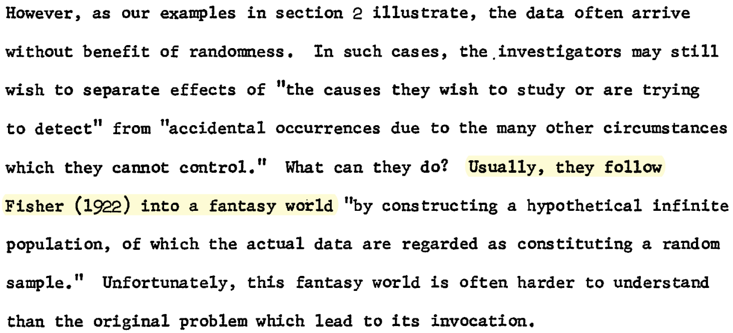
Invariant Inference via Residual Randomization

Panos Toulis
panos.toulis@chicagobooth.edu

Econometrics and Statistics
University of Chicago, Booth School of Business

(Freedman and Lane, ca 1980)

Freedman and Lane (1980?, 1983) criticized the classical framework of inference. This framework relies on iid samples from a hypothetical superpopulation, which they thought was problematic.



However, as our examples in section 2 illustrate, the data often arrive without benefit of randomness. In such cases, the investigators may still wish to separate effects of "the causes they wish to study or are trying to detect" from "accidental occurrences due to the many other circumstances which they cannot control." What can they do? Usually, they follow Fisher (1922) into a fantasy world "by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample." Unfortunately, this fantasy world is often harder to understand than the original problem which lead to its invocation.

Proposed procedure by F&L

Suppose we observe $\{(y_i, x_i, z_i)\}$, Y = house price, X = size, Z = #bathrooms. We ask whether Z is "significant for" Y.

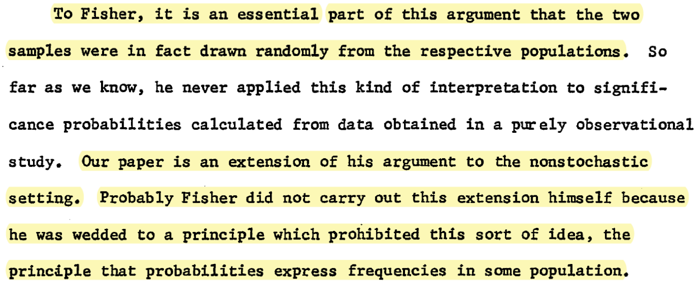
1. Fit the "null model" $Y \sim X$. Obtain $\hat{Y}_0$ and residuals $e = Y - \hat{Y}_0$.
2. Calculate the correlation coefficient, $r = r(Z, e)$.
3. Calculate a p-value according to permuted residuals:

$$\text{pval} = \frac{1}{n!} \sum_{\pi} 1\{r(Z, \pi e) > r\}.$$

According to F&L, this offers a "permutation interpretation" of significance levels that makes no assumption about the data generating mechanism.

A critique

- F&L never fleshed out what was the actual null being tested.
- They seemed to think that their approach was a logical extension of Fisherian randomization. Talking about Fisher's permutation test of independence between two random samples, they wrote:



> To Fisher, it is an essential part of this argument that the two samples were in fact drawn randomly from the respective populations. So far as we know, he never applied this kind of interpretation to significance probabilities calculated from data obtained in a purely observational study. Our paper is an extension of his argument to the nonstochastic setting. Probably Fisher did not carry out this extension himself because he was wedded to a principle which prohibited this sort of idea, the principle that probabilities express frequencies in some population.

They were right to think that significance levels need not only have a classical probabilistic interpretation.

But, ultimately, I believe they were wrong to think that you can do this with no assumptions on the data generating mechanism.

Overview of this work

- This paper also attempts an extension of Fisherian randomization to observational settings and inference (specifically, regression) following F&L.

- We impose invariance assumptions on the DGP. We thus use the framework of (Lehman and Romano, 2005) as a natural foundation.

