Recent developments in the design and analysis of clinical trials.

Javier Cabrera

Dept of Statistics, Dept of Medicine, Rutgers University

Outline

- Part I: Analysis of Biased and poorly randomized clinical studies
- Part II Alternative study designs.
- Part III. Machine learning and deep learning methods for data augmentation and matching populations.

Part I

Analysis of Biased and poorly randomized clinical studies

- Introduction. Poorly randomized and biased clinical studies.
- Super Learners for nonlinear function estimation.
- Propensity scores estimation using Super Learners and other methods.
- TMLE procedure

Introduction

In this lectures we will introduce the following topics:

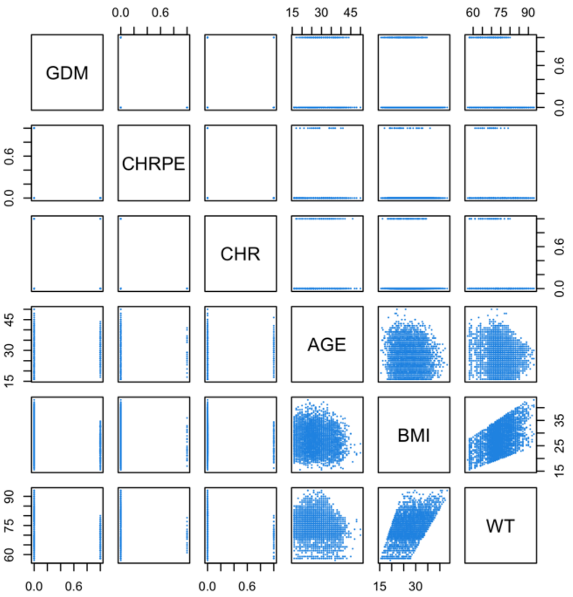
- Poorly randomized and biased clinical studies.
- SuperLearner. TMLE
- Indices: Propensity Scores, Differential Natural Hermite.
- Genetic Algorithm(GA)
- Augmenting Controls, Animal Studies, Rebalancing biased studies
- Datanuggets and Large Datasets/Big Data Matching
- Deep Learning, Auto Encoders/Decoders, Variational auto-encoder
- Expanding Applications to Large Datasets/Big Data.

A placenta abruption study

A study on preventing placenta abruption on women in NJ (PACER,2023) . Want to use real world controls from pregnancy database of New Jersey hospital births. We have 18 variables (25 features) in common between the clinical study and the real-world database. (Limitation)

We use synthetic data: *copydata* function on DNAMR

Scatter matrix of 6 of the 25 features in dataset.



Variables

MONTH
PE_MILD
PE_SEVERE
REGION
RACE
PRE_DM
OLIGO
MARITAL
MULTIPLE
HOSPBEDR
HOSPOWN
GES_HYP
GDM
CHRPE
CHR
AGE
BMI
WT

A placenta abruption study

- Angioedema disease of swelling of tissue all over the body. In some cases angioedema can be fatal.
- ACE inhibitors were discovered in Brazil from the poison of an Amazonian snake.
- Used for treating hypertension.
- Clinical study: ACE inhibitors may increase dead by angioedema patients?
- Outcome: Dead at 5 years follow up.
- Data: “acetrial.csv”.

Variables

The variables are:

Dead at 5 years:	0 censored, 1 dead
treat:	0 control, 1 ACE treated
age:	
drink, smoke,diab:	0 No, 1 Yes
gender:	0 Male, 1 Female
race:	0 White, 1 Black, 2 Other
sbp:	Systolic Blood Pressure.

Poorly randomized and biased clinical studies.

- Poorly randomized and biased clinical studies.

- Classical analysis: Linear models

  A: Treatment

  W: Confounders => model Misspecified

  No interactions

- Model estimation and inference (A-priori)

  Try many models and report the best: Bias and overfitting

  Treatment effect may pickup model missing interactions

Causal Inference: Estiment = Average Treatment Effect (ATE)
(Compared to Estimator, Estimate)

Statistics => Data Analysis => Data Mining => AI/ML

AI/ML + Stats => Good predictive models => SuperLearner

Super Learners for nonlinear function estimation

SuperLearner algorithm:

- Combines ML methods: GLM :: LASSO/ENET :: Random Forest :: GAM :: SVM :: NNET ::…
- Uses cross-validation creates an optimal weighted average of those models, aka an “ensemble”, using the test data performance. Possible to choose fold : 5, 10,20, leave-one out CV

