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 - Covariate adjustment
2. Notation
3. Focus of inference
4. Semiparametric model
5. Estimating functions using auxiliary covariates
6. Implementation
7. Simulations
8. Applications
9. Discussion
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/other-omic characteristics
- Prior treatment and medical history
- Baseline measure(s) of the outcome
1. Introduction – Covariate adjustment

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

**However:** Auxiliary baseline covariates

- May be *associated with outcome*
- May exhibit *chance imbalances*

**Covariate adjustment:** Incorporate *auxiliary baseline covariate* information in inference on treatment comparisons to

- Account for *chance imbalance*
- *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), ...
1. Introduction – Covariate adjustment

Considerable controversy:

• Potential *bias* due to post hoc (*subjective*) selection of covariates to use, and...

• ...temptation for a “*fishing expedition*” for *most dramatic* effect

• ⇒ *Triallists* and *regulatory authorities* reluctant to endorse

• Adjusted analyses must be *prespecified*

Standard approach to adjustment: *Direct regression modeling*

• Model outcome as a function of treatment assignment *and* covariates, e.g., via an *ANCOVA* model

• ⇒ *Inextricable link* between parameters involved in treatment comparisons and the “*adjustment*”
1. Introduction – Covariate adjustment

Our objective: General methodology for using auxiliary covariates that leads to more efficient estimators and tests of treatment effect

- Arises from applying theory of semiparametrics (e.g., Tsiatis, 2006)

- $k \geq 2$ arms, general outcome variables, general measures of treatment effect (e.g., difference of means, odds ratio, hazard ratio, difference of slopes, etc)

- Separates parameters involved in treatment comparisons from the “adjustment”...

- ...and hence leads to a principled approach to implementation that can obviate the usual concerns

For simplicity today: Restrict to $k = 2$ arms
2. Notation

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

- $Y =$ *outcome* (continuous, discrete, longitudinal, censored time to event, etc)
- $X =$ vector of *auxiliary baseline covariates*
- $Z =$ *indicator* of *treatment assignment*, e.g., $Z = 1$ experimental treatment, $Z = 0$ control
- $P(Z = 1) = \pi$, *known* randomization probability

**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:** *Continuous response Y*

\[ E(Y | Z) = \gamma + \beta Z \]

- \( \beta = E(Y | Z = 1) - E(Y | Z = 0) = \text{difference in treatment means} \)

**Example 2:** *Binary response* \((Y = 0, 1)\)

\[ \text{logit}\{E(Y | Z)\} = \text{logit}\{P(Y = 1|Z)\} = \gamma + \beta Z \]

- \( \beta = \text{Log-odds ratio} \) for treatment 1 relative to treatment 0
3. Focus of inference

Example 3: *Longitudinal continuous response*

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

\[ Y_{ij} = \gamma_1 + (\gamma_2 + \beta Z)t_{ij} + b_{0i} + b_{1i}t_{ij} + e_{ij}, \ j = 1, \ldots, m_i \]

\[ \gamma = (\gamma_1, \gamma_2, \sigma_e^2, D) \]

Example 4: *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**

\[ \lambda(t \mid Z) = \lambda_0(t) \exp(\beta Z) \]

- \(\lambda(t \mid Z)\) is the conditional hazard rate of failing at time \(t\) given \(Z\)

- \(\beta\) is the *log-hazard ratio* for treatment 1 relative to treatment 0
3. Focus of inference

Objectives:

- *Estimation* of \( \beta \)
- *Tests* regarding \( \beta \) (and other effect measures)
- *Here*: Consider *estimation*; parallel development for *testing*

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

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

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

- **ANCOVA model**
  $$E(Y|X, Z) = \alpha_0 + \alpha_1 X + \phi Z$$

- Contrast with
  $$E(Y|Z) = \gamma + \beta Z$$

- $\phi = \beta_x = \beta$ if the ANCOVA model is *correct*

- OLS estimator for $\phi$ is consistent for $\beta$ *regardless*

- ANCOVA model 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**
  \[
  \text{logit}\left\{ E(Y|Z) \right\} = \gamma + \beta Z
  \]

- **Conditional (on } \mathbf{X} \text{) model**
  \[
  \text{logit}\left\{ E(Y|X,Z) \right\} = \alpha_0 + \alpha^T X + \phi Z
  \]

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

- **Unconditional model**
  \[
  \lambda(t | Z) = \lambda_0(t) \exp(\beta Z)
  \]

- **Conditional (on } \mathbf{X} \text{) model**
  \[
  \lambda(t | X,Z) = \lambda_0(t) \exp(\alpha^T X + \phi Z)
  \]

