Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates

Anastasios (Butch) Tsiatis
Department of Statistics
North Carolina State University

http://www.stat.ncsu.edu/∼tsiatis

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

Outline

1. Introduction
2. Reasons for Covariate Adjustment
3. Conditional vs Unconditional Inference
4. Semiparametric Theory
5. Implementation
6. Simulations
7. Discussion

Introduction

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

- *Continuous response*: compare on the basis of *treatment means*
- *Binary response*: compare on the basis of *odds ratio*
- *Time to event*: compare on the basis of *treatment-specific hazard ratio*

In addition to outcome and treatment assignment: *Baseline auxiliary covariates*

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

Reasons for Covariate Adjustment

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

"Covariate adjustment:" with auxiliary baseline covariates has been advocated

- to account for chance imbalances in baseline covariates
- to gain efficiency
- *Extensive literature*: Senn (1989), Hauck et al. (1998), Koch et al. (1998), Tangen and Koch (1999), Pocock et al. (2002), ...
- *Extensive concerns*: Potential *bias* due to post hoc (*subjective*) selection of covariates to use, and...
- ...temptation to engage in a "*fishing expedition*" for the *most dramatic* effect
- $\Rightarrow$ *Trialists* and *regulatory authorities* reluctant to endorse

Covariate Adjustment

Standard approach to adjustment: *Direct regression modeling*

- Model outcome as a function of treatment assignment *and* covariates
- $\Rightarrow$ *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

- 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

Notation

- Data: $(Y_i, Z_i, X_i), i = 1, \ldots, n$, (iid) where for patient i
- Y_i response variable (discrete, continuous, longitudinal, censored)
- Z_i denotes treatment assignment (For simplicity we will consider only two treatments, but methods generalize easily to more than two treatments)
- Z_i (1=treatment, 0=control), $P(Z_i = 1) = \pi$
- X_i denotes other baseline covariates measured prior to randomization
- $X \perp\!\!\!\perp Z$

Unconditional Inference

Example 1: *continuous response* Y

$$E(Y \mid Z) = \alpha + \beta Z$$

- Here the parameter of interest is $\beta = E(Y|Z=1) - E(Y|Z=0) =$ *difference in treatment means*

Example 2: *binary response* $(Y = 0, 1)$

$$\text{logit}\{E(Y \mid Z\} = \text{logit}\{P(Y=1|Z)\} = \alpha + \beta Z$$

- Here the parameter of interest is $\beta =$*Log-odds ratio* for treatments 1 and 0

Unconditional Inference

Example 3: *Time to event (censored data)*

- Here the data are represented as $(U_i, \Delta_i, Z_i, X_i), i = 1, \ldots, n$
  - U_i is time to failure or censoring $=\min(T_i, C_i)$
  - Δ_i is failure indicator $=I(T_i \leq C_i)$
  - As before Z_i is treatment indicator and X_i denotes baseline covariates
- *Proportional hazards model*

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta Z),$$

  where $\lambda(t|Z)$ denotes the conditional hazard rate of failing at time t given treatment Z
- The parameter of interest is β=*Log-hazard ratio* for treatments 1 and 0

Conditional versus unconditional inference

Focus of inference: Comparisons based on β are *unconditional*

- Treatment effect *averaged across the population*
- E.g., $\beta = E(Y|Z=1) - E(Y|Z=0)$ in Example 1

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

$$\beta_x = E(Y|X=x, Z=1) - E(Y|X=x, Z=0)$$

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

Conditional versus unconditional inference

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

- *Unconditional model*

$$\text{logit}\{E(Y|Z\} = \alpha + \beta Z$$

- *Conditional* (*on* X) *model*

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

Similarly: *time to event outcome*

- *Unconditional model*

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta Z)$$

- *Conditional* (*on* X) *model*

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

- $\phi \neq \beta \Rightarrow$ *different focus*

Conditional versus unconditional inference

Debate: Which is more *clinically relevant*?

