The Wild Bootstrap

With heteroskedasticity of unknown form, the residual bootstrap is inappropriate. It forces bootstrap disturbances to be homoskedastic. The pairs bootstrap is asymptotically valid, but the wild bootstrap will usually provide more reliable inferences. It is designed for regression models with heteroskedasticity of unknown form.

The wild bootstrap conditions each value $y_i^*$ not only on $X_i$ but also on the residual for observation $i$, which is either $\tilde{u}_i$ or $\hat{u}_i$.

For a linear regression model with heteroskedastic disturbances, the wild bootstrap DGP is

$$y_i^* = X_i \hat{\beta} + u_i^*, \quad u_i^* = v_i^* \tilde{u}_i.$$

(1)

Here $\hat{\beta}$ denotes either restricted estimates $\tilde{\beta}$ or unrestricted estimates $\hat{\beta}$, and $v_i^*$ is an auxiliary random variable with mean 0 and variance 1.
Because \(v_i^*\) has variance 1, \(\text{Var}(u_i^*) = \text{Var}(\tilde{u}_i)\).
Thus \(\text{Var}(u_i^*)\) will, on average, be large for observations with large residuals and small for ones with small residuals.

We saw when discussing heteroskedasticity-robust inference that

\[
\hat{u}_i^2 = \omega_i^2 + v_i,
\]

(2)

where \(\omega_i^2\) is the variance of the \(i^{th}\) disturbance.

This suggests that \(\tilde{u}_i^2\) can be used to estimate \(\omega_i^2\). Of course, \(\tilde{u}_i^2\) is a very noisy estimator, but that does not matter asymptotically.

On average, the wild bootstrap DGP (1) preserves the variance of the original disturbances conditional on \(X\).

Ideally, auxiliary random variable \(v_i^*\) should have mean 0, variance 1, and all higher moments equal to 1. If so, \(u_i^*\) has same moments as \(\tilde{u}_i\).

Unfortunately, there exists no distribution with these properties!
The best choice usually seems to be the **Rademacher distribution**. It takes on just two values, 1 and $-1$, each with equal probability. Third and fourth moments of the Rademacher distribution are 0 and 1, respectively. Because the third moment is 0, the $u_i^*$ must be symmetric. Mammen (1993) suggested another two-point distribution that has a third moment of 1, but its fourth moment is 2. It is

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

Rademacher seems to outperform Mammen, even when the $u_i$ are asymmetric; see Davidson and Flachaire (2008) and Djogbenou, MacKinnon, and Nielsen (2019).

Another possibility is the $\text{N}(0, 1)$ distribution, which has first three moments 0, 1, 0. But its fourth moment is 3, which is much too large. A better choice is probably the uniform $\text{U}(-\sqrt{3}, \sqrt{3})$ distribution, which has fourth moment 1.8.
It is common to replace $\bar{u}_i$ in the bootstrap DGP (1) by $\psi(\bar{u}_i)$, where $\psi(\cdot)$ is a monotonically increasing transformation. This is to undo tendency for least squares residuals to be too small.

There are four natural transformations:

- $w_0$: $\psi(\bar{u}_i) = \hat{u}_i$ or $\tilde{u}_i$; \hspace{1cm} (4)
- $w_1$: $\psi(\bar{u}_i) = \sqrt{N/(N-k)}\hat{u}_i$ or $\sqrt{N/(N-k_1)}\tilde{u}_i$; \hspace{1cm} (5)
- $w_2$: $\psi(\bar{u}_i) = \frac{\hat{u}_i}{(1-h_i)^{1/2}}$ or $\frac{\tilde{u}_i}{(1-\tilde{h}_i)^{1/2}}$; \hspace{1cm} (6)
- $w_3$: $\psi(\bar{u}_i) = \frac{\hat{u}_i}{1-h_i}$ or $\frac{\tilde{u}_i}{1-\tilde{h}_i}$ . \hspace{1cm} (7)

The $i^{th}$ diagonal element of the “hat matrix” is denoted $h_i$ for the unrestricted model and $\tilde{h}_i$ for the restricted one.

Note that $w_0$ and $w_1$ yield identical bootstrap $t$ statistics.
Consider the hypothesis $\beta_2 = 0$ in the linear regression model

$$y = X_1 \beta_1 + \beta_2 x_2 + u, \quad \mathbb{E}(u) = 0, \quad \mathbb{E}(uu^\top) = \Omega,$$

where $\Omega$ is diagonal with typical diagonal element $\omega_i^2$.

We use the heteroskedasticity-robust $t$ statistic

$$t_2 = \frac{\hat{\beta}_2}{\text{s.e.}(\hat{\beta}_2)},$$

where s.e.$(\hat{\beta}_2)$ is the square root of the appropriate diagonal element of any valid HCCME $\hat{\text{Var}}_h(\hat{\beta})$.