- Two main contributions:

  1. Theoretical. Is it possible to conduct inference under only invariances?

     Under what conditions? Benefits compared to classical methods of inference? (e.g., bootstrap, "robust errors")

  2. Applied. A unified method of inference for different problem structures.

     In contrast, classical methods require a different variant for each different structure, which frequently confounds applied researchers

     This contrast will be more evident in complex structures such as double clustering.

I. Setup, main method and general results

II. Specialization to the linear model

Setup

Consider the 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; f may be unknown;
- $\varepsilon_i \in \mathbb{R}$ are unobserved errors.

The errors are invariant to $\mathbb{R}^n \to \mathbb{R}^n$ transformations from an algebraic group $\mathcal{G}_n$:

$$\varepsilon \overset{d}{=} \mathsf{g}\varepsilon \mid X, \text{ for all } \mathsf{g} \in \mathcal{G}_n. \tag{1}$$

<u>Notes:</u>

- (y_i, x_i) are not necessarily i.i.d.;
- ε could depend on X in complex ways (examples coming soon);
- No other assumptions (e.g., moment conditions) on the distribution of (X, ε).

* (Eq (1) extends the framework of (Lehman and Romano, 2005) in the observational setting.)

Examples of $\mathcal{G}_n$

Errors may have a complex structure. For example, they may be

- Exchangeable; e.g., $\varepsilon_i = g_{0,n} + g_{1,n}\epsilon_i,\ \epsilon_i \overset{\text{iid}}{\sim} F$. $g_{0,n}, g_{1,n}$ possibly unknown.
- Non-exchangeable but sign symmetric; e.g., $\varepsilon_i = g_{0,n} + g_{i,n}\epsilon_i$.
- Exchangeable only within certain clusters; e.g., $\varepsilon_{ck} = g_{0,n} + \eta_c + g_{1,n}\epsilon_{ck}$.
- Partially exchangeable in a dyadic structure; e.g., $\varepsilon_{rck} = g_{0,n} + \xi_r + \eta_c + g_{1,n}\epsilon_{rck}$.
- ...

These structures can be encoded in a parsimonious way through (1).

$\mathcal{G}_n$ is the inferential primitive because it encodes our main inferential assumption. This leads to the problem of invariant inference on β.

Throughout, we assume $\mathcal{G}_n$ known. It would be interesting (in future work) to consider a setting where $\mathcal{G}_n$ can be learned.

This assumption may be uncomfortable but it replaces the i.i.d. assumption. An alternative is “approximate symmetry” (Canay et al, 2017).

Testing the global null is easy

Suppose we want to test a global null hypothesis,

$$H_0 : \beta = \beta_0.$$

Take a test statistic T_n such that $T_n \overset{H_0}{=} t_n(\varepsilon)$ under the null for a known measurable function $t_n : \mathbb{R}^n \to \mathbb{R}$. Then:

1. Calculate $\varepsilon_0 = y - X\beta_0$, the errors under the null.
2. Calculate $T_n = t$ in the sample.
3. Calculate the p-value based on transformed errors:

$$\text{pval} = \frac{1}{|\mathcal{G}_n|} \sum_{\mathbf{g}} 1\{t_n(\mathbf{g}\varepsilon_0) > t\}. \tag{2}$$

Notes:

- Not hard to find an appropriate t_n in many problems.
- Validity of (2) is guaranteed by standard randomization theory (Lehman and Romano, 2005).

Properties

This test has some unique and highly desirable properties:

- Exact. The test is valid for any finite sample $n > 0$.
- Robust. No conditions on distribution of (ε, X); test is invariant to location-scale transformations of t_n.
- Simple. The test is easy to implement and communicate.

Of course, the downside here is that the null is too strong.

What if we wanted to test a simpler hypothesis, e.g., $H_0 : \beta_1 = 0$?

One approach is to approximate the exact test using residuals from an estimator.

Main Residual Randomization Procedure

Suppose we want to test a non-global H_0; e.g., $H_0 : a'\beta = a_0$.

