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.
- $\varepsilon \in \mathbb{R}^n$ are the errors (no assumption yet).

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

Standard approaches: parametric or nonparametric (e.g., normal OLS, bootstrap).

The standard approaches

Parametric approach:

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

Bootstrap approach:

- Resample $(y, X) \rightarrow$ bootstrap distribution of $\hat{\beta}$. Use CLT for inference.
- 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 (heterogeneity, clustered errors, autocorrelated errors, etc.)
- They rely on asymptotics.

What's wrong with bootstrap?

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.
- wild bootstrap.
- cluster wild bootstrap.
- block bootstrap.
- pigeonhole bootstrap.
- ...

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.

Statistical 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.

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

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).

Randomization Tests (Lehman and Romano, 2005)

Suppose that $D \in \mathbb{R}^n$ is our data, and $\mathcal{G}$ a group of transformations.
We are testing some H_0 under which:

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

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

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

Hence, we can test H_0 through, say, the p-value of T_n wrt to T_D.

$*$ This test is (i) exact, (ii) valid in finite samples and (iii) 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).

Two key definitions

Our approach relies on two key ideas:

1. **Inferential primitive.** We assume:

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

   where $\mathcal{G}$ is the *inferential primitive*, a group of $\mathbb{R}^n \to \mathbb{R}^n$ linear operators (e.g., permutations, random signs, cluster permutations, etc).
   [chosen by the analyst]

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

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

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

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 $\mathrm{Var}(G) = Id$, for random signs, where $G \sim \mathrm{Unif}(\mathcal{G})$.

Promises three 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. Not rely on good asymptotics (e.g., consistency, normality); and
3. Validity in finite samples.

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

The feasible procedure needs to rely on residuals.

Outline

1. Describe concrete procedure.
2. Main theoretical result on validity.
3. Cluster invariances: exchangeability, sign symmetry.
4. Two-way clustering.
5. Autocorrelated errors.
6. Conclusion.
7. **(if time) Finite-sample validity + High-Dimensional regression.*

Residual Randomization: Concrete procedure

1. Calculate the restricted OLS estimate:

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

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

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

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

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

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

Validity

Theorem

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

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

Then, the residual randomization procedure asymptotically valid:

$$\limsup_{n \to \infty} E(\phi(y; X)) \leq \alpha.$$

- 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., permute only 2 elements instead of n).
- Term [3] requires $\mathcal{G}$ not to change the "information structure".
- Consistency/normality of $\hat{\beta}$ is not necessary!

Example: Hormone data (Efron & Tibshirani, 1996)

