Life After Bootstrap: Residual Randomization Inference in Regression Models

Panagiotis (Panos) Toulis
panos.toulis@chicagobooth.edu

Econometrics and Statistics
University of Chicago, Booth School of Business

Setup

We focus on the quintessential regression model:

$$y = X\beta + \varepsilon.$$

- $y \in \mathbb{R}^n$ is the response; X is the $n \times p$ covariate matrix ($n > p$).
- $\varepsilon \in \mathbb{R}^n$ are unobserved errors (no assumption yet).

We wish to do inference on $\beta \in \mathbb{R}^p$ with minimal assumptions.

The standard approaches

Parametric approach:

- Posit a model for ε, derive some $\hat{\beta}$ (e.g., OLS). Use CLT.

Bootstrap approach:

- Resample $(y, X) \rightarrow$ bootstrap distribution of $\hat{\beta}$. Use CLT.
- Alternatively: fix X, and resample $\hat{\varepsilon}$ (residuals). This is known as residual bootstrap (Freedman and Lane, 1983).

Both approaches:

- Require some form of exchangeability.
- Cannot easily handle complex error structures.
- They rely on "nice asymptotic behavior" of $\hat{\beta}$.

What's wrong with bootstrap?

Nothing, per se. The bootstrap is one of the most important statistical tools.

However, it is based on uniform resampling, and so it does not work in cases with complex error structure without extensive modifications.

This is why we have the 'bootstrap zoo':

- residual bootstrap (Freedman and Lane, 1983).
- wild bootstrap (Wu, 1986).
- cluster wild bootstrap (Cameron, 2008).
- block bootstrap (Politis & Romano, 1992).
- pigeonhole bootstrap (Owen, 2007).
- ...

In other words, bootstrap starts with the procedure, then accommodates the particular error structure.

It should be the other way around!

Complex error structures

In practice, (regression) errors may have a complex dependency.

Inference is typically based on invariance assumptions on these errors.

Many forms of invariances. Errors may be:

- exchangeable (e.g., when generated under identical conditions).
- non exchangeable but sign-symmetric.
- clustered and independent across clusters but not within.
- doubly-clustered.
- autocorrelated.
- ...

Addressing complex error structures

Our proposal puts the error invariance assumption first. We assume, in particular, that there is a group of transformations $\mathcal{G}$ s.t.

$$\varepsilon \stackrel{d}{=} \mathrm{g}\varepsilon \mid X, \text{ for all } \mathrm{g} \in \mathcal{G}.$$

Reminiscent of the "structure of inference" (Fraser, 1960).

Naive bootstrapping no longer works because $\mathcal{G}$ may have a complex structure (e.g., clustering).

The framework of randomization tests is exactly what we need to proceed (Lehman and Romano, 2005). Field of active research: (Rosenbaum, 2010), (Imbens and Rubin, 2015), (Gerber and Green, 2012), (Athey et. al, 2019), (Ding et. al., 2014, 2017), (Canay et. al., 2017, 2019), (Wu and Ding, 2018), (Basse et al., 2019, 2020).

Some examples of $\mathcal{G}$

- Exchangeability: $(\varepsilon_1, \ldots, \varepsilon_n) \stackrel{d}{=} (\varepsilon_{\pi(1)}, \ldots, \varepsilon_{\pi(n)})$, where π denotes (random) permutation. Then,

$$\mathrm{g} = \sum_{i=1}^{n} 1_i 1'_{\pi(i)}, \quad \pi \sim \text{random permutation.}$$

- Sign symmetry: $(\varepsilon_1, \ldots, \varepsilon_n) \stackrel{d}{=} (\pm\varepsilon_1, \ldots, \pm\varepsilon_n)$. Then,

$$\mathrm{g} = \begin{bmatrix} \pm 1 & & 0 \\ & \ddots & \\ 0 & & \pm 1 \end{bmatrix} = \sum_{i=1}^{n} s_i 1_i 1'_i, \quad s_i \sim \text{random sign.}$$

We can easily derive their properties; e.g., $E(G) = 0$ and $\mathrm{Var}(G) = I$, for random signs, where $G \sim \mathrm{Unif}(\mathcal{G})$.

Randomization Tests (Lehman and Romano, 2005)

Let $D \in \mathbb{R}^n$ be the data, and $\mathcal{G}$ a group of $\mathbb{R}^n \times \mathbb{R}^n$ transformations.
We are testing some H_0 under which:

$$D \overset{d}{=} \mathsf{g}D, \text{ for all } \mathsf{g} \in \mathcal{G}.$$

Define a test statistic $T_n = t_n(D)$ and $\mathsf{T}_D = \{t_n(\mathsf{g}D) : \mathsf{g} \in \mathcal{G}\}$. Then,

$$T_n \mid \mathsf{T}_D = \text{ Uniform.}$$

To test H_0, we could take the p-value of T_n wrt to T_D.

$*$ This test is (i) exact in finite samples and (ii) works for any choice of T_n.

