

# Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates

---

Marie Davidian  
Department of Statistics  
North Carolina State University



<http://www.stat.ncsu.edu/~davidian>

(Joint work with M. Zhang, X. Lu, and A.A. Tsiatis)

# Outline

---

1. Introduction
2. Covariate adjustment
3. Focus of inference
4. Semiparametric model
5. Estimating functions using auxiliary covariates
6. Implementation
7. Simulations
8. Applications
9. Discussion

# 1. Introduction

---

**Primary objective of a randomized clinical trial:** *Compare treatments* with respect to some *outcome* of interest, for example

- *Continuous outcome*: Compare *treatment means*
- *Binary outcome*: Compare based on *odds ratios*
- *Longitudinal study*: Compare *treatment-specific slopes* in a *linear mixed model (continuous response)*
- *Time to event*: Compare based on *hazard ratios*

**In addition to outcome and treatment assignment:** *Auxiliary baseline covariates*

- *Demographic*, *physiologic*, *genetic/genomic* characteristics
- Prior *treatment* and *medical history*
- *Baseline* measure(s) of the outcome

# 1. Introduction

---

**PURSUIT trial:** “*Platelet Glycoprotein IIb/IIIa in Unstable Angina: Receptor Suppression Using Integrilin Therapy*”

- International multi-center clinical trial
- Compare anti-coagulant Integrilin+heparin and aspirin to heparin and aspirin (control) in subjects with ACS
- *Outcome*: Binary composite death or MI at 30 days  $\Rightarrow$  *log-odds ratio for Integrilin relative to control*
- Data from 5710 subjects
- 35 *auxiliary baseline covariates*, including age, height, weight, gender, race, geographic region, smoking status, diastolic and systolic blood pressure, creatine kinase and creatine kinase-MB ratios, disease history, treatment history, ...

# 1. Introduction

---

## AIDS Clinical Trials Group (ACTG) 175:

- 4 groups: ZDV monotherapy, ZDV+ddI, ZDV+zalcitabine, ddI monotherapy
- Data from 2139 subjects
- *Outcome*: CD4 count (cells/mm<sup>3</sup>) at 20±5 weeks
- 12 *auxiliary baseline covariates*: CD4 count (cells/mm<sup>3</sup>), CD8 count (cells/mm<sup>3</sup>), age (years), weight (kg), Karnofsky score (scale of 0-100), and indicator variables for hemophilia, homosexual activity, history of intravenous drug use, race (0=white, 1=non-white), gender (0=female), antiretroviral history (0=naive, 1=experienced), and symptomatic status (0=asymptomatic)

## 2. Covariate adjustment

---

**Ordinarily:** Inferences on treatment comparisons based *only on data on outcome and treatment assignment*

**“Covariate adjustment:”** Auxiliary baseline covariates may be *associated with outcome*

- Account for *chance imbalances*
- Potential to *gain efficiency*
- *Extensive literature:* Senn (1989), Hauck et al. (1998), Koch et al. (1998), Tangen and Koch (1999), Pocock et al. (2002), Lesaffre and Senn (2003), Grouin et al. (2004), ...
- *Extensive concerns:* Potential *bias* due to post hoc (*subjective*) selection of covariates to use, and...
- ...temptation for a “*fishing expedition*” for *most dramatic* effect
- ⇒ *Trialists* and *regulatory authorities* reluctant to endorse

## 2. Covariate adjustment

---

**Standard approach to adjustment:** *Direct regression modeling*

- Model outcome as a function of treatment assignment *and* covariates, e.g., via *ANCOVA model*
- $\Rightarrow$  *Inextricable link* between parameters involved in treatment comparisons and the “*adjustment*”

**Our objective:** A *general methodology* for using auxiliary covariates that leads to *more efficient* estimators and tests

- Based on the *theory of semiparametrics* (e.g., Tsiatis, 2006)
- *Separates* parameters involved in treatment comparisons from the “*adjustment*” . . .
- . . .and hence leads to a *principled approach* to implementation that can obviate the usual concerns

# 3. Focus of inference

---

**Setting:** a  $k$ -arm randomized trial,  $k \geq 2$ ,  $n$  subjects

**Data:**  $(Y_i, X_i, Z_i)$ ,  $i = 1, \dots, n$  iid

- $Y = \text{outcome}$
- $X =$  vector of *auxiliary baseline covariates*
- $Z = g$  is *indicator* of assignment to treatment group  $g = 1, \dots, k$
- $P(Z = g) = \pi_g$ ,  $\sum_{g=1}^k \pi_g = 1$ ,  $\pi = (\pi_1, \dots, \pi_k)^T$  *known*

**Key:** Randomization *guarantees*  $Z \perp\!\!\!\perp X$

- “ $\perp\!\!\!\perp$ ” means *independent of*

**Focus:** Parameter relevant to making *treatment comparisons*,  $\beta$

- $\beta$  defined in an appropriate *statistical model based on  $Y$  and  $Z$  only*