Option: **SuperLearner.CV.control(V = 5L)**

- It is asymptotically the best possible prediction algorithm that has been tested.
- It is computationally intensive

```
library(SuperLearner)
sl_libs <- c('SL.glmnet', 'SL.gam', 'SL.earth','SL.nnet','SL.glm')
Q <- SuperLearner(Y=d[,1,X=d[,-1],family=binomial(),SL.library=sl_libs,
                                        cvControl=SuperLearner.CV.control(V = 5L) )
PE=  predict(Q)$pred
```

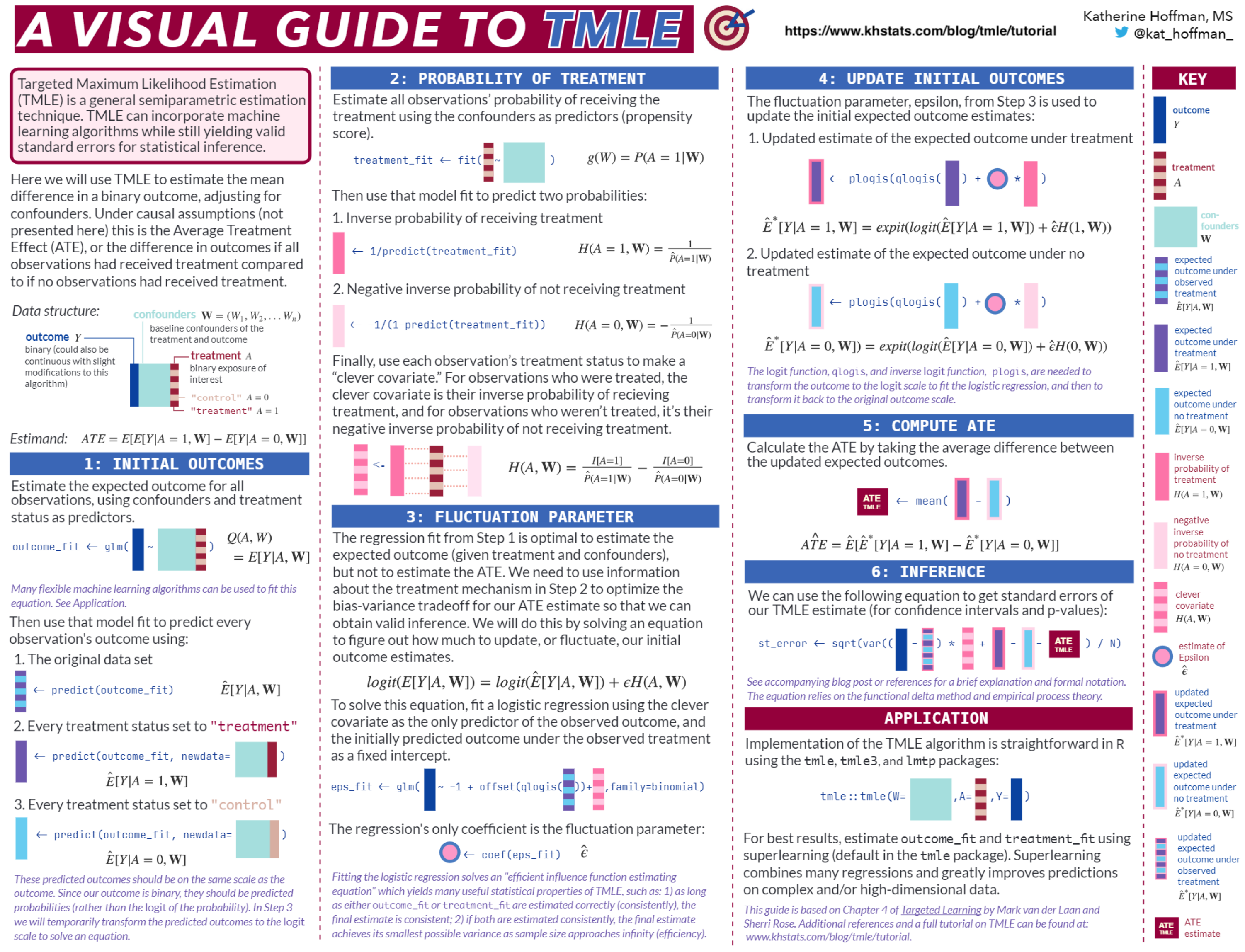
Propensity scores estimation using Super Learners