Example: permutation test (Fisher, 1935)

Suppose we have iid data $D_1 \sim P$ and $D_2 \sim Q$. We want to test

$$H_0 : P = Q.$$

Take $\mathcal{G}$ to be the set of permutations of (D_1, D_2), and choose a test statistic $T_n = t_n(D_1, D_2)$. [anything that quantifies distance of $D_1 - D_2$]

To test H_0, compare the observed value of T_n with $\mathsf{T}_D = \{t_n(D_1', D_2')\}$ where $(D_1', D_2') = \mathsf{g}(D_1, D_2)$ are derived from permutations of the combined dataset.

Note: The name "randomization test" is somewhat unfortunate. The method does not only work in randomized experiments, but is more generally applicable.

Taking this idea further

We extend the randomization test for inference in the model

$$y = X\beta + \varepsilon.$$

We focus on simple linear hypotheses:

$$H_0 : \lambda_1\beta_1 + \ldots \lambda_p\beta_p = \lambda_0, \text{ or } \lambda'\beta = \lambda_0, \text{ for short.}$$

This includes significance tests, $\beta_j = 0$. Also leads to confidence intervals via test inversion.

Our analysis will be conditional on X (as in residual bootstrap).

Main Idea

Our approach relies on two key ideas:

1. Inferential primitive. We assume:

$$\varepsilon \overset{d}{=} \mathsf{g}\varepsilon \mid X, \text{ for all } \mathsf{g} \in \mathcal{G},$$

where $\mathcal{G}$ is the inferential primitive, a group of $\mathbb{R}^n \to \mathbb{R}^n$ linear operators. [chosen by the analyst].

2. Invariant. Test statistic T_n such that for known function $t_n : \mathbb{R}^n \to \mathbb{R}$

$$T_n \overset{H_0}{=} t_n(\varepsilon).$$

Then, standard theory of (Lehman and Romano, 2005) suggests that we can test H_0 by comparing T_n with $\mathsf{T}_\varepsilon = \{t_n(\mathsf{g}\varepsilon) : \mathsf{g} \in \mathcal{G}\}$.

Promises four main benefits compared to the bootstrap:

1. Address the inference problem in a unified way, while bootstrap typically needs to be adapted to the task.
2. In particular, the same procedure will be applied to different problems. Only $\mathcal{G}$ needs to be defined for each problem.
3. Does not rely on "nice" asymptotic behavior; e.g., consistency of test statistic is not required. Leads to weaker conditions.
4. May be valid in finite samples.

BUT...this test is infeasible because ε are unknown.

The feasible procedure needs to rely on residuals, and is approximate.

Outline

1. Feasible procedure.
2. Main theoretical results on validity.
3. Clustered errors.
4. Two-way clustering.
5. Autocorrelated errors.
6. Conclusion.
7. (extra, if time) High-Dimensional regression. regression.

Residual Randomization: Concrete procedure

1. Calculate the restricted OLS estimate:

$$\hat{\beta}^{\mathrm{o}} = \arg\min_{b} ||y - Xb||^2, \quad \text{such that} \quad \lambda' b = \lambda_0.$$

   Calculate the corresponding restricted residuals, $\hat{\varepsilon}^{\mathrm{o}} = y - X\hat{\beta}^{\mathrm{o}}$.

2. Test statistic, $T_n = (\lambda'\hat{\beta} - \lambda_0)$, and let $T_n = t$ be the observed value. Implies $t_n(u) = \lambda'(X^\top X)^{-1} X^\top u$.

3. Generate $\mathsf{T}_R = \{t_n(G_r\hat{\varepsilon}^{\mathrm{o}}) : G_r \sim \mathsf{Unif}(\mathcal{G}), r = 1, \ldots, R\}$.

4. Calculate p-value: $\widehat{\mathsf{pval}} = E(\mathsf{T}_R \geq t)$.

At target level $\alpha \in (0, 1)$, the test decision is:

$$\phi_n(y; X) = \mathbb{I}\{\widehat{\mathsf{pval}} \leq \alpha\}.$$

Example: Hormone data (Efron & Tibshirani, 1996)

