Leveraging auxiliary covariates to improve efficiency for clinical trials: a general framework

Min Zhang

Department of Biostatistics

University of Michigan, Ann Arbor

Why Covariate Adjustment for Clinical Trials?

Extensive literature on theory and methods for robust covariate adjustment; still underutilized in current practice

Concerns and Responses:

- Unnecessary to adjust for imbalance in covariates due to randomization

  Response: Better to. Leveraging covariates always improves efficiency

- Interested in marginal treatment effect, not adjusted/conditional treatment effect

  Response: Will not change the estimand or the target

- Concerns about model misspecification

  Response: No assumptions on correct modeling required. Robustness guaranteed by randomization

Under the condition that covariate adjustment is carried out properly

General Framework

Results are based on:

- Zhang, M., Tsiatis, A.A., Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* 64,707-715
  - Study influence functions of all consistent and asymptotically normal estimators for parameters quantifying treatment effect
  - Provide a theoretical foundation and a general framework for covariate adjustment

Notation

Data from a K-arm randomized trial:

$$(Y_i, X_i, Z_i), i = 1, ..., n,$$

- Y_i: outcome of interest
- X_i: vector of auxiliary baseline covariates
- $Z_i = 1, ..., K$: indicator of treatment group
- $P(Z = g) = \pi_g$, $g = 1, ..., K$, and $\sum_{g=1}^{K} \pi_g = 1$
- $Z \perp\!\!\!\perp X$, ensured by randomization

When $K = 2$:

- $A_i = I(Z_i = 2) = 0, 1$: treatment indicator
- $P(A_i = 1) = \pi$: probability of receiving $A = 1$

Estimand

Estimand: β, a vector of parameters for making treatment comparison

Examples:

- **Continuous response:** $E(Y|Z) = \beta_1 + \beta_2 I(Z = 2) + \cdots + \beta_K I(Z = K)$

  where $\beta = (\beta_1, \beta_2, \cdots, \beta_K)^T$

- **Binary response:**
  - Odds ratio: logit $\{P(Y = 1|A)\} = \beta_1 + \beta_2 A$
  - Risk difference: $P(Y = 1|A) = \beta_1 + \beta_2 A$
  - Relative risk: $\log P(Y = 1|A) = \beta_1 + \beta_2 A$

  where $\beta = (\beta_1, \beta_2)^T$

Estimand

- **Continuous longitudinal response:**
  - $E(Y_{ij}|A_i) = \alpha + \{\beta_1 + \beta_2 A_i\}t_{ij}$
    where $\beta = (\beta_1, \beta_2)^T$; $\gamma = \alpha$
- **Survival response:**
  - Hazards ratio: $\lambda(t|A) = \lambda_0(t)\exp(\beta A)$
  - Difference in restricted mean lifetime by t
  - Difference in survival probability at t

Key Theoretical Results

- **Result 1**

  The class of *all unbiased estimating functions* for θ based on (Y, X, Z) are:

$$m^*(Y, X, Z; \theta) = m(Y, Z; \theta) - \sum_{g=1}^{K} \{I(Z = g) - \pi_g\} a_g(X)$$

  - $m(Y, Z; \theta)$ is any unbiased estimating function
  - $\theta = (\beta, \alpha)$: α other parameters estimated together with β

- **Result 2**

  Given $m(Y, Z; \theta)$, the *optimal choice of* $a_g(X)$ is

$$E\{m(Y, Z; \theta) \mid X, Z = g\}$$

Key Theoretical Results

Two treatment groups: $A_i = 0, 1$

- **Result 1**

  The class of *all unbiased estimating functions* for θ based on (Y, X, A) are:

  $$m^*(Y, X, A; \theta) = m(Y, A; \theta) - (A - \pi)a(X)$$

  where is $m(Y, A; \theta)$ is any unbiased estimating function, $\theta = (\beta, \alpha)$: α other parameters needed to be estimated together with β

- **Result 2**

  Given $m(Y, A; \theta)$, the *optimal choice of* $a(X)$ is

  $$E\{m(Y, A; \theta) \mid X, A = 1\} - E\{m(Y, A; \theta) \mid X, A = 0\}$$

Implementation: General Strategy

- **Step 1:** Obtain *"unadjusted" estimator* $\widehat{\theta}$ by solving

$$\sum_{i=1}^{n} m(Y_i, Z_i; \theta) = 0$$

- **Step 2:** For each group g, treating $m(Y_i, g; \widehat{\theta})$ as data *develop a regression model*

$$E\{m(Y, g; \widehat{\theta}) \mid X, Z = g\} = q_g(X, \zeta_g)$$

- **Step 3:** Obtain the *"adjusted" estimator* $\tilde{\theta}$ by solving

$$\sum_{i=1}^{n} \left[m(Y_i, Z_i; \theta) - \sum_{g=1}^{K} \{I(Z_i = g) - \pi_g\} q_g(X_i, \widehat{\zeta}_g) \right] = 0$$