### 3. Focus of inference

---

**Example 1:**  $k = 2$  arm trial, *continuous response*  $Y$

$$E(Y | Z) = \beta_1 + \beta_2 I(Z = 2), \quad \beta_1 = E(Y | Z = 1), \quad \beta = (\beta_1, \beta_2)^T$$

- $\beta_2 = E(Y | Z = 2) - E(Y | Z = 1) =$  *difference in treatment means*

**Example 2:**  $k = 3$  arm trial, *continuous response*  $Y$

$$E(Y | Z) = \beta_1 I(Z = 1) + \beta_2 I(Z = 2) + \beta_3 I(Z = 3), \quad \beta = (\beta_1, \beta_2, \beta_3)^T$$

- *Contrasts* of treatment means

**Example 3:**  $k = 2$  arm trial, *binary response* ( $Y = 0, 1$ )

$$\text{logit}\{E(Y | Z)\} = \text{logit}\{P(Y = 1|Z)\} = \beta_1 + \beta_2 I(Z = 2), \quad \beta = (\beta_1, \beta_2)^T$$

- $\beta_2 =$  *Log-odds ratio* for treatment 2 relative to treatment 1

### 3. Focus of inference

---

**Example 4:**  $k = 2$ , *longitudinal continuous response*

- *Linear mixed model*,  $(b_{0i}, b_{1i})^T \stackrel{\text{iid}}{\sim} \mathcal{N}(0, D)$ ,  $e_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$

$$Y_{ij} = \alpha + \{\beta_1 + \beta_2 I(Z_i = 2)\}t_{ij} + b_{0i} + b_{1i}t_{ij} + e_{ij}, \quad j = 1, \dots, m_i$$

$$\beta = (\beta_1, \beta_2)^T, \quad \gamma = \{\alpha, \sigma_e^2, \text{vech}(D)\}^T$$

- *Marginal model*  $E(Y_{ij} | Z_i) = \alpha + \{\beta_1 + \beta_2 I(Z_i = 2)\}t_{ij}$

$$j = 1, \dots, m_i, \quad \beta = (\beta_1, \beta_2)^T, \quad \gamma = \alpha$$

**Example 5:**  $k = 2$ , *censored time to event*

- Data  $(U_i, \Delta_i, X_i, Z_i)$ ,  $U_i = \min(T_i, C_i)$ ,  $\Delta_i = I(T_i \leq C_i)$
- *Proportional hazards model*  $\lambda(t | Z) = \lambda_0(t) \exp\{\beta I(Z = 2)\}$
- $\beta$  is the *log-hazard ratio* for treatment 2 relative to 1

# 3. Focus of inference

---

**In general:** Treatment comparisons may be formalized as *functions of elements* of  $\beta$

- *Estimation* of  $\beta$  (and functions thereof)
- *Tests* regarding elements/functions of  $\beta$
- *Here*: consider *estimation*; parallel development for testing

**Focus of inference:** Comparisons based on  $\beta$  are *unconditional*

- Treatment effect *averaged across the population*
- E.g.,  $\beta_2 = E(Y|Z = 2) - E(Y|Z = 1)$  in Example 1
- *Unconditional inference* is the usual focus of the *primary analysis* in most clinical trials

### 3. Focus of inference

---

**Alternative:** Comparison *conditional* on subset of the population with  $X = x$ ; e.g., in Example 1

$$\beta_x = E(Y|X = x, Z = 2) - E(Y|X = x, Z = 1)$$

- *ANCOVA model*  $E(Y|X, Z) = \alpha_0 + \alpha_1^T X + \phi I(Z = 2)$
- Contrast with  $E(Y|Z) = \beta_1 + \beta_2 I(Z = 2)$
- $\phi = \beta_x = \beta_2$  if ANCOVA model *correct*
- OLS estimator for  $\phi$  is consistent for  $\beta_2$  *regardless*
- ANCOVA is used for *covariate adjustment*  
(*direct regression modeling*)
- *Conditional* vs. *unconditional* not a *big deal*

### 3. Focus of inference

---

**Conditional vs. unconditional is a big deal:** E.g., *binary outcome*

- *Unconditional model*

$$\text{logit}\{E(Y|Z)\} = \beta_1 + \beta_2 I(Z = 2)$$

- *Conditional (on X) model*

$$\text{logit}\{E(Y|X, Z)\} = \alpha_0 + \alpha_1^T X + \phi I(Z = 2)$$

**Similarly:** *Time to event* outcome

- *Unconditional model*

$$\lambda(t | Z) = \lambda_0(t) \exp\{\beta I(Z = 2)\}$$

- *Conditional (on X) model*

$$\lambda(t | X, Z) = \lambda_0(t) \exp\{\alpha^T X + \phi I(Z = 2)\}$$

**Both:**  $\phi \neq \beta_2$  or  $\beta \Rightarrow$  *different focus*

# 3. Focus of inference

---

**Debate:** Which is more *clinically relevant*?

- Is a *scientific* and *philosophical* issue, not a *statistical* issue
- It is *not* our objective to resolve or enter into this debate!
- If interest focuses on *unconditional inference* . . .
- . . . we focus on making this inference (*inference on  $\beta$* ) as *efficient* as possible
- *Moderate to large  $n$  (asymptotic theory)*