Take an estimator $\hat{\beta}_n$ of β under the null (not necessarily "well behaved"). Then:

1. Calculate $\hat{\varepsilon}_0 = y - X\hat{\beta}_n$, the residuals under the null.
2. Calculate $T_n = t$ in the sample.
3. Calculate the p-value based on transformed errors:

$$\widehat{\text{pval}} = \frac{1}{m}\sum_{r=1}^{m} 1\{t_n(\mathrm{g}_r\hat{\varepsilon}_0) > t\}, \quad \mathrm{g}_r \overset{\text{iid}}{\sim} \text{Unif}(\mathcal{G}_n). \tag{3}$$

Notes:

- This is a form of residual randomization because we transform the residuals.
- We use this same procedure (3) for all problem structures (unified method), and just "plug-and-play" $\mathcal{G}_n$.
- We will ask: Under what conditions is (3) valid? Do we keep any of the aforementioned desirable properties of the exact test?

Validity

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 $G, G_1, G_2 \sim \text{Unif}(\mathcal{G}_n)$. Suppose that

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

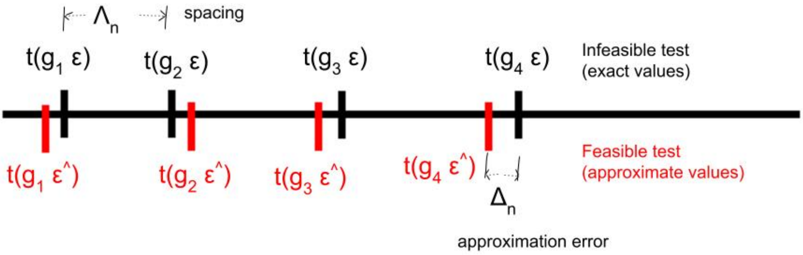
Then, residual randomization is asymptotically valid under H_0, that is,

$$\lim \sup_{n\to\infty} E(\phi_n | H_0) = \alpha.$$

Notes:

- Validity holds asymptotically if the variance of the spacings in the infeasible test (Λ_n) dominates the approximation errors of the feasible test (Δ_n).
- Notably, consistency of $\hat{\beta}$ or "$\sqrt{n}$-asymptotics" are not required; cf. (Chernozhukov et. al., 2021) on a permutation test using residuals in a panel data setting.

Proof sketch



The residual randomization test compares $T_n = t(\varepsilon)$ with the other approximate red values from the feasible test. It would be exact if it compared to the exact values.

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

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

This allows the residual randomization test to inherit some of the robustness properties of the idealized test, and our analysis is tight enough to capture that.

Finite sample rates

Theorem

Let $\frac{E(\Delta_n^2)}{E(\Lambda_n^2)} = \zeta_n^2 \to 0$, and $\bar{\Lambda}_n = \frac{\Lambda_n}{\mathrm{Var}(\Lambda_n)^{1/2}}$. Suppose also a γ-Hölder continuity property on $F_{\bar{\Lambda}_n}$ such that $F_{\bar{\Lambda}_n}(\epsilon) - F_{\bar{\Lambda}_n}(-\epsilon) = O(\epsilon^\gamma)$, $\epsilon > 0$. Then,

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

Notes:

- Parameter γ controls the tails of the spacings variable, Λ_n. Smaller $\gamma < 1$ correspond to heavy tailed (X, ε). Also, $A_\gamma = O(1)$.
- No further assumptions on the distribution/asymptotics of (X, ε).
- Under regularity conditions, $\zeta_n^2 = O(1/n)$, and $\gamma = 1$ (e.g., normal $\bar{\Lambda}_n$). This implies the rate $O(n^{-1/3})$, which may be viewed as the “price for robustness”.
- (Not shown here) The test is also consistent for all alternatives $\tau_n/\sigma_n \to \infty$, where τ_n is location shift under the alternative, and $\sigma_n^2 = E[t_n^2(\varepsilon)]$.

