Convergence for Individualizing TReatment Using Statistical approaches (CITRUS)

Max Welz[1,2] Kevin ten Haaf[2] Andreas Alfons[1]

[1] Erasmus University Rotterdam, Dept. of Econometrics

[2] Erasmus Medical Center, Dept. of Public Health

October 21, 2021

Tinbergen Institute PhD Seminar

Outline

1. Introduction

2. Data Generation
    - Setup
    - Simulation Scenarios

3. Estimation: Considered Methods

4. Discussion

Background

CITRUS =
Convergence for Individualizing TReatment Using Statistical approaches

- Project of the Econometric Institute and Erasmus Medical Center;
- Part of the *Convergence Alliance* between EUR, EMC, TU Delft;
- Fully funded by an *Open Mind Call* grant (08–12/2021).

Introduction

Evidence-based medicine: the *"conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients"* (Sackett et al., 1996).

- ➔ Identify *heterogeneous treatment effects* (HTE)!
- ➔ Methods from modern causal inference?
- ➔ CITRUS: Which methods are most eligible for medical data?
- ➔ Simulate common characteristics of medical data.

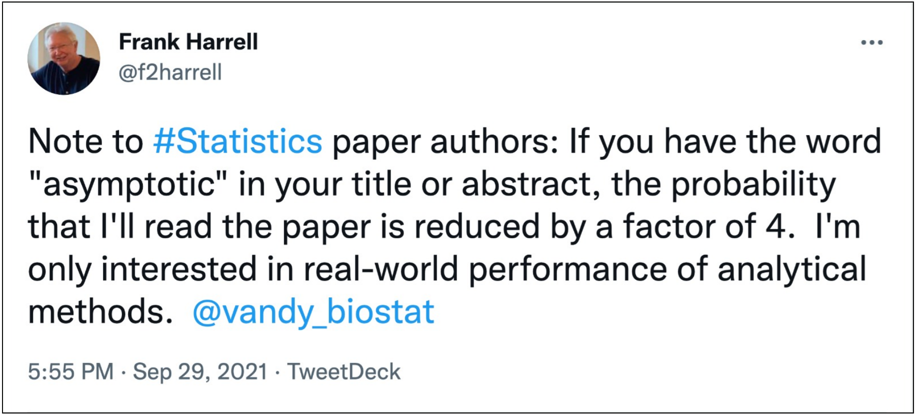
Introduction

In a perfect world, we'd have. . .

- *Many* observations: $n \to \infty$;
- Perfect information on relevant covariates (unconfoundedness);
- I.I.D. data from randomized experiments;
- Continuous data.

BUT: The world is not perfect (especially in medicine. . .)

Introduction



Frank Harrell
@f2harrell

Note to #Statistics paper authors: If you have the word "asymptotic" in your title or abstract, the probability that I'll read the paper is reduced by a factor of 4. I'm only interested in real-world performance of analytical methods. @vandy_biostat

5:55 PM · Sep 29, 2021 · TweetDeck

Introduction

In medical data, we typically have. . .

- Very finite sample sizes ($n \geq 500$ is rare);
- Noisy to incomplete representation of relevant covariates;
- Non-identically distributed samples;
- Improper randomization (sometimes. . .);
- Non-continuous data (e.g. categorical).

CITRUS research question:

- To what extent does this affect the performance of each HTE identification method?

Outline

1. Introduction

2. Data Generation
   - Setup
   - Simulation Scenarios

3. Estimation: Considered Methods

4. Discussion

Outline

1. Introduction
2. Data Generation
   - Setup
   - Simulation Scenarios
3. Estimation: Considered Methods
4. Discussion

Setup

Let there be n independent observations $(\mathbf{X}_i, Y_i, W_i, T_i)$.

- $\mathbf{X}_i$ is p-dimensional covariate vector;
- Y_i is outcome variable. Binary mortality indicator here: $Y_i = 1$ if i dies.
- W_i is binary treatment assignment variable: $W_i = 1$ if i is in treatment group. Assume RCT, so $\mathbb{P}[W_i = 1] = 0.5$.
- T_i is right-censored time at risk.