## 4. Semiparametric model

---

**In general:**  $\beta$  ( $p \times 1$ ) parameter relevant to making (*unconditional*) treatment comparisons in an assumed model for the *conditional distribution of  $Y$  given  $Z$*

- Possibly *additional* parameter  $\gamma$
- *Conditional density*  $p_{Y|Z}(y|z; \theta, \eta)$ ,  $\theta = (\beta^T, \gamma^T)^T$  ( $r \times 1$ )
- $\eta$  is an *additional nuisance parameter* needed to *describe fully* the class of densities being assumed
- $\eta$  *null* in *fully parametric models* (e.g., logistic, linear mixed models)
- $\eta$  *infinite-dimensional* in *nonparametric* or *semiparametric models* (e.g., treatment means, marginal longitudinal, proportional hazards models)

## 4. Semiparametric model

---

**Semiparametric model for all of  $(Y, X, Z)$ :** Class of joint densities

$$p_{Y,X,Z}(y, x, z; \theta, \eta, \psi, \pi) = p_{Y,X|Z}(y, x | z; \theta, \eta, \psi)p_Z(z; \pi),$$

$\theta = (\beta^T, \gamma^T)^T$  ( $r \times 1$ ), such that

- $\pi$  is *known*, so  $p_Z(z; \pi)$  is *completely specified*
- $Z \perp\!\!\!\perp X$  *by randomization*
- (i)  $\int p_{Y,X|Z}(y, x | z; \theta, \eta, \psi) dx = p_{Y|Z}(y|z; \theta, \eta)$
- (ii)  $\int p_{Y,X|Z}(y, x | z; \theta, \eta, \psi) dy = p_X(x)$
- $\psi$  needed to include all densities satisfying (i) and (ii)

**Goal:** *Consistent and asymptotically normal estimators* for  $\beta$  and *test procedures* based on  $(Y_i, X_i, Z_i)$ ,  $i = 1, \dots, n$ , iid making no assumptions beyond this *semiparametric model*

- Inclusion of  $X \Rightarrow$  “*covariate adjustment*”

# 5. Estimating functions using auxiliary covariates

---

**Approach:** Derive *estimators* by characterizing the class of all *estimating functions* for  $\theta$  (and hence  $\beta$ ) leading to estimators for  $\theta$  that are *consistent and asymptotically normal* under the semiparametric model

- *Estimating function*: Function of a single observation and parameters that can be used to construct *estimating equations* leading to *estimators* for the parameters
- $\Rightarrow$  We seek *unbiased estimating functions for  $\theta$*  depending on  $(Y, X, Z)$  (lead to *consistent and asymptotically normal estimators*), i.e.,

$$E\{m^*(Y, X, Z; \theta)\} = 0;$$

estimator is solution to

$$\sum_{i=1}^n m^*(Y_i, X_i, Z_i; \theta) = 0 \quad (r \times 1)$$

# 5. Estimating functions using auxiliary covariates

---

**Unbiased estimating functions depending on  $(Y, Z)$  only in models  $p_{Y|Z}(y|z; \theta, \eta)$  like those in our examples:**

$$m(Y, Z; \theta) \ (r \times 1) \Rightarrow \text{Solve } \sum_{i=1}^n m(Y_i, Z_i; \theta) = 0$$

- *Example 1*:  $E(Y | Z) = \beta_1 + \beta_2 I(Z = 2)$

$$m(Y, Z; \theta) = \{1, I(Z = 2)\}^T \{Y - \beta_1 - \beta_2 I(Z = 2)\}$$

yields *OLS estimator* for  $\beta \Rightarrow \hat{\beta}_{2,OLS} = \text{difference in sample means}$

- *Example 3*:  $\text{logit}\{E(Y | Z)\} = \beta_1 + \beta_2 I(Z = 2)$

$$m(Y, Z, ; \theta) = \{1, I(Z = 2)\}^T [Y - \text{expit}\{\beta_1 + \beta_2 I(Z = 2)\}]$$

yields *logistic regression MLE (log-odds ratio of sample proportions)*

# 5. Estimating functions using auxiliary covariates

---

**Main result:** For a given *semiparametric model*, members of the *class of all unbiased estimating functions for  $\theta$*  using *all of  $(Y, X, Z)$*  may be written

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

- $m(Y, Z; \theta)$  is a *fixed* ( $r \times 1$ ) unbiased estimating function for  $\theta$  in the specified model  $p_{Y|Z}(y|z; \theta, \eta)$
- $a_g(X)$ ,  $g = 1, \dots, k$ , are arbitrary  $r$ -dimensional functions of  $X$
- $a_g(X) \equiv 0 \ \forall g \Rightarrow$  “*unadjusted estimator*”  $\hat{\theta} = (\hat{\beta}^T, \hat{\gamma}^T)^T$
- “*Augmentation term*” effects the “*adjustment*”

