Semiparametric Estimation of Treatment Effect in a Pretest-Posttest Study

Marie Davidian
Department of Statistics
North Carolina State University

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

(Joint work with A.A. Tsiatis and S. Leon)

Outline

1. Introduction and review of popular methods
2. Influence functions
3. Robins, Rotnitzky, and Zhao (1994)
4. Estimation when full data are available
5. Estimation when posttest response is missing at random (MAR)
6. Full data influence functions, revisited
7. Simulation evidence
8. Application – ACTG 175
9. Discussion

1. Introduction and review

The "pretest-posttest" study: Ubiquitous in research in medicine, public health, social science, etc...

- Subjects are randomized to *two* treatments ("*treatment*" and "*control*")
- Response is measured at *baseline* ("*pretest*") and at a pre-specified *follow-up* time ("*posttest*")
- Focus of inference: "*Difference in change of mean response from baseline to follow-up between treatment and control*"

1. Introduction and review

For example: AIDS Clinical Trials Group 175

- 2139 patients randomized to ZDV, ZDV+ddI, ZDV+zalcitabine, ddI with equal probability (1/4)
- *Primary analysis* (time-to-event endpoint): ZDV inferior to other three (no differences)
- Two groups: ZDV alone ("*control*") and other three ("*treatment*")
- *Secondary analyses*: Compare change in CD4 count (immunologic status) from (i) baseline to 20±5 weeks and (ii) baseline to 96±5 weeks between control and treatment

1. Introduction and review

Formally: Define

Y_1	baseline (*pretest*) response (e.g., baseline CD4 count)
Y_2	follow-up (*posttest*) response (e.g., 20±5 week CD4 count)
Z	$=$ 0 if control, $=$ 1 if treatment, $P(Z=1)=\delta$

- By *randomization*, reasonable to assume

$$E(Y_1|Z=0) = E(Y_1|Z=1) = E(Y_1) = \mu_1$$

Effect of interest: β, where

$$\begin{aligned}
\{E(Y_2|Z=1) - E(Y_1|Z=1)\} - \{E(Y_2|Z=0) - E(Y_1|Z=0)\} \\
&= \{E(Y_2|Z=1) - \mu_1\} - \{E(Y_2|Z=0) - \mu_1)\} \\
&= E(Y_2|Z=1) - E(Y_2|Z=0) \\
&= \quad \mu_2^{(1)} \quad - \quad \mu_2^{(0)} \quad = \beta
\end{aligned}$$

1. Introduction and review

Basic data: (Y_{1i}, Y_{2i}, Z_i), $i = 1, \ldots, n$, iid

$$n_0 = \sum_{i=1}^{n}(1 - Z_i) = \sum_{i=1}^{n} I(Z_i = 0), \quad n_1 = \sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} I(Z_i = 1)$$

Popular estimators for β:

- *Two-sample t-test* estimator

$$\widehat{\beta}_{2samp} = n_1^{-1} \sum_{i=1}^{n} Z_i Y_{2i} - n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_{2i}$$

- "*Paired t-test*" estimator ("change scores")

$$\widehat{\beta}_{pair} = \overline{D}_1 - \overline{D}_0, \quad \overline{D}_c = n_c^{-1} \sum_{i=1}^{n} I(Z_i = c)(Y_{2i} - Y_{1i}), \quad c = 0, 1$$

1. Introduction and review

Popular estimators for β:

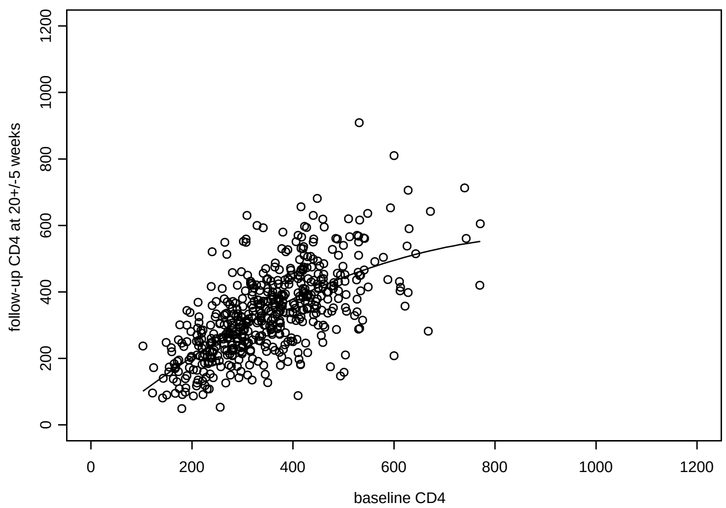
- *ANCOVA* – Fit the model

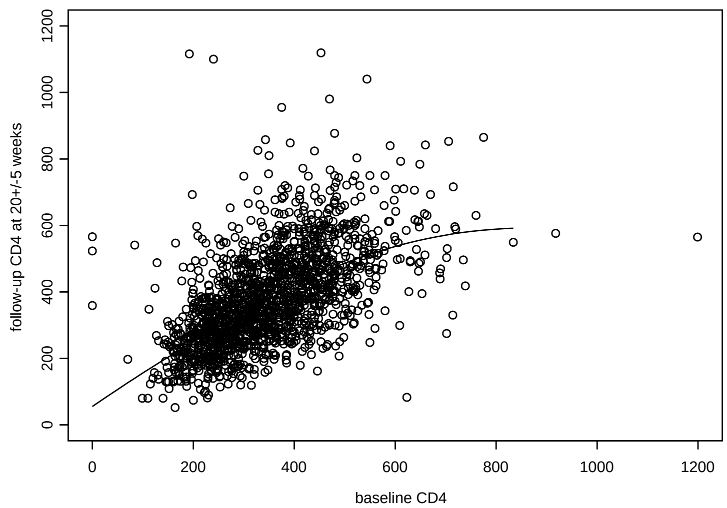
$$E(Y_2|Y_1, Z) = \alpha_0 + \alpha_1 Y_1 + \beta Z$$