There are both theoretical reasons and simulation evidence to suggest that we should impose the restriction $\beta_2 = 0$ on the bootstrap DGP (D&M ET 1999; D&F 2008; JGM 2013; DMN 2019).

Procedure for wild restricted (WR) bootstrap to test $\beta_2 = 0$ is:
1. Estimate (8) by OLS to obtain $\hat{\beta}_2$ and s.e.($\hat{\beta}_2$), which is hetero-robust, and use them to calculate $t_2$ in (9).

2. Re-estimate (8) subject to the restriction that $\beta_2 = 0$ in order to obtain restricted estimates $\tilde{\beta}$ and residuals $\tilde{u}_i$.

3. Generate $B$ bootstrap samples $y^*_b$ using the DGP

$$y^*_b = X_{1i} \tilde{\beta}_1 + v^*_b \psi(\tilde{u}_i). \quad (10)$$

4. For every bootstrap sample, calculate

$$t^*_b = \frac{\hat{\beta}^*_b}{\text{s.e.}(\hat{\beta}^*_b)}. \quad (11)$$

5. Calculate the symmetric bootstrap $P$ value

$$\hat{p}^*(t_2) = \frac{1}{B} \sum_{b=1}^{B} I(|t^*_b| > |t_2|). \quad (12)$$
Of course, the bootstrap $P$ value does not have to be symmetric.

- Can use any of HC$_1$ through HC$_3$ to calculate s.e.($\hat{\beta}_2$) and s.e.($\hat{\beta}^*_b$).
- Any of HC$_1$ through HC$_3$ can be combined with any of the $w_1$ through $w_3$ transformations.
- When some observations have much more leverage than others and $N$ is not quite large, it is wise to use $w_2$ or $w_3$ in the bootstrap DGP.

To obtain studentized bootstrap confidence intervals, we need to use an unrestricted wild bootstrap DGP, with $t^*_b = (\hat{\beta}^*_b - \hat{\beta}_2) / \text{s.e.}(\hat{\beta}^*_b)$.

Just which procedure works best in any given case depends on the $X$ matrix, the pattern of heteroskedasticity, and how they interact.

A very similar procedure can be used to bootstrap Wald tests. Asymptotic Wald tests often work badly when $r >> 1$.

For $N$ not small, asymptotic inference can work very well. If so, the wild bootstrap and asymptotic $P$ values should be very similar.
The Wild Cluster Bootstrap

With clustered disturbances, we can use either pairs cluster bootstrap or wild cluster bootstrap (CGM 2008, MW 2017, DMN 2019).

If data for $g^{th}$ cluster is $Z_g = [y_g \ X_g]$, each pairs cluster bootstrap sample consists of $G$ random drawings from the $Z_g$ matrices, stacked to form a matrix $Z^*$.

Unless $N_g$ is the same for all $g$, bootstrap samples will have varying numbers of observations. So $N^{*b}$ varies across $b$.

Wild cluster bootstrap uses same drawing of $v^*$ for every observation within each cluster. The bootstrap DGP is

$$y^*_{gi} = X_{gi} \hat{\beta} + u^*_{gi}, \quad u^*_{gi} = v^*_g \hat{u}_{gi},$$

where $g$ indexes clusters and $i$ indexes observations within each cluster. Entire vector of residuals for cluster $g$ is multiplied by $v^*_g$. 
Bootstrap samples mimic the pattern of intra-cluster correlation of the residuals as well as their conditional variances.

DMN (2019) proves that the WCB yields valid inferences asymptotically when $G \rightarrow \infty$, but perhaps more slowly than $N \rightarrow \infty$. There are limitations on the moments and heterogeneity of the scores $s_{gi} = X_{gi}^{\top}u_{gi}$, and cluster sizes are not allowed to vary too much. In particular, there cannot be one big cluster.

WCR (restricted) usually outperforms WCU (unrestricted) bootstrap.

- Number of distinct bootstrap samples if we are using any two-point distribution is only $2^G$. For $G = 8$, this is just 256.
- Webb (2014) proposed a symmetric six-point distribution with fourth moment $7/6$ for use in such cases.
- Clearly, $6^G$ is very much larger than $2^G$; for example, $6^8 = 1,679,616$.
- Using the six-point distribution solves the problem of too few distinct bootstrap samples for values of $G$ as small as about 5.
Cluster-robust inference can fail when $G_1$, the number of “treated” clusters, is small. Regressor of interest is a treatment dummy, say $d_{gi}$. $d_{gi}$ may be 1 only for some observations (DiD case) or all observations (pure treatment case) within each treated cluster.