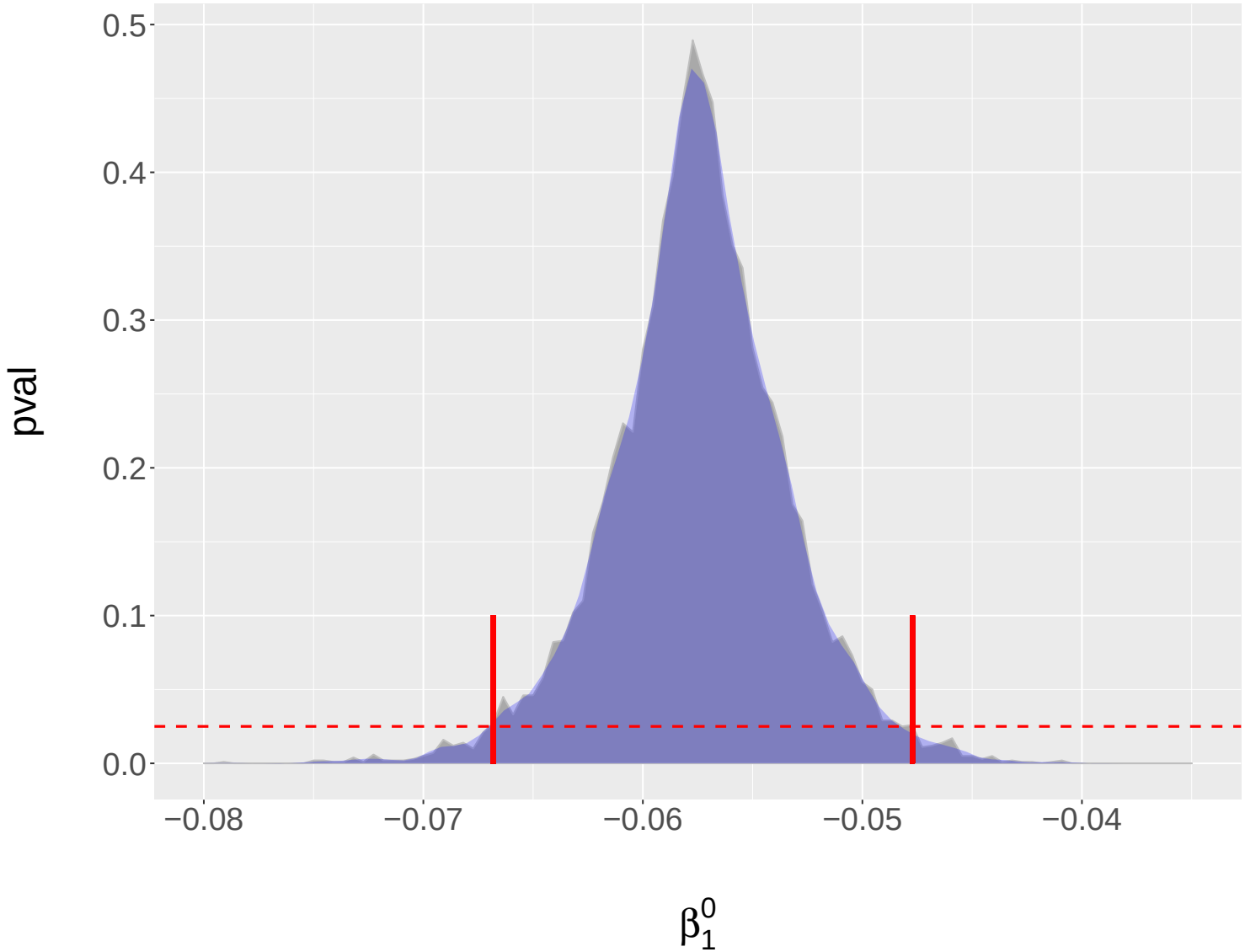
Consider the following regression model:

$$\underbrace{y_i}_{\text{hormone\_level}_i} = \beta_0 + \beta_1 \underbrace{x_i}_{\text{hrs\_device}_i} + \varepsilon_i.$$

Goal is to do inference on β_1 (suppose $\bar{x} = 0$).

To test $H_0 : \beta_1 = b$, the residual randomization method:

1. Calculates OLS estimates, $\hat{\beta}_0, \hat{\beta}_1$.
2. Uses $T_n = (\hat{\beta}_1 - b) \overset{H_0}{=} \frac{\sum_i (\varepsilon_i - \bar{\varepsilon}) x_i}{\sum_i x_i^2} \triangleq t_n(\varepsilon)$.
3. Calculates restricted residuals $\hat{\varepsilon}_i^{\text{o}} = \tilde{y}_i - \overline{\tilde{y}_i}$, where $\tilde{y}_i = y_i - b x_i$.
4. Compare T_n with $\{t_n(\text{g}\hat{\varepsilon}^{\text{o}}) : ...\}$.



Caption. Histogram of p-values for a sequence of tests, $H_0 : \beta_1 = \beta_1^0$. The horizontal dashed line marks the 0.025 threshold for the two-sides test. The two vertical lines mark the range of values for β_1 for which H_0 cannot be rejected.

	inference method	midpoint estimate	s.e.	95% interval
	OLS	-0.0574	$.0045$	$(-0.0665, -0.0482)$.
	bootstrap	-0.0574	$.0043$	$(-0.0660, -0.0488)$
	permutations	-0.0573	$.0048$	$(-0.0668, -0.0477)$
$\mathcal{G}$	random signs	-0.0595	$.0045$	$(-0.0686, -0.0504)$
	permutations, within	-0.0609	$.0043$	$(-0.0695, -0.0522)$
	signs, cluster	–	–	–
	double	-0.0582	0.0050	$(-0.0682, -0.0482)$

Also, a flexible way for sensitivity analysis by trying many different invariances.

$\mathcal{G}$="permutations" assumes exchangeable errors;
$\mathcal{G}$= "random signs" assumes error symmetry around zero;
$\mathcal{G}$="permutations, within" assumes exchangeable errors within clusters defined by the device manufacturer;
$\mathcal{G}$="sign, across" assumes error symmetry around zero on the cluster level;
$\mathcal{G}$="double" assumes both cluster invariances.