I. Setup, main method and general results

II. Specialization to the linear model

Linear model

In the linear setting, these results translate into simple and interpretable conditions.

From now on, we assume:

- $y = X\beta + \varepsilon$, linear model.
- $H_0 : a'\beta = a_0$, linear hypothesis.
- $T_n = a'\hat{\beta} - a_0$ with $\hat{\beta}$ being the OLS estimator.

  Under the null, $T_n = t_n(\varepsilon)$ for $t_n(u) = a'(X^\top X)^{-1}X^\top u$, as required.

Notes:

- We can also handle multiple linear hypotheses (not today).
- In the residual randomization test, we could use the residuals calculated under the null through a constrained OLS estimator. We analyze this test as well.

Exchangeable errors

Consider an error structure of the form

$$\varepsilon_i = g_{0,n} + g_{1,n}\epsilon_i,\ \ \epsilon_i \overset{\text{iid}}{\sim} F.$$

- Terms $g_{0,n}, g_{1,n}$ may depend on X but are common for all i.
- Then, (ε_i) are exchangeable; i.e., (1) holds with $\mathcal{G}_n =$ permutation group.

Theorem

In the linear model, suppose that the errors are exchangeable. Then,

$$a_1 = 0, \quad \text{and} \quad p/n = o(1)$$

are sufficient for asymptotic validity of residual randomization.

Notes:

- Basically no conditions on (X, ε) distribution. OLS could even be inconsistent.
- Validity even when $p \to \infty$, as long as $p = o(n)$.
- Exchangeability is a strong inferential assumption. This is line with the concept of exchangeability as a “mixture of i.i.d. sequences” from de Finetti’s theorem.

Comparison with “residual bootstrap”

In this setting, the residual randomization procedure is, operationally, almost identical to the “residual bootstrap” procedure.

To prove validity of the residual bootstrap, Freedman (1981) needed:

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. Thus, $p < \infty$.

See also (Bickel et al., 1981) and (Lopes, 2014) for the high dimensional regime.

These are stronger conditions than residual randomization.

The difference exists because bootstrap is trying to approximate the full sampling distribution of $\hat{\beta}$ and so it needs regularity conditions to ensure convergence in distribution at the appropriate rate.

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. Theory tends to be complicated, and the various methods have problems with small samples and non-normality.

"Cluster wild bootstrap" (Cameron et al, 2008) is a popular alternative.

Its theory is complicated, however, and cannot be easily extended to more complex structures.

Cluster invariances

Let $i = (c, k)$ in cluster c, replicate k. Consider an error structure of the form

$$\varepsilon_i = \varepsilon_{ck} = g_{0,n} + \eta_c + g_{c,n}\epsilon_{ck}.$$

Assume (η_c) and (ϵ_{ck}) are independent, and ϵ are iid.

- If η_c are exchangeable then $\mathcal{G}_n$ could be permutations within clusters;
- If η_c and ϵ_{ck} are sign symmetric then $\mathcal{G}_n$ could be sign flips on the cluster level.
- Or both invariances could hold.

Validity under clustered errors

Theorem (Summary)

Invariant	Condition
permutation within 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

Notes:

- Here, λ_x refers to a leverage condition on $X^\top X$; e.g., condition number, max leverage ratio across clusters, and so on.
- λ_ε represents similar leverage conditions on the error distribution; e.g., condition number of V_ε, max error variance ratio between clusters.
- Under standard cluster assumptions, $\lambda_x, \lambda_\varepsilon = O(1)$. In residual randomization, these quantities are allowed to increase.
- Also allows $p \to \infty$.

Finite number of clusters

The setting with a finite number of clusters ($J < \infty$) is tricky. Canay et.al. (2017) have showed that cluster wild bootstrap is asymptotically valid whenever

$$X_c^\top X_c \propto (X^\top X), \text{ in the limit.} \tag{HC}$$

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

In this setting, residual randomization is also asymptotically valid..