Suppose $\theta$ is the coefficient on a treatment dummy where $G_1$ is small. MW (2017, 2018) showed that:

- The CR standard error for $\hat{\theta}$ is much too small, so CR $t$ statistic grossly over-rejects the null that $\theta = 0$.
- When $G_1$ is very small, WCR bootstrap grossly under-rejects the null that $\theta = 0$. But it can over-reject for smallish values.
- When $G_1$ is small, WCU bootstrap severely over-rejects the null that $\theta = 0$. It is as bad as the CR $t$ statistic when $G_1 = 1$.
- The ordinary wild bootstrap (WR) can work much better than the WCR bootstrap when $G_1$ is very small.
- Randomization inference (MW, 2020) can also work well.
The figure shows rejection frequencies at .05 level for four tests in a pure treatment model. $G = 14$, $N/G = 200$ for all $g$, and $\rho = 0.10$.

- Number of treated clusters, $G_1$, varies from 1 to 13.
- Vertical axis has been subjected to a square root transformation.
- Rejection frequencies are based on 400,000 replications.
- Using hetero-robust (HC$_2$) $t$ statistics, plus $t(2798)$ distribution, results in massive over-rejection for all values of $G_1$.
- This would have been worse if $N/G$ or $\rho$ had been larger.
- Using cluster-robust (CV$_1$) $t$ statistics, plus $t(13)$ distribution, leads to severe over-rejection when $G_1 = 1$ and $G_1 = 13$.
- Over-rejection is much less severe when $G_1$ is not too far from $G/2$.
- WCU bootstrap over-rejects dreadfully when $G_1 = 1$ and $G_1 = 13$.
- WCR bootstrap under-rejects dreadfully when $G_1 = 1$ and $G_1 = 13$. 
WCR bootstrap usually works much better than CR \( t \) tests when \( G \) is small, and/or clusters are heterogeneous (including \( N_g \) varying a lot).

- The wild cluster bootstrap can be very cheap to compute; see Roodman, MacKinnon, Nielsen, and Webb (2019).
- Efficient computation is built into the `boottest` package for Stata.
- Cost of bootstrapping using `boottest` is \( O(BG^2) \) instead of \( O(BNk) \).
- This allows `boottest` to compute WCR bootstrap confidence intervals as well as WCR bootstrap \( P \) values.

In principle, we could transform the residuals before generating the bootstrap samples, as we should do with heteroskedasticity.

In particular, \( wc_2 \) is to \( CV_2 \) as \( w_2 \) is to \( HC_2 \), so that \( wc_2 \) can be expected to work somewhat better than \( wc_1 \).

But the transformation has to be applied to the entire vector \( \tilde{u}_g \) (or \( \hat{u}_g \)), and any procedure that creates and inverts \( M_{gg} \) matrices will be extremely expensive, or even infeasible, if \( N_g \) is large.
Power Loss from Bootstrapping

The power of a bootstrap test may, or may not, differ substantially from the power of the asymptotic test on which it is based.

A test based on the ideal bootstrap $P$ value ($B = \infty$) may reject more or less often than the corresponding asymptotic test (D&M 2006).

Generally, bootstrap tests appear to have less power, but only because the corresponding asymptotic test over-rejects.

A bootstrap test based on finite $B$ must reject less often than one based on $B = \infty$, although in many cases the power loss is negligible.

When $B$ is finite, $\hat{p}^*$ differs from $p^*$ because of random variation in bootstrap samples. Adding randomness to $p^*$ is equivalent to adding randomness to $\tau$, and so it necessarily reduces the power of the test. Power loss due to finite $B$ is $O(1/B)$; see D&M (2000).
Consider $z_{\beta_2}$ and $t_{\beta_2}$ for the classical normal linear model.

$z_{\beta_2}$ follows the $N(0, 1)$ distribution, because $\sigma$ is known, and $t_{\beta_2}$ follows the $t(N - k)$ distribution, because $\sigma$ is estimated.

$t_{\beta_2}$ is equal to $z_{\beta_2}$ times the random variable $\sigma / s$, which is independent of $z_{\beta_2}$ and the same for both $H_0$ and $H_1$.

- Multiplying $z_{\beta_2}$ by $\sigma / s$ adds independent random noise.
- This requires us to use a larger critical value, which in turn causes the test based on $t_{\beta_2}$ to be less powerful than the test based on $z_{\beta_2}$.

The figure illustrates power loss in going from $z_{\beta_2}$ to $t_{\beta_2}$, plus additional power loss from bootstrapping with finite $B$.

Power loss is very rarely a problem when $B = 999$, and it is never a problem when $B = 9,999$.

For confidence intervals, randomness due to finite $B$ shows up as intervals that are longer than necessary.
Power Loss from Bootstrapping

![Graph showing power loss from bootstrapping with curves for different distributions: N(0, 1), t(9), B = 99, B = 19. The x-axis represents \( \beta_2 / \sigma \) and the y-axis represents power.](https://via.placeholder.com/150)