# 5. Estimating functions using auxiliary covariates

---

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

- By  $Z \perp\!\!\!\perp X$ , *augmentation term* has *mean zero*  $\Rightarrow$  *unbiased*

**Adjusted estimator for  $\theta$ :** Solve

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

- *Judicious choice of  $a_g(X)$*   $\Rightarrow$  *improved efficiency* over the “*unadjusted*” estimator  $\hat{\theta}$

# 5. Estimating functions using auxiliary covariates

---

**Optimal estimating function in the class:** Elements of the estimator for  $\theta$  have *smallest asymptotic variance*

- Take  $a_g(X) = E\{m(Y, Z; \theta) \mid X, Z = g\}$ ,  $g = 1, \dots, k$
- *Optimal estimating function*

$$\sum_{i=1}^n \left[ m(Y_i, Z_i; \theta) - \sum_{g=1}^k \{I(Z_i = g) - \pi_g\} E\{m(Y, Z; \theta) \mid X_i, Z = g\} \right] = 0$$

- Yields *optimal “adjusted”* estimator for  $\beta$
- $E\{m(Y, Z; \theta) \mid X, Z = g\}$  are *unknown functions of  $X$*   $\Rightarrow$  *model them...*

# 6. Implementation

## Adaptive algorithm:

- (1) Solve  $\sum_{i=1}^n m(Y_i, Z_i; \theta) = 0 \Rightarrow \hat{\theta}$  and form  $m(Y_i, g; \hat{\theta})$ ,  $g = 1, \dots, k$ , for each subject  $i$  ( $r \times 1$ )
- (2) For *each group*  $g = 1, \dots, k$  *separately*, using the *r-variate “data”*  $m(Y_i, g; \hat{\theta})$  for  $i \in g$ , develop a *regression model*

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

$$q_{gu}(X, \zeta_{gu}) = \{1, c_{gu}^T(X)\}^T \zeta_{gu}, \quad u = 1, \dots, r,$$

and obtain  $\hat{\zeta}_g = (\hat{\zeta}_{g1}^T, \dots, \hat{\zeta}_{gr}^T)^T$  by *OLS separately* for  $u = 1, \dots, r$

- (3) For each  $i = 1 \dots, n$ , form *predicted values*  $q_g(X_i, \hat{\zeta}_g)$  for each  $g = 1, \dots, k$  and solve in  $\theta$  with  $\hat{\pi}_g = n^{-1} \sum_{i=1}^n I(Z_i = g)$

$$\sum_{i=1}^n \left[ m(Y_i, Z_i; \theta) - \sum_{g=1}^k \{I(Z_i = g) - \hat{\pi}_g\} q_g(X_i, \hat{\zeta}_g) \right] = 0 \Rightarrow \text{“adjusted” } \tilde{\theta}$$

## 6. Implementation

---

**Simplification:** When  $m(Y, Z; \theta) = A(Z, \theta)\{Y - f(Z; \theta)\}$  ( $r \times 1$ )

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

(2) Using data for  $i \in g$ , obtain  $\hat{\zeta}_g$ ,  $g = 1, \dots, k$ , by *OLS fit of*

$$E(Y \mid X, Z = g) = q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}\zeta_g$$

(3) Form for each  $i = 1, \dots, n$  *predicted values* for

$$E\{m(Y, Z; \theta) \mid X, Z = g\} \text{ as } q_g(X_i, \hat{\zeta}_g, \theta) = A(g, \theta)\{q_g^*(X_i, \hat{\zeta}_g) - f(g, \theta)\},$$

and *solve in*  $\theta$

$$\sum_{i=1}^n \left[ m(Y_i, Z_i; \theta) - \sum_{g=1}^k \{I(Z_i = g) - \hat{\pi}_g\} q_g(X_i, \hat{\zeta}_g, \theta) \right] = 0 \Rightarrow \text{“adjusted” } \tilde{\theta}$$

# 6. Implementation

---

**Special case:** *Example 1* ( $k = 2$  arm trial, *continuous response*  $Y$ )

$$E(Y | Z) = \beta_1 + \beta_2 I(Z = 2), \quad \beta_2 = E(Y | Z = 2) - E(Y | Z = 1)$$

- $\bar{Y}_g = n_g^{-1} \sum_{i=1}^n I(Z_i = g) Y_i$ ,  $n_g = \sum_{i=1}^n I(Z_i = g)$ ,  $g = 1, 2$
- All estimators for  $\beta_2$  are *asymptotically equivalent* to

$$\bar{Y}_2 - \bar{Y}_1 - \sum_{i=1}^n \{I(Z_i = 2) - \hat{\pi}_1\} \{n_1^{-1} a_1(X_i) + n_2^{-1} a_2(X_i)\}$$