Propensity Score Functions

- Treatment assignment variable: A { 1 Treatment , 0 Control }
- Confounders: $W = \{W_1,\ldots, W_{p-1}\}$
- PE model: PE = P(A=1| W) = E(A|W).
- This was estimated using logistic model, or ML methods, RF, SVM, gam or other
- Use SuperLearner To estimate PE, choose a set of methods that you prefer. Generally use Lasso :: Random Forest :: GAM :: Mars :: NNET

TMLE

https://www.khstats.com/blog/tmle/tutorial



A VISUAL GUIDE TO TMLE

https://www.khstats.com/blog/tmle/tutorial

Katherine Hoffman, MS
@kat_hoffman_

Targeted Maximum Likelihood Estimation (TMLE) is a general semiparametric estimation technique. TMLE can incorporate machine learning algorithms while still yielding valid standard errors for statistical inference.

Here we will use TMLE to estimate the mean difference in a binary outcome, adjusting for confounders. Under causal assumptions (not presented here) this is the Average Treatment Effect (ATE), or the difference in outcomes if all observations had received treatment compared to if no observations had received treatment.

Data structure:

confounders $\mathbf{W} = (W_1, W_2, \dots W_n)$
baseline confounders of the treatment and outcome

outcome Y
binary (could also be continuous with slight modifications to this algorithm)

treatment A
binary exposure of interest
"control" $A = 0$
"treatment" $A = 1$

Estimand: $ATE = E[E[Y|A=1, \mathbf{W}] - E[Y|A=0, \mathbf{W}]]$

1: INITIAL OUTCOMES

Estimate the expected outcome for all observations, using confounders and treatment status as predictors.

```
outcome_fit ← glm(  ~  )
```

$Q(A, W) = E[Y|A, \mathbf{W}]$

Many flexible machine learning algorithms can be used to fit this equation. See Application.

Then use that model fit to predict every observation's outcome using:

1. The original data set

```
← predict(outcome_fit)
```

$\hat{E}[Y|A, \mathbf{W}]$

2. Every treatment status set to "treatment"

```
← predict(outcome_fit, newdata=  )
```

$\hat{E}[Y|A=1, \mathbf{W}]$

3. Every treatment status set to "control"

```
← predict(outcome_fit, newdata=  )
```

$\hat{E}[Y|A=0, \mathbf{W}]$

These predicted outcomes should be on the same scale as the outcome. Since our outcome is binary, they should be predicted probabilities (rather than the logit of the probability). In Step 3 we will temporarily transform the predicted outcomes to the logit scale to solve an equation.

2: PROBABILITY OF TREATMENT

Estimate all observations' probability of receiving the treatment using the confounders as predictors (propensity score).

```
treatment_fit ← fit(  ~  )
```

$g(W) = P(A=1|\mathbf{W})$

Then use that model fit to predict two probabilities:

1. Inverse probability of receiving treatment

```
← 1/predict(treatment_fit)
```

$H(A=1, \mathbf{W}) = \frac{1}{\hat{P}(A=1|\mathbf{W})}$

2. Negative inverse probability of not receiving treatment

```
← -1/(1-predict(treatment_fit))
```

$H(A=0, \mathbf{W}) = -\frac{1}{\hat{P}(A=0|\mathbf{W})}$

Finally, use each observation's treatment status to make a "clever covariate." For observations who were treated, the clever covariate is their inverse probability of recieving treatment, and for observations who weren't treated, it's their negative inverse probability of not receiving treatment.

```
<-
```

$H(A, \mathbf{W}) = \frac{I[A=1]}{\hat{P}(A=1|\mathbf{W})} - \frac{I[A=0]}{\hat{P}(A=0|\mathbf{W})}$

3: FLUCTUATION PARAMETER

The regression fit from Step 1 is optimal to estimate the expected outcome (given treatment and confounders), but not to estimate the ATE. We need to use information about the treatment mechanism in Step 2 to optimize the bias-variance tradeoff for our ATE estimate so that we can obtain valid inference. We will do this by solving an equation to figure out how much to update, or fluctuate, our initial outcome estimates.

$$logit(E[Y|A, \mathbf{W}]) = logit(\hat{E}[Y|A, \mathbf{W}]) + \epsilon H(A, \mathbf{W})$$

To solve this equation, fit a logistic regression using the clever covariate as the only predictor of the observed outcome, and the initially predicted outcome under the observed treatment as a fixed intercept.

```
eps_fit ← glm(  ~ -1 + offset(qlogis(  ))+  ,family=binomial)
```

The regression's only coefficient is the fluctuation parameter:

```
← coef(eps_fit)
```

$\hat{\epsilon}$

Fitting the logistic regression solves an "efficient influence function estimating equation" which yields many useful statistical properties of TMLE, such as: 1) as long as either `outcome_fit` or `treatment_fit` are estimated correctly (consistently), the final estimate is consistent; 2) if both are estimated consistently, the final estimate achieves its smallest possible variance as sample size approaches infinity (efficiency).

