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

This paper considers the problem of doing inference in a model

$$y_i = f(x_i, \beta) + \varepsilon_i,$$

where

- $y_i \in \mathbb{R}$ is the response; $x_i \in \mathbb{R}^p$ are covariates.
- $\varepsilon_i \in \mathbb{R}$ are the unobserved errors.

We insist on using only structural assumptions on the errors.

Specifically, we don't assume independence of ε_i, or any limit on the distr of X.

Structural assumptions

Errors may share a complex dependency. For example, they may be

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

These can be encoded in a parsimonious way through a group of transformations $\mathcal{G}$ and assuming:

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

This will be our main inferential (structural) assumption.

Randomization inference (Fisher, 1935); (Lehman and Romano, 2005) works with such assumptions in experimental settings.

We extend this to observational settings through "residual randomization" (RR).

Main contributions

Residual Randomization (RR) directly compares to bootstrap as a general method for inference. The main points of comparison:

1. RR generally requires mild conditions on the (ε, X)-distr (e.g., leverage structure).

   Bootstrap relies on $\sqrt{n}$-asymptotic normality, which is usually much stronger.

2. RR uses one single procedure for each $\mathcal{G}$. Unified way of inference.

   Bootstrap requires a different variant for each setting. [the "bootstrap zoo": wild, residual, cluster, block, pigeonhole,...]

3. In most cases, the same RR procedure is valid also in high-dimensional settings where $p < n$ but $p \to \infty$.

   Bootstrap needs completely new theory and new variants for that.

4. In some (limited) settings, RR can even be finite-sample valid.

RR is a robust and principled way of inference. There is life after bootstrap!

Outline

1. Residual randomization inference.
2. Main condition for validity.
3. Exchangeability in linear model.
4. Clustered error structures in linear model.

Main Idea

Suppose we want to test a null hypothesis, H_0, on β.

We assume the existence of a test statistic T_n for which there is a known function $t_n : \mathbb{R}^n \to \mathbb{R}$ satisfying:

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

Example: Suppose linear f and $H_0 : \beta_1 = 0$. Take $T_n = \hat{\beta}_1$. Then, under H_0,

$$\hat{\beta}_1 = (0, 1, 0 \ldots)'(X^\top X)^{-1} X^\top \varepsilon.$$

Because of invariance, $t_n(\mathsf{g}\varepsilon) \overset{d}{=} T_n$ under the null. So, we can easily construct a randomization test for H_0.

Classical randomization test

In particular, standard randomization theory (Lehman and Romano, 2005) suggests that we can test H_0 through

$$\text{Reject } H_0 \text{ if } (1/|\mathcal{G}|) \sum_{g \in \mathcal{G}} 1\{t_n(g\varepsilon) \geq T_n\} \leq \alpha. \tag{1}$$

Many excellent (and unique) properties of such tests:

- Finite-sample exact. They work great for any finite $n > 0$. Robustness.
- Assumption-free. No conditions on (ε, X)-distribution.
- Easy to implement and communicate.

Of course, the downside here is that the standard test is infeasible because ε are unobserved.

The idea is to approximate it through residuals.

Main Residual Randomization Procedure

Suppose we have an estimator $\hat{\beta}$ of β. [not necessarily consistent, or "well behaved".]

Then, the feasible procedure is to use $t_n(\mathsf{g}\hat{\varepsilon})$ from residuals $\hat{\varepsilon}_i = y_i - f(x_i; \hat{\beta})$ in the standard procedure; i.e., we do

1. Calculate the residuals, $\hat{\varepsilon} = y - X\hat{\beta}$.
2. Compute T_n .
3. Set $\mathsf{T}_R = \{t_n(\mathsf{g}_r\hat{\varepsilon}) : \mathsf{g}_r \sim \mathsf{Unif}(\mathcal{G}), r = 1, \ldots, R\}$. [no need to consider entire $\mathcal{G}$]
4. Calculate p-value: $\widehat{\text{pval}} = (1/R) \sum_{t \in \mathsf{T}_R} 1(t \geq T_n)$.

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

$$\phi_n = 1(\widehat{\text{pval}} \leq \alpha).$$

Procedure is simple and general.

But we are losing finite-sample validity. How about robustness?

Validity

i.e., “Under what conditions is residual randomization valid?”