- *In this class:* ANCOVA, ANCOVA with *treatment-covariate interaction*, Koch et al. (1998)'s "*nonparametric*" estimator,...
- *Optimal estimator* takes  $a_g(X) = E(Y|X, Z = g)$ ,  $g = 1, 2$
- For  $g = 1, 2$ , substitute *OLS fits* of

$$E(Y|X, Z = g) = q_g^*(X, \zeta_g) = q_g(X, \zeta_g) = (\{1, c_g^T(X)\}) \zeta_g,$$

See Tsiatis et al. (2008)

## 6. Implementation

---

**Standard errors:** For  $\tilde{\theta}$  and hence  $\tilde{\beta}$

- $\tilde{\theta}$  is an *M-estimator*
- $\Rightarrow$  *Sandwich method* for asymptotic covariance matrix for  $\tilde{\beta}$

# 6. Implementation

---

**Properties:** From *semiparametric theory*

- With the (*linear*) *regression models*  $q_g$  as above,  $\tilde{\theta}$  is *guaranteed relatively more efficient* than  $\hat{\theta}$ , even if  $q_g$  *incorrect*
- $\tilde{\theta}$  has *same asymptotic properties* as if limits in probability of  $\hat{\zeta}_g$  were *known* and the true  $\pi_g$  substituted, even if  $q_g$  *incorrect*
- *Nonlinear models*  $q_g$  and *not OLS*  $\Rightarrow$  *approximately true*
- $\tilde{\theta}$  is *consistent and asymptotically normal* regardless of  $q_g$
- If the  $q_g$  models are *exactly correct*  $\Rightarrow$   $\tilde{\theta}$  is *asymptotically equivalent* to the *optimal estimator* if we *knew*  $E\{m(Y, Z; \theta) \mid X, Z = g\}$
- *In general*: the closer the  $q_g$  are to the *true functions of  $X$* , the *closer to optimal*

# 6. Implementation

---

## By-product:

- The “*adjustment*” for  $X$  is determined *separately by treatment group*...
- ... *and* regression modeling is carried out *independently of  $\tilde{\beta}$*
- $\Rightarrow$  Can develop models *without concerns* over *subjectivity*

## “Principled” strategy:

- *Regression modeling* for each  $g = 1, \dots, k$  based on data for  $i \in g$  *only* may be carried out by *separate analysts for each  $g$* ...
- ... *different from* those who calculate  $\tilde{\theta}$  (and hence  $\tilde{\beta}$ )
- $\Rightarrow$  A sponsor could retain *different CROs* to build the models for each treatment

# 6. Implementation

---

**Hypothesis tests:** *More powerful tests* of “*no treatment effects*” via “*augmentation*”

- For  $H_0$  involving  $s$  degrees of freedom, “*unadjusted*” test statistics  $T_n$  *asymptotically equivalent to*

$$\left\{ n^{-1/2} \sum_{i=1}^n \ell(Y_i, Z_i) \right\}^T \Sigma^{-1} \left\{ n^{-1/2} \sum_{i=1}^n \ell(Y_i, Z_i) \right\},$$

$$\ell(Y, Z) \ (s \times 2), \ \Sigma = E_{H_0} \{ \ell(Y, Z) \ell(Y, Z)^T \}$$

- Tests regarding  $\beta$ , other tests (e.g., Kruskal-Wallis)
- *Idea*: Replace  $\ell(Y, Z)$  by suitably *augmented* version  $\ell^*(Y, X, Z)$  and  $\Sigma$  by  $\Sigma^* = E_{H_0} \{ \ell^*(Y, X, Z) \ell^*(Y, X, Z)^T \}$
- *Analogous* implementation

# 7. Simulations

---

**Binary response,  $k = 2$ :** 5000 Monte Carlo data sets,  $n = 600$

$$\text{logit}\{E(Y|Z)\} = \beta_1 + \beta_2 I(Z = 2), \quad \theta = (\beta_1, \beta_2)^T$$

- $P(Z = 1) = P(Z = 2) = 0.5$
- $X = (X_1, \dots, X_8)^T$ ,  $X_1, X_3, X_8 \sim \mathcal{N}(0, 1)$ ,  $P(X_4 = 1) = 0.3$ ,  
 $P(X_6 = 1) = 0.5$ ,

$$X_2 = 0.2X_1 + 0.98U_1, \quad X_5 = 0.1X_1 + 0.2X_3 + 0.97U_2$$

$$X_7 = 0.1X_3 + 0.99U_3, \quad U_\ell \sim \mathcal{N}(0, 1), \quad \ell = 1, 2, 3$$

- $(X_1, \dots, X_4)$  “*important*,”  $(X_5, \dots, X_8)$  “*unimportant*”
- Generate  $Y$  as Bernoulli with

$$\text{logit}\{P(Y = 1|Z = g, X)\} = \alpha_{0g} + \alpha_g^T X, \quad g = 1, 2$$

