen (1993) also suggested the continuous distribution:

$$v_i^* = u_i/\sqrt{2} + \tfrac{1}{2}(w_i^2 - 1), \tag{12}$$

where u_i and w_i are independent standard normals. [There is a serious typo in the article, which makes it look as if $u_i = w_i$.] For this distribution,

$$\mathrm{E}(v_i^*) = 0,\ \ \mathrm{E}(v_i^{*2}) = 1,\ \ \mathrm{E}(v_i^{*3}) = 1,\ \ \mathrm{E}(v_i^{*4}) = 6. \tag{13}$$

This gets the third moment right, but the fourth moment is extremely large.

Mammen also suggests another, similar, distribution that is more complicated than (12) and has a slightly smaller fourth moment.

Estimating Covariance Matrices

Bootstrap methods are sometimes used to estimate standard errors and covariance matrices. If $\hat{\boldsymbol{\beta}}_j^*$ is the estimate for the j^{th} bootstrap sample, and $\bar{\boldsymbol{\beta}}^*$ denotes the average of the $\hat{\boldsymbol{\beta}}_j^*$, then the usual estimator is

$$\widehat{\mathrm{Var}}^*(\hat{\beta}) = \sum_{j=1}^{B} (\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)(\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)^\top. \tag{14}$$

Evidently,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^* &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{u}_j^*) - \bar{\boldsymbol{\beta}}^* \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}_j^* + (\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^*).\end{aligned} \tag{15}$$

If the OLS estimator is unbiased, then $\mathrm{E}(\hat{\boldsymbol{\beta}}_j^*) = \hat{\boldsymbol{\beta}}$. Thus we can ignore $\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^*$ if B is large enough.

The first term in the last line of (15) times itself transposed is

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}_j^*\boldsymbol{u}_j^{*\top}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{16}$$

This looks like a sandwich covariance matrix, but with $\boldsymbol{u}_j^*\boldsymbol{u}_j^{*\top}$ instead of a diagonal matrix.

Because $\mathrm{E}(v_i^*)^2 = 1$, diagonal elements of $\boldsymbol{u}_j^* \boldsymbol{u}_j^{*\top}$ have expectation $f^2(\hat{u}_i)$. For Rademacher, these diagonal elements are precisely $f^2(\hat{u}_i)$.

Off-diagonal elements must have expectation zero because $\mathrm{E}(v_i^* v_j^*) = 0$. For Rademacher, each off-diagonal element is the product of the same two transformed residuals multiplied by $+1$ or -1.

Thus, as B becomes large, the matrix $\boldsymbol{X}^\top \boldsymbol{u}_j^* \boldsymbol{u}_j^{*\top} \boldsymbol{X}$ should converge to the matrix $\boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X}$, where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ diagonal matrix with the squares of the $f(\hat{u}_i)$ on the diagonal.

When the transformation $f(\cdot)$ is w1, w2, or w3, the bootstrap covariance matrix estimator (14) converges to HC1, HC2, or HC3 as $B \to \infty$.

Conclusion: Using the wild bootstrap to estimate covariance matrices is just an expensive way to approximate various HCCMEs, with unnecessary simulation randomness.

- Pointless for making inferences about linear regression models.
- Might be useful for obtaining covariance matrices for nonlinear functions of those coefficients.
- Might be useful for nonlinear regression models.

Similar arguments apply to using the pairs bootstrap.

Bootstrap Testing

Consider the heteroskedasticity-robust t statistic

$$\tau(\hat{\beta}_l - \beta_l^0) = \frac{\hat{\beta}_l - \beta_l^0}{\sqrt{\big[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\big]_{ll}}}. \tag{17}$$

To calculate wild bootstrap P value, estimate (1) under the null hypothesis to obtain $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$. Then generate B bootstrap samples, using the DGP

$$y_i^* = \boldsymbol{X}_i\tilde{\boldsymbol{\beta}} + f(\tilde{u}_i)v_i^*. \tag{18}$$

As in (3), there are several choices for the transformation $f(\cdot)$.

For each bootstrap sample, calculate $\tau(\hat{\beta}_{lj}^*)$, the bootstrap analog of (17):

$$\tau(\hat{\beta}_{lj}^* - \beta_l^0) = \frac{\hat{\beta}_{lj}^* - \beta_l^0}{\sqrt{\big[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}_j^*\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\big]_{ll}}}. \tag{19}$$

$\hat{\beta}_{lj}^*$ is the OLS estimate for the $j^{\,\text{th}}$ bootstrap sample. $\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}_j^*\boldsymbol{X}$ is computed in the same way as $\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}$, but uses residuals from bootstrap regression.

It is easier, especially if performing several tests, to use unrestricted estimates rather than restricted ones in the bootstrap DGP.

- If so, use (3) instead of (18) to generate the bootstrap data.
- Compute bootstrap statistics $\tau(\hat{\beta}^*_{lj} - \hat{\beta}^l)$ instead of $\tau(\hat{\beta}^*_{lj} - \beta^l_0)$. That is, replace β^l_0 by $\hat{\beta}^l$ in (19). This is essential, since bootstrap test statistics must test a hypothesis that is true for the bootstrap data.

It is almost always better to use restricted estimates in the bootstrap DGP, because the DGP is estimated more efficiently when true restrictions are imposed; see Davidson and MacKinnon (1999).

However, using restricted estimates is a lot more work.

- With unrestricted estimates, we simply generate B bootstrap samples and use them for all tests and confidence intervals.
- With restricted estimates, we have to generate B bootstrap samples for each restriction we wish to test.
- With restricted estimates, each confidence interval requires $(m_1 + m_2)B$ bootstrap samples, where m_1 and m_2 depend on how many iterations it takes to locate each end of the interval.

Bootstrap *P* Values

Many authors talk about bootstrap critical values. In practice, there is no reason ever to compute one. Bootstrap P values yield exactly the same test results and normally provide far more information.