Validity

Theorem

Suppose that $X^\top X$ is invertible, and let $G \sim \text{Unif}(\mathcal{G})$. Suppose also that

$$\frac{E\left(||\hat{\beta} - \beta||^2 \mid X\right)^{[1]}}{E\left(\text{Var}(t_n(G\varepsilon) \mid \varepsilon, X) \mid X\right)^{[2]}} E(||(X^\top X)^{-1} X^\top G X - cI||^2)^{[3]} \to 0, \quad (1)$$

for some constant $c \in \mathbb{R}$. Then, residual randomization is asymptotically valid under H_0, that is,

$$\lim_{n\to\infty} \sup E(\phi_n(y; X)|H_0) \leq \alpha.$$

- Under standard conditions, Equation (1) is $O(1/n)$.
- Term [2] depends on the "complexity" of $\mathcal{G}$. Inference will break down if $\mathcal{G}$ is not "complex enough" (e.g., permutes only 2 elements instead of n).
- Term [3] requires $\mathcal{G}$ not to change the "information structure" by much.
- In Term [1], consistency/normality of $\hat{\beta}$ is not necessary!

Comparison to bootstrap

The "bootstrap principle" is that $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ has same distribution as $\sqrt{n}(\hat{\beta} - \beta)$. So, "good asymptotic" behavior of $\hat{\beta}$ is required, in general.

For instance, to do valid residual bootstrap in our setting, Freedman (1981) requires:

1. $\varepsilon_1, \ldots, \varepsilon_n \sim$ i.i.d., with mean 0 and finite variance σ^2.
2. $(1/n)X^\top X \to V$, where V is positive definite.

In residual randomization, validity depends on the interaction between $\mathcal{G}$ and the estimator $\hat{\beta}$. Indeed,

- $\hat{\beta}$ may not be $\sqrt{n}$-consistent.
- $(1/n)X^\top X$ may not even converge.

Robustness

Theorem

Let $G, G_1, G_2 \sim \text{Uniform}(\mathcal{G})$, i.i.d. Suppose that:

1. $E\left(||\hat{\beta} - \beta||^2 \mid X\right) \asymp E\left(\text{Var}(t_n(G\varepsilon) \mid \varepsilon, X) \mid X\right)$. [Terms [1] and [2].]
2. $F_{\bar{\Lambda}_n}(\epsilon) = O(\epsilon^\gamma)$, where $\bar{\Lambda}_n = \frac{|\Lambda_n|}{\text{Var}(\Lambda_n)^{1/2}}$ with $\Lambda_n = t_n(G_1\varepsilon) - t_n(G_2\varepsilon)$.
3. $E(||(X^\top X)^{-1}X^\top GX - cI||^2) = O(c_n^2)$ with $c_n \downarrow 0$. [Term [3].]

Then,

$$E(\phi_n(y; X) \mid H_0) = \alpha + A_\gamma O(c_n^{2\gamma/(2+\gamma)}),$$

where $A_\gamma = O(R^{4/(2+\gamma)})$.

- Condition 2 requires that the distribution of the test statistic does not "degenerate too fast"; e.g., it would be a problem if $t_n(G_1\varepsilon) = t_n(G_2\varepsilon)$ w.h.p.
- Under regular conditions, rate to validity is $O(n^{-1/3})$.
- Term A_γ shows that increasing the number of randomization samples (R) has a negative impact on the rate.

Clustered errors

In many problems the datapoints are clustered. Usually, the errors are assumed independent across clusters, but possibly correlated within.

There are numerous "cluster-robust" error methods but they rely heavily on asymptotics, and have problems with small samples and non-normality.

"Cluster wild bootstrap" (Cameron et al, 2008) is an alternative but works only under strict conditions; cannot be easily extended (e.g., to "two-way clustering").

With J clusters and m units/cluster, the idea is to:

1. Split the residuals in clusters: $\{(e_{11}, \ldots, e_{1m}), \ldots, (e_{J1}, \ldots, e_{Jm})\}$.
2. In a wild bootstrap scheme, flip the signs on the cluster level:

$$\begin{aligned} &\{+(e_{11}, \ldots, e_{1m}), \ldots, -(e_{J1}, \ldots, e_{Jm})\} \\ &\{+(e_{11}, \ldots, e_{1m}), \ldots, +(e_{J1}, \ldots, e_{Jm})\} \\ &\{-(e_{11}, \ldots, e_{1m}), \ldots, -(e_{J1}, \ldots, e_{Jm})\} \\ &\ldots \end{aligned} \tag{1}$$

Which invariance works here?

Residual randomization offers a natural way of inference.

Just assume an invariance on the cluster level.

i.e., define $\mathcal{G}$ as:

- permutations within clusters.
- sign flips across clusters.
- both operations; etc.

Example: Clustered errors

Consider the following regression model:

$$y_i = -1 + 0.2x_i + \varepsilon_i,$$

where $x_i = i/n$, and

$$\varepsilon_i \sim \begin{cases} N(0, 0.1^2) & \text{if } x_i \leq 0.9 \\ N(0, 5^2) & \text{if } x_i > 0.9. \end{cases}$$

Two clusters for the errors. One has much higher variance than the other.