- Most trials: *unconditional primary analysis*
- $\Rightarrow$ We focus on *unconditional* inference

Semiparametric model

In general: β is the parameter relevant to making (*unconditional*) treatment comparisons in an assumed model for the *conditional distribution of* Y *given* Z

- Possibly *additional* parameter α
- *Conditional density* $p_{Y|Z}(y|z;\theta,\eta),\ \theta=(\beta,\alpha)$
- η is an *additional nuisance parameter* needed to *describe fully* the class of densities being assumed
- η *null* in *fully parametric models*
- η *infinite-dimensional* in *nonparametric* or *semiparametric models*

Semiparametric model

- *Fully parametric model* (e.g., logistic model for binary response)

$$\text{logit}\{E(Y|Z)\} = \alpha + \beta Z$$

- *Nonparametric model* (e.g., for continuous response Y)

$$E(Y|Z) = \alpha + \beta Z$$

- *Semiparametric model* (e.g., proportional hazards model for time to event outcome)

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta Z)$$

2. Semiparametric model

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

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

$\theta = (\beta, \alpha)$, such that

- π is *known*, so $p_Z(z; \pi)$ is *completely specified*
- $Z \perp\!\!\!\perp X$ *by randomization*
- $\int p_{Y,X|Z}(y, x \mid z; \theta, \eta, \psi)\, dx = p_{Y|Z}(y|z; \theta, \eta)$
- $\int p_{Y,X|Z}(y, x \mid z; \theta, \eta, \psi)\, dy = p_X(x)$

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

- Inclusion of $X \Rightarrow$ "*covariate adjustment*"

Semiparametric theory

Approach: Derive *estimators* by characterizing the class of all *estimating functions* for θ (and hence β) leading to estimators for θ 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* θ depending on (Y, Z, X) (lead to *consistent and asymptotically normal estimators*);

$$E_\theta\{m(Y, Z, X; \theta)\} = 0.$$

- Corresponding estimator is solution to

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

Estimating functions without auxiliary covariates

Start by considering unbiased estimating functions depending on (Y, Z) only:

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

- *Example 1*: $E(Y \mid Z) = \alpha + \beta Z$

  $$m(Y, Z; \theta) = (1, Z)^T (Y - \alpha - \beta Z)$$

  yields *OLS estimator* for $\beta \Rightarrow \widehat{\beta}_{OLS} =$ *difference in sample means*

- *Example 2*: $\text{logit}\{E(Y \mid Z)\} = \alpha + \beta Z$

  $$m(Y, Z, ; \theta) = (1, Z)^T \{Y - \text{expit}(\alpha + \beta Z)\}$$

  yields *logistic regression MLE, also log-odds ratio of sample proportions*

Estimating functions without auxiliary covariates

For the *Proportional hazards model* of Example 3, the parameter β is estimated by maximizing the partial likelihood or solving the estimating equation

$$\sum_{i=1}^{n} \int \{Z_i - \bar{Z}(u, \beta)\} dN_i(u) = 0,$$

where $N_i(u) = I(U_i \leq u, \Delta_i = 1)$ and

$$\bar{Z}(u, \beta)\} = \frac{\sum Z_i \exp(\beta Z_i) I(U_i \geq u)}{\sum \exp(\beta Z_i) I(U_i \geq u)}$$

Estimating functions using auxiliary covariates

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

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

- $m(Y, Z; \theta)$ is a *fixed* unbiased estimating function for θ without auxiliary covariates
- $a(X)$ is an arbitrary function of X
- $a(X) \equiv 0 \Rightarrow$ "*unadjusted estimator*" $\widehat{\theta} = (\widehat{\beta}, \widehat{\alpha})$
- "*Augmentation term*" effects the "*adjustment*"

Estimating functions using auxiliary covariates

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

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

Adjusted estimator for θ: Solve

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

- *Judicious choice of* $a(X)$ $\Rightarrow$ *improved efficiency* over the "*unadjusted*" estimator $\widehat{\theta}$

Estimating functions using auxiliary covariates

Optimal estimating function in the class: Elements of the estimator have *smallest asymptotic variance*

- Take $a(X) = E\{m(Y, Z; \theta) \,|\, X, Z = 1\} - E\{m(Y, Z; \theta) \,|\, X, Z = 0\}$
- *Optimal estimating equation*

$$\sum_{i=1}^{n} \Big(m(Y_i, Z_i; \theta) - (Z_i - \pi)\big[E\{m(Y, Z; \theta) \,|\, X_i, Z = 1\} - E\{m(Y, Z; \theta) \,|\, X_i, Z = 0\}\big] \Big) = 0$$

- $E\{m(Y, Z; \theta) \,|\, X, Z = g\}, g = 0, 1$ are *unknown functions of* X $\Rightarrow$ *model them*...