Setup

Rubin causal model: The DGP can generally be viewed as

$$\pi_i(W) = F_{logistic}\left(\theta(\mathbf{X}_i)W + \nu(\mathbf{X}_i) + \varepsilon_i\right) = \mathbb{P}[Y_i(W) = 1|W, \mathbf{X}_i],$$

$$Y_i(W) \sim \text{Binomial}\Big(\pi_i(W)\Big),$$

$$T_i(W) = g\left(\mathbf{X}_i, Y_i(W)\right).$$

- $Y_i(W)$, $T_i(W)$ are the potential outcome and time at risk, resp.;
- $\theta(\mathbf{X}_i)$ is the HTE function: causal effect of W on Y_i;
- $\nu(\mathbf{X}_i)$ is the nuisance function: raw effect of $\mathbf{X}_i$ on Y_i;
- ε_i is zero-mean noise term;
- $g : \mathbb{R}^p \times \{0, 1\} \to [0, \infty)$ is the survival function.

Setup

$$Y_i(W) \sim \text{Binomial}\Big(\pi_i(W)\Big), \qquad T_i(W) = g\big(\mathbf{X}_i, Y_i(W)\big).$$

We only ever observe one of the potential outcomes/times at risk:

$$(Y_i, T_i) = \begin{cases} \big(Y_i(1), T_i(1)\big) & \text{if } W_i = 1, \\ \big(Y_i(0), T_i(0)\big) & \text{if } W_i = 0. \end{cases}$$

Using $(\mathbf{X}_i, Y_i, W_i, T_i)$, estimate the HTE

$$\tau_i = \pi_i(1) - \pi_i(0),$$

that is, treatment-induced reduction of mortality probability.

Quantities of interest

- HTE parameters $\theta_i = \theta(\mathbf{X}_i)$ and 95% coverage thereof;
- HTE τ_i and 95% coverage thereof;
- Within-group average treatment effect of group $\mathcal{G} \subseteq \{1, \ldots, n\}$:

$$ATE_{\mathcal{G}} = |\mathcal{G}|^{-1} \sum_{i \in \mathcal{G}} \Big(\pi_i(1) - \pi_i(0) \Big);$$

- Within-group average *relative* treatment effect of $\mathcal{G}$:

$$ARTE_{\mathcal{G}} = |\mathcal{G}|^{-1} \sum_{i \in \mathcal{G}} \pi_i(1) \Big/ \pi_i(0).$$

Outline

1 Introduction

2 Data Generation
- Setup
- Simulation Scenarios

3 Estimation: Considered Methods

4 Discussion

Overview: Simulation Scenarios

We use simulation to emulate a medical DGP. We model...

- Various sample sizes $n \in \{100,\ 250,\ 500,\ 1,000,\ 10,000\}$;
- Categorical representation of relevant covariates;
- Undersampling of relevant subgroups;
- Various degrees of nonlinearity and sparsity;
- Anomalous individuals.

Scenario 1: Categorical Representation

- Let variable X^* be a major driver of treatment effect heterogeneity.
- We don't observe X^*, but just a categorical version thereof, X.
- Motivation: Cancer stage.

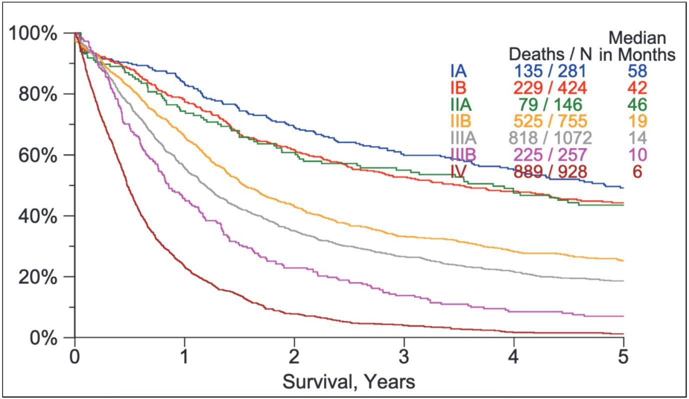


Figure: 5-year-survival probability by lung cancer stage; taken from Figure 6A in Groome et al. (2007)

Scenario 1: Categorical Representation

- Let variable X^* be a major driver of treatment effect heterogeneity.
- We don't observe X^*, but a categorical version thereof, X.
- The important variable *Stage* in cancer data is categorical!

How to simulate this?

- Generate continuous X^* measuring cancer severity;
- Group X^* into groups based on its quantiles;
- The group membership X (= Stage) is observed.

Procedure generalizes to non-cancer applications.