- *ANCOVA II* – Include *interaction* and estimate β as coefficient of $Z - \overline{Z}$ in regression of $Y_2 - \overline{Y}_2$ on $Y_1 - \overline{Y}_1$, $Z - \overline{Z}$, and $(Y_1 - \overline{Y}_1)(Z - \overline{Z})$
- *GEE* – $(Y_1, Y_2)^T$ is multivariate response with mean $(\mu_1, \mu_2 + \beta Z)^T$ and (2×2) unstructured covariance matrix
- Assume *linear* relationship between Y_2 and Y_1

1. Introduction and review

ACTG 175: Y_2 = CD4 at 20±5 weeks vs. Y_1 = baseline CD4 (control and treatment groups)





 NC STATE UNIVERSITY

1. Introduction and review

Additional data: Baseline and intermediate *covariates*

X_1 Baseline (pre-treatment) characteristics

X_2 Characteristics observed after pretest but before posttest, including intermediate responses

In ACTG 175:

- X_1 includes weight, age, gender, Karnofsky score, prior ARV therapy, CD8 count, sexual preference,...
- X_2 includes off treatment indicator, intermediate CD4, CD8

1. Introduction and review

Additional estimators:

- *Fancier* regression models, e.g.

$$E(Y_2|Y_1, Z) = \alpha_0 + \alpha_1 Y_1 + \alpha_2 Y_1^2 + \beta Z$$

- Adjustment for *baseline covariates*, e.g.,

$$E(Y_2|X_1, Y_1, Z) = \alpha_0 + \alpha_1 Y_1 + \alpha_2 X_1 + \beta Z$$

- *Both*

- *Intuitively*, adjustment for *intermediate covariates* buys nothing without some assumptions (formally exhibited shortly. . .) and could be *dangerous*

1. Introduction and review

Which estimator?

- In many settings, no *consensus*
- Is *normality* required?
- What if the relationship isn't *linear*?
- What if a model for $E(Y_2|X_1, Y_1, Z)$ is *wrong*?

1. Introduction and review

Which estimator?

- In many settings, no *consensus*
- Is *normality* required?
- What if the relationship isn't *linear*?
- What if a model for $E(Y_2|X_1, Y_1, Z)$ is *wrong*?

Further complication: *Missing* posttest response Y_2

- In ACTG 175, no missing CD4 for any subject at 20±5 weeks...
- ...but 797 (*37%*) of subjects were *missing* CD4 at 96±5 weeks (almost entirely due to dropout from study)
- Common in practice – *complete case analysis*, which yields possibly *biased* inference on β unless Y_2 is *missing completely at random*

Introduction and review

Missing at random (MAR) assumption: Posttest missingness associated with (X_1, Y_1, X_2, Z) but not Y_2

- Often reasonable, but is an *assumption*

Full data: If *no missingness*, observe (X_1, Y_1, X_2, Y_2, Z)

Ordinarily: Models for (X_1, Y_1, X_2, Y_2, Z) may involve *assumptions*

- If Y_2 not missing, *widespread belief* that normality of (Y_1, Y_2) is required for validity of "popular" estimators
- When Y_2 is MAR, *maximum likelihood*, *imputation* approaches require assumptions on aspects of the *joint distribution* of (X_1, Y_1, X_2, Y_2, Z)
- *Consequences*?

Introduction and review

Semiparametric models:

- May contain *parametric* and *nonparametric* components
- *Nonparametric components* – unable or unwilling to make specific assumptions on aspects of (X_1, Y_1, X_2, Y_2, Z)

Introduction and review

Semiparametric models:

- May contain *parametric* and *nonparametric* components
- *Nonparametric components* – unable or unwilling to make specific assumptions on aspects of (X_1, Y_1, X_2, Y_2, Z)

Here: Consider a *semiparametric model* for (X_1, Y_1, X_2, Y_2, Z)

- No assumptions on joint distribution of (X_1, Y_1, X_2, Y_2, Z) beyond *independence* of (X_1, Y_1) and Z induced by *randomization* (nonparametric)
- Interested in the *functional* of this distribution

$$\beta = \mu_2^{(1)} - \mu_2^{(0)} = E(Y_2|Z=1) - E(Y_2|Z=0)$$

Introduction and review

Where we are going: Under this semiparametric model

- Find a class of *consistent and asymptotically normal* (*CAN*) estimators for β when *full data* are available and identify the "*best*" (*efficient*) estimator in the class
- As a by-product, show that "*popular*" estimators are potentially *inefficient* members of this class – can do *better*!
- When Y_2 is *MAR*, find a class of *CAN* estimators for β and identify the "*best*"
- In both cases, translate the theory into *practical techniques*

