OBSERVATIONAL DATA ANALYSIS COMPETITION:
HETEROGENEOUS RESPONSE CHALLENGE

Some Relatively Simple Ways to Get Good Answers

Xiaochun Li, Ph.D.
Associate Professor
Division of Biostatistics
Indiana University School of Medicine

What do we want to estimate?

The population average treatment effect (ATE), i.e.,

$$E(Y_1) - E(Y_0)$$

Y_1 and Y_0 are conterfactual outcomes

In plain words: what if scenarios

The expected response if treatment had been assigned to the entire study population minus the expected response if control had been assigned to the entire study population

What do we want to estimate?

If suspect treatment heterogeneity, may want to estimate *local* average treatment difference instead,

$$E(Y_1|L) - E(Y_0|L)$$

where L is a vector of covariates.

Issues

- Patients were NOT Randomized to Treatment in any known way …i.e. potential treatment selection bias
- We do not observe both Y_1 and Y_0 for a patient – only one of them depending on the patient's treatment assignment
- Unknown confounders are present
  - Not much can be done if a confounder is not measured or included in the study
  - Could do sensitivity analysis to assess their impact on the causal estimates

Baseline covariate balance assessment

Variable	Arm 0	Arm 1	P value	Std.diff
age	43.7	45.6	<0.001	0.16
female	72.0%	71.8%	0.53	0.00
pain	0.62	0.77	<0.001	0.20
Hosp. count	0.14	0.21	<0.001	0.16
ER count	0.38	0.51	<0.001	0.16
Office visits	6.1	7.7	<0.001	0.28
Psyc pcnt	51.3%	54.0%	<0.001	0.11
w. Prev cost ($)	6755	9805	<0.001	0.33

Standardized difference

- A statistic not depending on the sample size
  - Exclude hypothesis testing
- d>0.25, better balance is needed (Cochran, 1968)

$$d = \frac{|\bar{l}_{treatment} - \bar{l}_{control}|}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

6

Simple ideas to deal with treatment selection bias

- if can assume 'no unmeasured confounders' (NUC):

$$\begin{aligned} E[Y_a] &= E[E(Y_a|L)] \\ &\overset{NUC}{=} E[E(Y_a|A=a,L)] \\ &\overset{consistency}{=} E[E(Y|A=a,L)] \end{aligned} \qquad (2.1)$$

- we can impute counter-factual response (underlined quantity) by regression
  - parametric, or non-parametric, with exact matching as a special case