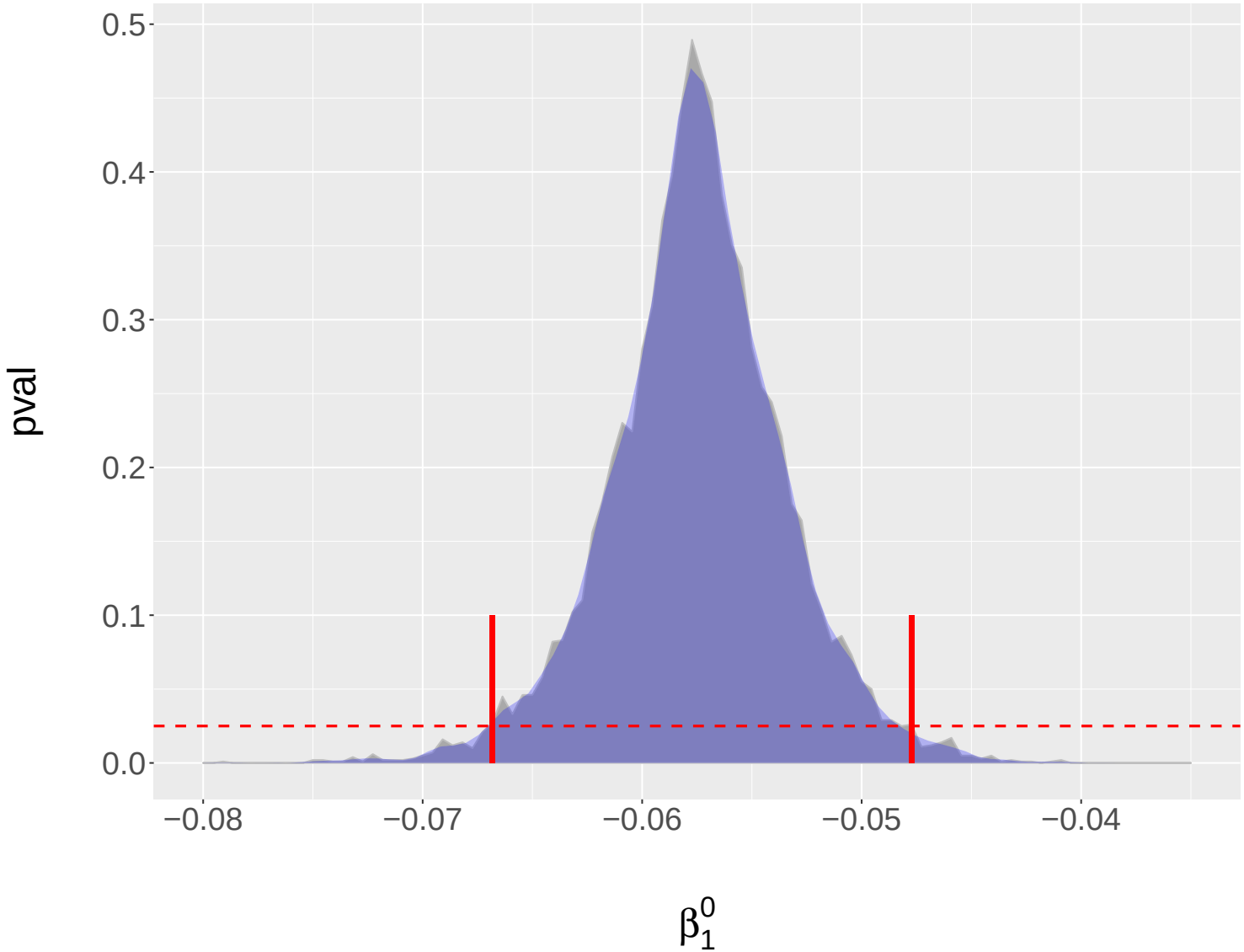
Consider the following regression model:

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

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 $\hat{\varepsilon}_i^{\mathsf{o}} = \tilde{y}_i - \overline{\tilde{y}_i}$, where $\tilde{y}_i = y_i - b x_i$.
4. Compare T_n with $\{t_n(\mathrm{g}\hat{\varepsilon}^{\mathsf{o}}) : ...\}$.



$*$ 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	$(-0.0682, -0.0482)$

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.

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 or non-normality.

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

Which invariance works here?

Residual randomization offers a natural way of inference.

Just assume invariance *on the cluster level*.