Introduction and review

Where we are going: Under this semiparametric model

- Find a class of *consistent and asymptotically normal* (*CAN*) estimators for β when *full data* are available and identify the "*best*" (*efficient*) estimator in the class
- As a by-product, show that "*popular*" estimators are potentially *inefficient* members of this class – can do *better*!
- When Y_2 is *MAR*, find a class of *CAN* estimators for β and identify the "*best*"
- In both cases, translate the theory into *practical techniques*

What we will exploit: Theory in a *landmark* paper by Robins, Rotnitzky, and Zhao (1994)

2. Influence functions

Definition: For functional β in a parametric or semiparametric model, an estimator $\widehat{\beta}$ based on iid random vectors W_i, $i = 1, \ldots, n$, is *asymptotically linear* if

$$n^{1/2}(\widehat{\beta} - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(W_i) + o_p(1) \quad \text{for some } \varphi(W)$$

$\beta_0 =$ true value of β $(p \times 1)$, $\quad E\{\varphi(W)\} = 0, \quad E\{\varphi^T(W)\varphi(W)\} < \infty$

2. Influence functions

Definition: For functional β in a parametric or semiparametric model, an estimator $\widehat{\beta}$ based on iid random vectors W_i, $i = 1, \ldots, n$, is *asymptotically linear* if

$$n^{1/2}(\widehat{\beta} - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(W_i) + o_p(1) \quad \text{for some } \varphi(W)$$

$$\beta_0 = \text{ true value of } \beta \ (p \times 1), \quad E\{\varphi(W)\} = 0, \quad E\{\varphi^T(W)\varphi(W)\} < \infty$$

- $\varphi(W)$ is called the *influence function* of $\widehat{\beta}$
- If $\widehat{\beta}$ is also *regular* (not "pathological"), $\widehat{\beta}$ is *CAN* with *asymptotic covariance matrix* $E\{\varphi(W)\varphi^T(W)\}$
- *Efficient influence function* $\varphi^{eff}(W)$ has "*smallest*" covariance and corresponds to the *efficient*, *regular asymptotically linear* (*RAL*) *estimator*

2. Influence functions

For example: It may be shown directly by manipulating the expression for $n^{1/2}(\widehat{\beta}_{2samp} - \beta)$ and using

$$n_0/n \rightarrow 1-\delta,\ \ n_1/n \rightarrow \delta \ \ \text{as } n \rightarrow \infty$$

that $\widehat{\beta}_{2samp}$ has influence function of the form

$$\frac{Z(Y_2-\mu_2^{(1)})}{\delta} - \frac{(1-Z)(Y_2-\mu_2^{(0)})}{1-\delta} = \frac{Z(Y_2-\mu_2^{(0)}-\beta)}{\delta} - \frac{(1-Z)(Y_2-\mu_2^{(0)})}{1-\delta}$$

[depends on $W = (X_1, Y_1, X_2, Y_2, Z)$]

Why is this useful? There is a *correspondence* between *CAN, RAL estimators* and *influence functions*

- By identifying influence functions, one can deduce estimators

2. Influence functions

General principle: Solve $\sum_{i=1}^{n} \varphi(W_i) = 0$ for β

For example: Influence function for $\widehat{\beta}_{2samp}$

$$0 = \sum_{i=1}^{n} \left\{ \frac{Z_i(Y_{2i} - \mu_2^{(0)} - \beta)}{\delta} - \frac{(1 - Z_i)(Y_{2i} - \mu_2^{(0)})}{1 - \delta} \right\}$$

- Substituting $\mu_2^{(0)} = n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_{2i}$ and solving for β yields

$$\beta = n_1^{-1} \sum_{i=1}^{n} Z_i Y_{2i} - n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_{2i}$$

- In general, *closed form* may not be possible

3. Robins, Rotnitzky, and Zhao (1994)

What did RRZ do? Derived *asymptotic theory* based on influence functions for inference on functionals in *general semiparametric models* where some components of the *full data* are possibly *MAR*

Observed data: Data observed when some components of the *full data* are *potentially missing*

3. Robins, Rotnitzky, and Zhao (1994)

What did RRZ do, more specifically? For the functional of interest, distinguished between

- *Full-data influence functions* – correspond to RAL estimators calculable if full data were available; functions of the full data
- *Observed-data influence functions* – correspond to RAL estimators calculable from the observed data under MAR; functions of the observed data
- RRZ characterized the class of *all observed-data influence functions* for a general semiparametric model, including the *efficient* one, . . .
- . . . and showed that observed-data influence functions may be expressed in terms of full-data influence functions

3. Robins, Rotnitzky, and Zhao (1994)

The main result: Generic *full data* $D = (O, M)$, *semiparametric model* for D, *functional* β

O	Part of D that is *always observed* (*never missing*)
M	Part of D that may be *missing*
R	$= 1$ if M is observed, $= 0$ if M is missing

- *Observed data* are (O, R, RM)
- Let $\varphi^F(D)$ be a *full data influence function*
- Let $\pi(O) = P(R = 1|D) = P(R = 1|O) > \epsilon$ (*MAR assumption*)
- All *observed-data influence functions* have form

$$\frac{R\varphi^F(D)}{\pi(O)} - \frac{R - \pi(O)}{\pi(O)} g(O), \quad g(O) \text{ square-integrable}$$