Validity

i.e., "Under what conditions is residual randomization valid?"

Theorem

Let $\Lambda_n = t_n(G_1\varepsilon) - t_n(G_2\varepsilon)$ and let $\Delta_n = t_n(G\hat{\varepsilon}) - t_n(G\varepsilon)$, where $G, G_1, G_2 \sim \text{Unif}(\mathcal{G})$. Suppose that

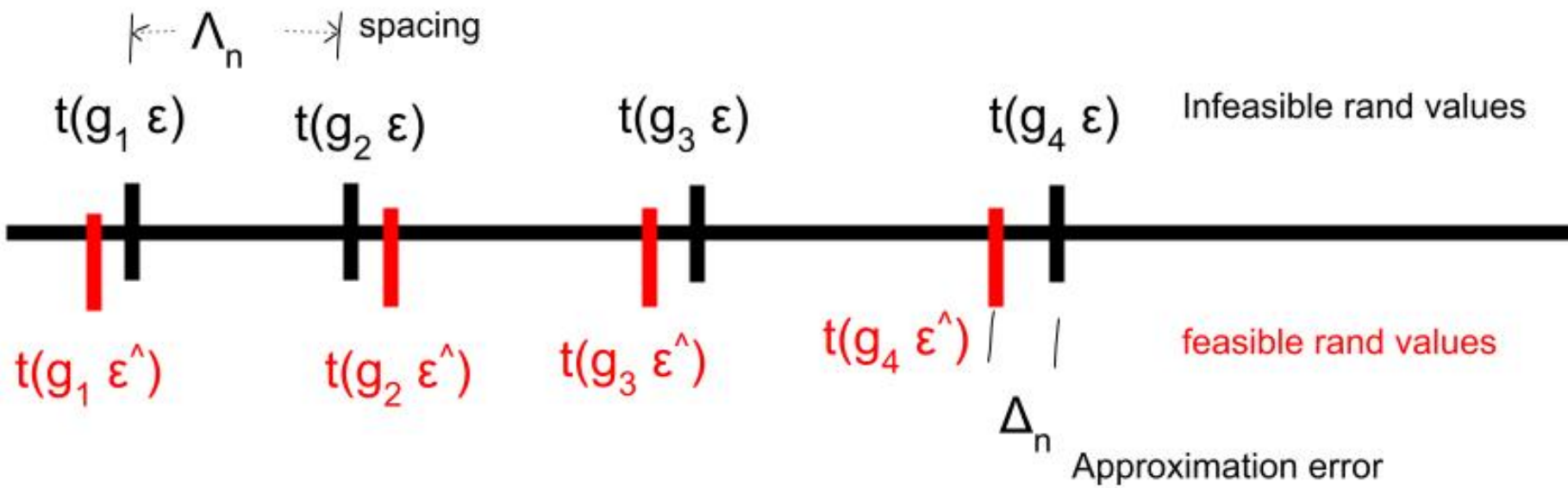
$$P(\Lambda_n = 0) \to 0, \text{ and } \frac{E(\Delta_n^2)}{E(\Lambda_n^2)} \to 0.$$

Then, residual randomization is asymptotically valid under H_0, that is,

$$\limsup_{n\to\infty} E(\phi_n | H_0) \leq \alpha.$$

- Λ_n is the "spacing" between the randomization values of the feasible test.
- Δ_n is the approximation error of the infeasible test.
- The theorem shows that residual randomization is valid asymptotically if the variance of the spacings in the feasible test dominates the variance of the approximation error.
- Notably: consistency of $\hat{\beta}$ or "$\sqrt{n}$-conditions" are not necessary!

Intuition



A randomization test relies on the quantiles of the randomization distribution.

The quantiles of the infeasible test (and so the test decision) are very similar to the infeasible test as long as $\Delta_n \preceq \Lambda_n$.

They are actually identical when $\max \Delta_n \leq \min \Lambda_n$.

This shows why RR inherits (partially) the robustness of the infeasible test.

Finite-sample rates

Theorem

Let $\frac{E(\Delta_n^2)}{E(\Lambda_n^2)} = \zeta_n^2 \to 0$, and $\bar{\Lambda}_n = \frac{|\Lambda_n|}{\text{Var}(\Lambda_n)^{1/2}}$. Suppose also that $F_{\bar{\Lambda}_n}(\epsilon) = O(\epsilon^\gamma)$. Then,

$$E(\phi_n \mid H_0) = \alpha + A_\gamma O(\zeta_n^{2\gamma/(2+\gamma)}).$$

- Parameter γ controls how small the spacings are.