Implementation

Approach: *Adaptive algorithm*

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

(2) For *each group* $g = 0, 1$ *separately*, using the "*data*" $m(Y_i, Z_i; \widehat{\theta})$ for $Z_i = g$, develop a *regression model*

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

$$q_g(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g,$$

and obtain $\widehat{\zeta}_g$ by *OLS separately*

(3) For each $i = 1 \ldots, n$, form *predicted values* $q_g(X_i, \widehat{\zeta}_g)$ for each $g = 0, 1$ and solve in θ with $\widehat{\pi} = n^{-1} \sum_{i=1}^{n} Z_i$

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

Implementation

Properties: From *semiparametric theory*

- With the *regression models* q_g as above, $\widetilde{\theta}$ is *guaranteed relatively more efficient* than $\widehat{\theta}$, even if q_g *incorrect*
- $\widetilde{\theta}$ is *consistent and asymptotically normal regardless* of q_g
- If the q_g models are *exactly correct* $\Rightarrow$ $\widetilde{\theta}$ is *asymptotically equivalent* to the *optimal estimator* if we *knew* $E\{m(Y, Z; \theta) \mid X, Z = g\}$

Implementation

By-product:

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

"Principled" strategy:

- *Regression modeling* for each $g = 0, 1$ based on data for $i \in g$ *only* may be carried out by *separate analysts for each* g...
- ...*different from* those who calculate $\widetilde{\theta}$ (and hence $\widetilde{\beta}$)

Implementation

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

- $\widetilde{\theta}$ is an *M-estimator*
- $\Rightarrow$ *Sandwich method* for asymptotic variance for $\widetilde{\beta}$

Implementation

Special case: *Example 1* (*continuous response* Y)

- All estimators for β are *asymptotically equivalent* to

$$\overline{Y}_1 - \overline{Y}_0 - \sum_{i=1}^{n} (Z_i - \widehat{\pi}) \left\{ n_1^{-1} a_1(X_i) + n_0^{-1} a_0(X_i) \right\},$$

  where $\overline{Y}_g$ denotes treatment-specific sample average for treatment $g = (0, 1)$

- *In this class*: ANCOVA, ANCOVA with *treatment-covariate interaction*, Koch et al. (1998)'s "*nonparametric*" estimator,...

- *Optimal estimator* takes

$$a_g(X) = E(Y|X, Z = g), \quad g = 0, 1$$

See Tsiatis et al. (2008)

Simulations

1. Binary response: 5000 Monte Carlo data sets, $n = 200$

$$\text{logit}\{E(Y|Z)\} = \alpha + \beta Z$$

- $P(Z = 1) = P(Z = 0) = 0.5$
- $X = (X_1, \ldots, X_4)^T$, $(X_1, X_2, X_3)^T \sim \mathcal{N}(0, G)$, $P(X_4 = 1) = 0.3$
- Generate Y as Bernoulli with

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

  α_g chosen to yield *mild*, *moderate*, or *strong* association between Y and X for each g ($R^2 = 0.12, 0.25, 0.34$)

- *Unadjusted estimate* via logistic regression MLE
- *Adjusted estimates* via "*direct approach*" with different choices for $E(Y|X, Z = g) = q_g^*(X, \zeta_g)$

Simulations

"Augmentations:"

- Aug 1: $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$, $c_g(X) = (X_1, X_2, X_3, X_4)^T$, fit by OLS
- Aug 2: $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$, $c_g(X) =$ "*true covariates only*," fit by OLS
- Aug 3, 4: Like Aug 1, 2 but logit$\{q_g^*(X, \zeta_g)\} = \{1, c_g^T(X)\}^T \zeta_g$, fit by MLE

Simulations

Method	True	MC Bias	MC SD	Ave. SE	Cov. Prob	Rel. Eff.
			Mild Correlation			
Unadjusted	-0.218	0.008	0.171	0.170	0.949	1.00
Aug. 1	-0.218	0.006	0.165	0.163	0.947	1.07
			Moderate Correlation			
Unadjusted	-0.150	0.015	0.167	0.168	0.952	1.00
Aug. 1	-0.150	0.013	0.158	0.158	0.945	1.11
			Strong Correlation			
Unadjusted	0.078	-0.001	0.166	0.165	0.950	1.00
Aug. 1	0.078	-0.001	0.154	0.153	0.947	1.16

Aug 2, 3, 4 virtually identical

Simulations

Censored survival data: *Proportional hazards model*

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta Z)$$

In order to generate data where

- the distribution of T given Z follows a proportional hazards model
- T and X are correlated
- X and Z are independent

1. We generate bivariate data (V, X) from a bivariate normal density with mean zero, variance 1 , and correlation ρ
2. Independently generate treatment indicator Z as a Bernoulli(π)
3. Let $T = -\exp(-\beta Z)\log\{1 - \Phi(V)\}$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal
4. Censoring was generated as an independent exponential distribution $C \sim Exp(c)$.