3. Robins, Rotnitzky, and Zhao (1994)

Result: Strategy for deriving estimators for a *semiparametric model*

1. Characterize the class of *full-data influence functions* (which correspond to *full-data estimators*)
2. Characterize the *observed data* under the particular MAR mechanism and the class of *observed-data influence functions*
3. Identify *observed-data estimators* with influence functions in this class

Our approach: Follow these steps for the *semiparametric pretest-posttest model*

- Joint distribution of (X_1, Y_1, X_2, Y_2, Z) *unspecified* except (X_1, Y_1) independent of Z

4. Estimation with full data

Full-data influence functions: Can show (later) under the *semiparametric pretest-posttest model* that all full-data influence functions are of the form

$$\left\{ \frac{Z(Y_2-\mu_2^{(1)})}{\delta} - \frac{(Z-\delta)}{\delta} h^{(1)}(X_1,Y_1) \right\} - \left\{ \frac{(1-Z)(Y_2-\mu_2^{(0)})}{1-\delta} + \frac{(Z-\delta)}{1-\delta} h^{(0)}(X_1,Y_1) \right\},$$

for arbitrary $h^{(c)}(X_1,Y_1)$, $c=0,1$ with var$\{h^{(c)}(X_1,Y_1)\} < \infty$

- *Difference* of influence functions for estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$
- *Full-data estimators* may depend on X_1 (but not X_2)

4. Estimation with full data

Full-data influence functions: Can show (later) under the *semiparametric pretest-posttest model* that all full-data influence functions are of the form

$$\left\{ \frac{Z(Y_2 - \mu_2^{(1)})}{\delta} - \frac{(Z-\delta)}{\delta} h^{(1)}(X_1, Y_1) \right\} - \left\{ \frac{(1-Z)(Y_2 - \mu_2^{(0)})}{1-\delta} + \frac{(Z-\delta)}{1-\delta} h^{(0)}(X_1, Y_1) \right\},$$

for arbitrary $h^{(c)}(X_1, Y_1)$, $c = 0, 1$ with var$\{h^{(c)}(X_1, Y_1)\} < \infty$

- *Difference* of influence functions for estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$
- *Full-data estimators* may depend on X_1 (but not X_2)

Efficient full-data influence function: Corresponding to *efficient full-data estimator*; takes

$$h^{(c)}(X_1, Y_1) = E(Y_2 | X_1, Y_1, Z = c) - \mu_2^{(c)}, \ c = 0, 1$$

4. Estimation with full data

"Popular" estimators: *Influence functions* of $\widehat{\beta}_{2samp}$, $\widehat{\beta}_{pair}$, ANCOVA, ANCOVA II, and GEE have

$$h^{(c)}(X_1, Y_1) = \eta_c(Y_1 - \mu_1), \quad c = 0, 1, \text{ for constants } \eta_c$$

- E.g., $\eta_c = 0$, $c = 0, 1$ for $\widehat{\beta}_{2samp}$
- So popular estimators are in the class $\Longrightarrow$ are *CAN* even if (Y_1, Y_2) are not normal
- Regression estimators incorporating *baseline covariates* are also in the class, e.g., $E(Y_1|X_1, Y_1, Z) = \alpha_0 + \alpha_1 Y_1 + \alpha_2 X_1 + \beta Z$
- Popular estimators are potentially *inefficient* among class of RAL estimators for semiparametric model

4. Estimation with full data

"Popular" estimators: *Influence functions* of $\widehat{\beta}_{2samp}$, $\widehat{\beta}_{pair}$, ANCOVA, ANCOVA II, and GEE have

$$h^{(c)}(X_1, Y_1) = \eta_c(Y_1 - \mu_1), \quad c = 0, 1, \text{ for constants } \eta_c$$

- E.g., $\eta_c = 0$, $c = 0, 1$ for $\widehat{\beta}_{2samp}$
- So popular estimators are in the class $\Longrightarrow$ are *CAN* even if (Y_1, Y_2) are not normal
- Regression estimators incorporating *baseline covariates* are also in the class, e.g., $E(Y_1|X_1, Y_1, Z) = \alpha_0 + \alpha_1 Y_1 + \alpha_2 X_1 + \beta Z$
- Popular estimators are potentially *inefficient* among class of RAL estimators for semiparametric model

How to use all this? *Efficient estimator* is "*best*!"