- ...under (HC) and sign symmetric cluster errors.
- ...under exchangeable errors within clusters.
- ... or under both conditions.

Moreover, whenever (HC) holds in the sample, then the residual randomization test using restricted OLS residuals is finite sample valid. Example: Behrens-Fisher problem

Simulation: One-way clustered data

The data generating model is

$$y_i = \beta_0 + x_i\beta + \varepsilon_i, \;\; x_i = \mathsf{x}_c + \mathsf{x}_{ic}, \;\; \varepsilon_i = \eta_c + \epsilon_{ic}, \;\; i \in [c]. \tag{4}$$

We consider the following simulation settings:

- $\eta_c = 0$ or $\eta_c \sim N(0,1)$; and $\epsilon_{ic} \sim N(0,1)$, all i.i.d..
- $\mathsf{x}_c \sim N(0,1)$ or $\mathsf{x}_c \sim .5LN(0,1)$, the log-normal distribution.
- For heteroskedasticity, we scale the errors by $3|x_i|$.
- $J_n \in \{10, 15, 20\}$ with 30 units per cluster; thus, $n \in \{300, 450, 600\}$.
- $(\beta_0, \beta_1) = (0,0)$ with homoskedasticity; and $(\beta_0, \beta_1) = (1,0)$ with heteroskedasticity.

(A) Homoskedastic												
	cluster effect (η_c)											
	$\eta_c = 0$ (no clustered effects)						$\eta_c \sim N(0,1)$ (clustered effects)					
	number of clusters (J)											
	$J = 10$		$J = 15$		$J = 20$		$J = 10$		$J = 15$		$J = 20$	
	cluster cov. (x_c)											
	N(0,1)	LN(0,1)										
OLS	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Standard	0.057	0.051	0.054	0.051	0.050	0.053	0.490	0.382	0.480	0.394	0.493	0.421
Cluster robust	0.086	0.090	0.076	0.079	0.061	0.077	0.103	0.110	0.081	0.089	0.081	0.090
Rands.												
RR-c-sign	0.059	0.047	0.054	0.049	0.047	0.054	0.053	0.055	0.056	0.048	0.055	0.050
RR-c-double	0.061	0.054	0.056	0.052	0.051	0.056	0.055	0.052	0.054	0.046	0.051	0.050

(B) Heteroskedastic												
	$\eta_c = 0$						$\eta_c \sim N(0,1)$					
	$J = 10$		$J = 15$		$J = 20$		$J = 10$		$J = 15$		$J = 20$	
OLS	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Standard	0.228	0.249	0.244	0.264	0.239	0.286	0.278	0.274	0.301	0.303	0.301	0.309
Cluster robust	0.095	0.140	0.085	0.116	0.074	0.114	0.100	0.126	0.091	0.124	0.075	0.121
Rands.												
RR-c-sign	0.055	0.084	0.055	0.072	0.052	0.072	0.049	0.065	0.059	0.071	0.056	0.072
RR-c-double	0.205	0.194	0.198	0.174	0.183	0.170	0.166	0.167	0.168	0.163	0.150	0.155

Table: Rejection rates under the null, $H_0 : \beta_1 = 0$. Method “RR-c-sign” is the residual randomization test with cluster sign symmetry. “RR-c-double” also permutes within clusters. Column (1) corresponds to normal covariates; and (2) corresponds to lognormal covariates.

Two-way clustered errors

In many problems there is more structure; e.g., (school, classroom), (state, city), (firm, department), etc.

Certain variants of “cluster-robust” error methods that have been extended to two-way clustering (Cameron et al, 2011). These methods (including their implementations) have the same problems as before, and may even give invalid variance estimates.

Official implementations (at least in `R`) are highly problematic.

Bootstrap approaches have only recently appeared (Davezies et.al., 2018), (Menzel, 2017), (McKinnon et.al., 2021).

Their theory is extremely complex so far, and remains limited. For example, it requires that both clusters increase in size. Additional restrictions in the DGP.