  Smaller γ implies more "concentration" of Λ_n on zero (=small spacings in feasible test). Larger γ implies larger spacings.

- Under regularity conditions, $\zeta_n^2 = O(1/n)$, and $\gamma = 1$ (e.g., normal $\bar{\Lambda}_n$). Then, the rate to validity is $O(n^{-1/3})$.

- Best rate is $O(n^{-2/3})$.["price of robustness"]

Linear model

These results translate into simple conditions when we move to the linear model:

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

Consider testing a linear hypothesis $H_0 : a'\beta = a_0$.

In this setting, for residual randomization we can simply use

$$T_n = (a'\hat{\beta} - a_0)$$

as the test statistic. Here, $\hat{\beta}$ is the OLS estimate (could also be restricted OLS). Thus, under the null, $t_n(u) = a'(X^\top X)^{-1}X^\top u$, and $T_n = t_n(\varepsilon)$ as required.

The test can be "inverted" to also construct RR-based CIs for each individual β.

Linear model : exchangeability

In this setting, we assume that $\mathcal{G}$ represents permutations of n-length vectors.

In other words, for any permutation π,

$$(\varepsilon_1, \ldots, \varepsilon_n) \stackrel{d}{=} (\varepsilon_{\pi(1)}, \ldots, \varepsilon_{\pi(n)})$$

Theorem

In the linear model, suppose that $P(\Lambda_n = 0) \to 0$ and that the errors are exchangeable. Then,

$$p/n = o(1)$$

is sufficient for asymptotic validity of residual randomization.

- No additional conditions on X! Or the error distribution. OLS could even be inconsistent.
- Validity even when $p \to \infty$, as long as $p = o(n)$.
- Theorem also shows that exchangeability is actually a strong inferential assumption.

Comparison with 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.

In this setting, residual bootstrap 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.

Much stronger conditions than residual randomization.

Also requires $p < \infty$.

Illustration

Assume $x_i \sim t_5$ and $\varepsilon_i = \text{Weibull}(0.6)$. We let $p = \sqrt{n}$.

Rejection rates of RR over 2,000 replications:

n	p	rejection rate %
25	5	10.6
50	7	7.2
80	8	7.8
100	10	7.3
120	10	6.4
150	12	6.3
200	14	6.1
300	17	5.7
500	22	6.0

Clustered errors

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

e.g., observations clustered by state, or school/grade, and so on.

There are numerous analytic "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").

Which invariance works here?

Residual randomization offers a natural way of inference.

Work with invariances on the cluster level.

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

- permutations within clusters.
- sign flips across clusters (equivalent to "cluster wild bootstrap").
- 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:

