Covariate adjustment in multi-armed, possibly factorial experiments

Peng Ding

UC Berkeley, Statistics

November 3, 2022

Joint work with Anqi Zhao from the National University of Singapore

ArXiv version: https://arxiv.org/abs/2112.10557

Journal of the Royal Statistical Society, Series B

Randomized experiment

- Simple starting point with a binary treatment $Z_i \in \{1, 0\}$
- Outcomes for unit i
  - two potential outcomes $Y_i(1)$ and $Y_i(0)$: all fixed
  - one observed outcome $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$
- Complete randomization: Z_i's are random permutations
- OLS $Y_i \sim 1 + Z_i$ (Neyman 1923)
  - coefficient of Z_i: difference in means of outcomes
  - unbiased for average treatment effect $\tau = N^{-1} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}$
  - EHW/sandwich standard error is conservative for true standard error

Randomized experiment with covariates

- Additional pretreatment covariate vector x_i for unit i
- Center $\bar{x} = N^{-1} \sum_{i=1}^{N} x_i = 0$ to simplify the presentation
- Fisher's ANCOVA $Y_i \sim 1 + Z_i + x_i$
  - textbook recommendation
  - coefficient of Z_i is not unbiased for τ, but consistent
  - expected to improve efficiency, but may harm efficiency
  - efficiency gain if "slope of $Y_i(1) \sim 1 + x_i$" = "slope of $Y_i(0) \sim 1 + x_i$"
  - EHW standard error is conservative for true standard error
- Criticized by David A. Freedman (2008)
- Fisher's ANCOVA is not ideal

Lin (2013)'s fully interacted regression

- Lin's ANCOVA $Y_i \sim 1 + Z_i + x_i + Z_i x_i$
  - coefficient of Z_i is not unbiased for τ, but consistent
  - no harm to efficiency, asymptotically
  - superior to $Y_i \sim 1 + Z_i$ and $Y_i \sim 1 + Z_i + x_i$, asymptotically
  - EHW standard error is conservative for true standard error
- Theory for ANCOVA's does not assume correct linear model
  - so we can even run OLS for binary outcomes
- Design-based theory: random permutation of Z_i's drives inference
  - e.g. permutational central limit theorem (Li and Ding 2017 for review)

Extension to treatment with multiple levels

- Lin's theory shows the importance of the interaction term $Z_i x_i$
  - especially for heterogeneous slopes $Y_i(z) \sim 1 + x_i$ for $z = 0, 1$
- What if Z_i has multiple levels?
  - treatment $Z_i \in \{1, \dots, Q\}$
  - indicators $t_i = (1(Z_i = 1), \dots, 1(Z_i = Q))^\mathsf{T}$
  - potential outcomes: $(Y_i(1), \dots, Y_i(Q))$
  - mean vector $\bar{Y} = (\bar{Y}(1), \dots, \bar{Y}(Q))$
- Seems immediate to extend Neyman, Fisher and Lin

$$
\begin{aligned}
\hat{Y}_{\mathrm{N}} &: \quad Y_i \sim t_i \\
\hat{Y}_{\mathrm{F}} &: \quad Y_i \sim t_i + x_i \\
\hat{Y}_{\mathrm{L}} &: \quad Y_i \sim t_i + t_i \otimes x_i
\end{aligned}
$$

The story seems complete

$$\begin{aligned} \hat{Y}_{\mathrm{N}} &: \quad Y_i \sim t_i \\ \hat{Y}_{\mathrm{F}} &: \quad Y_i \sim t_i + x_i \\ \hat{Y}_{\mathrm{L}} &: \quad Y_i \sim t_i + t_i \otimes x_i \end{aligned}$$

- Coefficient of t_i estimates $\bar{Y} = (\bar{Y}(1), \ldots, \bar{Y}(Q))$ consistently
- EHW standard error is conservative for true standard error
- Asymptotic efficiency theory favors $\hat{Y}_{\mathrm{L}}$