➔ Does categorical representation affect model's performance?

Scenario 2: Undersampling of Subgroups

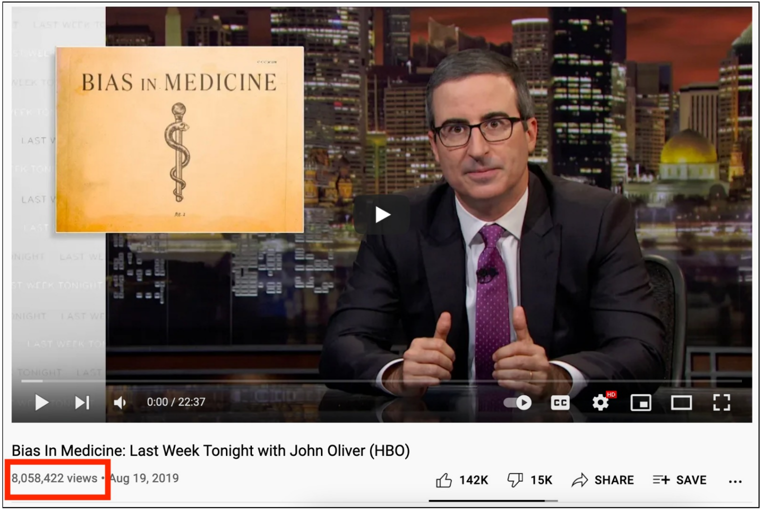
Motivation: Females and non-Whites are often underrepresented in trials.

- 2010 US Census: 72% of Americans are White and 13% African-American.
- Case study: In the largest clinical trial on lung cancer screening:
  - 90% of participants are White and 4.5% African-American.
  - 59% are male, 41% female.
  - Source: The National Lung Screening Trial Research Team (2011).
  - Possible heterogeneity along ethnicity (Blom et al., 2020).

➔ Does undersampling affect model's ability to capture heterogeneity?

Scenario 2: Undersampling of Subgroups

Luckily, this issue has recently started to attract public attention:



Bias In Medicine: Last Week Tonight with John Oliver (HBO)

8,058,422 views • Aug 19, 2019

Scenario 3: Nonlinearity

Recall the potential mortality probabilities:

$$\pi_i(W) = \mathbb{P}[Y_i(W) = 1 | W_i, \mathbf{X}_i] = F_{logistic}\left(\theta(\mathbf{X}_i)W + \nu(\mathbf{X}_i) + \varepsilon_i\right).$$

The HTE and nuisance functions ($\theta(\cdot)$ and $\nu(\cdot)$) may be nonlinear.

We consider various degrees of nonlinearity:

- E.g. quadratic, exponential, logarithmic, a mix thereof;
- Nonlinear interactions of variables;
- Baseline is linearity.

➔ Does degree of nonlinearity affect model's performance?

Scenario 4: Sparsity

The HTE function $\theta(\mathbf{X}_i)$ may effectively only depend on a subset of $\mathbf{X}_i$.
➔ Not all variables affect treatment effectiveness.

Easiest example: Assume linearity in parameters, i.e.

$$\theta(\mathbf{X}_i) = \mathbf{X}_i^\top \boldsymbol{\beta} = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p},$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is fixed and sparse.

➔ Only variables with nonzero coefficient matter for HTE.
➔ Can the model identify these variables?

Scenario 5: Anomalous Individuals

- There may be individuals with extreme characteristics.
- For example: extreme smokers ($\geq$ 100 cigarettes/day!) or extremely obese individuals.
- Often, such individuals are at extreme mortality risk, regardless of treatment assignment status.

→ Can a few extreme individuals affect the model's fit?

Outline

1 Introduction

2 Data Generation
- Setup
- Simulation Scenarios

3 Estimation: Considered Methods

4 Discussion

Overview: Methods for Estimating HTE

Methods from medicine:

- Rate ratios;
- Predictive modeling: risk & effect models.

Methods from stats/metrics:

- Double Machine Learning (Chernozhukov et al., 2018);
- Generic Machine Learning (Chernozhukov et al., 2020);
- Causal random forests (Athey et al., 2019).

There exist many more; e.g. survival-based models.

You will notice that medical methods are different to causal inference.