e.g., 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

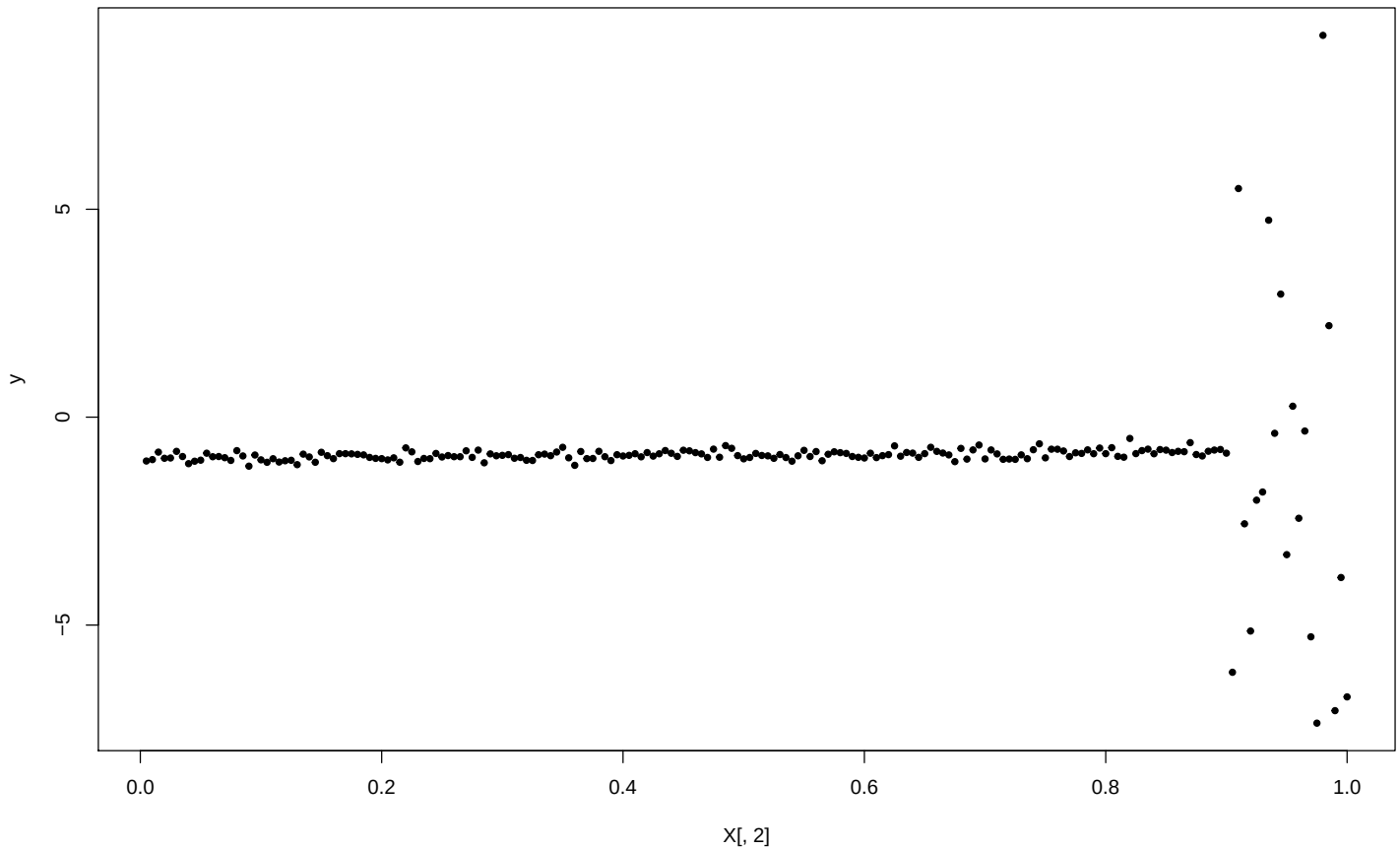
$$\varepsilon_i \sim \begin{cases} N(0, 0.1^2) & \text{if } x_i \leq 0.9 \\ N(0, 5^2) & \text{if } x_i > 0.9. \end{cases}$$