4. Estimation with full data

Efficient estimator: Setting sum over i of *efficient influence function* $=$ 0 and replacing δ by $\widehat{\delta} = n_1/n$ yields

$$\begin{aligned} \beta &= n_1^{-1}\left\{\sum_{i=1}^{n} Z_i Y_{2i} - \sum_{i=1}^{n}(Z_i - \widehat{\delta})E(Y_{2i}|X_{1i}, Y_{1i}, Z_i = 1)\right\} \\ &- n_0^{-1}\left\{\sum_{i=1}^{n}(1 - Z_i)Y_{2i} + \sum_{i=1}^{n}(Z_i - \widehat{\delta})E(Y_{2i}|X_{1i}, Y_{1i}, Z_i = 0)\right\} \end{aligned}$$

- *Practical use* – *replace* $E(Y_2|X_1, Y_1, Z = c)$ by *predicted values* $\widehat{e}_{h(c)i}$, say, $c = 0, 1$, from *parametric* or *nonparametric* regression modeling
- Can lead to substantial *increase in precision* over popular estimators
- *Advantage* – even if $E(Y_2|X_1 Y_1, Z = c)$ are modeled *incorrectly*, $\widehat{\beta}$ is still *consistent*

5. Estimation with posttest MAR

Observed data: (X_1, Y_1, X_2, Z) are *never missing*, Y_2 *may* be missing for some subjects

- $R = 1$ if Y_2 observed, $R = 0$ if Y_2 missing
- *Observed data* are $(X_1, Y_1, X_2, Z, R, RY_2)$
- *MAR assumption*

$$\begin{aligned} P(R = 1 | X_1, Y_1, X_2, Y_2, Z) &= P(R = 1 | X_1, Y_1, X_2, Z) \\ &= \pi(X_1, Y_1, X_2, Z) \geq \epsilon > 0 \end{aligned}$$

$$\pi(X_1, Y_1, X_2, Z) = Z\pi^{(1)}(X_1, Y_1, X_2) + (1 - Z)\pi^{(0)}(X_1, Y_1, X_2),$$

$$\pi^{(c)}(X_1, Y_1, X_2) = \pi(X_1, Y_1, X_2, c), \;\; c = 0, 1$$

5. Estimation with posttest MAR

Recall: Generic form of *observed-data influence functions*

$$\frac{R\varphi^F(D)}{\pi(O)} - \frac{R - \pi(O)}{\pi(O)} g(O)$$

For simplicity: Focus on influence functions for estimators for $\mu_2^{(1)}$

- Those for estimators for $\mu_2^{(0)}$ *similar*
- Influence functions for estimators for β: take the *difference*

5. Estimation with posttest MAR

Recall: Generic form of *observed-data influence functions*

$$\frac{R\varphi^F(D)}{\pi(O)} - \frac{R-\pi(O)}{\pi(O)} g(O)$$

For simplicity: Focus on influence functions for estimators for $\mu_2^{(1)}$

- Those for estimators for $\mu_2^{(0)}$ *similar*
- Influence functions for estimators for β: take the *difference*

Full-data influence functions for estimators for $\mu_2^{(1)}$: Have form

$$\frac{Z(Y_2 - \mu_2^{(1)})}{\delta} - \frac{(Z-\delta)}{\delta} h^{(1)}(X_1, Y_1), \quad \text{var}\{h^{(1)}(X_1, Y_1)\} < \infty$$

5. Estimation with posttest MAR

Thus: *Observed-data influence functions* for estimators for $\mu_2^{(1)}$ have form

$$\frac{R\{Z(Y_2-\mu_2^{(1)})-(Z-\delta)h^{(1)}(X_1,Y_1)\}}{\delta\pi(X_1,Y_1,X_2,Z)}-\frac{R-\pi(X_1,Y_1,X_2,Z)}{\pi(X_1,Y_1,X_2,Z)}g^{(1)}(X_1,Y_1,X_2,Z)$$

$$\text{var}\{h^{(1)}(X_1,Y_1)\}<\infty, \quad \text{var}\{g^{(1)}(X_1,Y_1,X_2,Z)\}<\infty$$

- Choice of $h^{(1)}$ leading to the *efficient* observed-data influence function *need not be the same* as that leading to the *efficient full-data* influence function *in general*
- Turns out that the optimal $h^{(1)}$ *is* the same in the *special case* of the pretest-posttest problem…

5. Estimation with posttest MAR

Re-writing: Equivalently, *observed-data influence functions* are

$$\frac{RZ(Y_2-\mu_2^{(1)})}{\delta\pi(X_1,Y_1,X_2,Z)} - \frac{(Z-\delta)}{\delta}h^{(1)}(X_1,Y_1) - \frac{R-\pi(X_1,Y_1,X_2,Z)}{\delta\pi(X_1,Y_1,X_2,Z)}g^{(1)'}(X_1,Y_1,X_2,Z)$$

- *Optimal* choices (*efficient influence function*) are

$$h^{eff(1)}(X_1,Y_1) = E(Y_2|X_1,Y_1,Z=1) - \mu_2^{(1)}$$

$$\begin{aligned} g^{eff(1)'}(X_1,Y_1,X_2,Z) &= Z\{E(Y_2|X_1,Y_1,X_2,Z) - \mu_2^{(1)}\} \\ &= Z\{E(Y_2|X_1,Y_1,X_2,Z=1) - \mu_2^{(1)}\} \end{aligned}$$

- *Efficient influence function* is of form

$$\frac{RZ(Y_2-\mu_2^{(1)})}{\delta\pi^{(1)}(X_1,Y_1,X_2)} - \frac{(Z-\delta)}{\delta}h^{(1)}(X_1,Y_1) - \frac{\{R-\pi^{(1)}(X_1,Y_1,X_2)\}Z}{\delta\pi^{(1)}(X_1,Y_1,X_2)}q^{(1)}(X_1,Y_1,X_2)$$