Heuristics based on the derived linear model

- Design-based theory does not assume the linear model is correct
- Then why does OLS work?
- Derived linear model (termed by Oscar Kempthorne)
  - OLS decomposition: $Y_i(q) = \bar{Y}(q) + x_i^{\mathsf{T}} \gamma_q + \varepsilon_{\mathrm{L},i}(q)$
  - observed outcome $Y_i = \sum_{q=1}^{Q} 1(Z_i = q) Y_i(q)$ decomposes into

$$Y_i = t_i^{\mathsf{T}} \bar{Y} + (t_i \otimes x_i)^{\mathsf{T}} \gamma + \varepsilon_{\mathrm{L},i} \text{ where}$$

$$t_i = \begin{pmatrix} 1(Z_i = 1) \\ \vdots \\ 1(Z_i = Q) \end{pmatrix}, \quad \varepsilon_{\mathrm{L},i} = \sum_{q=1}^{Q} 1(Z_i = q) \varepsilon_{\mathrm{L},i}(q), \quad \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_Q \end{pmatrix}$$

  - residuals have complex mean and covariance

With moderate sample size, asymptotics can be misleading

- Consider an example with $N = 1000$, $Q = 16$ and $J = \dim(x) = 9$
  - $\hat{Y}_{\text{N}}$ requires estimating $Q = 16$ coefficients
  - $\hat{Y}_{\text{F}}$ requires estimating $Q + J = 25$ coefficients
  - $\hat{Y}_{\text{L}}$ requires estimating $Q(J+1) = 160$ coefficients

- A leading motivating example: factorial experiment
  - with K binary factors, $Q = 2^K$
  - e.g., if $K = 10$ then $Q = 2^{10} > 1000$

- Lessons learned
  - asymptotics for $\hat{Y}_{\text{L}}$ can be misleading in finite samples
  - going back to $\hat{Y}_{\text{F}}$ may not be a bad idea
  - even $\hat{Y}_{\text{N}}$ and $\hat{Y}_{\text{F}}$ can be demanding for sample size with large K

OLS is too demanding for data, so we use RLS

- RLS for restricted least squares
  - RLS = OLS under restrictions on the coefficients
  - classic in statistics (Rao 1973) and econometrics (Theil 1971)

- $Y_i \sim t_i + t_i \otimes x_i$ under linear restriction

$$R\theta = r$$

  - user-specified R and r for coefficient θ

- Restriction reflects our belief on the data-generating process
  - coefficient of t_i corresponds to the mean of the potential outcomes
  - coefficient of $t_i \otimes x_i$ corresponds to coefficient of x_i in $Y_i(q) \sim 1 + x_i$

Examples of RLS

- RLS is a theoretical device
  - $\hat{Y}_{\mathrm{L}}$ is unrestricted
  - $\hat{Y}_{\mathrm{N}}$ and $\hat{Y}_{\mathrm{F}}$ are RLS: some coefficients are simply 0 or equal
- Another illustrating example from 2^2 factorial experiment
  - assume prior belief of no interaction (may be wrong)

$$Y_i(++) - Y_i(+-) - Y_i(-+) + Y_i(--) = 0$$

  - restricts means of the potential outcomes and coefficients of x_i
- RLS is more relevant for 2^K factorial experiment with $K > 2$
  - often plausible to assume no higher order interactions
  - discuss later

Type I of RLS: correlation-only restriction

- No restriction on coefficient of t_i
- Only restricts coefficient of $t_i \otimes x_i$: correlation between $Y_i(q)$ and x_i
- Fisher's ANCOVA as special case: coef of $1(Z_i = q)x_i$'s are the same
- Can be more general: some combination of coefficients are zero
- Design-based theory of RLS under correlation-only restriction?

Design-based theory for correlation-only RLS

- $\hat{Y}_{\mathrm{r}} =$ coefficient of t_i
- $\hat{Y}_{\mathrm{r}}$ versus $\hat{Y}_{\mathrm{L}}$: Fisher's versus Lin's ANCOVA with binary Z
- $\hat{Y}_{\mathrm{r}}$ is consistent for potential outcome means $(\bar{Y}(1), \ldots, \bar{Y}(Q))$
- Loses asymptotic efficiency when restriction is wrong
- $\hat{Y}_{\mathrm{r}}$ achieves the same efficiency as $\hat{Y}_{\mathrm{L}}$ when restriction is correct
- $\hat{Y}_{\mathrm{r}}$ can be better than $\hat{Y}_{\mathrm{L}}$ when restriction is slightly wrong (finite N)
- Fisher's ANCOVA is not too bad: Schochet (2008) gave examples

Type II of RLS: separable restriction

- R is diagonal:

$$R = \begin{pmatrix} \rho_Y & 0 \\ 0 & \rho_\gamma \end{pmatrix}$$

- Separable restrictions on $\bar{Y}$ and γ
  - $\bar{Y}$: means of potential outcomes, or coef of t_i
  - γ: coef of x_i in $Y_i(q) \sim 1 + x_i$, or coef of $t_i \otimes x_i$
  - assume ρ_Y is not empty but ρ_γ can be empty

- Example: 2^K factorial experiment without higher order interactions

- Design-based theory of RLS under separable restriction?