For $K = 2$, equivalently

$$\sum_{i=1}^{n} \left[m(Y_i, Z_i; \theta) - (A_i - \pi)\{q_1(X_i, \widehat{\zeta}_1) - q_0(X_i, \widehat{\zeta}_0)\} \right] = 0$$

Implementation: General Strategy

Recommendations for step 2:

- Build regression models:

$$E\{m(Y, g; \widehat{\theta}) \mid X, Z = g\} = q_g(X, \zeta_g) = \{q_{g1}(X, \zeta_{g1}), \ldots, q_{gr}(X, \zeta_{gr})\}^T$$

  - Specify a *linear* regression model, including intercept and *basis functions* in X (eg, polynomials terms in X, interaction terms, splines)
  - Obtain estimates $\widehat{\zeta}_g = (\widehat{\zeta}_{g1}^T, ..., \widehat{\zeta}_{gr}^T)^T$ via *OLS*
  - Guaranteed to be more efficient (unless covariates are not predictive at all)

- **Neither validity nor efficiency gain require assumption of correct modeling**
- In practice, often $E\{m(Y, Z; \theta) = B(Z; \theta)\{Y - f(Z; \theta)\}$, equivalently, step 2 becomes modeling for $E(Y|X, Z = g)$
  - Popular models for Y exist and can be used

Implementation: Examples

$K = 2$

X^*: a vector including 1, X, and other basis functions of X

Difference in Mean/risk

- Unadjusted estimating equation

  $\sum_{i=1}^n m(Y_i, A_i; \beta) = \sum_{i=1}^n \begin{bmatrix} 1-A_i \\ A_i \end{bmatrix} (Y_i - \beta_1 - \beta_2 A_i) = 0$

- Build regression models for $Y|X, A = a, a = 0, 1$
  - Continuous Y: $E(Y|X, A = a) = \zeta_a X^*$
  - Alternatively for binary Y: $\text{logit} P(Y = 1|X, A = a) = \zeta_a X^*$

  And obtain predicted value for Y_i under each a, denoted by $\widehat{Y}_i^a$.

- Solve augmented estimating equation:

  $\sum_{i=1}^n \left\{ \begin{bmatrix} 1-A_i \\ A_i \end{bmatrix} (Y_i - \beta_1 - \beta_2 A_i) - (A_i - \widehat{\pi}) \begin{bmatrix} -(\widehat{Y}_i^0 - \beta_1) \\ \widehat{Y}_i^1 - \beta_1 - \beta_2 \end{bmatrix} \right\} = 0$

Implementation: Examples

Log odds ratio when Y is binary

- Unadjusted estimating function

  $\sum_{i=1}^{n} m(Y_i, Z_i; \beta) = \sum_{i=1}^{n} \left[\begin{smallmatrix} 1-A_i \\ A_i \end{smallmatrix} \right] \{Y_i - \text{expit}(\beta_1 + \beta_2 A_i)\}$

- Build regression models for $Y|X, A = a, a = 0, 1$
  - $\text{logit} P(Y = 1|X, A = a) = \zeta_a X^*$
  - Alternatively, $E(Y|X, A = a) = \zeta_a X^*$

  And obtain predicted value for Y_i under each a: $\widehat{Y}_i^a$.

- Solve augmented estimating equation:

$$\sum_{i=1}^{n} \left\{ \left[\begin{smallmatrix} 1-A_i \\ A_i \end{smallmatrix} \right] \{Y_i - \text{expit}(\beta_1 + \beta_2 A_i)\} \right.$$
$$\left. -(A_i - \widehat{\pi}) \left[\begin{smallmatrix} -\{\widehat{Y}_i^0 - \text{expit}(\beta_1)\} \\ \widehat{Y}_i^1 - \text{expit}(\beta_1 + \beta_2) \end{smallmatrix} \right] \right\} = 0$$

Implementation: Examples

Difference in slopes when Y is longitudinal

- Unadjusted analysis using SAS *proc mixed*

  $Y_{ij} = \alpha + (\beta_1 + \beta_2 A_i)t_{ij} + b_{0i} + b_{1i}t_{ij} + e_{ij}$, where
  $(b_{0i}, b_{1i}) \sim N(0, D), e_{ij} \sim N(0, \sigma_e^2)$ to obtain $\widehat{D}, \widehat{\sigma}_e^2$ for calculating $\widehat{V}_i$