**Both:** \( \phi \neq \beta \Rightarrow \text{different focus} \)
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

Model for the data \((Y_i, Z_i)\) only: Class of all probability densities

\[
p_{Y,Z}(y, z; \theta, \eta, \pi) = p_{Y \mid Z}(y \mid z; \theta, \eta) p_Z(z; \pi), \quad \theta = (\beta, \gamma)
\]

- \(\pi\) is known, so \(p_Z(z; \pi)\) is completely specified
- \(p_{Y \mid Z}(y \mid z; \theta, \eta)\) is a density consistent with the situation of interest
- E.g., a fully parametric model (e.g., logistic, linear mixed model)
- E.g., a nonparametric model (treatment means) or semiparametric model (proportional hazards)
4. Semiparametric model

Model for all data \((Y_i, X_i, Z_i)\): Class of all probability densities

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

- \(\pi\) is known, so \(p_Z(z; \pi)\) is completely specified
- \(Z \perp X\) by randomization
- \(p_{Y,X|Z}(y, x | z; \theta, \eta, \psi)\) is consistent with \(p_{Y|Z}(y | z; \theta, \eta)\)

Goal: Consistent and asymptotically normal estimators for \(\beta\) under this semiparametric model for \((Y, X, Z)\)

- Inclusion of \(X\) \(\Rightarrow\) covariate adjustment
- Find the most precise such estimator

Approach: Use semiparametric theory to find all unbiased estimating functions for \(\theta\) (and hence \(\beta\)) under the semiparametric model
5. Estimating functions using auxiliary covariates

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

\[
m(Y, Z; \theta) \Rightarrow \text{solve the estimating equation } \sum_{i=1}^{n} m(Y_i, Z_i; \theta) = 0
\]

- **Example 1**: \(E(Y | Z) = \gamma + \beta Z\)

\[
m(Y, Z; \theta) = (1, Z)^T (Y - \gamma - \beta Z)
\]

**OLS estimator** for \(\beta \Rightarrow \hat{\beta}_{OLS} = \text{difference in sample means}

- **Example 2**: \(\logit\{E(Y | Z)\} = \gamma + \beta Z\)

\[
m(Y, Z, ; \theta) = (1, Z)^T \{Y - \expit(\gamma + \beta Z)\}
\]

**logistic regression MLE** (log-odds ratio of sample proportions)
5. Estimating functions using auxiliary covariates

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

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

- $m(Y, Z; \theta)$ is a fixed unbiased estimating function for $\theta$ using $(Y, Z)$ only in the specified model $p_{Y|Z}(y|z; \theta, \eta)$

- $a(X)$ is an arbitrary function of $X$

- $a(X) \equiv 0 \Rightarrow \text{“unadjusted estimator” } \hat{\theta} = (\hat{\beta}, \hat{\gamma})$

- “Augmentation term” effects the “adjustment”

Adjusted estimator for $\theta$: Solve the estimating equation

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

- Judicious choice of $a(X) \Rightarrow$ improved efficiency over the “unadjusted” estimator $\hat{\theta}$
5. Estimating functions using auxiliary covariates

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

**Optimal estimating function:** Elements of the estimator for \( \theta \) have *smallest asymptotic variance*

- Take \( a(X) = E\{m(Y, Z; \theta) \mid X, Z = 1\} - E\{m(Y, Z; \theta) \mid X, Z = 0\} \)

- **Optimal estimating equation**

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

- Yields optimal "adjusted" estimator for \( \beta \)

- \( E\{m(Y, Z; \theta) \mid X, Z = g\} \) are *unknown functions of X* ⇒ *model them...*
6. Implementation

Adaptive algorithm:

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

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

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

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

and obtain $\hat{\zeta}_g$ by OLS separately

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

$$\sum_{i=1}^{n} \left[ m(Y_i, Z_i; \theta) - (Z_i - \hat{\pi}) \{q_1(X_i, \hat{\zeta}_1) - q_0(X_i, \hat{\zeta}_0)\} \right] = 0 \Rightarrow \text{“adjusted” } \tilde{\theta}$$
6. Implementation

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

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

- $\Rightarrow$ Model $E(Y \mid X, Z = g)$, the regression relationship in each treatment group

**In general:** *Standard errors* for $\tilde{\theta}$ and hence $\tilde{\beta}$

- $\tilde{\theta}$ is an *M-estimator*

- $\Rightarrow$ *Sandwich method* for asymptotic variance $\tilde{\beta}$
6. Implementation

**Special case:** *Example 1, continuous response* $Y$

$$E(Y | Z) = \gamma + \beta Z, \quad \beta = E(Y | Z = 1) - E(Y | Z = 0)$$

- All estimators for $\beta$ are *asymptotically equivalent* to

$$\bar{Y}_1 - \bar{Y}_0 - \sum_{i=1}^{n} (Z_i - \hat{\pi}) \left\{ n_0^{-1} h_0(X_i) + n_1^{-1} h_1(X_i) \right\}$$

$\bar{Y}_g$ and $n_g$ are sample average and sample size for group $g$, $h_g(X)$ are arbitrary functions of $X$.