Two-way cluster invariance

Let $i = (r, c, k)$ in “row cluster” r, “column cluster” c, replicate k. Suppose

$$\varepsilon_i = \varepsilon_{rck} = g_{0,n} + \xi_r + \eta_c + g_{1,n}\epsilon_{rck}.$$

Assume (ξ_r), (η_c) and (ϵ_{ck}) are mutually independent, and ϵ_{rck} are iid.

- If ξ_r, η_c are exchangeable then ε have a “partial exchangeability” property (Aldous, 1981). Then, $\mathcal{G}_n$ denotes row-wise or column-wise permutations in a two-array representation of ε.
- If $\xi = \eta$ and there is one replication per cell, then the same $\mathcal{G}_n$ holds as above with the row-wise and column-wise permutations coupled.

Partial exchangeability

We arrange ε in a two-dim array:

$$\text{Let } \mathcal{E} = \begin{pmatrix} \{\varepsilon_1, \varepsilon_2\} & \{\varepsilon_5, \varepsilon_6\} \\ \{\varepsilon_3, \varepsilon_4\} & \{\varepsilon_7, \varepsilon_8\} \end{pmatrix} \equiv \begin{pmatrix} \{\varepsilon_{11(1)}, \varepsilon_{11(2)}\} & \{\varepsilon_{12(1)}, \varepsilon_{12(2)}\} \\ \{\varepsilon_{21(1)}, \varepsilon_{21(2)}\} & \{\varepsilon_{22(1)}, \varepsilon_{22(2)}\} \end{pmatrix}.$$

Consider the following error transformations.

$$\begin{aligned} \mathcal{E} &\overset{d}{=} \begin{pmatrix} \{\varepsilon_{21(1)}, \varepsilon_{21(2)}\} & \{\varepsilon_{22(1)}, \varepsilon_{22(2)}\} \\ \{\varepsilon_{11(1)}, \varepsilon_{11(2)}\} & \{\varepsilon_{12(1)}, \varepsilon_{12(2)}\} \end{pmatrix} \overset{d}{=} \begin{pmatrix} \{\varepsilon_{22(1)}, \varepsilon_{22(2)}\} & \{\varepsilon_{21(1)}, \varepsilon_{21(2)}\} \\ \{\varepsilon_{12(1)}, \varepsilon_{12(2)}\} & \{\varepsilon_{11(1)}, \varepsilon_{11(2)}\} \end{pmatrix} \\ &\overset{d}{=} \begin{pmatrix} \{\varepsilon_{22(2)}, \varepsilon_{22(1)}\} & \{\varepsilon_{21(1)}, \varepsilon_{21(2)}\} \\ \{\varepsilon_{12(1)}, \varepsilon_{12(2)}\} & \{\varepsilon_{11(2)}, \varepsilon_{11(1)}\} \end{pmatrix}. \end{aligned}$$

- In the first step, we permuted the rows of $\mathcal{E}$.
- In the second step we permuted its columns.
- In the third step, we permuted the observations within the diagonal cells of $\mathcal{E}$.

Partial exchangeability implies that the distribution of $\mathcal{E}$ remains invariant throughout all these steps.

In dyadic exchangeability the row-column permutations are coupled.

Validity under partial exchangeability

Theorem

In the linear model, suppose that the errors are partially exchangeable. Then,

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

are sufficient for asymptotic validity of residual randomization.

Notes:

- Leverage quantities are similar to clustered case, i.e., condition number of $X^\top X$, max leverage ratio, and condition number of V_ε.
- Again, the test allows for $p \to \infty$.
- The test is valid (asymptotically) when both clusters grow.
- (Not shown above) Moreover, when one cluster size stays finite, then the test can still be valid if we center the data of the other cluster.