4: UPDATE INITIAL OUTCOMES

The fluctuation parameter, epsilon, from Step 3 is used to update the initial expected outcome estimates:

1. Updated estimate of the expected outcome under treatment

```
← plogis(qlogis(  ) +  *  )
```

$\hat{E}^*[Y|A=1, \mathbf{W}] = expit(logit(\hat{E}[Y|A=1, \mathbf{W}]) + \hat{\epsilon}H(1, \mathbf{W}))$

2. Updated estimate of the expected outcome under no treatment

```
← plogis(qlogis(  ) +  *  )
```

$\hat{E}^*[Y|A=0, \mathbf{W}]) = expit(logit(\hat{E}[Y|A=0, \mathbf{W}]) + \hat{\epsilon}H(0, \mathbf{W}))$

The logit function, `qlogis`, and inverse logit function, `plogis`, are needed to transform the outcome to the logit scale to fit the logistic regression, and then to transform it back to the original outcome scale.

5: COMPUTE ATE

Calculate the ATE by taking the average difference between the updated expected outcomes.

```
ATE TMLE ← mean(  -  )
```

$\hat{ATE} = \hat{E}[\hat{E}^*[Y|A=1, \mathbf{W}] - \hat{E}^*[Y|A=0, \mathbf{W}]]$

6: INFERENCE

We can use the following equation to get standard errors of our TMLE estimate (for confidence intervals and p-values):

```
st_error ← sqrt(var((  -  ) *  +  -  - ATE TMLE ) / N)
```

See accompanying blog post or references for a brief explanation and formal notation. The equation relies on the functional delta method and empirical process theory.

APPLICATION

Implementation of the TMLE algorithm is straightforward in R using the `tmle`, `tmle3`, and `lmtp` packages:

```
tmle::tmle(W=  ,A=  ,Y=  )
```

For best results, estimate `outcome_fit` and `treatment_fit` using superlearning (default in the `tmle` package). Superlearning combines many regressions and greatly improves predictions on complex and/or high-dimensional data.

This guide is based on Chapter 4 of Targeted Learning by Mark van der Laan and Sherri Rose. Additional references and a full tutorial on TMLE can be found at: www.khstats.com/blog/tmle/tutorial.

KEY

outcome Y

treatment A

con- founders $\mathbf{W}$

expected outcome under observed treatment $\hat{E}[Y|A, \mathbf{W}]$

expected outcome under treatment $\hat{E}[Y|A=1, \mathbf{W}]$

expected outcome under no treatment $\hat{E}[Y|A=0, \mathbf{W}]$

inverse probability of treatment $H(A=1, \mathbf{W})$

negative inverse probability of no treatment $H(A=0, \mathbf{W})$

clever covariate $H(A, \mathbf{W})$

estimate of Epsilon $\hat{\epsilon}$

updated expected outcome under treatment $\hat{E}^*[Y|A=1, \mathbf{W}]$

updated expected outcome under no treatment $\hat{E}^*[Y|A=0, \mathbf{W}]$

updated expected outcome under observed treatment $\hat{E}^*[Y|A, \mathbf{W}]$

ATE TMLE — ATE estimate

TMLE

Step 1: Estimate the Outcome

$$Q(A, \mathbf{W}) = \mathrm{E}[Y|A, \mathbf{W}]$$

Q_A: all observations

Q_1: If every observation received the treatment.

$$\hat{\mathrm{E}}[Y|A=1, \mathbf{W}]$$

Q_0: If every observation received the control.

$$\hat{\mathrm{E}}[Y|A=0, \mathbf{W}]$$

Causal Inference: Estiment = Average Treatment Effect
(Compare to Estimator, Estimate)

$$A\hat{T}E_{G-comp} = \hat{\Psi}_{G-comp} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mathrm{E}}[Y|A=1, \mathbf{W}] - \hat{\mathrm{E}}[Y|A=0, \mathbf{W}])$$

TMLE

Step 2: Estimate the Probability of Treatment

$$g(\mathbf{W}) = \Pr(A = 1|\mathbf{W})$$

The inverse probability of receiving treatment.