Provided B is chosen so that $\alpha(B+1)$ is an integer, we can estimate the level α critical value for a test in the upper tail as number $(1-\alpha)(B+1)$ in the (ascending) sorted list of the $\tau_j^* \equiv \tau(\hat{\beta}_{lj}^* - \beta_0^l)$.

Example: If $B = 999$ and $\alpha = 0.05$ for a one-tailed test, the critical value we want is number 950 in the sorted list.

Upper-tail bootstrap *P* value:

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\tau_j^* > \hat{\tau}). \tag{20}$$

Use this for hetero-robust F tests or any other test that rejects only in the upper tail. Choose B such that $\alpha(B+1)$ is an integer.

Notice that $\hat{p}^*(\hat{\tau}) < 0.05$ whenever $\hat{\tau}$ is greater than number $(1-\alpha)(B+1)$ in the sorted list of the τ_j^*.

Equal-tail bootstrap *P* value:

$$\hat{p}^*(\hat{\tau}) = 2\min\biggl(\frac{1}{B}\sum_{j=1}^{B}\mathrm{I}(\tau_j^* \le \hat{\tau}),\ \ \frac{1}{B}\sum_{j=1}^{B}\mathrm{I}(\tau_j^* > \hat{\tau})\biggr). \tag{21}$$

Use this whenever we want to reject in both tails, and the distribution of τ is not symmetric around zero. Choose B such that $\alpha(B+1)/2$ is an integer.

Symmetric bootstrap *P* value:

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B}\mathrm{I}\bigl(|\tau_j^*| > |\hat{\tau}|\bigr). \tag{22}$$

Use this if the distribution of $\hat{\tau}$, and hence also of τ_j^*, is (approximately) symmetric around zero. It should yield slightly better finite-sample properties when that is the case. Choose B such that $\alpha(B+1)$ is an integer.

Simultaneous Equations and the Wild Bootstrap

Just about the simplest simultaneous equations model is

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}_1 \tag{23}$$

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{u}_2. \tag{24}$$

Here (23) is the structural equation of interest, and (24) is an unrestricted reduced form equation.

Davidson and MacKinnon (2010) discusses several wild bootstrap procedures for testing the hypothesis that $\beta = \beta_0$.

The best of these methods is **wild restricted efficient** (or **WRE**) bootstrap. Bootstrap DGP:

$$y_{1i}^* = \beta_0 y_{2i}^* + \boldsymbol{Z}_i\tilde{\boldsymbol{\gamma}} + f_1(\tilde{u}_{1i})v_i^* \tag{25}$$

$$y_{2i}^* = \boldsymbol{W}_i\tilde{\boldsymbol{\pi}} + f_2(\tilde{u}_{2i})v_i^*, \tag{26}$$

where $\tilde{\boldsymbol{\gamma}}$ and the residuals $\tilde{u}_{1i}$ come from an OLS regression of $\boldsymbol{y}_1 - \beta_0\boldsymbol{y}_2$ on $\boldsymbol{Z}$, $\tilde{\boldsymbol{\pi}}$ comes from an OLS regression of $\boldsymbol{y}_2$ on $\boldsymbol{W}$ and $\tilde{\boldsymbol{u}}_1$, and $\tilde{\boldsymbol{u}}_2 \equiv \boldsymbol{y}_2 - \boldsymbol{W}\tilde{\boldsymbol{\pi}}$.

The WRE bootstrap DGP has three important features:

- Reduced-form equation (26) is estimated efficiently, by including structural residuals as an additional variable.
- The same random variable v_i^* multiplies the transformed residuals for both equations. Thus correlation between structural and reduced-form residuals is retained by bootstrap error terms.
- Structural equation (25) uses restricted (OLS) estimates instead of unrestricted (2SLS) ones, which are not necessarily too small.

Bootstrap tests of hypotheses about β based on the WRE bootstrap perform remarkably well, whenever the sample size is not too small (400 seems to be sufficient) and the instruments are not very weak.

What mostly causes asymptotic tests to perform poorly is simultaneity combined with weak instruments, not heteroskedasticity. The main reason to use the WRE bootstrap is to compensate for the weak instruments.

It is also asymptotically valid to use a nonrobust test statistic together with the wild bootstrap, or a robust test statistic together with a bootstrap method that does not take account of heteroskedasticity.

Cluster-Robust Covariance Matrices

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_m \end{bmatrix} \equiv \boldsymbol{X\beta} + \boldsymbol{u}.$$

There are m clusters, indexed by j, stacked into a vector $\boldsymbol{y}$ and a matrix $\boldsymbol{X}$. If $\hat{\boldsymbol{u}}_j$ denotes the vector of OLS residuals for the j^{th} cluster, a **cluster-robust covariance matrix estimator** has the form

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{j=1}^{m} \boldsymbol{X}_j^\top \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^\top \boldsymbol{X}_j\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{27}$$

- Sandwich form similar to HCCME, but with more complicated filling.
- Robust to heteroskedasticity and within-cluster correlation.

Cameron, Gelbach, and Miller (2008) propose the wild bootstrap DGP:

$$y_{ji}^* = \boldsymbol{X}_{ji}\hat{\boldsymbol{\beta}} + f(\hat{u}_{ji})v_j^*, \tag{28}$$

where j indexes clusters, i indexes observations, and the v_j^* come from the Rademacher distribution.

Problem! For any two-point distribution, number of distinct bootstrap samples is just 2^m; see Webb (2012).

Here are some values of 2^m:

$$2^5 = 32;\ 2^6 = 64;\ 2^7 = 128;\ 2^8 = 256;\ 2^9 = 512;\ 2^{10} = 1024.$$

When $m = 5$, each set of bootstrap samples will contain each of the 32 possible ones, repeated various numbers of times.

