Estimating treatment effects in online experiments

Media in Context and the 2015 General Election: How Traditional and Social Media Shape Elections and Governing (ES/M010775/1)

University of Exeter
A brief intro to the potential outcomes framework

- Typical case: binary treatment:
  - (relatively) easy to generalize to more complex treatment regimes (see references)
- \( D_i = 1 \) if subject \( i \) receives treatment, 0 otherwise
- \( Y_i(1) \) is the outcome for a subject who received the treatment
- \( Y_i(0) \) is the outcome if \( i \) was assigned to control
- Treatment effect for \( i \) is \( \beta_{D_i} : Y_i(1) - Y_i(0) \)
- Obvious problem:
A brief intro to the potential outcomes framework

- Typical case: binary treatment:
  - (relatively) easy to generalize to more complex treatment regimes (see references)

- $D_i = 1$ if subject $i$ receives treatment, 0 otherwise

- $Y_i(1)$ is the outcome for a subject who received the treatment

- $Y_i(0)$ is the outcome if $i$ was assigned to control

- Treatment effect for $i$ is $\beta_{D_i} : Y_i(1) - Y_i(0)$

- Obvious problem: we only get to observe $Y_i(1)$ OR $Y_i(0)$
  - fundamental problem of causal inference

- “Solution”: under random assignment to treatment conditions, we take averages: we estimate the ATE
A brief intro to the potential outcomes framework (cont.)

- \( ATE = E(\beta_{Di}) = E(E(Y_i(1) - Y_i(0))) = E(Y_i(1)) - E(Y_i(0)) \)

- That is, simply take the average of \( Y \) for those treated/not treated, and take the difference
  - Again, random assignment to treatment is important here: on average, no difference between treated and control beyond treatment condition \( \rightarrow \) differences in outcome are explained by \( D \)

- This is what we typically do when we compute difference in means (e.g., via t-tests) or differences in proportions across treatment conditions, or when we estimate parametric regression models like

\[
Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i
\]
From ATE to CATE

- In practice, equation 1 assumes that the treatment effect is constant across subjects.

- This is a very restrictive and potentially unrealistic assumption in some settings.

- For instance, in the media-related survey experiment conducted in our project, it is reasonable to assume that several factors may intervene between treatment and response (e.g., Druckman and Chong, 2007).
  - e.g., media consumption habits, partisan affiliation, interest in politics, etc.

- A more flexible approach is to allow treatment effects to vary with relevant background (pre-treatment) characteristics.
From ATE to CATE (cont.)

- This takes us from the estimation of ATE(s) to CATE(s)
  - CATE: conditional average treatment effects
  - i.e., average treatment effects among subgroups defined by baseline covariates
- The usual way of doing is to simply interact these relevant covariates with $D$:

\[
Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i X_i
\]

\[
= \beta_0 + \beta_2 X_i + (\beta_1 + \beta_3 D_i X_i) D_i
\] (2)

- Example from our research: “script ATE-CATE.R”
Average and conditional treatment effects
From ATE to CATE (cont.)

- Problems with the standard “interactive” approach?

  - Difficult to interpret & understand beyond 2-way interactions
    - Many interactions also lower statistical power and lead to imprecise estimates
  
  - So we typically use a few relevant mediators that need to be selected a priori
    - Bypassing alternative explanations
  
  - Model mis-specification and sensitivity to functional forms (especially when the mediator is continuous)
  
  - Assumes a deterministic relationship between mediator and treatment

More recent/sophisticated strategies:

1. Mixture models/latent class regression analysis
2. Non-parametric approaches: Bayesian Trees, LASSO regressions, Machine Learning, Ensemble Methods
From ATE to CATE (cont.)

- Problems with the standard “interactive” approach?
  - Difficult to interpret & understand beyond 2-way interactions
    - many interactions also lower statistical power and lead to imprecise estimates
  - So we typically use a few relevant mediators that need to be selected a priori
    - bypassing alternative explanations
  - Model mis-specification and sensitivity to functional forms (especially when the mediator is continuous)
  - Assumes a deterministic relationship between mediator and treatment

- More recent/sophisticated strategies:
  1. Mixture models/latent class regression analysis
  2. Non-parametric approaches: Bayesian Trees, LASSO regressions, Machine Learning, Ensemble Methods
Latent Class Models of Treatment Effect Heterogeneity