Design-based theory for separable RLS

- $\hat{Y}_{\mathsf{r}} =$ coefficient of t_i
- $\hat{Y}_{\mathsf{r}}$ can be inconsistent if restriction on $\bar{Y}$ is wrong
- $\hat{Y}_{\mathsf{r}}$ is consistent if restriction on $\bar{Y}$ is correct
- $\hat{Y}_{\mathsf{r}}$ is as efficient as $\hat{Y}_{\mathrm{L}}$ if both restrictions are correct
- Best linear consistent estimator if both restrictions are correct and treatment effects are constant across units
  - $\hat{Y}_{\mathsf{r}}$ can outperform $\hat{Y}_{\mathrm{L}}$ in efficiency
  - extension of the classic Gauss–Markov theorem for RLS

A brief summary: roles of different restrictions

- Restriction on $\bar{Y}$:

  trade-off between asymptotic bias and variance

- Restriction on γ:

  trade-off between finite-sample performance and asymptotic efficiency

Inference with RLS

- Consistency (with proper restriction) and asymptotic normality $\hat{Y}_{\mathsf{r}}$
- Robust covariance estimation
  - we were unware of the discussion even in the classic literature
  - motivated by the algebraic fact:

$$\hat{\theta}_{\mathsf{r}} = (I - M_{\mathsf{r}} R)\hat{\theta}_{\mathrm{L}} + M_{\mathsf{r}} r$$

    where $M_r = (\chi_{\mathrm{L}}^{\mathsf{T}} \chi_{\mathrm{L}})^{-1} R^{\mathsf{T}} \{R(\chi_{\mathrm{L}}^{\mathsf{T}} \chi_{\mathrm{L}})^{-1} R^{\mathsf{T}}\}^{-1}$ with design matrix χ_{L}
  - $\hat{\Psi}_{\mathsf{r}} =$ upper $Q \times Q$ submatrix of double-decker-taco (DDT) covariance:

$$(I - M_{\mathsf{r}} R)(\chi_{\mathrm{L}}^{\mathsf{T}} \chi_{\mathrm{L}})^{-1} \chi_{\mathrm{L}}^{\mathsf{T}} \mathrm{diag}(\hat{\varepsilon}_{\mathsf{r},i}^2) \chi_{\mathrm{L}} (\chi_{\mathrm{L}}^{\mathsf{T}} \chi_{\mathrm{L}})^{-1} (I - M_{\mathsf{r}} R)^{\mathsf{T}}$$

    where $\hat{\varepsilon}_{\mathsf{r},i}$'s are residuals from RLS

Practical implications of RLS

- RLS = OLS with transformed regressors
  - some coefficients are 0
  - some coefficients are equal: collapse regressors
  - some linear combination of coefficients are 0: re-parametrization
- DDT from RLS = EHW from the corresponding OLS
  (a useful algebraic fact)
- Example: Fisher's ANCOVA + EHW standard error directly
- Theoretically: RLS is a tool to unify the results
- Practically: enough to use "OLS + EHW" with proper specification

Back to the leading motivation: 2^K factorial experiments

- K binary factors $z_1, \ldots, z_K \in \{-1, +1\}$
- Main effect of factor k:

$$\tau_{\{k\}} = \frac{1}{2^{K-1}} \sum_{q: z_k = +1} \bar{Y}(q) - \frac{1}{2^{K-1}} \sum_{q: z_k = -1} \bar{Y}(q)$$

- Interaction between factors k and k':

$$\tau_{\{k,k'\}} = \frac{1}{2^{K-1}} \sum_{q: z_k z_{k'} = +1} \bar{Y}(q) - \frac{1}{2^{K-1}} \sum_{q: z_k z_{k'} = -1} \bar{Y}(q)$$

- Can define higher order interactions similarly
- A subset of $\{1, \ldots, K\}$ indicates a factorial effect: $\tau_{\mathcal{K}}$

Factor-saturated regression for 2^K factorial experiments

- Can simply run treatment-based regression
  - OLS $Y_i \sim t_i$; linear transformation of coefficient