The 95% confidence interval from OLS (with $n = 200$), is:

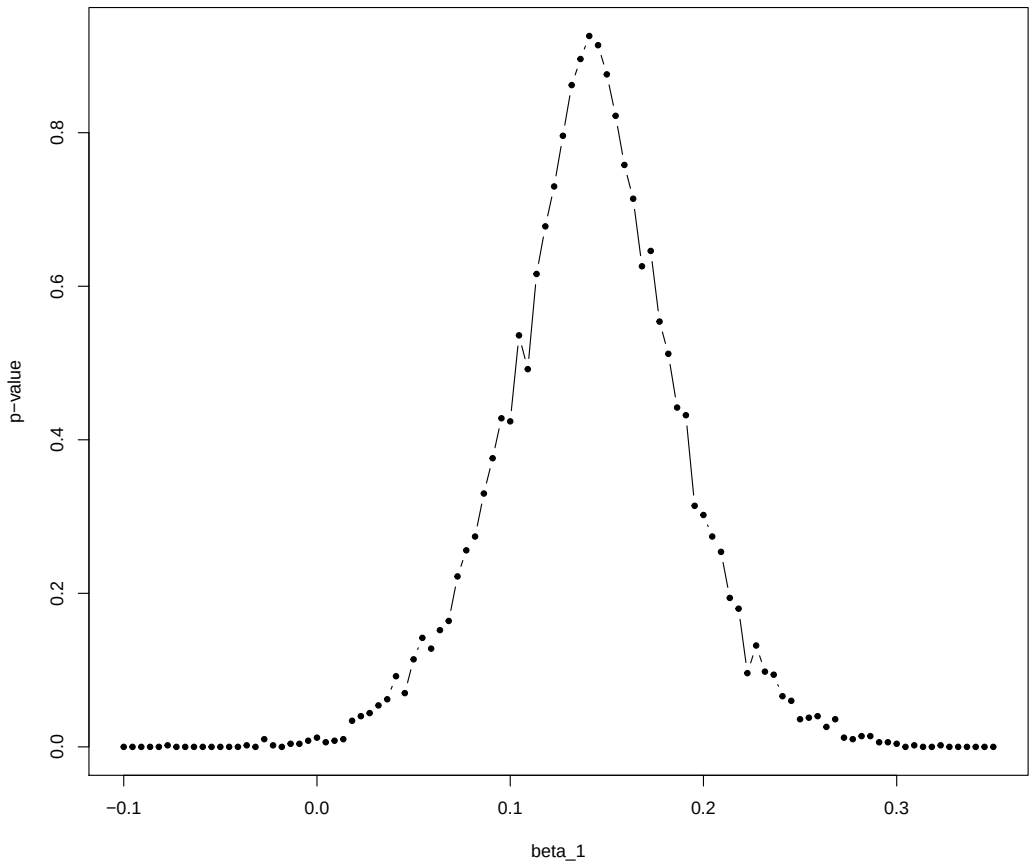
```
> confint(lm(y ~ x))
        2.5 %     97.5 %
X   -0.88       0.50
```

OLS fails badly to detect significance.

Residual randomization with two clusters

Define $\mathcal{G} = \{$"as permutations within each cluster"$\}$. (assume known clustering)

The p-value plot is shown below:



The (inverted) 95% CI is much better centered than regular OLS.

Validity under clustered errors

	Cluster size, $J_n = \lvert \mathrm{C}^n \rvert$	
inferential primitive	$J_n = J < \infty$	$J_n \to \infty$
exchangeability within clusters ($\mathcal{G}^{\mathsf{p}}$)	$X^\top \mathbb{1}/n = 0$, or HA	$X^\top \mathbb{1}/n = 0$, or HA
sign symmetry across clusters ($\mathcal{G}^{\mathsf{s}}$)	HA	HA, or $\sum_{c=1}^{J_n} n_c^2/n^2 \to 0$.
double invariance ($\mathcal{G}^{\mathsf{p}+\mathsf{s}}$)	$X^\top \mathbb{1}/n = 0$, or HA	$X^\top \mathbb{1}/n = 0$, or HA, or $\sum_{c=1}^{J_n} n_c^2/n^2 \to 0$.

"HA" is a homogeneity condition that, in the limit, for every cluster c:

$$(X^\top X)^{-1} X^\top \mathrm{D}_c X \to b_c I,$$

where D_c is diagonal with 1 only for the units in cluster c.

Canay et.al. (2017) showed that under HA the cluster wild bootstrap is asymptotically valid when $J_n < \infty$. But this is generally a strong assumption.

Residual randomization does not require HA for validity.

Finite-sample validity

Suppose that:

1. The errors are sign symmetric across clusters (as defined earlier).
2. Finite-sample HA holds; i.e., for every cluster c it holds that

$$X_c^\top X_c \propto X^\top X,$$

where X_c is covariate matrix in the cluster.
[Important: The clustering may be chosen by the analyst.]

Theorem (Summary)

The cluster-sign residual randomization test (based on $\mathcal{G}^s$) is finite-sample exact under conditions 1 and 2.

Proof sketch: The difference between $t_n(\mathrm{g}\hat{\varepsilon}^0) - t_n(\varepsilon)$ is average of terms

$$\lambda'(X_c^\top X_c)^{-1}(X^\top X)(\hat{\beta} - \beta) \propto \lambda'\hat{\beta} - \lambda_0 \overset{H_0}{=} 0,$$

Example: Behrens-Fisher problem

Angrist and Pischke (2009) and Imbens and Kolesar (2016) studied the following problem:

$$y_i = \beta_0 + \beta_1 d_i + \varepsilon_i,$$

where d_i is binary (treatment or control), and $\text{Var}(\varepsilon_i) = d_i\sigma_1^2 + (1 - d_i)\sigma_0^2$.