- Different sub-populations of experimental subjects respond differently to treatment
- The number of heterogeneous groups is not known a priori, but selected based on statistical criteria (e.g., AIC, BIC, DIC)
- Accommodates several mediating factors
- Accounts for unobserved heterogeneity in treatment-covariate interaction
- Basic idea:

\[ Y_i = \beta_j \text{Treatment}_i + \alpha_j X_i, \quad i = 1, \ldots, N; j = 1, \ldots, J \]  

- Each subject is classified into 1 of \( J \) “classes”
  - Within each class, treatment effects are simply given by \( \beta_j \)
  - Variations in \( \beta_j \) across classes capture differences in responsiveness to treatment across sub-populations
How do we assign subjects into classes?

\[ Pr(Class_i = j) = \frac{\exp(\gamma_j W_i)}{\sum_k \exp(\gamma_k W_i)} \]  

- \( W_i \) contains relevant moderating variables (potentially including some of the \( X_i \))

Example: Impact of reasons to back down from EU referendum promise on government evaluation

- Treatment: EU referendum was just a campaign promise to attract UKIP voters
  - Control: government will not renege on its promise

- Outcome: Approve or disapprove of government action

- Possible moderators: Identification with UKIP, political interest and knowledge, media consumption and trust, socio-demographic characteristics (e.g., age, education, income)
  \( \rightarrow \) too many for a full-interactive approach
So, we fit a mixture model

- does heterogeneity exist? (i.e., do we distinguish classes of experimental subjects?)
- how many classes?
- what is driving heterogeneity?

We use a Bayesian estimation approach - Markov chain Monte Carlo (MCMC) simulations

- no asymptotic approximations: suitable for typical experimental samples
- flexibility to explore posterior distribution of parameters

However, we could fit the same model using ML-based methods (e.g., EM algorithm)
Basic rationale behind estimation

- Basic estimation steps:
  1. Start by randomly assigning an individual to a “class”
  2. Regress $Class_i$ on $W_i$ to see which variables determine class membership
  3. Estimate the outcome model $Y_i = \beta_j Treatment_i$ separately for each class
  4. Repeat until convergence
     - check using standard Bayesian convergence diagnostics (e.g., Gelman-Rubin, Geweke, Heidel)

- Let’s try a very simple example: “script LCR.R”
Class-specific effects

Determinants of Class 2

Intercept
Prior Exposure
Political Knowledge
Media Use
Media Trust
Interest Politics
Partisan: Conservative
Partisan: Labour
Partisan: Libdem
Partisan: UKIP
Independents
University Education
Extension to multiple outcomes

- The finite mixture modeling approach to estimating CATE is also easy to extend to multiple outcome variables
  - and categorical outcomes
  - not so easy to accomplish using some of the other approaches we will see later today

- Example: experiment on media framing and attitudes towards the new government majority
  - treatment: media report on the “decisiveness” of the majority
  - control: business news piece
  - outcomes: several attitudes about governments’ ability to exert power and accountability (agree/disagree)
    - The government will be able to fulfill its campaign promises
    - It is important to command a majority in parliament to govern
    - The government has little effect on economic performance
    - The government’s ability to improve life in Britain depends on the support from other parties
We can fit an ordered probit mixture model:

\[
\prod_{i}^{N} \sum_{j}^{J} \pi_{i,j} \prod_{k=1}^{5} \prod_{m=1}^{M} p_{j,k}(Y_{i,k}=m) \tag{5}
\]

where

\[
p(y_{i,k,j} = m) = P(\tau_{m-1,k,j} - \beta_{k,j} T_i < \epsilon_{i,k} < \tau_{m,k,j} - \beta_{k,j} T_i) \tag{6}
\]

i.e.,

- the treatment effect $\beta$ varies across classes $j = 1, \ldots, J$ and outcomes $k = 1, \ldots, 5$

- and $\pi_{i,j} = Pr(Class_i = j) = \exp(\gamma_j W_i) / \left( \sum_k \exp(\gamma_k W_i) \right)$
So,

1. Subjects are classified into “classes” based on $W_i$ and the responses to $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,5}$