- Factor-based regression is more straightforward and popular

$$
\begin{aligned}
& Y_i \sim 1 + \sum_{k=1}^{K} Z_{ik} + \sum_{k \neq k'} Z_{ik} Z_{ik'} + \cdots + \prod_{k=1}^{K} Z_{ik} \\
\text{or} \quad & Y_i \sim 1 + \sum_{\mathcal{K} \in \mathcal{P}_K} Z_{i,\mathcal{K}}
\end{aligned}
$$

  - regressor $Z_{i,\mathcal{K}} = \prod_{k \in \mathcal{K}} Z_{ik}$
  - $\mathcal{P}_K$ is the set containing all non-empty subsets of $\{1, \ldots, K\}$

Factor-based regression has many advantages

- Coefficient of $Z_{i,\mathcal{K}}$ as estimates of factorial effects
- Obtain EHW standard errors directly
- Can incorporate covariates

$$
\begin{aligned}
\tilde{\tau}_{\mathrm{N}} &: \quad Y_i \sim 1 + \sum_{\mathcal{K} \in \mathcal{P}_K} Z_{i,\mathcal{K}} \\
\tilde{\tau}_{\mathrm{F}} &: \quad Y_i \sim 1 + \sum_{\mathcal{K} \in \mathcal{P}_K} Z_{i,\mathcal{K}} + x_i \\
\tilde{\tau}_{\mathrm{L}} &: \quad Y_i \sim 1 + \sum_{\mathcal{K} \in \mathcal{P}_K} Z_{i,\mathcal{K}} + x_i + \sum_{\mathcal{K} \in \mathcal{P}_K} Z_{i,\mathcal{K}} \cdot x_i
\end{aligned}
$$

  - $\tilde{\tau}_{\mathrm{N}}$ and $\tilde{\tau}_{\mathrm{F}}$: RLS with correlation only restriction
  - $\tilde{\tau}_{\mathrm{L}}$ is asymptotically the most efficient, but demanding for sample size

Factor-unsaturated regression for 2^K factorial experiments

- Higher order interactions are small or unimportant
- Simpler regression specifications (RLS with separable restriction)

$$
\begin{aligned}
\tilde{\tau}_{\mathrm{N,r}} &: \quad Y_i \sim 1 + \sum_{\mathcal{K} \in \mathcal{F}_+} Z_{i,\mathcal{K}} \\
\tilde{\tau}_{\mathrm{F,r}} &: \quad Y_i \sim 1 + \sum_{\mathcal{K} \in \mathcal{F}_+} Z_{i,\mathcal{K}} + x_i \\
\tilde{\tau}_{\mathrm{L,r}} &: \quad Y_i \sim 1 + \sum_{\mathcal{K} \in \mathcal{F}_+} Z_{i,\mathcal{K}} + x_i + \sum_{\mathcal{K} \in \mathcal{F}_+} Z_{i,\mathcal{K}} \cdot x_i
\end{aligned}
$$

  - recall $Z_{i,\mathcal{K}} = \prod_{k \in \mathcal{K}} Z_{ik}$
  - $\mathcal{F}_+$ is a subset of $\mathcal{P}_K$
  - examples of $\mathcal{F}_+$: all main effects; or up to second order interactions

Factor-unsaturated regression for 2^K factorial experiments

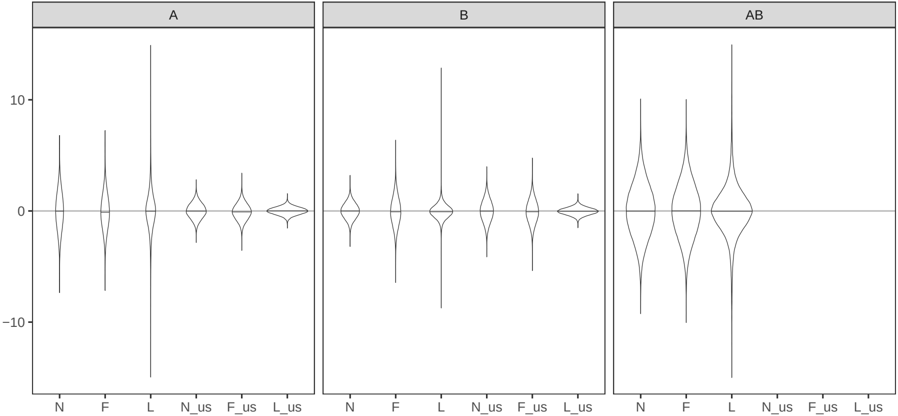
- Consistent if nuisance interactions are 0
- Consistent if treatment groups have equal sizes
- Can be inconsistent if nuisance interactions are not 0 and treatment groups have varying sizes
- Asymptotically normal and EHW standard error is conservative
- More efficient than estimates from factor-saturated regression if nuisance interactions are 0 and treatment effects are constant