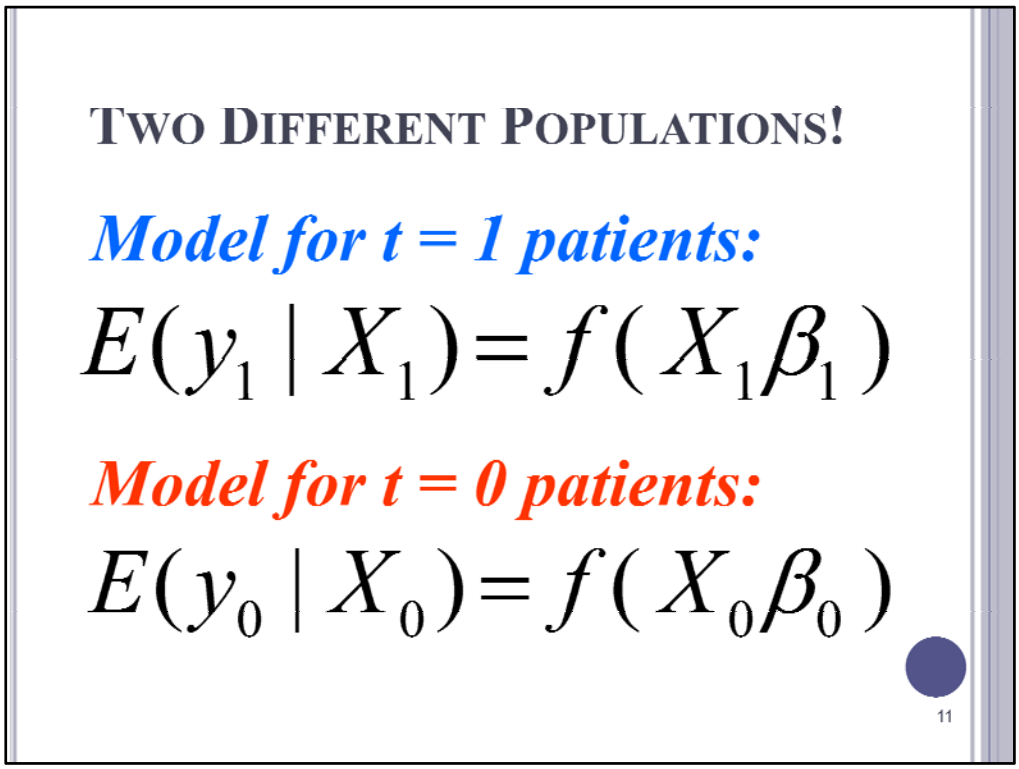
Implementation of simple ideas

Parametric regressions

- Linear regression with common β for both arms 0 and 1
- Linear regression with separate βs for arms 0 and 1
- Main effects + 2-way interactions
- Polynomial terms with the same or separate βs for arms 0 and 1
- More complicated models

Non-parametric regressions

- Matching
- Matching with imputation (mean or parametric regression)
- gam, trees, gbm etc



Y1 and Y0 are the vectors of observed outcomes for treated and control patients, respectively. These vectors are usually of different length. In fact, treated and control patients are probably best visualized as possibly coming from two distinctly different populations. For any number of reasons, outcomes Y1 and Y0 may be very different.

X1 and X0 are observed matrices of patient characteristics; both have the same number of columns in the same order. But X1 and X0 may also represent very different patient characteristics.

Beta1 and Beta0 are vectors of unknown regression coefficients corresponding to the columns of X1 and X0, respectively. There is absolutely no reason why these two vectors should be the same!

OVER-SIMPLIFICATION?

Model with only a "Main Effect" for Treatment = 2α (scalar)

12

$$E\begin{bmatrix} y_1 \mid X_1 \\ y_0 \mid X_0 \end{bmatrix} = f\left\{\begin{pmatrix} +1 & X_1 \\ -1 & X_0 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right\}$$

…IFF one assumes $\beta = \beta_1 = \beta_0$ (vectors.)

Implicit Mean Notation: The model contains an intercept term, so the above notation assumes that "overall means" have been subtracted from each full column of y and X …i.e the columns end up summing to ZERO. NOTE that the LS counter-factual-difference for every patient is thus the same, namely 2 times alpha-hat.

DIFFERENTIAL PATIENT RESPONSE…

"one degree-of-freedom" for interaction of treatment with x:

$$E(y \mid t, x) = \mu + t\alpha + x\beta + (t \times x)\gamma$$

Least Squares Counterfactual Difference

for t = 1 minus that for t = –1 is $2(\alpha + x\gamma)$.

13

For models with an intercept and with treatment choice coded as t = +1 or -1, the LScfd estimate would then be 2 times [alpha-hat plus (gamma-hat times X)].

NOTE: Many more than just one-degree-of-freedom for "interaction" effects may be badly needed because, although "t" is a discreet variable with only 2 levels, the X-variable may be CONTINUOUS.

For models with several interactions between treatment choice and patient x-characteristics, the LScfd would be the sum of several such terms.

Blank filler slide.

CEM
coarsened exact matching

- Make Treatment Comparisons only within "Clusters" of Patients who are Well-Matched on their X-characteristics
- Luckily, the X-values in the competition dataset have already been COARSENED in the sense that there are "only" 39,788 unique X-vectors observed among the 249,958 patients. Thus, **on overall average**, about 6 patients are observed with each of these distinct X-vectors.
  - Demonstration using SAS proc MEANS.
  - Demonstration using R-code …with work-around when your system doesn't have enough memory to aggregate( ) 8 vars.