Rate Ratios

1. Partition sample in groups $\mathcal{G}_0, \mathcal{G}_1$ you want to compare.
2. The cumulative times-at-risk of each group are

$$N_0 = \sum_{i \in \mathcal{G}_0} T_i \quad \text{and} \quad N_1 = \sum_{i \in \mathcal{G}_1} T_i.$$

3. Define fatality counting variables

$$P_0 = \#\{i \in \mathcal{G}_0 : Y_i = 1\} \quad \text{and} \quad P_1 = \#\{i \in \mathcal{G}_1 : Y_i = 1\}.$$

4. Assume

$$P_0 \sim Poisson(N_0 \lambda_0) \quad \text{and} \quad P_1 \sim Poisson(N_1 \lambda_1)$$

for fixed, but unknown $\lambda_0, \lambda_1 > 0$.

Rate Ratios

5. We are interested in inference on the *rate ratio*

$$\xi = \lambda_0 / \lambda_1 .$$

6. Test $H_0 : \xi = 1$ against $H_1 : \xi \neq 1$ by UMP test (e.g. Lehmann and Romano, 1986).
7. If H_0 is rejected, there is evidence for systematic mortality differences between the two groups: means that treatment effect is different between them.

Rate Ratios

- Rate ratios are also called “one-variable-at-a-time” analyses.
- Heavily criticized by recent literature (Kent et al., 2020): e.g. low power, multiplicity.
- Nevertheless, still common method for HTE identification in medicine.
- But: Literature admits that better methods are required (e.g. Kent et al., 2020).

Predictive Risk Models

Risk models (Kent et al., 2020) are a two stage approach to identify HTE.

1. Stage 1: Fit logistic regression model (w/o treatment variable!)

$$\log\left(\frac{\mathbb{P}[Y_i = 1|\mathbf{X}_i]}{1-\mathbb{P}[Y_i = 1|\mathbf{X}_i]}\right) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}$$

   and calculate $\widehat{\eta}_i = \widehat{\beta}_0 + \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}$.

2. Stage 2: Fit the logistic regression model

$$\log\left(\frac{\mathbb{P}[Y_i = 1|\mathbf{X}_i, W_i, \widehat{\eta}_i]}{1-\mathbb{P}[Y_i = 1|\mathbf{X}_i, W_i, \widehat{\eta}_i]}\right) = \gamma_0 + \gamma_1 W_i + \gamma_2 \widehat{\eta}_i + \gamma_3 \widehat{\eta}_i W_i.$$

3. Calculate predicted risk

$$risk_i(W) = \widehat{\mathbb{P}}[Y_i = 1|\mathbf{X}_i, W, \widehat{\eta}_i].$$

Predictive Risk Models

$$\log\left(\frac{\mathbb{P}[Y_i = 1|\mathbf{X}_i, W_i, \widehat{\eta}_i]}{1 - \mathbb{P}[Y_i = 1|\mathbf{X}_i, W_i, \widehat{\eta}_i]}\right) = \gamma_0 + \gamma_1 W_i + \gamma_2 \widehat{\eta}_i + \gamma_3 \widehat{\eta}_i W_i.$$

- Idea behind two stages: Separate the explanatory power for Y_i into a part that is due to $\mathbf{X}_i$ and a part due to W_i.
- If treatment is effective, then $H_0 : \gamma_1 = 0$ should be rejected.
- If there is heterogeneity, then $H_0 : \gamma_3 = 0$ should be rejected.
- Predicted risk $risk_i(W) = \widehat{\mathbb{P}}[Y_i = 1|\mathbf{X}_i, W, \widehat{\eta}_i]$ can be used for HTE identification (more on this later).

Predictive Effect Models

Effect models (Kent et al., 2020) rely on variable selection to identify heterogeneity.

1. Specify set $\mathcal{I} \subseteq \{1, \ldots, p\}$ of covariates to interact W with.
2. Consider the logistic regression model

$$\log\left(\frac{\mathbb{P}[Y_i = 1|\mathbf{X}_i, W_i]}{1 - \mathbb{P}[Y_i = 1|\mathbf{X}_i, W_i]}\right) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta} + \gamma_0 W_i + \sum_{j \in \mathcal{I}} \gamma_j W_i X_{i,j}.$$

3. Fit the model using a regularization penalty on coefficient size (e.g. elastic net; Zou and Hastie, 2005).
4. Calculate predicted risk

$$risk_i(W) = \widehat{\mathbb{P}}[Y_i = 1|\mathbf{X}_i, W].$$

Predictive Effect Models

$$\log\left(\frac{\mathbb{P}[Y_i = 1|\mathbf{X}_i, W_i]}{1-\mathbb{P}[Y_i = 1|\mathbf{X}_i, W_i]}\right) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta} + \gamma_0 W_i + \sum_{j\in\mathcal{I}} \gamma_j W_i X_{i,j}.$$

- If treatment is effective, then W_i should be selected;
- If there is heterogeneity along the variables in $\mathcal{I}$, these variables should be selected.
- We recommend to use a regularization penalty akin to Bien et al. (2013) to account for special hierarchical structure of interaction effects.