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

- *Optimal estimator* takes $h_g(X) = E(Y|X, Z = g), g = 0, 1$

See Tsiatis et al. (2008)
6. Implementation

Properties: From *semiparametric theory*

- With the \((linear)\) regression models \(q_g(X, \zeta_g)\) as above, \(\tilde{\theta}\) is guaranteed relatively more efficient than \(\hat{\theta}\), even if \(q_g\) incorrect
- \(\tilde{\theta}\) 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 \(E\{m(Y, Z; \theta) \mid X, Z = g\}\)
6. Implementation

By-product:

- The “adjustment” for $X$ is determined separately by treatment group.

  - ... and regression modeling is carried out independently of $\tilde{\beta}$

  - ⇒ 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 $\tilde{\theta}$ (and hence $\tilde{\beta}$)

  - ⇒ A sponsor could retain different CROs to build the models for each treatment
7. Simulations

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

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

- $P(Z = 1) = P(Z = 2) = 0.5$
- $X = (X_1, \ldots, X_8)^T$, combination of continuous and discrete
- $(X_1, \ldots, X_4)$ “important,” $(X_5, \ldots, X_8)$ “unimportant”
- Generate $Y$ as Bernoulli with
  $$\text{logit}\{P(Y = 1|Z = g, X)\} = \alpha_0 g + \alpha^T g X, \ g = 0, 1$$

$\alpha_g$ 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 \( E(Y \mid X, Z = g) \) developed as

Aug. 1 \( q_g(X, \zeta_g) = \) linear model using only “important” covariates, fit by OLS

Aug. 2 \( q_g(X, \zeta_g) = \) linear model using all covariates \( X \), fit by OLS

Aug. 3,4 Like Aug. 1,2 but use a logistic model and fit by ML

Aug. 5 Like Aug 1,2 but use *forward selection* with OLS

Aug. 6 Like Aug 3,4 but use *forward selection* with ML

**Competitor:** “Usual” – fit

\[
\text{logit}\{E(Y \mid X, Z)\} = \alpha_0 + \alpha_1^T X + \phi Z
\]

by ML using only “important” covariates
### 7. Simulations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mild Association</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>−0.494</td>
<td>0.002</td>
<td>0.168</td>
<td>0.166</td>
<td>0.948</td>
<td>1.00</td>
</tr>
<tr>
<td>Aug. 1</td>
<td>−0.494</td>
<td>0.000</td>
<td>0.156</td>
<td>0.153</td>
<td>0.948</td>
<td>1.16</td>
</tr>
<tr>
<td>Usual</td>
<td>−0.494</td>
<td>−0.091</td>
<td>0.185</td>
<td>0.182</td>
<td>0.922</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>Moderate Association</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>−0.490</td>
<td>0.001</td>
<td>0.165</td>
<td>0.165</td>
<td>0.948</td>
<td>1.00</td>
</tr>
<tr>
<td>Aug. 1</td>
<td>−0.490</td>
<td>−0.002</td>
<td>0.140</td>
<td>0.139</td>
<td>0.950</td>
<td>1.39</td>
</tr>
<tr>
<td>Usual</td>
<td>−0.490</td>
<td>−0.218</td>
<td>0.203</td>
<td>0.201</td>
<td>0.813</td>
<td>0.31</td>
</tr>
<tr>
<td><strong>Strong Association</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>−0.460</td>
<td>0.004</td>
<td>0.164</td>
<td>0.165</td>
<td>0.954</td>
<td>1.00</td>
</tr>
<tr>
<td>Aug. 1</td>
<td>−0.460</td>
<td>0.000</td>
<td>0.132</td>
<td>0.131</td>
<td>0.952</td>
<td>1.55</td>
</tr>
<tr>
<td>Usual</td>
<td>−0.460</td>
<td>−0.321</td>
<td>0.223</td>
<td>0.220</td>
<td>0.695</td>
<td>0.18</td>
</tr>
</tbody>
</table>

**Aug 1–6 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)
9. Discussion

- General approach to using *auxiliary baseline covariates* to *improve efficiency* of *estimators* and *tests* (increased *power*)

- General measures of *treatment effect*

- Arises naturally via *semiparametric theory*

- Incorporation of covariate information *separated from* evaluation of treatment effects

- Effects of *model selection* deserve further study

- Can be extended to handle *missing outcome*
References