Simulations

- Treatment was assigned with $\pi = .5$
- the correlation of V and X was $\rho = .7$ which resulted in roughly a correlation of 0.6 between T and X
- We took $\beta = 0$ (null hypothesis) and $\beta = .25$
- The value c for the exponential distribution of the censoring variable that would result in roughly 25% of the data being censored
- Sample sizes of 250 and 600 were considered

Simulations

"Estimators considered:"

- $\widehat{\beta}_{\mathrm{PH}}$: Unadjusted estimator using MPLE from unconditional model

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta Z)$$

- $\widehat{\beta}_{\mathrm{AUG}}$: Augmentation term used $q_g(X, \zeta_g) = \{1, X, X^2\}^T \zeta_g$, fit by OLS

- $\widehat{\beta}_{\mathrm{REG}}$: We also considered the estimator $\widehat{\phi}$ obtained by considering the Cox regression model

$$\lambda(t|X, Z) = \lambda_0(t)\exp(\alpha_1 X + \alpha_2 X^2 + \phi Z)$$

Note: *This is not the true conditional model*

Simulations $\beta = 0$

	n	$\widehat{\beta}_{\mathrm{PH}}$	$\widehat{\beta}_{\mathrm{AUG}}$	$\hat{\beta}_{\mathrm{REG}}$
Bias	250	0.002	-0.004	-0.003
	600	0.001	-0.002	-0.001
SE	250	0.148	0.117 (1.60)	0.150 (NA)
	600	0.095	0.075 (1.59)	0.095 (NA)
MCSE	250	0.146	0.120 (1.48)	0.170 (0.74)
	600	0.095	0.076 (1.56)	0.107 (0.79)

Simulations $\beta = .25$

	n	$\widehat{\beta}_{\mathrm{PH}}$	$\widehat{\beta}_{\mathrm{AUG}}$	$\hat{\beta}_{\mathrm{REG}}$
Bias	250	0.004	−0.002	0.092
	600	−0.008	−0.008	0.091
SE	250	0.149	0.118 (1.60)	0.152 (NA)
	600	0.095	0.076 (1.58)	0.097 (NA)
MCSE	250	0.147	0.121 (1.47)	0.171 (0.74)
	600	0.096	0.077 (1.55)	0.107 (0.80)

Example

- Considered ACTG 175
- A randomized study of 2139 patients with HIV disease to four antiretroviral regimes
- treatment 0 (Zidovudine, 532 patients) treatment 1 (Zidovudine and didanosine, 522 patients), treatment 2 (Zidovudine and zalcitabine, 524 patients) and treatment 3 (Didanosine, 561 patients)
- The primary endpoint was a combined endpoint corresponding to the first time that a patient had a ≥ 50 percent decline in their CD4 cell count, an event indicating progression to the acquired immunodeficiency syndrome (AIDS), or death.
- Roughly 76% of the data were censored, almost all administrative censoring.

Example

- A comparison was made between treatment 0 (control) versus treatments 1,2, and 3 respectively
- We also considered several prognostic baseline auxiliary covariates including CD4, CD8, age (years), weight (kg), history of IV drug use (0=no, 1=yes), Karnofsky score (on a scale of 0-100), Zidovudine in the 30 days prior to 175 (0=no, 1=yes), number of days pre-175 antiretroviral therapy and symptomatic indicator (0=asymp, 1=symp)