Obtaining HTE estimates

Given: Predicted risk $risk_i(W)$ in risk or effect model.

Goal: Estimate HTE $\tau_i = \pi_i(1) - \pi_i(0)$ to identify heterogeneity.

- Define the *reversed* treatment assignment

$$W_i^{rev} = \begin{cases} 1 & \text{if } W_i = 0; \\ 0 & \text{if } W_i = 1. \end{cases}$$

- Estimate τ_i via

$$\widehat{\tau}_i = \begin{cases} risk_i(W_i) - risk_i(W_i^{rev}) & \text{if } W_i = 1, \\ risk_i(W_i^{rev}) - risk_i(W_i) & \text{if } W_i = 0. \end{cases}$$

- Thereupon, one can obtain estimates of $ATE_{\mathcal{G}}, ARTE_{\mathcal{G}}$.

Outline

1. Introduction

2. Data Generation
   - Setup
   - Simulation Scenarios

3. Estimation: Considered Methods

4. Discussion

Outlook

In CITRUS, we point out problems. We do not (yet) propose methodological solutions.

➔ Potential for many valuable novel contributions! For example,

- Derive valid confidence intervals for risk model coefficients (2SLS literature?)
- Derive probability of selecting all correct variables in effect model (build on Bien et al. (2013)?)
- Derive valid confidence intervals for effect model coefficients (build on Dezeure et al. (2015) and Van de Geer (2016)?)
- Valid subgroup-level inference (build on Guo and He (2021)?)
- Robustify regression (also in survival models!) (build on Lecué and Lerasle (2020)?)

Outlook

We might want to start working on some of these extensions in the future (EUR and EMC plan to intensify their collaboration).

Let us know if you are interested!

(It's fine if you are in AMS)

Thank you for your attention! Any questions?

Slides: https://mwelz.github.io/publications/ti2021.pdf

References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized Random Forests. *Annals of Statistics*, 47(2):1148–1178.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A Lasso for Hierarchical Interactions. *Annals of Statistics*, 41(3):1111 – 1141.

Blom, E., ten Haaf, K., Arenberg, D., and de Koning, H. (2020). Disparities in Receiving Guideline-Concordant Treatment for Lung Cancer in the United States. *Annals of the American Thoracic Society*, 17(2):186–194.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2020). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv preprint: arXiv1712.04802*.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-Dimensional Inference: Confidence Intervals, *p*-Values and R-Software hdi. *Statistical Science*, 30(4):533–558.

Groome, P. A., Bolejack, V., Crowley, J. J., Kennedy, C., Krasnik, M., Sobin, L. H., and Goldstraw, P. (2007). The IASLC Lung Cancer Staging Project: Validation of the Proposals for Revision of the T, N, and M Descriptors and Consequent Stage Groupings in the Forthcoming (Seventh) Edition of the TNM Classification of Malignant Tumours. *Journal of Thoracic Oncology*, 2(8):694–705.

References (cont.)

Guo, X. and He, X. (2021). Inference on selected subgroups in clinical trials. *Journal of the American Statistical Association*, 116(535):1498–1506.

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., et al. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine*, 172(1):35–45.

Lecué, G. and Lerasle, M. (2020). Robust Machine Learning by Median-of-Means: Theory and Practice. *Annals of Statistics*, 48(2):906–931.

Lehmann, E. L. and Romano, J. P. (1986). *Testing Statistical Hypotheses*. Wadsworth & Brooks/Cole, Pacific Grove, California, 2nd edition.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence-Based Medicine: What it is and What it isn't. *The British Medical Journal*, 312(7023):71–72.

The National Lung Screening Trial Research Team (2011). Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine*, 365(5):395–409.

Van de Geer, S. (2016). *Estimation and Testing Under Sparsity*. Springer.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.