5. Estimation with posttest MAR

Result: With the *optimal* $h^{(1)}$, $q^{(1)}$, algebra yields

$$\mu_2^{(1)} = (n\delta)^{-1}\left\{\sum_{i=1}^{n}\frac{R_i Z_i Y_{2i}}{\pi^{(1)}(X_{1i}, Y_{1i}, X_{2i})} - \sum_{i=1}^{n}(Z_i - \delta)E(Y_{2i}|X_{1i}, Y_{1i}, Z_i = 1)\right.$$
$$\left. - \sum_{i=1}^{n}\frac{\{R_i - \pi^{(1)}(X_{1i}, Y_{1i}, X_{2i})\}Z_i}{\pi^{(1)}(X_{1i}, Y_{1i}, X_{2i})}E(Y_{2i}|X_{1i}, Y_{1i}, X_{2i}, Z_i = 1)\right\}$$

- *Similarly* for $\mu_2^{(0)}$ depending on $\pi^{(0)}$, $E(Y_2|X_1, Y_1, Z = 0)$, $E(Y_2|X_1, Y_1, X_2, Z = 0)$
- *Estimator* for β – *take the difference*
- *Practical use* – *replace* these quantities by *predicted values* from *regression modeling* (coming up)

5. Estimation with posttest MAR

Complication 1: $\pi^{(c)}(X_1, Y_1, X_2)$ are *not known*, $c = 0, 1$

- Common strategy: adopt *parametric models* (e.g. *logistic regression*) depending on parameter $\gamma^{(c)}$

$$\pi^{(c)}(X_1, Y_1, X_2; \gamma^{(c)})$$

- Imposes an *additional assumption* on semiparametric model for (X_1, Y_1, X_2, Y_1, Z)
- *Substitute* the MLE $\widehat{\gamma}^{(c)}$ for $\gamma^{(c)}$, obtain *predicted values* $\widehat{\pi}_i^{(c)}$
- As long as *this model* is *correct*, resulting estimators will be CAN

5. Estimation with posttest MAR

Complication 2: Modeling $E(Y_2|X_1, Y_1, Z = c)$, $E(Y_2|X_1, Y_1, X_2, Z = c)$, $c = 0, 1$

- *MAR* $\Longrightarrow E(Y_2|X_1, Y_1, X_2, Z) = E(Y_2|X_1, Y_1, X_2, Z, R = 1)$ (can base modeling/fitting on *complete cases* only)
- Obtain *predicted values* $\widehat{e}_{q(c)i}$, $c = 0, 1$
- However, *ideally* require compatibility, i.e.

  $$E(Y_2|X_1, Y_1, Z) = E\{E(Y_2|X_1, Y_1, X_2, Z)|X_1, Y_1, Z\}$$

  and *no longer valid* to fit using only complete cases
- *Practically* – go ahead and *model directly* and fit using *complete cases*, obtain *predicted values* $\widehat{e}_{h(c)i}$
- Estimation of parameters in these models *does not affect* (asymptotic) *variance* of $\widehat{\beta}$ as long as $\pi^{(c)}$ models are *correct*

5. Estimation with posttest MAR

Estimator: With $\widehat{\delta} = n_1/n$

$$\widehat{\beta} = n_1^{-1}\left\{\sum_{i=1}^{n}\frac{R_i Z_i Y_{2i}}{\widehat{\pi}_i^{(1)}} - \sum_{i=1}^{n}(Z_i - \widehat{\delta})\widehat{e}_{h(1)i} - \sum_{i=1}^{n}\frac{(R_i - \widehat{\pi}_i^{(1)})Z_i\widehat{e}_{q(1)i}}{\widehat{\pi}_i^{(1)}}\right\}$$

$$-n_0^{-1}\left\{\sum_{i=1}^{n}\frac{R_i(1-Z_i)Y_{2i}}{\widehat{\pi}_i^{(0)}} + \sum_{i=1}^{n}(Z_i - \widehat{\delta})\widehat{e}_{h(0)i} - \sum_{i=1}^{n}\frac{(R_i - \widehat{\pi}_i^{(0)})(1-Z_i)\widehat{e}_{q(1)i}}{\widehat{\pi}_i^{(0)}}\right\}$$

- *Efficient* if modeling done *correctly*; otherwise, *close to optimal* performance
- Taking $\widehat{e}_{h(c)i} = \widehat{e}_{q(c)i} = 0$ yields the *simple inverse-weighted complete case* estimator (*inefficient*)
- Modeling $E(Y_2|X_1, Y_1, Z = c)$, $E(Y_2|X_1, Y_1, X_2, Z = c)$ "*augments*" this, taking advantage of relationships among variables to *improve precision*

5. Estimation with posttest MAR

"Double Robustness:" Still *consistent* if

- $\pi^{(c)}$ are *correctly modeled* but $E(Y_2|X_1, Y_1, Z = c)$ and $E(Y_2|X_1, Y_1, X_2, Z = c)$ *aren't*
- $E(Y_2|X_1, Y_1, Z = c)$ and $E(Y_2|X_1, Y_1, X_2, Z = c)$ are *correctly modeled* but $\pi^{(c)}$ *aren't*
- No longer *efficient*