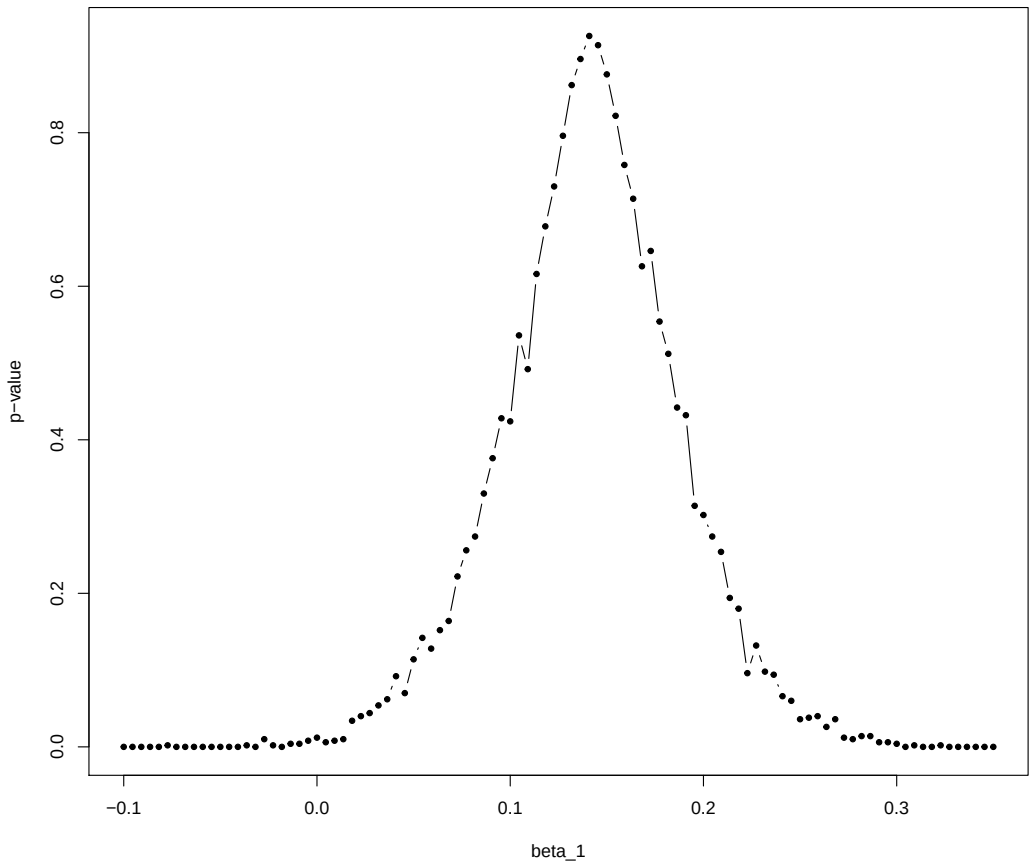
```
> confint(lm(y ~ x))
        2.5 %     97.5 %
X   -0.88       0.50
```

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

invariant	condition
permutation with clusters	$\lambda_x \lambda_\varepsilon p/n = o(1)$
sign symmetry across clusters	$\lambda_x \lambda_\varepsilon p^3 \sum_{c=1}^{J} (n_c^2/n^2) = o(1)$
both	either

Here, λ_x refers to a leverage condition on $X^\top X$ (e.g., condition number), and λ_ε to a leverage condition on the error distribution.

Under the standard regularity conditions in the literature, $\lambda_x = O(1)$ and $\lambda_\varepsilon = O(1)$. In RR, these quantities are allowed to increase.

These results also significantly extend Canay et.al. (2017) who showed that cluster wild bootstrap is asymptotically valid when $J_n < \infty$ and

$$X_c^\top X_c \propto (X^\top X).$$

This is a cluster homogeneity condition, and is generally strong.

Finite-sample validity

Under a strict homogeneity condition, RR can be finite-sample exact!

Theorem (Summary)

Suppose that

1. The errors are sign symmetric across clusters (as defined earlier).
2. The homogeneity condition holds for some clustering.

Then, cluster-sign residual randomization test (based on $\mathcal{G}^s$) and restricted OLS is finite-sample exact.

Proof sketch: The approximation error $t_n(\mathbf{g}\hat{\varepsilon}^{\mathrm{o}}) - t_n(\varepsilon)$ satisfies

$$a'(X_c^\top X_c)^{-1}(X^\top X)(\hat{\beta}^{\mathrm{o}} - \beta) \propto a'\hat{\beta}^{\mathrm{o}} - a_0 \overset{H_0}{=} 0,$$

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. [standard assumption]
2. For every cluster c, $X_c^\top X_c$ only depends on proportion of treated units in c, which is the same for every $c = 1, 2, 3$, by construction!

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

The resulting randomization test is a cluster sign test with 3 clusters, and is exact.

[However, because of few clusters, minimum p-value is $1/8 = 0.125$, and so we need to randomize the test decision to bring it down to 0.05. Still valid, but loses power.]

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).

Again, 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”.

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".

Residual randomization is valid as long as

$$p^4 \lambda_x \lambda_\varepsilon (1/|\text{row clusters}| + |\text{col clusters}|) = o(1).$$

- Leverage quantities are very similar to clustered case (e.g., condition number of $X^\top X$).
- Test allows $p \to \infty$!
- Test is valid (asymptotically) when both clusters grow.
- When one cluster size stays finite, then the test can still be valid if we center the data of the other cluster.

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 $H_0 : \beta_1 = 1.0$.

Concluding remarks

- Residual randomization addresses inference in regression models with complex error structure.