There are $n_1 = \sum_i d_i = 3$ treated units, and $n_0 = 27$ controls.

This is an instance of the Behrens–Fisher problem. Standard t-test does not work here because σ_0^2, σ_1^2 are unknown.

No good methods available. Also, very small sample creates problems.

Here, an exact residual randomization test is possible!

Example: Behrens-Fisher problem

Split units in three clusters, each cluster 1 treated unit and 9 controls:
(treated, control) = $(1, 9), (1, 9), (1, 9)$.

1. Assume sign-symmetric errors across clusters.
2. For every cluster c, matrix $X_c^\top X_c$ only depends on proportion of treated units, which is the same for every $c = 1, 2, 3$, by construction!

   So, $X_c^\top X_c \propto X^\top X$ as required.

Note: The resulting randomization test is a cluster sign test with 3 clusters.
Thus, minimum p-value is $1/8 = 0.125$, and so we need to tweak the test (randomize the decision sometimes) to bring it down to 0.05.

Panel (A). True $\beta_1 = 0.0$

	Error type, ε_i											
	normal				t_3				mixture			
	Control standard deviation, σ_0											
Method	0.5	1	2	5	0.5	1	2	5	0.5	1	2	5
BM	0.050	0.028	0.010	0.002	0.034	0.015	0.004	0.000	0.252	0.225	0.034	0.003
r-sign	0.095	0.012	0.000	0.000	0.067	0.012	0.001	0.000	0.213	0.010	0.001	0.000
r-exact	0.048	0.052	0.052	0.050	0.055	0.057	0.054	0.049	0.050	0.046	0.058	0.049

Panel (B). True $\beta_1 = 1.0$

Method	0.5	1	2	5	0.5	1	2	5	0.5	1	2	5
BM	0.215	0.161	0.069	0.008	0.146	0.086	0.028	0.003	0.122	0.130	0.119	0.009
r-sign	0.448	0.149	0.007	0.000	0.270	0.065	0.003	0.000	0.214	0.122	0.004	0.000
r-exact	0.124	0.116	0.111	0.073	0.098	0.101	0.081	0.062	0.094	0.083	0.093	0.073

Panel (C). True $\beta_1 = 2.0$

Method	0.5	1	2	5	0.5	1	2	5	0.5	1	2	5
BM	0.553	0.511	0.359	0.049	0.418	0.332	0.166	0.016	0.326	0.310	0.183	0.055
r-sign	0.899	0.632	0.090	0.000	0.655	0.290	0.032	0.000	0.978	0.673	0.070	0.001
r-exact	0.172	0.177	0.168	0.119	0.147	0.145	0.131	0.089	0.197	0.197	0.173	0.127

Table: Rejection rates of cluster sign test (r-sign), and exact randomization test (r-exact) for the Behrens–Fisher problem. "BM" refers to an adjusted t-test proposed by Imbens and Kolesar (2016) based on the bias correction method of McCaffrey and Bell (2002).

Two-way (or multi-way) clustering

In many problems there are more than two clusters; e.g., (school, classroom), (state, city), (firm, department), etc. "Dyadic regression" falls in this setting.

There are certain variants of "cluster-robust" error methods that have been extended to two-way clustering (Cameron et al, 2011). Other approaches include (Davezies et.al., 2018), (Menzel, 2017), (McKinnon et.al., 2017).

These methods heavily rely on asymptotics, and may give invalid estimates (e.g., non-positive definite covariance estimates).

In addition, the underlying assumptions are restrictive; e.g.,

- McKinnon et.al. (2017) require that the majority of "cells" become empty in the limit.
- (Davezies et.al., 2018) requires that the number of both types of clusters tends to infinity.
- (Menzel, 2017) focuses on estimating marginal expectations and not regression.

Which invariance works here?

Residual randomization can be applied naturally in this setting.

A reasonable assumption is "exchangeability within each individual cluster".

> i.e., define $\mathcal{G}$ = "permutations of entire rows or entire columns".

Example: Dyadic regression

Suppose that datapoint i is in "row-cluster" $r(i)$ and in "column-cluster" $c(i)$.

Consider the dyadic regression model:

$$y_i = \beta_0 + \beta_1 |x_{r(i)} - x_{c(i)}| + \varepsilon_i.$$

For the residual randomization test:

1. Fit constrained OLS and calculate restricted residuals $\hat{\varepsilon}^{\text{o}}$.
2. Arrange the residuals in rows and columns.
3. At every resampling, permute $\hat{\varepsilon}^{\text{o}}$ row-wise and/or column-wise.
4. Use new set of residuals to generate new y and re-fit OLS.
5. Produce the p-value as usual.