Webb (2012) suggests a six-point distribution. Here are some values of 6^m:

$$6^5 = 7,776;\ 6^6 = 46,656;\ 6^7 = 279,936;\ 6^8 = 1,679,616.$$

This would evidently solve the problem, at least for $m \geq 7$. Webb's distribution has six mass points:

$$-\sqrt{1.5},\ -1,\ -\sqrt{0.5},\ \sqrt{0.5},\ 1,\ \sqrt{1.5},$$

each of which has probability 1/6. It is easy to see that:

$$\mathrm{E}(v_i^*) = 0,\ \ \mathrm{E}(v_i^{*2}) = 1,\ \ \mathrm{E}(v_i^{*3}) = 0,\ \ \mathrm{E}(v_i^{*4}) = 7/6. \tag{29}$$

Simulation Experiments

Past simulation experiments in Davidson and Flachaire (2008), MacKinnon (2011), and other papers collectively suggest that:

- Rademacher distribution outperforms Mammen's two-point distribution, even when the error terms are not symmetric.
- We should always generate bootstrap samples using parameter estimates and residuals under the null.
- All methods work well when the h_i are similar in magnitude, that is, when there are no observations with high leverage.
- All methods improve rapidly as $n \to \infty$ when the largest h_i converge rapidly as the sample size increases.
- Some methods can work quite poorly even for quite large samples if there remain high-leverage observations in those samples.
- Some methods can be considerably more powerful than others. Use H1? Hausman and Palmer (2012) suggests that wild bootstrap methods may less powerful than a technique they propose.

New experiments are based on the ones in MacKinnon (2011), with two changes. The error terms are now skewed (rescaled, recentered $\chi^2(5)$) rather than standard normal, and there are more choices for the v_i^*.

The DGP is

$$y_i = \beta_1 + \sum_{k=2}^{5} \beta_k X_{tk} + u_i, \quad u_i = \sigma_i \varepsilon_i, \quad \varepsilon_i \sim \mathrm{IID}(0,1), \tag{30}$$

where all regressors are drawn randomly from the **standard lognormal distribution**, parameters are $\beta_k = 1$ for $k \leq 4$, $\beta_5 = 0$, and

$$\sigma_i = z(\gamma)\big(\beta_1 + \sum_{k=2}^{5} \beta_k X_{tk}\big)^{\gamma}. \tag{31}$$

Here $z(\gamma)$ is a scaling factor chosen so that the average variance of u_i is equal to 1.

In the experiments, $0 \leq \gamma \leq 2$. Note that $\gamma = 0$ implies homoskedasticity, and $\gamma >> 1$ implies rather extreme heteroskedasticity.

Features of the DGP:

- Chosen to make heteroskedasticity-robust inference difficult.
- Lognormal regressors imply that many samples will contain a few observations on the X_{tj} that are quite extreme.
- The most extreme observation in each sample will tend to become more so as the sample size increases.
- Largest value of h_i tends to be large and to decline very slowly as $n \to \infty$.

Alternative approach:

- Choose a fixed or random $\boldsymbol{X}$ matrix for a small sample size and form larger samples by repeating it as many times as necessary.
- Only as many distinct values of h_i as number of observations in the original sample. All of those values must be proportional to $1/n$.
- Since $h_i^{\max}$ declines like $1/n$, heteroskedasticity-robust inference improves rapidly as n increases.

My approach is probably too pessimistic, but approach that constructs $\boldsymbol{X}$ by repetition, which was used by MacKinnon and White (1985), Davidson and Flachaire (2008), and many other papers, is much too optimistic.