Numeric examples: 2^2 factorial experiment with $J = 20$

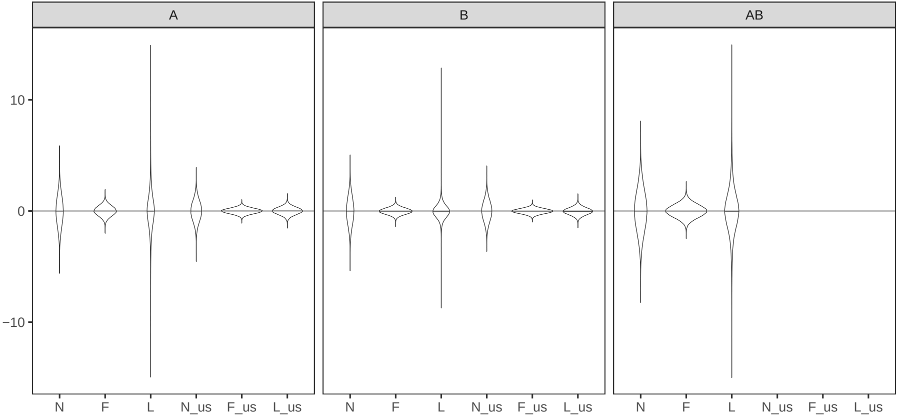
- $N = 100$ with $(N_{--}, N_{-+}, N_{+-}, N_{++}) = (22, 23, 24, 31)$
- $Y_i(q)$ normal linear models, without individual interactions
- Many possible regressions:

	regression equation
N	$Y_i \sim 1 + A_i + B_i + A_iB_i$
F	$Y_i \sim 1 + A_i + B_i + A_iB_i + x_i$
L	$Y_i \sim 1 + A_i + B_i + A_iB_i + x_i + A_ix_i + B_ix_i + A_iB_ix_i$
N_us	$Y_i \sim 1 + A_i + B_i$
F_us	$Y_i \sim 1 + A_i + B_i + x_i$
L_us	$Y_i \sim 1 + A_i + B_i + x_i + A_ix_i + B_ix_i$

Example 1: no interaction, heterogeneous correlations



Example 2: no interaction, homogeneous correlations



Discussion I: data-dependent restrictions

- Assumed data-independent restriction on coefficients $R\theta = r$
- Important next steps: data-dependent restrictions $\hat{R}\theta = \hat{r}$
  - model selection and post-selection inference
  - principles for factorial effects: sparsity, hierarchy, heredity (Wu and Hamada 2009)
- Theory for design-based model selection and inference is quite sparse (Bloniarz et al 2016)
- Ongoing research

Discussion II: rerandomization in design

Section S5

- Regression adjustment and rerandomization are duals
- Rerandomization enforces covariate balance in the design stage
  - accept treatment allocations with small variability of $\hat{x}(q)$'s across q
  - many choices of contrasts with $Q \geq 2$
- Theoretical guarantees of rerandomization
  - $\hat{Y}_{\mathrm{L}}$: no additional asymptotic efficiency gain
  - $\hat{Y}_{\mathrm{r}}$: no additional asymptotic efficiency gain if restriction on γ is correct; additional asymptotic efficiency gain otherwise

Discussion III: fractional factorial experiment

- If we believe no higher order interactions
  - option 1: full factorial experiment + unsaturated regression
  - option 2: fractional factorial experiment + unsaturated regression
- e.g. factors A, B and C, assuming no interactions

A	B	C=AB
$-$	$-$	$+$
$-$	$+$	$-$
$+$	$-$	$-$
$+$	$+$	$+$

- The choice is an estimation-exploration trade-off
  - fractional factorial experiments improves estimation efficiency
  - full factorial experiments allows for exploring interactions

Related papers

- Zhao, A. and Ding, P. (2022+). Regression-based causal inference with factorial experiments: estimands, model specifications, and design-based properties. *Biometrika*
- Zhao, A. and Ding, P. (2022+). Covariate adjustment in multi-armed, possibly factorial experiments. *Journal of the Royal Statistical Society, Series B*