Panel (A). True $\beta_1 = 1.0$

	Error-covariate, (ε_i, x_i)											
	(normal, normal)			(normal, lognormal)			(mixture, normal)			(mixture, lognormal)		
	Sample size, n											
	100	625	2500	100	625	2500	100	625	2500	100	625	2500
HC	.320	.167	.118	.392	.330	.238	.322	.172	.131	.437	.414	.311
2way robust	.090	.061	.046	.114	.091	.062	.080	.041	.050	.101	.091	.057
RR	.060	.057	.052	.047	.055	.045	.053	.037	.057	.053	.057	.050

Panel (B). True $\beta_1 = 1.2$

	100	625	2500	100	625	2500	100	625	2500	100	625	2500
HC	.363	.279	.359	.488	.616	.734	.360	.267	.279	.470	.543	.675
2way robust	.150	.537	.981	.301	.788	.997	.144	.524	.983	.286	.775	.997
RR	.075	.134	.252	.155	.372	.609	.079	.144	.245	.157	.390	.601

Table: Rejection rates for HC2 robust errors, two-way robust errors (bootstrap), and the double permutation test in dyadic regression study. Null hypothesis is $H_0 : \beta_1 = 1.0$.

Autocorrelated errors

In panel data, the errors may be autocorrelated:

$$y_t = x_t'\beta + \varepsilon_t.$$

For example, we may have $\varepsilon_t = \rho_t \varepsilon_{t-1} + u_t$, where u_t is iid noise, and $\rho_t \in (0, 1)$ may be non-stationary.

There are several "HAC" methods in the literature for such models (White et al, 1980; Andrews, 1991). Generally they are not robust as they are extensions of "HC" methods with stronger assumptions.

Problems with heavy-tailed data, non-normality, and/or small samples.

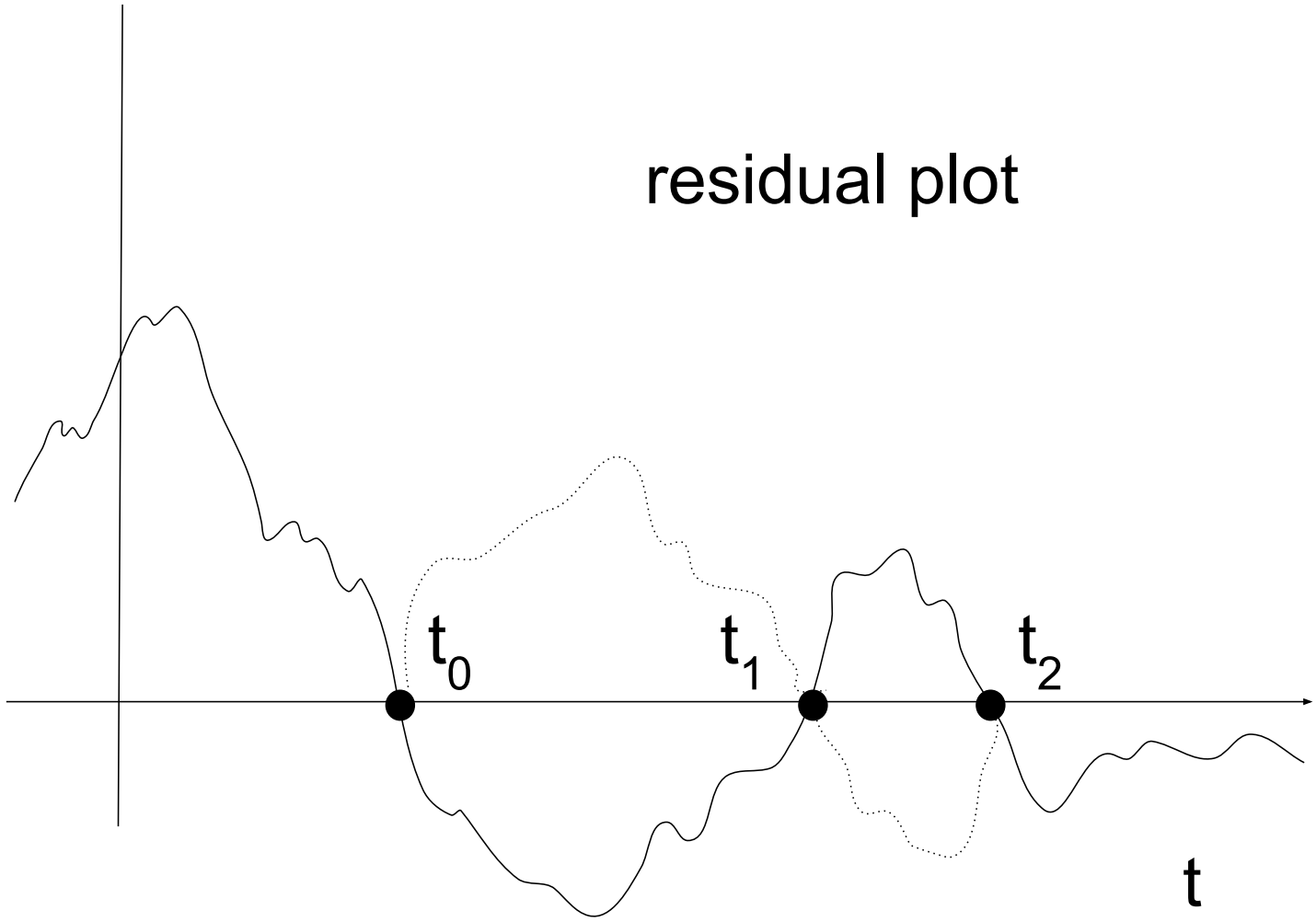
Which invariance works here?