$\alpha_g$  chosen to yield *mild*, *moderate*, or *strong* association between  $Y$  and  $X$  for each  $g$  ( $R^2 = 0.16, 0.32, 0.41$ )

# 7. Simulations

---

**Several ways:** Models  $q_g^*(X, \zeta_g)$  for  $E(Y | X, Z = g)$  developed as

Aug. 1  $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$ ,  $c_g(X) =$  “*true*,” fit by OLS

Aug. 2  $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$ ,  $c_g(X) = X$ , fit by OLS

Aug. 3  $\text{logit}\{q_g^*(X, \zeta_g)\} = \{1, c_g^T(X)\}^T \zeta_g$ ,  $c_g(X) =$  “*true*,” fit by IRWLS

Aug. 4  $\text{logit}\{q_g^*(X, \zeta_g)\} = \{1, c_g^T(X)\}^T \zeta_g$ ,  $c_g(X) = X$ , fit by IRWLS

Aug. 5  $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$ ,  $c_g(X)$  fit by OLS with *forward selection*

Aug. 6  $\text{logit}\{q_g^*(X, \zeta_g)\} = \{1, c_g^T(X)\}^T \zeta_g$ ,  $c_g(X)$  by IRWLS with *forward selection*

- “*true*” =  $c_g(X)$  contains only “*important*” covariates  $(X_1, \dots, X_4)$

**Competitors:** Fit  $\text{logit}\{E(Y|X, Z)\} = \alpha_0 + \alpha_1^T X + \phi I(Z = 2)$  by IRWLS

Usual 1 Include  $(X_1, \dots, X_4)$  only

Usual 2 Subset of  $X$  to include by *forward selection*

# 7. Simulations

Method	True	MC Bias	MC SD	Ave. SE	Cov. Prob	Rel. Eff.
Mild Association						
Unadjusted	-0.494	0.002	0.168	0.166	0.948	1.00
Aug. 1	-0.494	0.000	0.156	0.153	0.948	1.16
Usual 1	-0.494	-0.091	0.185	0.182	0.922	0.66
Moderate Association						
Unadjusted	-0.490	0.001	0.165	0.165	0.948	1.00
Aug. 1	-0.490	-0.002	0.140	0.139	0.950	1.39
Usual 1	-0.490	-0.218	0.203	0.201	0.813	0.31
Strong Association						
Unadjusted	-0.460	0.004	0.164	0.165	0.954	1.00
Aug. 1	-0.460	0.000	0.132	0.131	0.952	1.55
Usual 1	-0.460	-0.321	0.223	0.220	0.695	0.18

*Aug 1–6 and Usual 1,2 virtually identical*

# 7. Simulations

---

## Additional simulations qualitatively similar:

- *Continuous response*, difference of  $k = 2$  means based on ACTG 175 (Tsiatis et al., 2008)
- *Continuous longitudinal response*, difference of  $k = 2$  slopes, linear mixed model (Zhang et al., 2008)
- *Censored time to event*:  $k = 2$ , log-hazard ratio (Lu and Tsiatis, 2008)

## 8. Applications

---

**PURSUIT trial:** “*Platelet Glycoprotein IIb/IIIa in Unstable Angina: Receptor Suppression Using Integrilin Therapy*”

- International multi-center clinical trial
- Compare anti-coagulant Integrilin+heparin and aspirin to heparin and aspirin (control) in subjects with ACS
- *Outcome:* Binary composite death or MI at 30 days ⇒ *log-odds ratio for Integrilin relative to control*
- Data from 5710 subjects
- 35 *auxiliary baseline covariates*, including age, height, weight, gender, race, geographic region, smoking status, diastolic and systolic blood pressure, creatine kinase and creatine kinase-MB ratios, disease history, treatment history, . . .

# 8. Applications

---

## Results: Log-odds ratio

- *Unadjusted estimate and SE*:  $\hat{\beta} = -0.174 (0.073)$
- *Adjusted procedure*:  $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$ ,  $g = 1, 2$ ,  $c_g(X)$  includes *main effects* of all 35 covariates, fit by OLS
- *Adjusted estimate and SE*:  $\tilde{\beta}_2 = -0.163 (0.071)$
- SE unadjusted/SE adjusted = 1.06

# 8. Applications

---

## AIDS Clinical Trials Group (ACTG) 175:

- $k = 4$  groups: ZDV monotherapy ( $g = 1$ ), ZDV+ddl, ( $g = 2$ ), ZDV+zalcitabine ( $g = 3$ ), and ddl monotherapy ( $g = 4$ ),  $\pi_g = 1/4$ ,  $g = 1, \dots, 4$
- Data from 2139 subjects
- *Outcome*: CD4 count (cells/mm<sup>3</sup>) at 20±5 weeks
- 12 *auxiliary baseline covariates*: CD4 count (cells/mm<sup>3</sup>), CD8 count (cells/mm<sup>3</sup>), age (years), weight (kg), Karnofsky score (scale of 0-100), and indicator variables for hemophilia, homosexual activity, history of intravenous drug use, race (0=white, 1=non-white), gender (0=female), antiretroviral history (0=naive, 1=experienced), and symptomatic status (0=asymptomatic)