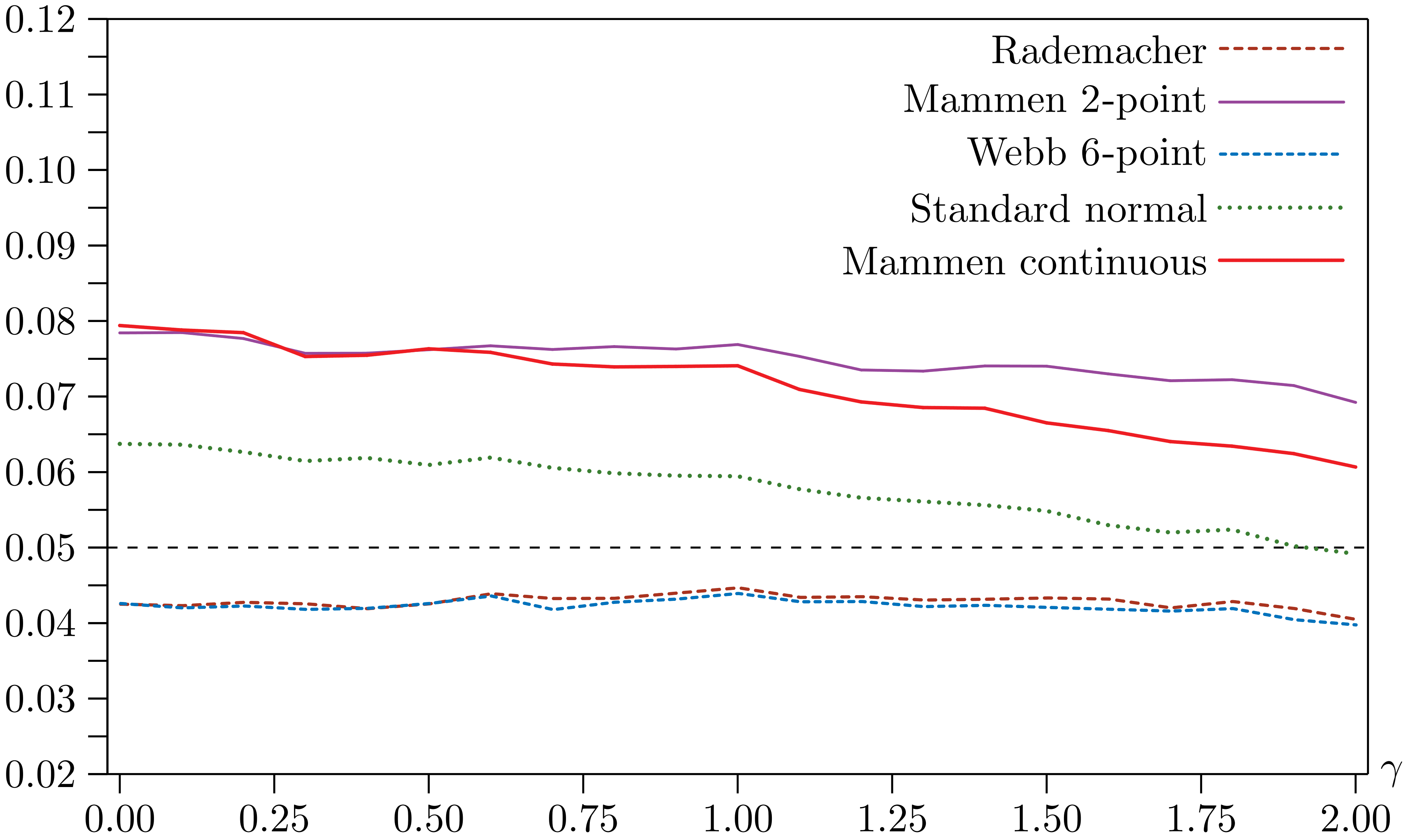

Figure 1. Rejection frequencies for wild (w3) bootstrap HC1 t tests, $n = 40$

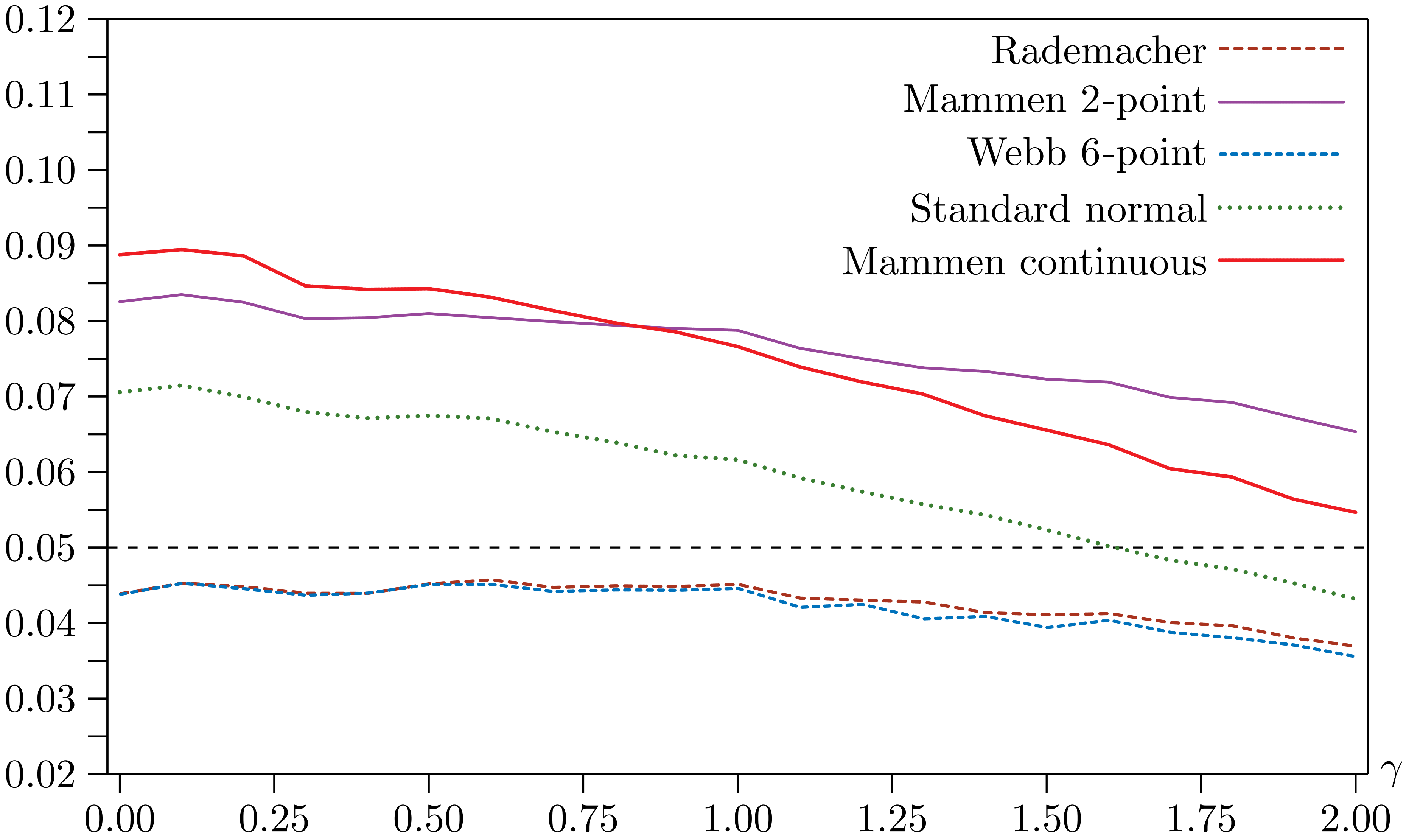

Figure 2. Rejection frequencies for wild (w3) bootstrap HC3 t tests, $n = 40$

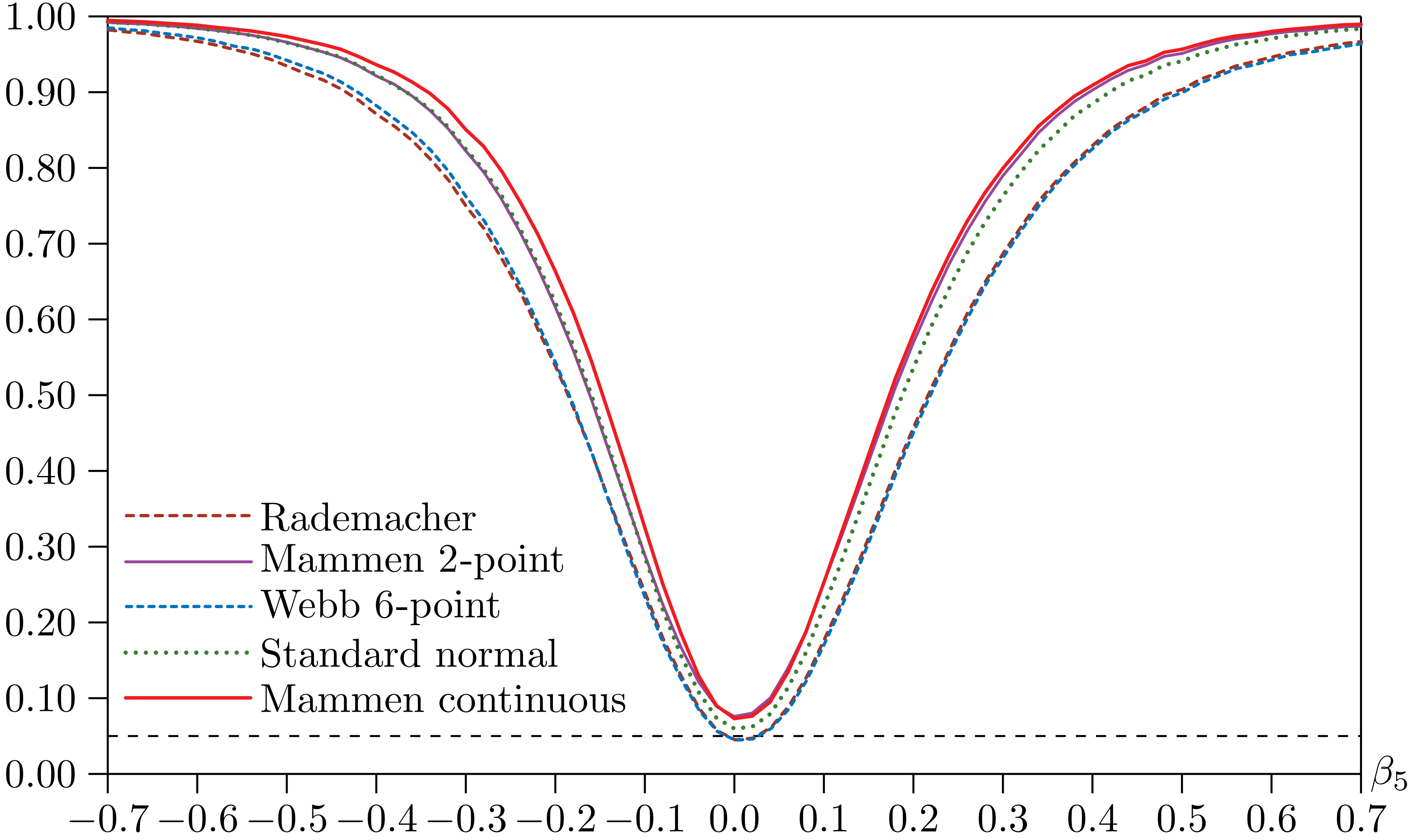

Figure 3. Power of wild bootstrap hetero-robust HC1 t tests, $\gamma = 1$, $n = 40$

The Score Wild Bootstrap

For any Type 2 maximum likelihood estimator,

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\iota}^\top \boldsymbol{G}(\hat{\boldsymbol{\theta}}) = \mathbf{0},$$

where $\boldsymbol{g}(\hat{\boldsymbol{\theta}})$ is the **score vector** and the $n \times k$ matrix $\boldsymbol{G}(\hat{\boldsymbol{\theta}})$ contains the contributions to the scores. Hu and Kalbfleisch (2000) and Klein and Santos (2011) suggested bootstrapping the rows of $\boldsymbol{G}(\hat{\boldsymbol{\theta}})$. The former proposed the **estimating function bootstrap**, which resamples the rows like the pairs bootstrap. The latter proposed the **score wild bootstrap**.

In many cases, the score wild bootstrap can be implemented very easily via an **artificial regression**. In general, such a regression has the form

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{b} + \boldsymbol{Z}(\boldsymbol{\theta})\boldsymbol{c} + \text{residuals}, \tag{32}$$

where $\boldsymbol{\theta}$ is a k–vector, $\boldsymbol{r}(\boldsymbol{\theta})$ is a column vector of length n (or maybe $2n$), and $\boldsymbol{R}(\boldsymbol{\theta})$ is a matrix with as many rows as $\boldsymbol{r}(\boldsymbol{\theta})$ and k columns. The $n \times r$ matrix $\boldsymbol{Z}(\boldsymbol{\theta})$ depends on the same data and parameters as $\boldsymbol{r}(\boldsymbol{\theta})$ and $\boldsymbol{R}(\boldsymbol{\theta})$.

Crucial properties: Score vector is $\boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta})$, and $\frac{1}{n}\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta})$ converges to the information matrix $\mathcal{I}(\boldsymbol{\theta})$. See Davidson and MacKinnon (2001).