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

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

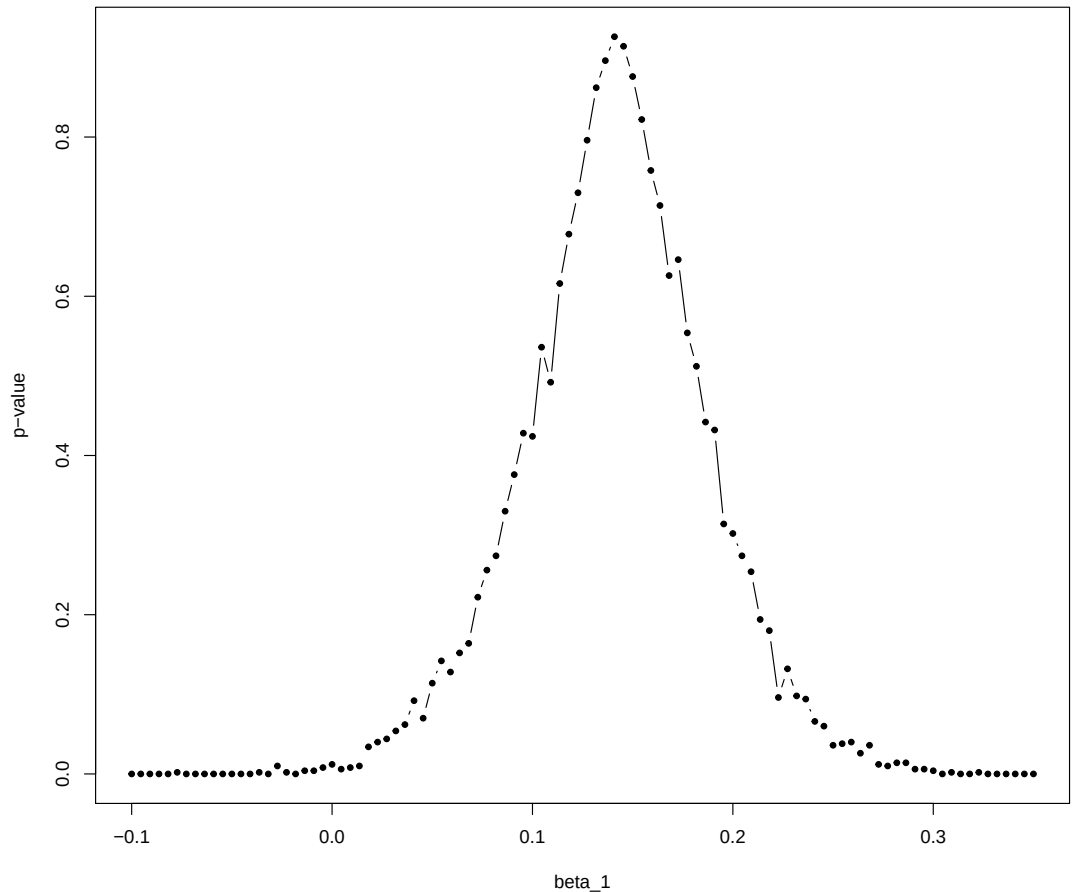
Two clusters for the errors. One has much higher variance than the other. OLS fails badly to detect significance.

Illustration: data with two heterogeneous clusters



Residual randomization with two clusters

Define $\mathcal{G}$ as permutations *within* each cluster (assume clusters known). The p-value plot is shown below:



The 95% CIs are much better centered than regular OLS.

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).

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

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 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
bootstrap	.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
bootstrap	.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.