See code on next four pages.

```
/* Program: nestanova.sas, perform nested anova within patient subgroups via proc means. */
/* Author: Bob Obenchain */

libname mdd_sas "C:\HR Challenge\SAS";

title 'Sort observed data on trtm choice and 8 baseline X-covariates.';

data hrc_temp;
  set mdd_sas.mdd250k;
run;

proc sort data=hrc_temp;
  by trtm age gender pain hospcount ercount offcount psycpcnt wprevcost;
run;

title 'Calculate mean y-outcomes for each non-empty cell.';

proc means noprint data=hrc_temp;
  var wyrcost;
  by trtm age gender pain hospcount ercount offcount psycpcnt wprevcost;
  output out=mdd_sas.hrc_tab mean(wyrcost) = mwyrc;
run;

data hrc_trtm1;
  set mdd_sas.hrc_tab;
  mwyrc1 = mwyrc;
  pats1 = _freq_;
  if trtm eq 1 then output;
  drop trtm mwyrc;
run;

data hrc_trtm0;
  set mdd_sas.hrc_tab;
  mwyrc0 = mwyrc;
  pats0 = _freq_;
  if trtm eq 0 then output;
  drop trtm mwyrc;
run;

title 'Calculate Local Treatment Differences and total patients per Cell/Cluster.';

data hrc_tab10;
  merge hrc_trtm1 hrc_trtm0;
  by age gender pain hospcount ercount offcount psycpcnt wprevcost;
  ltd = mwyrc1 - mwyrc0;
  pats = pats1 + pats0;
  clus = _n_;
run;

proc sort data=hrc_temp;
  by age gender pain hospcount ercount offcount psycpcnt wprevcost;
run;

title 'Merge Cluster/Cell-Level with Patient-Level Results.';
```

```
data mdd_sas.hrc_join;
  merge hrc_temp hrc_tab10;
  by age gender pain hospcount ercount offcount psycpcnt wprevcost;
run;

proc sort data=mdd_sas.hrc_join;
  by seqno;
run;
```

```
## Aggregate on all eight X-variables...

setwd("H:/HR Challenge")
dat <- read.csv("mdd250k.csv")

## Form cells/clusters using trtm=1 patients...
mdd1 <- dat[dat$trtm==1, -c(1,3)]

## sort eight X-covariates so that cells consist of consecutive patients (rows)
hrc1 <- aggregate(mdd1$wyrcost, list(mdd1$age,mdd1$gender,mdd1$pain,mdd1$hospcount,
          mdd1$ercount,mdd1$offcount,mdd1$psycpcnt,mdd1$wprevcost), mean);
hrc1$x1 <- hrc1$x
hrc1$x <-NULL
rm(mdd1)

mdd0 <- dat[dat$trtm==0, -c(1,3)]
hrc0 <- aggregate(mdd0$wyrcost, list(mdd0$age,mdd0$gender,mdd0$pain,mdd0$hospcount,
          mdd0$ercount,mdd0$offcount,mdd0$psycpcnt,mdd0$wprevcost), mean);
hrc0$x0 <- hrc0$x
hrc0$x <-NULL
rm(mdd0)

## merge 0 and 1 and compute LTDs...

hrc01 <- merge(hrc0,hrc1)
hrc01$ltd <- hrc01$x1-hrc01$x0

## join back to dat

names(hrc01)[1:8] <- c("age","gender","pain","hospcount",
          "ercount","offcount","psycpcnt","wprevcost")
hrc_join <- merge(hrc01, dat, all=TRUE)

hrc_join <- hrc_join[order(hrc_join$seqno),]
```

```r
## R-code for computers without enough memory to aggregate on eight X-variables...

setwd("H:/HR Challenge/R")
dat <- read.csv("mdd250k.csv")

ind <- apply(dat[,4:11],1,function(x) paste(x, sep="", collapse=""))

mdd0 <- dat[dat$trtm==0, -c(1,3)]

ind0 <- ind[dat$trtm==0]

hrc0 <- tapply(mdd0$wyrcost, list(ind0), mean)

mdd1 <- dat[dat$trtm==1,-c(1,3)]

ind1 <- ind[dat$trtm==1]

hrc1 <- tapply(mdd1$wyrcost, list(ind1), mean)

## merge 0 and 1 and compute LTDs...

hrc0.df <- data.frame(cov.id=names(hrc0), x0=hrc0)
hrc1.df <- data.frame(cov.id=names(hrc1), x1=hrc1)

hrc01 <- merge(hrc0.df,hrc1.df)
hrc01$ltd <- hrc01$x1-hrc01$x0

## join back to dat

dat$cov.id <- unclass(ind)
hrc_join <- merge(hrc01, dat, all=TRUE)
hrc_join <- hrc_join[order(hrc_join$seqno),]
```