When evaluated at any root-n consistent estimator $\acute{\boldsymbol{\theta}}$, (32) becomes

$$\acute{\boldsymbol{r}} = \acute{\boldsymbol{R}}\boldsymbol{b} + \acute{\boldsymbol{Z}}\boldsymbol{c} + \text{residuals}, \tag{33}$$

where $\acute{\boldsymbol{r}} \equiv \boldsymbol{r}(\acute{\boldsymbol{\theta}})$ and so on. The obvious test statistic is the reduction in the SSR associated with $\acute{\boldsymbol{Z}}$, which is

$$\acute{\boldsymbol{r}}^\top \boldsymbol{M}_{\acute{\boldsymbol{R}}} \acute{\boldsymbol{r}} - \acute{\boldsymbol{r}}^\top \boldsymbol{M}_{[\acute{\boldsymbol{R}}\ \ \acute{\boldsymbol{Z}}]} \acute{\boldsymbol{r}}. \tag{34}$$

As usual, this can be written as

$$\acute{\boldsymbol{r}}^\top \boldsymbol{M}_{\acute{\boldsymbol{R}}} \acute{\boldsymbol{Z}} (\acute{\boldsymbol{Z}}^\top \boldsymbol{M}_{\acute{\boldsymbol{R}}} \acute{\boldsymbol{Z}})^{-1} \acute{\boldsymbol{Z}}^\top \boldsymbol{M}_{\acute{\boldsymbol{R}}} \acute{\boldsymbol{r}}, \tag{35}$$

and it is asymptotically distributed as $\chi^2(r)$. In the case of LM tests, we evaluate everything at $\tilde{\boldsymbol{\theta}}$. Now $\tilde{\boldsymbol{r}}$ is orthogonal to $\tilde{\boldsymbol{R}}$, so the first term in (34) is just $\tilde{\boldsymbol{r}}^\top \tilde{\boldsymbol{r}}$, and the test statistic (34) reduces to

$$\tilde{\boldsymbol{r}}^\top \boldsymbol{P}_{[\tilde{\boldsymbol{R}}\ \ \tilde{\boldsymbol{Z}}]} \tilde{\boldsymbol{r}}, \tag{36}$$

which is just the ESS from regression (33).

Artificial Regression Wild and Pairs Bootstraps

We hold $\acute{\boldsymbol{R}}$ and $\acute{\boldsymbol{Z}}$ fixed and draw wild bootstrap samples by multiplying $\acute{r}_i$, the $i^{\rm th}$ element of $\acute{\boldsymbol{r}}$, by v_i^*. In the case of the Rademacher distribution, this just means randomly changing the sign of $\acute{r}_i$ with probability one-half.

Note that $\boldsymbol{R}_j^{*\top}\boldsymbol{R}_j^*$, $\boldsymbol{R}_j^{*\top}\boldsymbol{Z}_j^*$, and $\boldsymbol{Z}_j^{*\top}\boldsymbol{Z}_j^*$ are identical in all bootstrap samples. Only $\boldsymbol{r}_j^{*\top}\boldsymbol{R}_j^*$ and $\boldsymbol{r}_j^{*\top}\boldsymbol{Z}_j^*$ vary. This should make computation of the **wild artificial regression bootstrap** extremely inexpensive.

We can also resample from the rows of $[\acute{\boldsymbol{r}}\;\;\acute{\boldsymbol{R}}\;\;\acute{\boldsymbol{Z}}]$ to obtain the **pairs artificial regression bootstrap**.

In either case, we run the two artificial regressions

$$\boldsymbol{r}_j^* = \boldsymbol{R}_j^*\boldsymbol{b} + \text{residual, and} \tag{37}$$

$$\boldsymbol{r}_j^* = \boldsymbol{R}_j^*\boldsymbol{b} + \boldsymbol{Z}_j^*\boldsymbol{c} + \text{residual,} \tag{38}$$

and the bootstrap statistic is

$$\boldsymbol{r}_j^{*\top}\boldsymbol{M}_{\boldsymbol{R}_j^*}\boldsymbol{r}_j^* - \boldsymbol{r}_j^{*\top}\boldsymbol{M}_{[\boldsymbol{R}_j^*\;\;\boldsymbol{Z}_j^*]}\boldsymbol{r}_j^*. \tag{39}$$

This all seems very simple, elegant, and inexpensive.

Problem: It does not actually work!

Example: Nonnested binary response models.

Suppose we have two competing binary response models:

$$H_1: \;\; \mathrm{E}(y_t \,|\, \Omega_t) = F_1(\boldsymbol{X}_{1t}\boldsymbol{\beta}_1) \quad \text{and} \tag{40}$$

$$H_2: \;\; \mathrm{E}(y_t \,|\, \Omega_t) = F_2(\boldsymbol{X}_{2t}\boldsymbol{\beta}_2), \tag{41}$$

One way to nest H_1 and H_2 in an artificial compound model is:

$$H_C: \;\; \mathrm{E}(y_t \,|\, \Omega_t) = (1-\alpha)F_1(\boldsymbol{X}_{1t}\boldsymbol{\beta}_1) + \alpha F_2(\boldsymbol{X}_{2t}\boldsymbol{\beta}_2). \tag{42}$$

To test H_1 against H_C, we first replace $\boldsymbol{\beta}_2$ by its ML estimate $\hat{\boldsymbol{\beta}}_2$ and then construct an artificial regression to test the null hypothesis that $\alpha = 0$. This artificial regression is

$$\hat{V}_t^{-1/2}(y_t - \hat{F}_{1t}) = \hat{V}_t^{-1/2}\hat{f}_{1t}\boldsymbol{X}_{1t}\boldsymbol{b} + a\hat{V}_t^{-1/2}(\hat{F}_{2t} - \hat{F}_{1t}) + \text{residual}, \tag{43}$$

where

$$V_t(\boldsymbol{\beta}) \equiv F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big). \tag{44}$$