## 8. Applications

---

**Model:** *Treatment means*  $\beta = (\beta_1, \dots, \beta_4)^T$

$$E(Y | Z) = \beta_1 I(Z = 1) + \beta_2 I(Z = 2) + \beta_3 I(Z = 3) + \beta_4 I(Z = 4)$$

- $\hat{\beta} = \text{sample averages} = (336.14, 403.17, 372.04, 374.32)^T$ ,  
SEs =  $(5.68, 6.84, 5.90, 6.22)^T$
- *Adjusted procedure*:  $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$ ,  $g = 1, 2$ ,  $c_g(X)$  includes *main effects* of all 12 covariates, fit by OLS
- $\tilde{\beta} = \text{adjusted estimates} = (333.85, 403.83, 370.43, 376.45)^T$ ,  
SEs =  $(4.61, 5.93, 4.89, 5.11)^T$
- SE unadjusted/SE adjusted =  $(1.51, 1.33, 1.46, 1.48)^T$

# 8. Applications

---

**Hypothesis tests:** 3 degrees of freedom

- *Wald test* for  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ :  
Unadjusted      59.40  
Adjusted        109.58
- *Kruskal-Wallis* for  $H_0 : F_1(u) = \dots = F_4(u) = F(u)$ , *adjusted version* using linear, quadratic terms in  $X$   
Unadjusted      49.04  
Adjusted        100.53
- *Overwhelming evidence* in either case, but proposed statistics *considerably larger*

# 9. Discussion

---

- General approach to using *auxiliary baseline covariates* to *improve efficiency* of *estimators* and *tests*
- General measures of *treatment effect*
- Arises naturally via *semiparametric theory*
- Incorporation of covariate information *separated from* evaluation of treatment effects
- We *do not* identify the optimal estimating function *over all possible*, only for a *given fixed*  $m(Y, Z; \theta)$  – *very hard*, gains likely *minimal*
- For *differences of means* and *binary outcomes* the estimating function is optimal
- Effects of *model selection* deserve further study
- Can be extended to handle *missing outcome*

# References

---

- Gilbert, P. B., Sato, M., Sun, X., and Mehrotra, D. V. (2009). Efficient and robust method for comparing the immunogenicity of candidate vaccines in randomized clinical trials. *Vaccine* **27**, 396–401.
- Lu, X. and Tsiatis, A. A. (2008). Improving the efficiency of the logrank test using auxiliary covariates. *Biometrika* **95**, 679–694 .
- Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.
- Tsiatis, A.A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27**, 4658–4677.
- Zhang, M., Tsiatis, A.A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64**, 707–715.

# Appendix: Hypothesis tests

---

**Same idea:** *More powerful tests* of “no treatment effects” via “*augmentation*”

- For  $H_0$  involving  $s$  degrees of freedom, “*unadjusted*” test statistics  $T_n$  *asymptotically equivalent to*

$$\left\{ n^{-1/2} \sum_{i=1}^n \ell(Y_i, Z_i) \right\}^T \Sigma^{-1} \left\{ n^{-1/2} \sum_{i=1}^n \ell(Y_i, Z_i) \right\},$$

$$\ell(Y, Z) \text{ (} s \times 2 \text{)}, \Sigma = E_{H_0} \{ \ell(Y, Z) \ell(Y, Z)^T \}$$

- E.g.,  $H_0 : C\beta = 0$  for  $C$  ( $s \times p$ ), *Wald test*

$$T_n = (C\hat{\beta})^T (n^{-1}\hat{\Sigma})^{-1} C\hat{\beta}, \quad \hat{\beta} \text{ solves } \sum_{i=1}^n m(Y_i, Z_i; \hat{\theta}) = 0$$

$\Rightarrow \ell(Y, Z) = CA m(Y, Z, \theta_0)$  for a matrix  $A$ ,  $\theta_0$  the value under  $H_0$

# Appendix: Hypothesis tests

---

$$\left\{ n^{-1/2} \sum_{i=1}^n \ell(Y_i, Z_i) \right\}^T \Sigma^{-1} \left\{ n^{-1/2} \sum_{i=1}^n \ell(Y_i, Z_i) \right\}$$

- E.g.,  $H_0 : F_1(u) = \dots = F_k(u) = F(u)$ ,

$$F_g(u) = P(Y \leq u | Z = g), \quad F(u) = P(Y \leq u)$$

*Kruskal-Wallis test*  $T_n = 12 \sum_{g=1}^k n_g \{ \bar{R}_g - (n+1)/2 \}^2 / \{ n(n+1) \}$

$$\Rightarrow \ell(Y, Z) = \left[ \{ I(Z=1) - \pi_1 \} \{ 1/2 - F(Y) \}, \dots, \{ I(Z=k) - \pi_k \} \{ 1/2 - F(Y) \} \right]^T$$