Standard invariance concepts do not work here due to serial dependence.

However, for the AR(1) process:

$$\varepsilon_t \stackrel{d}{=} -\varepsilon_t \mid \{\varepsilon_{t-1} = 0\}.$$

The error series can be reflected around the time axis!



We can reflect the residuals between the endpoints t_j. Call this $\mathcal{G}^{\text{ref}}$.

The "reflection" randomization test

1. Calculate the restricted residuals, $\hat{\varepsilon}^{o}$.
2. Order their absolute values, $|\hat{\varepsilon}^{o}|$, and select the $J+1$ smallest values. Denote the corresponding timepoints as $t_0, ..., t_J$.
3. Define the clustering, $\{\{t_0, ..., t_1\}, \{t_1+1, ..., t_2\}, ..., \{t_{J-1}+1, t_J\}\}$.
4. Perform the cluster sign test based on the clustering from step 3.

+ Does not rely on normality.
+ Can work with non-stationary series.
+ Good empirical performance.

Panel (A): $\rho = 0.3$								
	Error $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, $u_t = \ldots$							
	normal				mixture			
	Covariates x_t							
	iid		autocorrelated		iid		autocorrelated	
Method	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
OLS	0.052	0.054	0.073	0.078	0.053	0.050	0.073	0.071
HAC	0.066	0.112	0.065	0.112	0.066	0.145	0.070	0.130
reflection test, uncond.	0.031	0.030	0.034	0.034	0.045	0.048	0.042	0.042
reflection test, cond.	0.051	0.048	0.054	0.055	0.053	0.057	0.050	0.049

Panel (B): $\rho = 0.8$								
	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
OLS	0.048	0.048	0.341	0.339	0.049	0.050	0.336	0.346
HAC	0.050	0.087	0.104	0.128	0.053	0.097	0.102	0.141
reflection test, uncond.	0.022	0.023	0.024	0.027	0.031	0.029	0.032	0.030
reflection test, cond.	0.049	0.052	0.055	0.061	0.053	0.050	0.052	0.051

Table: Rejection rates for OLS, HAC errors, and the reflection test.

Example: Complex panel data

Suppose panel data

$$y_{it} = x'_{it}\beta + \varepsilon_{it},$$

where

- $x_{it} = \rho_1 x_{i,t-1} +$ LN is autocorrelated with log-normal errors.
- $\varepsilon_{it} = \rho_2 \varepsilon_{i,t-1} + \eta_i +$ N is autocorrelated with random "firm effect" and normal errors.

What method to use here?

Example: Complex panel data

Suppose panel data

$$y_{it} = x'_{it}\beta + \varepsilon_{it},$$

where

- $x_{it} = \rho_1 x_{i,t-1} +$ LN is autocorrelated with log-normal errors.
- $\varepsilon_{it} = \rho_2 \varepsilon_{i,t-1} + \eta_i +$ N is autocorrelated with random "firm effect" and normal errors.

What method to use here?

- We can use $\mathcal{G}^{\text{ref}}$ on $(\varepsilon_{it} - \bar{\varepsilon}_{i\cdot})$ because of AR structure.
- We can also permute ε_{it} with respect to i if η_i are exchangeable.

Simulated study with 5 firms, 200 timepoints, $\rho_1 = \rho_2 = 0.8, \eta_i \sim t_5$:

OLS	HAC	RR(uncond.)	RR(cond.)
37.55	11.95	2.56	4.83

Code: https://www.dropbox.com/s/kaegbx29bgwc9k4/temple_WF.zip?dl=0

Concluding remarks

- Residual randomization addresses inference in regression models with complex error structure.
- It does so in a unified structure. Good practice: first think about invariances, then do inference. The method is valid (asymptotically) in many settings.
- In extensive simulations, the method performs favorably to established bootstrap variants, and "robust error" methods.
- Extensions to models with autocorrelated errors (and high-dimensional regression) are also considered with notable empirical success.

Thank You.

"Life After Bootstrap: Residual Randomization Inference in Regression Models"
(working paper, 2019)

"Introduction to Residual Randomization: The R Package `RRI`"
(Technical report, 2019)

`https://www.ptoulis.com/residual-randomization`

Extensions: High-dimensional regression

Consider the ridge estimator, $\hat{\beta}^{\text{ridge}}$. We can show that:

$$\lambda' \hat{\beta}^{\text{ridge}} - \lambda_0 + \mu \lambda' P_\mu^{-1} \beta = \lambda' P_\mu^{-1} X^\top \varepsilon,$$

where $P_\mu = X^\top X + \mu I$ is the ridge matrix.

1. Thus, we can isolate the right term as our invariant:

$$t_n(\varepsilon) = \lambda' P_\mu^{-1} X^\top \varepsilon,$$

2. and consider the left term as our test statistic,

$$T_n = \lambda' \hat{\beta}^{\text{ridge}} - \lambda_0 + \mu \lambda' P_\mu^{-1} \hat{\beta}$$

For $\hat{\beta}$ we can either plug-in the ridge estimate or some LASSO estimate.

The rest of the procedure remains the same, and can handle (ostensibly) complex error structures. See paper for detailed experiments.