2. Within each class $j$, for each outcomes $k = 1, \ldots, 5$, the treatment effect is given by $\beta_{j,k}$

3. Heterogeneity in responsiveness to treatment can be gauged by comparing $\beta_{j,k}$ and $\beta_{j',k}$

Example: “script LCR - oprobit.R”
Alternative approaches: Bayesian trees

- Mixture modeling is a “semi-parametric” approach
- Main drawback: model mis-specification
- Fully non-parametric methods are less sensitive to choice of specific functional form
- On the other hand, typically require larger samples and can be sometimes difficult to interpret
- One example of a non-parametric method: BART
  - useful for high-dimensional data
  - less sensitive to specification of functional forms than parametric models
  - more robust to the choice of tuning parameters than other statistical learning techniques
  - existing off the shelf software (in R) minimizes the need for programming (and statistical) expertise
Basic idea behind BART

- Repeatedly split the sample into ever more homogeneous groups based on the values of each of the covariates. E.g.: is $X_i \geq X_0$?
  - Yes: Node 1; No: Node 2
  - Repeat this process for each variable until each unit of analysis is assigned to one terminal node
Basic idea behind BART (cont)

- Now assume that you fit a non-parametric model to the observations in each of these terminal nodes

\[ Y_i = f(X_i) + \epsilon_i \]  

(7)

where \( f(X_i) \) is an unknown - non-parametric- function of the covariate vector \( X_i \), and \( \epsilon \) is an error term

- Some trees contained treated units, other non-treated units: comparison of \( \hat{Y} \) provides an estimate of the treatment effect

- Furthermore, the units in the different terminal nodes (N1-N6 in the Figure) differ in the values of the covariates.

- So we can compare how the treatment effects varies with values of the covariate
  - e.g., compare the difference in \( \hat{Y} \) in N1-N2 against the difference in N3-N4 and against N5-N6
Basic idea behind BART (cont)

- Now, the Figure displayed only one possible tree.

- There are potentially infinite trees defined by different threshold values/decision rules $X_0$ for different combinations of covariates.

- So, essentially, BART repeats this procedure for $J$ trees, with $J$ typically in the hundreds.

\[
Y = \sum_{j=1}^{J} f(X, T_j) + \epsilon \quad (8)
\]

where $T_j$, $j = 1, \ldots, J$ is a given tree with a particular set of nodes and decision rules connecting the nodes.

- $T_j$ is treated as an additional parameter in the model, which is typically estimated by MCMC.
Basic idea behind BART (cont)

- Example: “script BART.R”

![CATE estimate for AGE graph](image)
BART models can also be used when the outcome variable is binary, rather than continuous. Simply replace equation 8 with

$$P(Y = 1|X) = \Phi(\sum_{j=1}^{J} f(X, T_j))$$  \hspace{1cm} (9)

Extensions to other outcomes - ordinal, count - is not so straightforward (mixture models are “better” in that regard)

Also, sample sizes need to be quite large - larger than the typical sample sizes in experimental politics
Alternative approaches: Support Vector Machine (Imai and Ratkovic)

- Imai and Ratkovic propose another non-parametric approach that can be used to estimate heterogeneous treatment effects.

- The method is a Squared Loss Support vector Machine with LASSO constraints over pre-treatment and causal heterogeneity parameters.
  - sounds Chinese, I know! But don’t worry too much
  - The bottom line is: the method simultaneously selects a subset of the relevant pre-treatment covariates that moderate the treatment, and estimate these interaction

- So, advantage over mixture models and BART: no need to include all possible moderators (mixture model) or define the relevant set of high-order interactions/criteria for trees (BART).
  - Imai and Ratkovic’s method “does it for you” in a single step
Alternative approaches: Support Vector Machine (cont.)

- Example: “script Findit.R”
Alternative approaches: Support Vector Machine (cont.)

- Drawback: no easy way to compute uncertainty measures (i.e., can only compare differences in point estimates)

- More generally, non-parametric approaches are rather difficult to extend to non-binary treatment regimes
  - we have typically 4 treatment conditions in our experiments
  - we could apply these tools for pairwise comparisons
  - however, no easy way out in the case of - say - ordered treatment regimes

- The mixture modeling approach does not face this problem
Some useful references to get you started


Some useful references to get you started (cont.)


- And some useful - and freely available - R packages
  - “flexmix: Flexible Mixture Modeling.” https://cran.r-project.org/web/packages/flexmix/index.html
  - “BayesTree: Bayesian Additive Regression Trees.” https://cran.r-project.org/web/packages/BayesTree/index.html.