Imputation of Missing Values from Uninformative (Pure) Subgroups

- Impute missing values with the overall Mean LTD.
- Impute missing values with their predicted LTD from a (parametric) model that produces no missing values.
- Impute missing values with their predicted values from an LTD distribution that has been "shifted" so that it has the same overall mean LTD as that of the non-missing (non-parametric) LTD estimates.

root MSE Loss

$$\sqrt{\text{Mean}\left[\left(\hat{\Delta}_i - \Delta_i\right)^2\right]}$$

Δ_i = True LTD for i^{th} Patient

= E(*y* | *x*, trtm=1) – E(*y* | *x*, trtm=0)

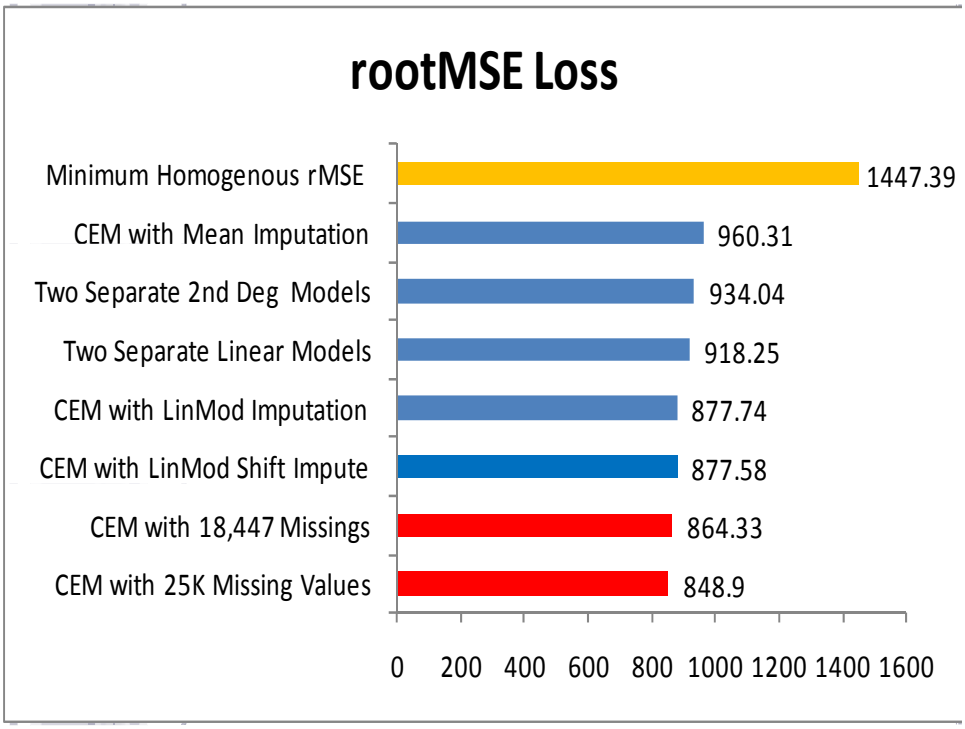
Absolute mean error

$$\text{abs}\left(\text{Mean}(\hat{\Delta}_i) - \text{Mean}(\Delta_i)\right)$$

$$\text{True Main} - \text{Effect}\, E(\Delta_i) \;=\; -\$\,650.42$$

Δ_i = True LTD for i^{th} Patient

= E(*y* | *x*, trtm=1) – E(*y* | *x*, trtm=0)



True Main-Effect = $ –650.42