- – For simplicity fit $E(Y|X, A = a) = \zeta_a X^*$ using OLS
  – Obtain predicted value for $Y_i = (Y_{i1}, \ldots, Y_{in_i})$ under each a: $\widehat{Y}_i^a$

- Solve augmented estimating equation:

$$\sum_{i=1}^{n} \Big\{ \big[1_{n_i}, t_i, A_i t_i\big]^T \widehat{V}_i^{-1} \{Y_i - (1_{n_i}, t_i, A_i t_i)[\begin{smallmatrix}\alpha\\ \beta_1\\ \beta_2\end{smallmatrix}]\}$$

$$-(A_i - \widehat{\pi})\{Q_i^1(Y_i^1 - \mu_i^1) - Q_i^0(Y_i^0 - \mu_i^0)\Big\} = 0,$$

where $Q^a = \Big[[1_{n_i}, t_i, at_i]^T \widehat{V}_i^{-1}, \mu^a = (1_{n_i}, t_i, at_i)[\begin{smallmatrix}\alpha\\ \beta_1\\ \beta_2\end{smallmatrix}], a = 0, 1$

Variance Estimator/Hypothesis Testing

- For making inference, one needs to estimate the variance of $\widetilde{\beta}$ consistently
  - Adjusted estimator $\widetilde{\beta}$ is obtained by solving estimating equations
  - By standard M estimation theory, the variance can consistently estimated by sandwich variance estimator
- Within the same framework, on can derive robust, more powerful tests through covariate adjustment
  - Wald test based on adjusted $\widetilde{\beta}$
  - For example, covariate adjusted more powerful Wilcoxon-rank sum test/Kruskal-Wallis

Simulations

Estimate difference in slopes in a linear mixed model

$$\text{Model: } Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}, j = 1, \ldots, m_i$$

Data generation:

- Baseline variables: 3 continuous variables
- Outcome:
  - Subject-specific intercept: $\beta_{0i} = 0.5 + 0.2X_{1i} + 0.5X_{2i} + b_{0i}$
  - Subject-specific slope: $\beta_{1i} = \alpha_{0g} + \alpha_{1g}X_{1i}^2 + \alpha_{2g}X_{2i} + \alpha_{3g}X_{3i} + b_{1i}$, for $g = 1, 2$

Methods:

- Aug 1: augmented method; augmentation term represents the true form
- Aug 3: augmented method; augmentation term involves quadratic term in X
- Usual: fit a linear mixed model adjusting for covariates

Simulations

Estimate difference in slopes in a linear mixed model

Method	True	MC Bias	MC SD	Ave. SE	Cov. Prob	Rel. Eff.
			Mild Association			
Unadjusted	0.300	0.000	0.100	0.099	0.951	1.00
Aug. 1	0.300	-0.001	0.095	0.094	0.951	1.10
Aug. 3	0.300	-0.001	0.096	0.094	0.950	1.08
			Moderate Association			
Unadjusted	0.300	0.000	0.107	0.106	0.949	1.00
Aug. 1	0.300	-0.001	0.097	0.095	0.951	1.22
Aug. 3	0.300	-0.001	0.097	0.095	0.952	1.21
			Strong Association			
Unadjusted	0.300	0.000	0.116	0.115	0.950	1.00
Aug. 1	0.300	-0.001	0.098	0.096	0.951	1.41
Aug. 3	0.300	-0.001	0.098	0.096	0.951	1.39

Monte Carlo size = 5000, $n = 200, m_i \approx 10$

Final Remarks

- Theory and a general framework in:

  Zhang, M., Tsiatis, A.A., Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* 64,707-715

- Other methods and implementations are available in the literature, with influence functions belonging to the class studied in Zhang, Tsiatis and Davidian (2008). Therefore, the general statements regarding robustness and efficiency gain still apply

- Other work within the framework or based on similar augmentation ideas:
  - Tsiatis, A.A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials. *Statistics in Medicine* 27, 4658-4677
  - Lu, X. and Tsiatis, A.A. (2008). Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika* 95, 679-694
  - Zhang, M. and Gilbert, B. P. (2010). Increasing the efficiency of prevention trials by incorporating baseline covariates. *Statistical Applications in Infectious Diseases*. Vol. 2: Iss. 1, Article 1
  - Zhang, M. (2015). Robust methods to improve efficiency and reduce bias due to chance imbalance in estimating survival curves in randomized clinical trials. *Lifetime Data Analysis*, 21(1),119-137