- It does so in a unified way. Good practice: first think about invariances, then do inference. The method is valid (asymptotically) in many settings.
- Conditions are considerably weaker than bootstrap. And robust to (ε, X)-distribution (e.g., heavy tails).

- In extensive simulations, the method performs favorably to established bootstrap variants, and "robust error" methods.

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)

"Robust inference for high-dimensional linear models via residual randomization"
(with Wang et al, ICML 2021).

`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.

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.

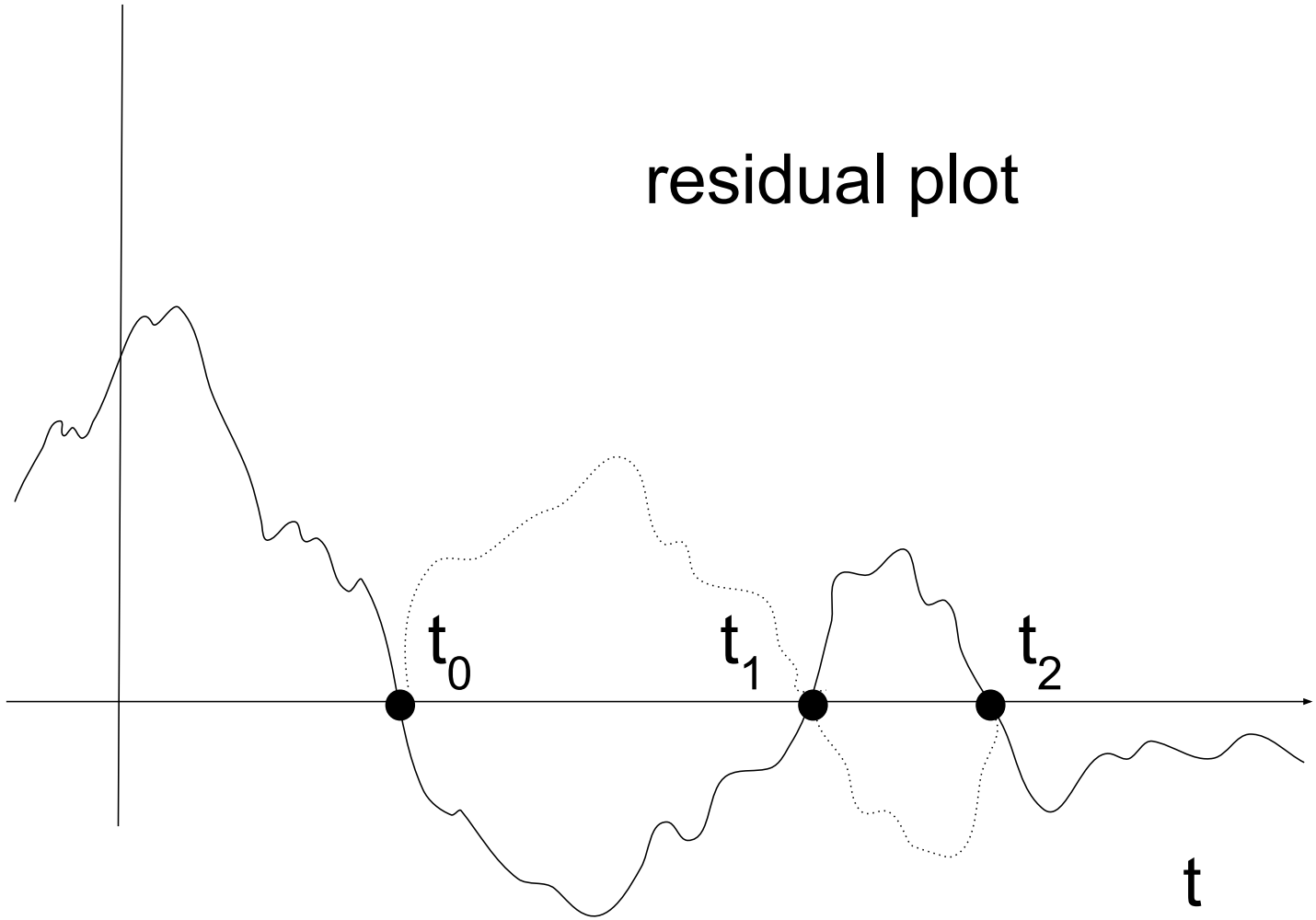
Which invariance works here?

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

However, for the AR(1) process:

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

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