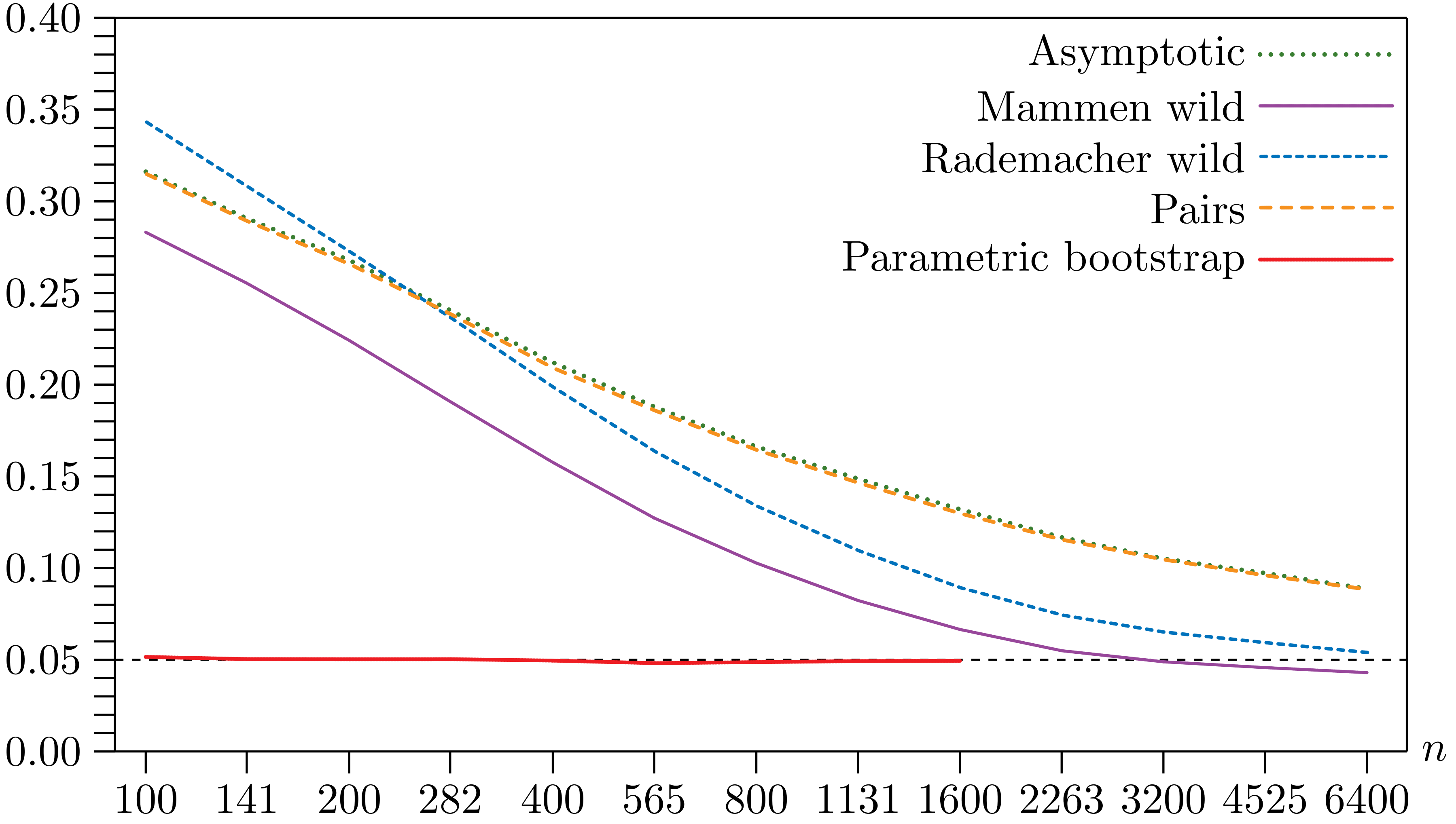

Figure 4. Rejection frequencies for tests of nonnested logit models, $\beta = 1$

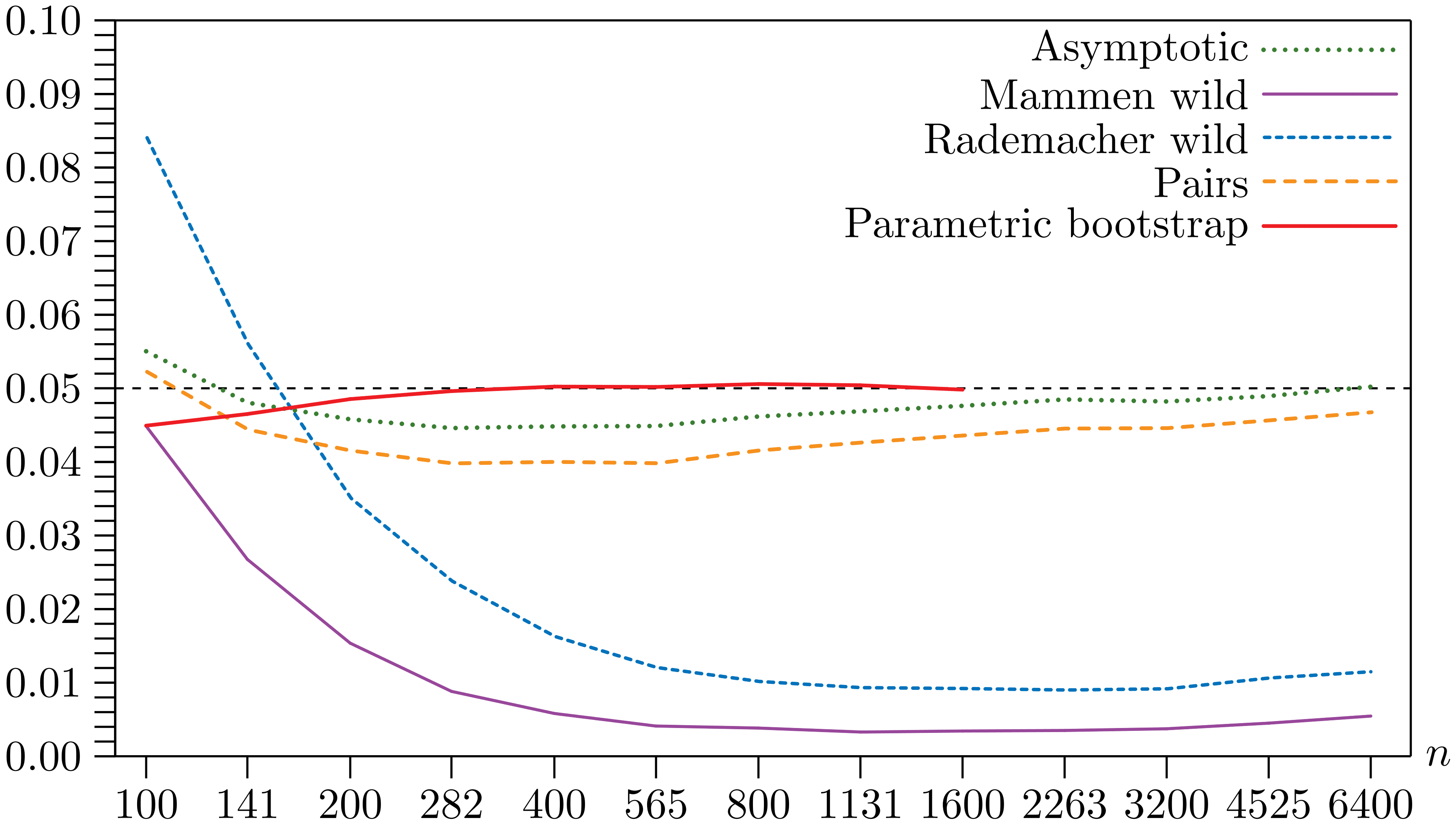

Figure 5. Rejection frequencies for tests of nonnested logit models, $\beta = 2$

- Asymptotic test can overreject or underreject, often severely, depending on number of regressors in each model and how well H_1 fits.
- Overrejection when H_1 fits poorly is expected. Underrejection when H_1 fits well is a mystery.
- Parametric bootstrap works extremely well. Note that graphs stop at $n = 1600$ to save computer time.
- Wild artificial regression bootstrap tests can work better or worse than asymptotic tests.
- Pairs artificial regression bootstrap tests mimic asymptotic tests.
- Mammen wild bootstrap rejects less frequently than Rademacher wild bootstrap, whether asymptotic test is overrejecting or underrejecting.
- If we use the Rademacher distribution, we are assuming that the $\acute{r}_i$ are symmetrically distributed, which is surely not true. This seems to be a case where we might want to preserve the third moment.

Tentative conclusion: The score wild bootstrap, the artificial regression wild bootstrap, and the estimating function bootstrap are completely useless.

Explanation:

These are all just computationally expensive ways to compute approximations to asymptotic quantities! If the resulting bootstrap tests happen to perform better than asymptotic tests, it is just coincidence.

Consider the wild artificial regression bootstrap based on the Rademacher distribution. It conditions on $\acute{\boldsymbol{\theta}}$, on all of the regressors in (33), and on the absolute values of the elements of $\acute{\boldsymbol{r}}$.

The only random elements of the bootstrap DGP are the signs of the elements of the $\boldsymbol{r}_j^*$. Thus the bootstrap expectations of $\acute{\boldsymbol{R}}^\top\boldsymbol{r}_j^*$ and $\acute{\boldsymbol{Z}}^\top\boldsymbol{r}_j^*$ are indeed zero.

The distribution of the bootstrap statistic (39), conditional on the regressors $\acute{\boldsymbol{R}}$ and $\acute{\boldsymbol{Z}}$, is that of any chi-squared test with exogenous regressors. It fails to be exactly chi-squared only because the $\acute{\boldsymbol{r}}$, and hence the $\boldsymbol{r}_j^*$, are not Gaussian.