Simulation: Dyadic regression

The data generating model is defined as follows:

$$y_i = \beta_0 + \beta_1 |x_r - x_c| + \varepsilon_i, \ \ \varepsilon_i = \eta_r + \eta_c + \epsilon_{rc}, \ \ i = (r, c).$$

Simulation settings:

- N row clusters, N column clusters.
- $\epsilon_{jj'} \sim N(0,1)$; and $\eta_j \sim N(0,1)$ or $\eta_j \sim .5N(-1, .25^2) + .5N(1, .25^2)$, $j = 1, \ldots, N$.
- $x_j \sim N(0,1)$ or $x_j \sim LN(0,1)$, the standard log-normal distribution, $j = 1, \ldots, N$.
- $N \in \{10, 20, 35\}$ so that $n \in \{45, 190, 595\}$ (increases quadratically with N).
- $(\beta_0, \beta_1) = (1, 1)$. We test $H_0 : \beta_1 = 1$ (true) and $H_0 : \beta_1 = 1.3$ (false).

Next table reports results over 40,000 replications on four methods:

I. A HC2 robust error method as a strawman.
II. A standard two-way clustered error method, implemented with function `vcovCL` from the `sandwich` `R` package. In this method, we include dyad fixed effects.
III. A linear mixed model with dyad RE through `lmer` in the `lme4` `R` package.
IV. The residual randomization test based on dyadic exchangeability.

Panel (A). $H_0 : \beta_1 = 1.0$

	Error-covariate, (ε_i, x_i)											
	(N,N)			(N, LN)			(LN, N)			(LN, LN)		
	Sample size, n											
Method	45	190	595	45	190	595	45	190	595	45	190	595
(I) HC2	18.07	29.21	41.50	22.90	40.16	53.65	14.03	25.87	38.14	15.77	30.27	46.06
(II) 2-clust.	11.32	9.62	8.20	13.55	12.34	10.83	11.27	9.71	7.93	13.28	12.19	11.00
(III) RE	9.37	6.07	5.80	11.91	8.71	7.68	9.00	6.72	5.63	11.13	8.72	7.45
(IV) RR	5.11	4.63	5.09	5.00	5.09	4.91	4.89	5.17	5.04	4.85	4.94	4.97

Panel (B). $H_0 : \beta_1 = 1.3$

	45	190	595	45	190	595	45	190	595	45	190	595
RE	25.57	67.05	98.12	25.46	49.68	84.51	28.61	70.11	98.24	29.12	55.20	85.70
RR	12.31	21.93	34.78	11.80	20.31	31.46	19.68	42.19	58.02	19.15	42.84	61.44

Table: Rejection rates (%) for HC2 errors, two-way cluster robust errors (“2-clust”), a random effects (RE) model, and residual randomization (“RR-dyadic”) under dyadic exchangeability using 2,500 resamples.

Application: International trade and currency unions

Rose and Engel (2002) studied whether currency unions are associated with increased economic integration. They work on a gravity model of trade:

$$\log(\text{TRADE}_{rc}) = \beta_0+\beta_1\text{CU}_{rc}+\beta_2\log(\text{GDP}_r)+\beta_3\log(\text{GDP}_c)+\beta_4\text{LANG}_{rc}+...+\varepsilon_{rc}.$$

In this context, Cameron and Miller (2014) discuss several standard error estimates: regular OLS, HC, two-way clustered, and dyadic.

The standard errors increase as more structure is imposed, but their analysis maintains the main result of Rose and Engel (2002) that β_1 is positive and significant.

(There is an important issue with missing data that remains unaddressed.)

Here, we can apply dyadic exchangeability along with other invariances as well.

Specification/Model				95% CI	for β_1
Panel (A)			estimate	lower	upper
(Rose and Engel, 2002)					
OLS			1.054 (0.38)	0.300	1.808
OLS, centered X			1.038 (0.33)	0.392	1.684
(Cameron and Miller, 2014)					
OLS, clustered by country 1			1.484 (0.28)	0.931	2.038
OLS, clustered by country 2			1.484 (0.62)	0.262	2.706
Panel (B)					
Res. Randomization	Invariant	Filter			
RR-perms.	$\mathcal{G}_n^{\mathrm{p}}$	{}		0.126	1.828
RR-signs	$\mathcal{G}_n^{\mathrm{s}}$	$*$		0.139	1.955
RR-double	$\mathcal{G}_n^{\mathrm{p+s}}$	$*$		0.076	1.941
RR-dyadic	$\mathcal{G}_n^{\mathrm{P}*}$	$*$		0.0519	0.2660
$*$	$*$	{continent}		-0.0626	0.3617
$*$	$*$	{language}		-0.0814	0.2303
$*$	$*$	{continent, language}		-0.0438	0.1965

