Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates

Min Zhang, Anastasios A. Tsiatis and Marie Davidian

(2008, Biometrics)

Presented by Rui Zhang

April 17, 2012

Randomized Clinical Trials: introduction

- Clinical Trials:
  - Outcome Y: continuous, binary or categorical, usually recorded before (Y_0) and after treatment (Y_1), possibly multiple times after treatment carry-out, forming longitudinal data
  - Baseline covariates X: such as demographic information, recorded before treatment starts. Y_0 is also a baseline covariates.
  - Treatment arm Z: treatment group assignment
- Randomized Clinical Trials: patients assigned to treatment arms randomly: $X, Y_0 \perp Z$

Randomized Clinical Trials: analyse

How do we compare outcomes from different treatment arms?
One answer: comparing outcome means across treatment arms

- Y is continuous: ANOVA, t-test, linear regression
- Y is binary: odds ratio, logistic regression
- In summary, (Generalized) Linear regression:
  $g(\mathbb{E}(Y|Z)) \sim \alpha + \beta Z$
  β here represents treatment effect(s), can either be a number or a vector.

Then why do researchers also record baseline variable X and Y_0?

- Someone argued that baseline variable X and Y_0 can also contribute to β estimation efficiency, when they are correlated with Y.
- A more efficient method (looking quite suspicious): regress Y against treatments Z, adjusting for X or Y_0:

$$g(\mathbb{E}(Y|Z,X)) \sim \alpha + \beta' Z + \gamma X \tag{1}$$

Yang (2001) *Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Trial*

$$g(\mathbb{E}(Y|Z,X)) \sim \alpha + \beta' Z + \gamma Y_0 \tag{2}$$

After adjusting for X or Y_0

- Interpretation of results changed:
  - So now we are estimating treatment effects conditioning on X or Y_0
  - Instead of estimating the population-specific treatment effect, we are estimating subject-specific effect.
- Making some assumptions:
  In both two models, we are assuming linear regression relationship between outcome and baseline variables.

After adjusting for X or Y_0: do we get a correct result and gain efficiency?

- Y is continuous:
  - β' and β coincide in magnitude
  - $X \perp Z$, X serves as a precision variable (the same argument for Y_0)
  - Adjusting for X or Y_0 reduces treatment effect

  Regression adjustment is a good solution.
- Y is binary and we used a log/logit link
  - β' is moved further away from NULL, i.e., 0
  - $s.d.(\hat{\beta}')$ is larger
  - It is hard to tell if power of Wald test increases or decreases

Simulation results for X adjustment

		Adjusted	Un-adjusted
Linear Model	MC mean in $\hat{\beta}'$	2.001	2.000
$\beta = 2.000$	MC mean in $s.e.(\hat{\beta}')$	0.04161	0.008160
Logistic Model	MC mean in $\hat{\beta}'$	1.924	2.018
$\beta = 1.910$	MC mean in $s.e.(\hat{\beta}')$	0.2089	0.2165

Another solution: semi-parametric model

- What is a semi-parametric model?
  Under certain circumstances, we are only interested in part of the distribution characteristics (such as first moment), but do not want to estimated the whole distribution explicitly.
  For example, we want to estimate $g(X)$ from $\mathbb{E}(Y|X) = g(X)$ without assuming any higher moment informations.
- Why they choose it?
  In this paper the estimating equation was proved to be

$$\sum_{i=1}^{n}\{m(Y_i, Z_i; \theta) - \sum_{g=1}^{k}\{I(Z_i = g) - \pi_g\} \times \mathbb{E}\{m(Y, Z; \theta)|X_i, Z = g\}\} = 0$$

"separates estimation of the treatment effect from the adjustment"

Semi-parametric model Construction

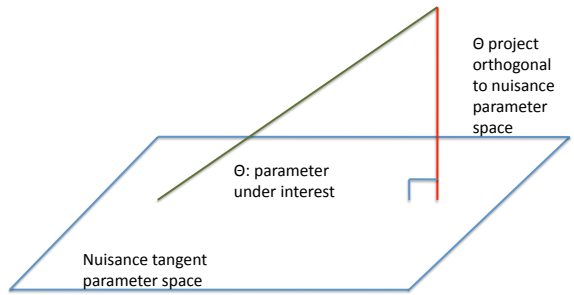
Suppose we can write out the set of density functions representing our data (Y, Z, X) as

$$\begin{aligned}\mathscr{P} = \quad & \{p(x, y, z, \theta, \pi, \eta, \psi) : \theta \in \Theta \subset R^K, \\ & \pi \in B_1 \subset R^K, \alpha \in B_2 \subset R^s, \psi \in B_3 \subset R^r\}\end{aligned}$$

- $\theta = (\alpha, \beta)^T$, the parameter of interest
- π is treatment arm assignment probability, a known constant vector
- η is nuisance parameters used to describe joint distribution between Y and Z.
- ψ is nuisance parameter used to describe joint distribution between Y, X and Z.
- Assume s and r to be infinite, that is, η and ψ, nuisance parameters are of infinite dimension

Finding an estimation equation: a sketch

Starting from a parametric model, in which parameters θ_1, θ_2 coming from finite Euclidean Space, in which we are only interested in estimating θ_1



- Nuisance Tangent Space: spanned by score function of θ_2
- We do not want to estimate θ_2 nor have any information of it
- So we will focus on the part of θ_1 orthogonal to plane spanned by θ_2

Sketch continued

- In semi-parametric models, we start with a parametric sub-model, which
  1. is contained in semi-parametric model
  2. contains the truth
- After finding some estimating equation, we can generalize it into semi-parametric model case (hopefully)

Next...

- Simulations using binary outcome and logit link
- And more proof details on semi-parametric model?