The distribution of the bootstrap test statistic is valid asymptotically, but it tells us nothing about the finite-sample properties of the true DGP!

Summary

The wild bootstrap can be a very valuable addition to the econometrician's toolkit.

Whenever error terms are independent but not identically distributed, resampling residuals is not valid.

Instead, the wild bootstrap conditions on the observed (rescaled) residuals. In the case of Rademacher, it conditions on their absolute values.

With samples of moderate size, we can make very accurate inferences by combining an HCCME with a suitably chosen wild bootstrap DGP.

There is limited evidence in favour of combining H1, which is equivalent to H0 when we bootstrap, with w3 based on restricted estimates.

The wild bootstrap can also be useful with clustered data and for simultaneous equations models.

Bootstrap methods based on contributions to the scores, whether pairs or wild, are fundamentally misguided.

References

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

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). “The size distortion of bootstrap tests,” *Econometric Theory*, 15, 361–376.

Davidson, R., and J. G. MacKinnon (2001). “Artificial regressions,” in *Companion to Theoretical Econometrics*, ed. B. Baltagi, Oxford, Blackwell, 16–37.

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

Hausman, J. A., and C. Palmer (2012). “Heteroskedasticity-robust inference in finite samples,” *Economics Letters*, 116, 232–235.

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

Kline, P. M., and A. Santos (2011). “A score based approach to wild bootstrap inference,” *Journal of Econometric Methods*, forthcoming.

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

MacKinnon, J. G. (2011). “Thirty years of heteroskedasticity-robust inference,” QED Working Paper No. 1268.

MacKinnon, J. G., and H. White (1985). “Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties,” *Journal of Econometrics*, 29, 305–325.

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

Webb, M. (2012). “Reworking wild bootstrap based inference for clustered errors,” paper presented at the 2012 Annual Meeting of the Canadian Economics Association.