Example

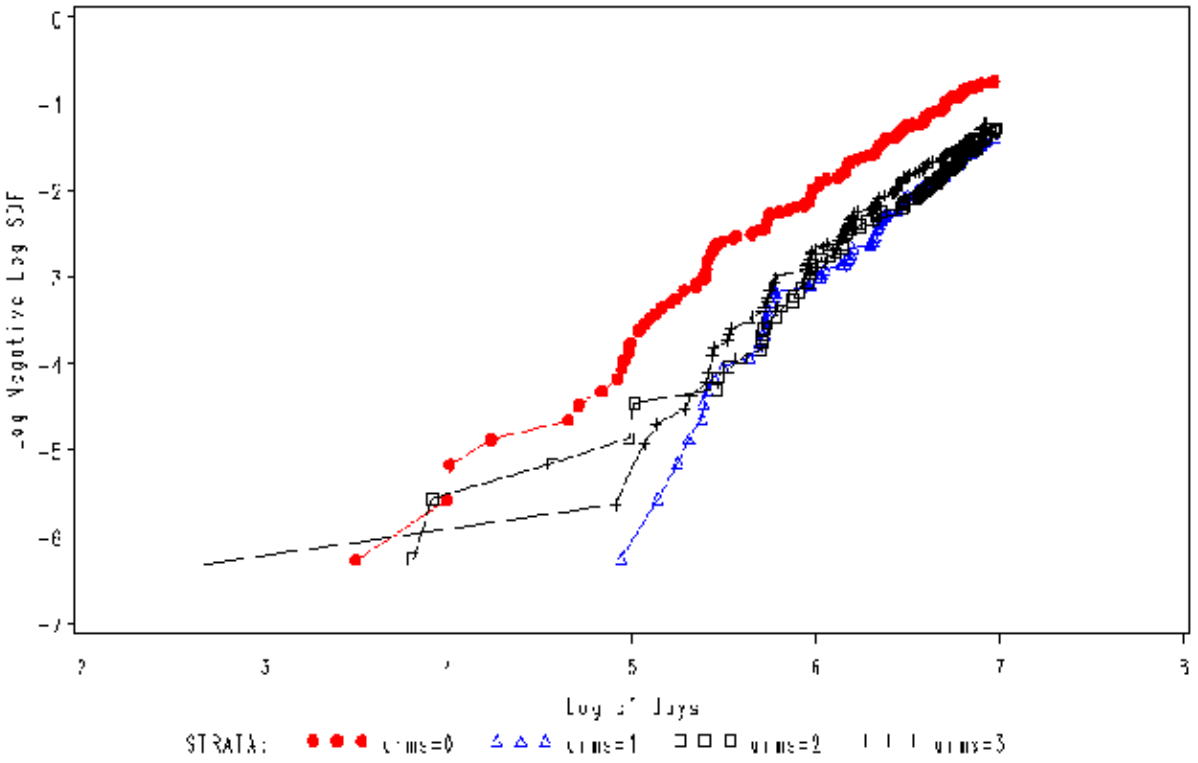

Figure 1: Log negative log survival function of time to death for each treatment

Example

Table 1: Estimates of $\hat{\beta}_{\mathrm{PH}}$ and $\hat{\beta}_{\mathrm{AUG}}$ on the ACTG 175 data (*RE is the relative efficiencies with respect to* $\hat{\beta}_{\mathrm{PH}}$.)

		Estimates	Standard Errors	RE
Treatment 0 and 1	$\hat{\beta}_{\mathrm{PH}}$	-0.703	0.124	1.00
	$\hat{\beta}_{\mathrm{AUG}}$	-0.723	0.110	1.25
Treatment 0 and 2	$\hat{\beta}_{\mathrm{PH}}$	-0.640	0.121	1.00
	$\hat{\beta}_{\mathrm{AUG}}$	-0.555	0.104	1.36
Treatment 0 and 3	$\hat{\beta}_{\mathrm{PH}}$	-0.528	0.116	1.00
	$\hat{\beta}_{\mathrm{AUG}}$	-0.627	0.105	1.21

Discussion

- General approach to using *baseline auxiliary covariates* to *improve efficiency* of *estimators* and theory can also be applied to *tests*
- General measures of *treatment effect*
- Arises naturally via *semiparametric theory*
- Even when regression adjustment leads to improved estimators of unconditional treatment effect (i.e., linear models) there is a tension between gains in efficiency and compromised analysis
- Incorporation of covariate information *separated from* evaluation of treatment effects
- Impact of model selection
- Can be extended to *k-arm trials* and *missing data*

References

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

Leon, S. Tsiatis, A.A., and Davidian, M. (2003). Semiparametric efficiency estimation of treatment effect in a pretest-posttest study. *Biometrics* **59,** 1046–1055.

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.

Lu, X., and Tsiatis, A.A. (2008). Improving efficiency of the log-rank test using auxiliary covariates. In Press, *Biometrika*.