Table: Panel (A) reports OLS-based results. Panel (B) reports results from residual randomization. The "filter" indicates that the dyadic permutation test uses only units sharing the same value for the variables included in the filter.

Concluding remarks

- Statistical inference is possible under invariance assumptions.
- Residual randomization offers a general implementation that leverages and extends randomization theory. The same underlying procedure is employed for any data structure ("plug-and-play").
- With this approach the problem structure dictates the analysis in a simple and unified manner.
- The resulting procedures are valid under weak conditions on the leverage of (X, ε) while allowing for $p \to \infty$ and heavy tailed data.

Thank You.

“Invariant Inference via Residual Randomization” (2022)
https://arxiv.org/pdf/1908.04218.pdf

“Introduction to Residual Randomization: The R Package `RRI`”
(Technical report, 2019)
https://cran.r-project.org/package=RRI

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

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

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).

Back

Missing data

Let $\mu_{ij} = \{0,1\}$ denote whether pair (i,j) is observed ($\mu_{ij} = 1$), and define $\mathsf{M} = (\mu_{ij}) \in \{0,1\}^{J\times J}$.

Any dependence between M and the data generating mechanism could significantly affect the results from statistical inference.

However, this issue is ignored by standard error methods. Moreover, the bootstrap methods effectively consider M as a random variable, and impose on M the same partial exchangeability structure as the observations. This is unrealistic because missingness is usually dyad-specific; e.g., whether the trade flow between two countries is missing or not may depend on their geographic and cultural affinity, both of which are not partially exchangeable.

Under residual randomization, let P denote a clustering of $\mathbb{N}$ such that the trade flow between any dyad (i,j) in the same cluster is observed. Suppose that

$$\varepsilon \overset{d}{=} \mathbf{g}\varepsilon \mid X, \mathsf{M}, \text{ for any } \mathbf{g} \in \mathcal{G}_n^{\mathsf{P}} \subseteq \mathcal{G}_n, \mathsf{M}. \tag{5}$$

Then, residual randomization is valid.

Back

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.

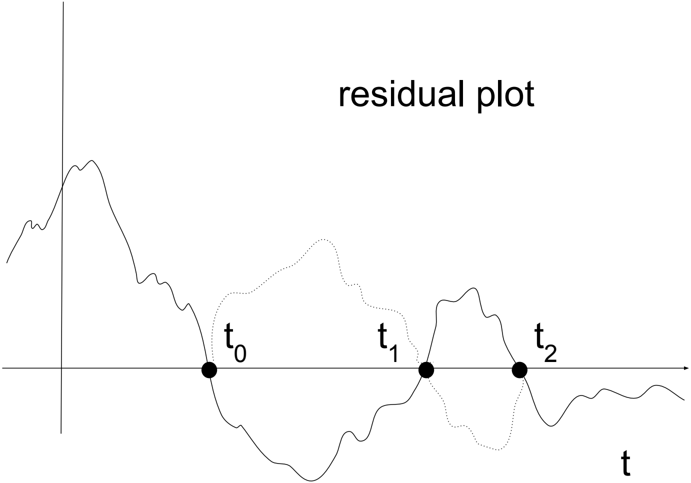
Which invariance works here?

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 $\mathcal{G}_n^{\text{ref}}$.

The "reflection" randomization test

1. Calculate the restricted residuals, $\hat{\varepsilon}^r$.

2. Order their absolute values, $|\hat{\varepsilon}^r|$, 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.