$$H(1, \mathbf{W}) = \frac{1}{g(\mathbf{W})} = \frac{1}{\Pr(A = 1|\mathbf{W})}$$

The negative inverse probability of not receiving treatment.

$$H(0, \mathbf{W}) = -\frac{1}{1 - g(\mathbf{W})} = -\frac{1}{\Pr(A = 0|\mathbf{W})}$$

“Clever” Covariate:

$$H(A, \mathbf{W}) = \frac{\mathrm{I}(A = 1)}{\Pr(A = 1|\mathbf{W})} - \frac{\mathrm{I}(A = 0)}{\Pr(A = 0|\mathbf{W})}$$

TMLE

Steps 3- 4: Update the Initial Estimates of the Expected Outcome

$$\hat{\mathrm{E}}^*[Y|A,\mathbf{W}] = expit(logit(\hat{\mathrm{E}}[Y|A,\mathbf{W}]) + \hat{\epsilon}H(A,\mathbf{W}))$$

$$\hat{\mathrm{E}}^*[Y|A=1,\mathbf{W}] = expit(logit(\hat{\mathrm{E}}[Y|A=1,\mathbf{W}]) + \hat{\epsilon}H(1,A))$$

$$\hat{\mathrm{E}}^*[Y|A=0,\mathbf{W}] = expit(logit(\hat{\mathrm{E}}[Y|A=0,\mathbf{W}]) + \hat{\epsilon}H(0,W))$$

Step 5: Compute the Statistical Estimand of Interest

$$A\hat{T}E_{TMLE} = \hat{\Psi}_{TMLE} = \frac{1}{N}\sum_{i=1}^{N}[\hat{E}^*[Y|A=1,\mathbf{W}] - \hat{E}^*[Y|A=0,\mathbf{W}]]$$

TMLE

Step 6: Calculate the Standard Errors for Confidence Intervals and P-values

$$\hat{IF} = (Y - \hat{E}^*[Y|A,\mathbf{W}])H(A,\mathbf{W}) + \hat{E}^*[Y|A=1,\mathbf{W}] - \hat{E}^*[Y|A=0,\mathbf{W}] - A\hat{T}E$$

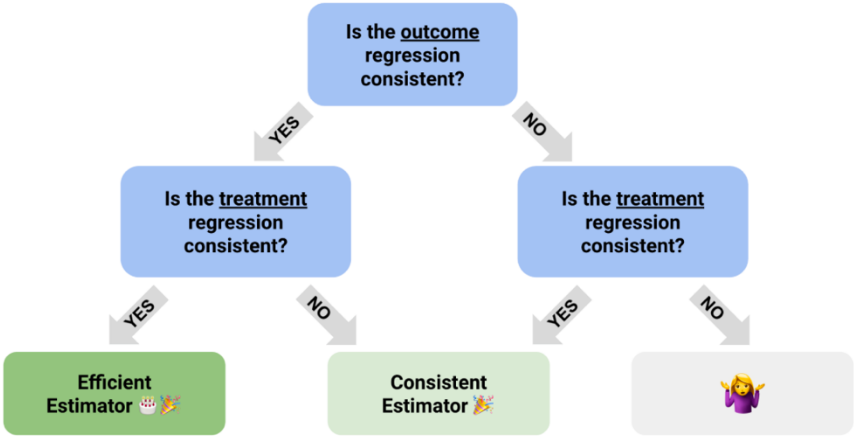
$$\hat{SE} = \sqrt{\frac{var(\hat{IF})}{N}}$$

Properties of TMLE

It allows the use of machine learning algorithms while still yielding asymptotic properties for inference.

TMLE is a **doubly robust** estimator: If both models for Y or for the probability of treatment are consistent then TMLE is consistent

If both regressions are consistent, the **final estimate will reach the smallest possible variance at a rate of root n**


Is the outcome regression consistent?
YES
NO
Is the treatment regression consistent?
Is the treatment regression consistent?
YES
NO
YES
NO
Efficient Estimator 🎂🎉
Consistent Estimator 🎉
🤷‍♀️

khstats.com/blog/tmle/tutorial

TMLE

Practice:

Problem 1. Practice with the file “TMLE.Rmd”

Problem 2. Use the acetrial dataset in the file “acetrial.csv”
response = death
treatment = treat
Minimum covariates = Gender, Smoke, BMI and SBP

- Estimate the propensity scores function using SuperLearner
- Fit TLME model