# Appendix: Hypothesis tests

---

**In general:** Under *local alternatives*  $H_{1n} \rightarrow H_0$  at rate  $n^{-1/2}$

$$n^{-1/2} \sum_{i=1}^n \ell(Y_i, Z_i) \xrightarrow{\mathcal{L}} \mathcal{N}(\tau, \Sigma)$$

- $T_n$  asymptotically noncentral  $\chi_s^2$  with *noncentrality parameter*

$$\tau^T \Sigma^{-1} \tau$$

- More powerful test  $\Rightarrow$  noncentrality parameter *as large as possible*

# Appendix: Hypothesis tests

---

**“Augmented” test statistic:**

$$T_n^* = \left\{ n^{-1/2} \sum_{i=1}^n \ell^*(Y_i, X_i, Z_i) \right\}^T \Sigma^*{}^{-1} \left\{ n^{-1/2} \sum_{i=1}^n \ell^*(Y_i, X_i, Z_i) \right\}$$

$$\ell^*(Y, X, Z) = \ell(Y, Z) - \sum_{g=1}^k \{I(Z = g) - \pi_g\} a_g(X)$$

- $\Sigma^* = E_{H_0} \{ \ell^*(Y, X, Z) \ell^*(Y, X, Z)^T \}$
- $n^{-1/2} \sum_{i=1}^n \ell^*(Y_i, X_i, Z_i) \xrightarrow{\mathcal{L}} \mathcal{N}(\tau, \Sigma^*)$  (*same*  $\tau$ )
- $T_n^*$  *asymptotically noncentral*  $\chi_s^2$  with *noncentrality parameter*

$$\tau^T \Sigma^*{}^{-1} \tau$$

**Optimal test:** Noncentrality parameter *as large as possible* when

$$a_g(X) = E\{\ell(Y, Z) | Z = g, X\}, \quad g = 1, \dots, k$$

# Appendix: Hypothesis tests

---

**Implementation:** *Adaptive algorithm*

- (1) Obtain  $\hat{\ell}(Y_i, Z_i)$ ,  $i = 1, \dots, n$  (e.g., *substitute  $\hat{\theta}$* )
- (2) For each  $g = 1, \dots, k$  *separately*, using data from  $i \in g$ , develop *regression models*

$$E\{\hat{\ell}(Y, g)|X, Z = g) = q_g(X, \zeta_g) = \{q_{g1}(X, \zeta_{g1}) \dots, q_{gs}(X, \zeta_{gs})\}^T$$

as before; obtain  $\hat{\zeta}_g$  by OLS

- (3) Form *predicted values*  $q_g(X_i, \hat{\zeta}_g)$ , obtain

$$\hat{\ell}^*(Y_i, X_i, Z_i) = \hat{\ell}(Y_i, Z_i) - \sum_{g=1}^k \{I(Z_i = g) - \hat{\pi}_g\} q_g(X_i, \hat{\zeta}_g),$$

form  $\hat{T}_n^*$ , and compare to  $\chi_s^2$  distribution

**Properties:** *Analogous* to estimators

# Appendix: Hypothesis tests

---

**Simulation,  $k = 3$ :**  $H_0 : F_1(u) = F_2(u) = F_3(u) = F(u)$

- 10,000 MC data sets,  $n = 200, 400$
- $P(Z = g) = 1/3, g = 1, 2, 3$
- $(Y, X)^T | Z$  *bivariate normal*

$$E\{(Y, X)^T | Z\} = \{\beta_1 I(Z = 1) + \beta_2 I(Z = 2), 0\}^T$$

$$\text{var}\{(Y, X)^T | Z\} = \text{vech}(1, \rho, 1)$$

- $\rho = 0.25, 0.50, 0.75$ , *mild*, *moderate*, *strong* correlation
- $H_0 : \beta_1 = \beta_2 = 0, H_1 : \beta_1 = 0.25, \beta_2 = 0.4$
- $T_n =$  *Kruskal-Wallis*
- $T_n^*$  with  $q_g(X, \zeta_g)$ :  $q_{gu}(X, \zeta_{gu}) = \{1, c_{gu}^T(X)\}^T \zeta_{ug}, u = 1, 2,$   
 $c_{gu}(X) = (X, X^2)^T$

# Appendix: Hypothesis tests

---

$\rho$	$n$	Null		Alternative	
		$T_n$	$\hat{T}_n^*$	$T_n$	$\hat{T}_n^*$
0.25	200	0.05	0.05	0.51	0.54
	400	0.05	0.05	0.83	0.85
0.50	200	0.05	0.05	0.51	0.64
	400	0.05	0.05	0.83	0.92
0.75	200	0.05	0.05	0.51	0.85
	400	0.05	0.05	0.83	0.99

---