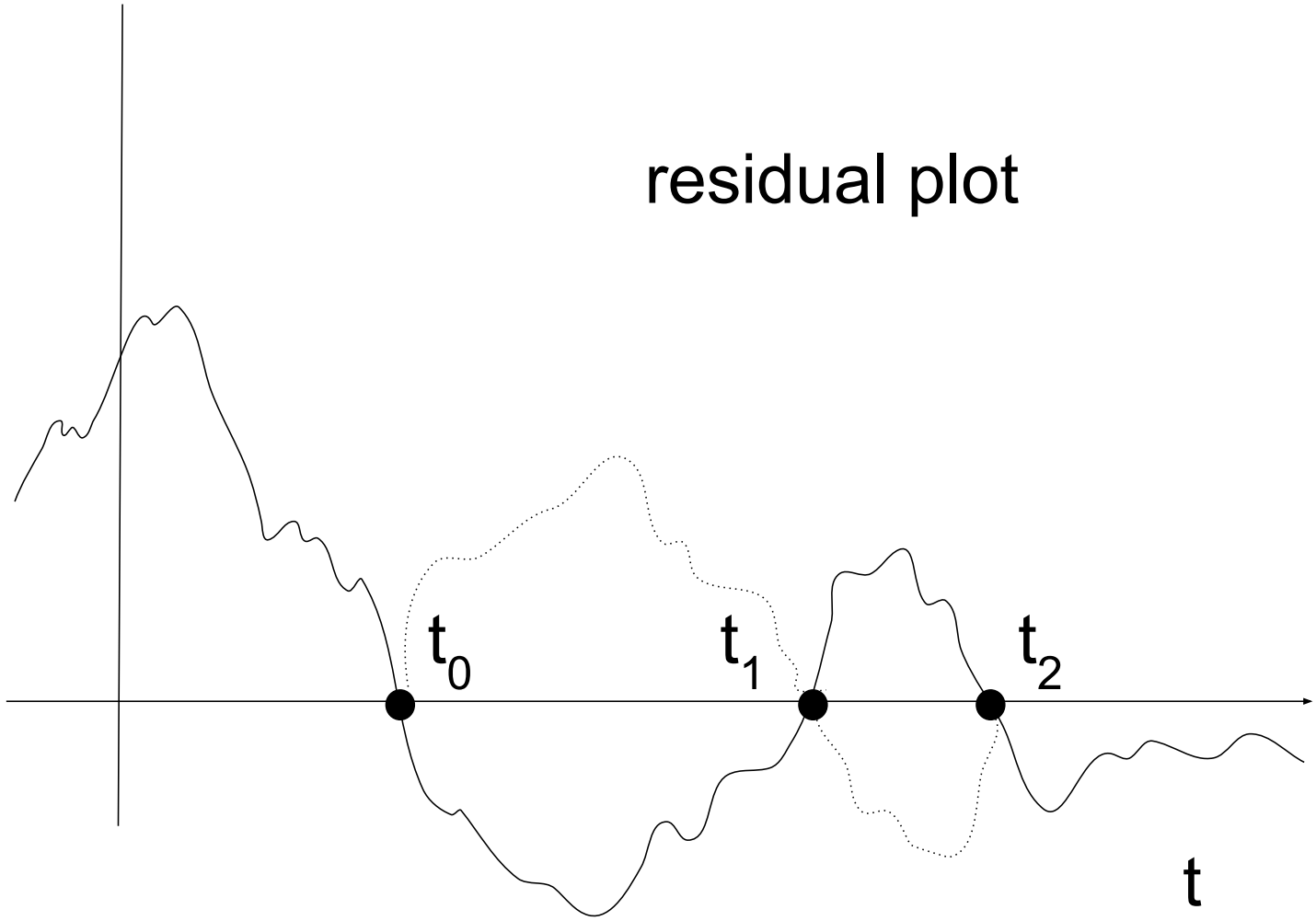
Which invariance works here?

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.

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.

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`

Appendix: Finite-sample validity

Under certain conditions, the residual randomization tests can even be *exact* (i.e., valid for any finite n).

One interesting setting is when there is a clustering of the data such that:

1. The errors are sign symmetric across clusters (as defined earlier).
2. 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.

Theorem (Summary)

If these conditions hold then the cluster-sign residual randomization test is exact.

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

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

The clustering may be chosen by the analyst.

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 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 satisfactory method is 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 has one treated unit and the rest are controls: $(1, 9), (1, 9), (1, 9)$.

1. Assume sign-symmetric errors across clusters.
2. For every cluster c it holds that

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

since $X_c X_c^\top$ depends only on the proportion of treated units in c. This is fixed across clusters by construction.

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

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.