If *both* sets of models *incorrect*, *inconsistent* in general

5. Estimation with posttest MAR

"Double Robustness:" Still *consistent* if

- $\pi^{(c)}$ are *correctly modeled* but $E(Y_2|X_1, Y_1, Z = c)$ and $E(Y_2|X_1, Y_1, X_2, Z = c)$ *aren't*
- $E(Y_2|X_1, Y_1, Z = c)$ and $E(Y_2|X_1, Y_1, X_2, Z = c)$ are *correctly modeled* but $\pi^{(c)}$ *aren't*
- No longer *efficient*

If *both* sets of models *incorrect*, *inconsistent* in general

Standard errors: Use the *sandwich* formula (follows from influence function)

5. Estimation with posttest MAR

Recap: This approach requires one to make an assumption about $\pi^{(c)}(X_1, Y_1, X_2)$, $c = 0, 1$

- *No assumption* is made about $E(Y_2|X_1, Y_1, X_2, Z = c)$, $E(Y_2|X_1, Y_1, Z = c)$
- Model is *still semiparametric*
- ... and *double robustness* holds

5. Estimation with posttest MAR

Recap: This approach requires one to make an assumption about $\pi^{(c)}(X_1, Y_1, X_2)$, $c = 0, 1$

- *No assumption* is made about $E(Y_2|X_1, Y_1, X_2, Z = c)$, $E(Y_2|X_1, Y_1, Z = c)$
- Model is *still semiparametric*
- . . . and *double robustness* holds

Alternative approach: Make an assumption *instead* about the $E(Y_2|X_1, Y_1, X_2, Z = c)$, $E(Y_2|X_1, Y_1, Z = c)$

- *Efficient estimator* is *maximum likelihood*
- Don't need to even worry about $\pi^{(c)}(X_1, Y_1, X_2)$
- But *no double robustness property!*

6. Full data, revisited

How did we get the full-data influence functions?

- One way – use *classical semiparametric theory*
- Another way – View as a "*fake missing data problem*" by casting the full-data problem in terms of *counterfactuals*

6. Full data, revisited

How did we get the full-data influence functions?

- One way – use *classical semiparametric theory*
- Another way – View as a "*fake missing data problem*" by casting the full-data problem in terms of *counterfactuals*

Counterfactual representation:

- $Y_2^{(1)}$, $Y_2^{(0)}$ are *potential posttest responses* if a subject were assigned to control or treatment
- We *observe* $Y_2 = ZY_2^{(1)} + (1-Z)Y_2^{(0)}$
- "*Fake full data*" $(X_1, Y_1, X_2, Y_2^{(0)}, Y_2^{(1)}, Z)$
- "*Fake observed data*" $(X_1, Y_1, X_2, Z, ZY_2^{(1)}, (1-Z)Y_2^{(0)})$
- Apply the *RRZ theory*

7. Simulation evidence

Full-data problem:

- Substantial *gains in efficiency* over "popular" methods, especially when there are *nonlinear* relationships among variables
- *Parametric* and *nonparametric* regression modeling work well
- *Valid* standard errors, confidence intervals

7. Simulation evidence

Full-data problem:

- Substantial *gains in efficiency* over "popular" methods, especially when there are *nonlinear* relationships among variables
- *Parametric* and *nonparametric* regression modeling work well
- *Valid* standard errors, confidence intervals

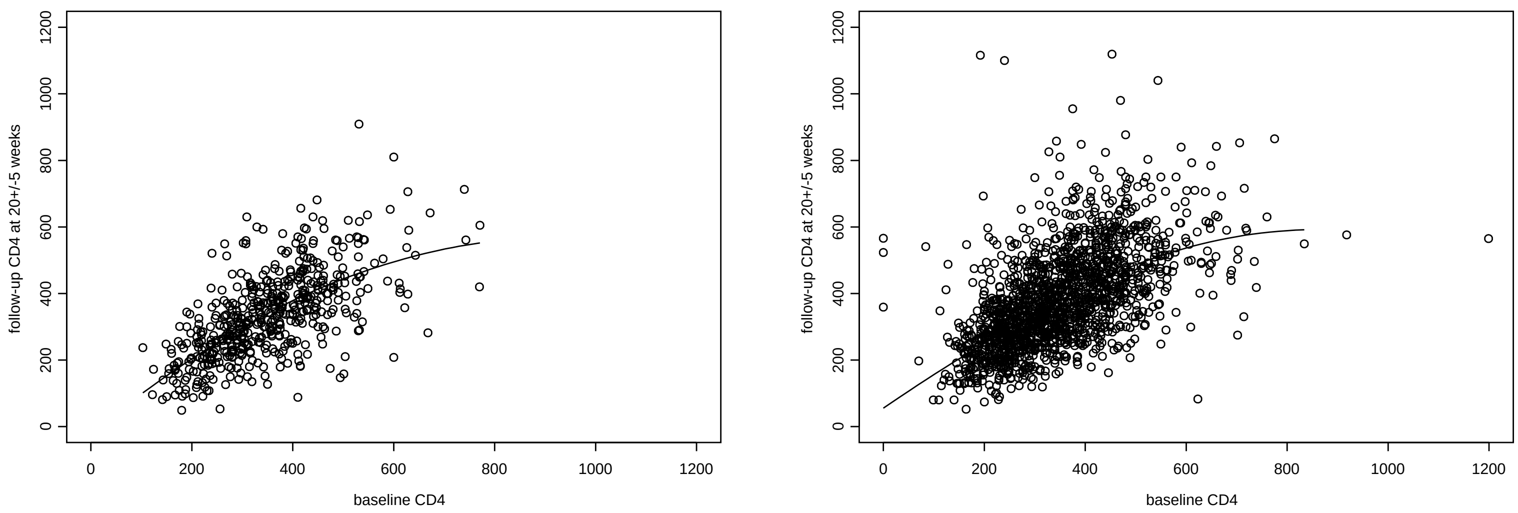
Observed-data problem:

- "Popular" methods with complete cases can exhibit *substantial* biases
- *Inverse-weighted complete case estimator* is unbiased but *inefficient*
- Substantial *gains in efficiency* possible through modeling
- *Valid* standard errors, confidence intervals

8. Application – ACTG 175

Recall: Y_2 = CD4 at 20±5 weeks vs. Y_1 = baseline CD4 (control and treatment groups)

- Apparent *curvature*



8. Application – ACTG 175

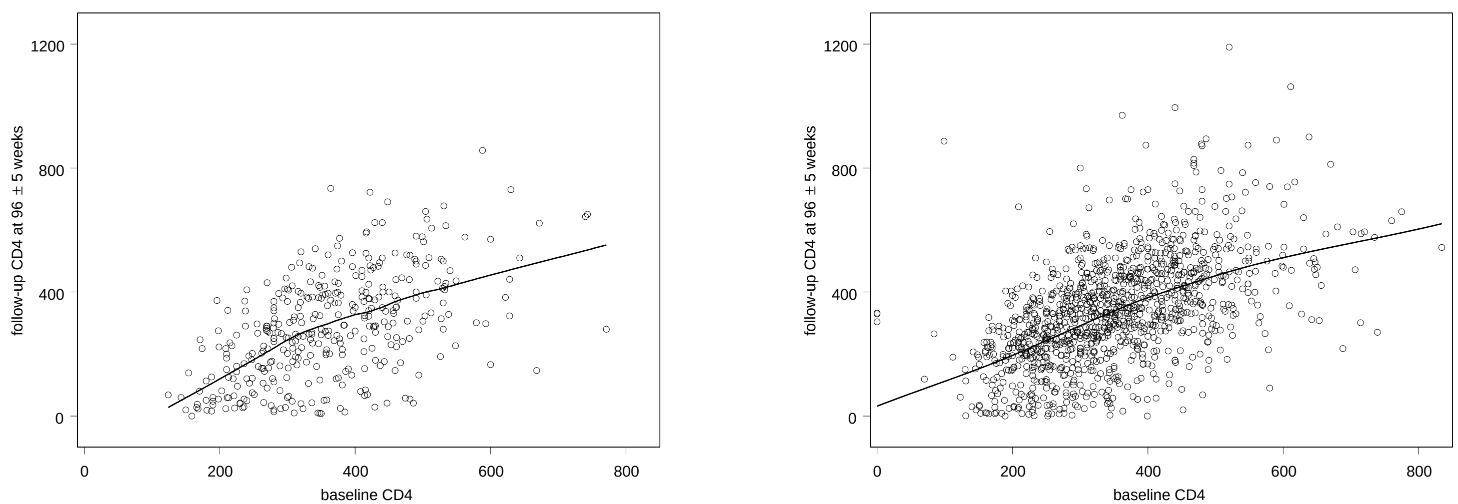
Results: Models for $E(Y_2|X_1, Y_1, Z = c)$, $c = 0, 1$

Estimator	$\widehat{\beta}$	SE
Parametric modeling (quadratic in Y_1)	50.8	5.0
Nonparametric modeling (GAM)	50.0	5.1
ANCOVA	49.3	5.4
Paired t	50.1	5.7
Two-sample t	45.5	6.8

8. Application – ACTG 175

Complete cases: Y_2 = CD4 at 96±5 weeks vs. Y_1 = baseline CD4 (control and treatment groups)

- 37% *missing* Y_2



8. Application – ACTG 175

Results: Logistic regression for $\pi^{(c)}$, $c = 0, 1$; parametric regression modeling of $E(Y_2|X_1, Y_1, X_2, Z = c)$, $E(Y_2|X_1, Y_1, Z = c)$

Estimator	$\widehat{\beta}$	SE
Parametric modeling (quadratic in Y_1)	57.2	10.2
Simple inverse-weighting	54.7	11.8
ANCOVA	64.5	9.3
Paired t	67.1	9.3

9. Discussion

- *RRZ theory* applied to a standard problem
- *General framework* for pretest-posttest analysis illuminating how relationships among variables may be fruitfully *exploited*
- *Practical estimators*
- Can be *extended* to *censored covariate information*
- Results are *equally applicable* to *baseline covariate adjustment* in comparison of *two means* (Y_1 is just another baseline covariate)
- Lots of methods for this problem (*likelihood*, *imputation combinations thereof*, ...); *semiparametric theory* provides a framework for understanding *commonalities* and *differences* among them

9. Discussion

References: *Gory details* available in

Leon, S., Tsiatis, A.A., and Davidian, M. (2003) Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics* 59, 1048–1057.

Davidian, M., Tsiatis, A.A., and Leon, S. (2005) Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science*, to appear.

Forthcoming:

Tsiatis, A.A. (200X) *Semiparametrics and Missing Data*. New York: Springer.