Causal Inference for ATM

An Example of GDP Counterfactual Estimation

Akhil Shah

RAND Corporation

June, 2017

Example of ATM Question Requiring Counterfactuals

- Consider situations that required a GDP
- Can we statistically quantify resulting outcomes without GDP
- **How much airborne delay if GDP is not implemented?**
- Causal Inference methods can help estimate this counterfactual
  - Machine Learning enables flexible and scalable modeling

Other domains have applied causal inference

- “How does a job training program affect salaries?"
- “How does smoking cessation counseling affect mortality?"
- “Is an ISP violating Net Neutrality?"
- Commonality: Many important applications are not amenable to a Randomized Control Trial (RCT)

A statistical framework: Rubin potential outcomes

- $X, Y, T = \{\text{Predictors, Outcome, Treatment indicator}\}$
- Each unit has a potential outcome:
  - $Y_i(T_i = 0)$ and $Y_i(T_i = 1)$
  - But only one is observed: either $T_i = 1$ or $T_i = 0$!
- In a RCT: T_i would be randomized, so...
  - Average Treatment Effect (ATE): $E[Y(1) - Y(0)]$
  - Estimate ATE simply by: $1/N \sum Y_i(1) - Y_i(0)$
  - This works since random assignment balances covariates of treatment and control groups

The fundamental issue: counterfactuals must be estimated

- Many studies are Observational
  - units are not assigned randomly to treatment/control group
- Confounding: some predictors determine outcome and treatment assignment
  - Difference in covariates between treatment and control group can be statisticaly significant
- Assumption: Treatment assignment T indep. of potential outcomes Y given X
- Challenge: Covariate space $\dim(X)$ can be large
  - How do you match on a vector X_i to simulate a counteractual from the other group?
- Key result: $\pi_i = P(T_i | X_i)$ is a (scalar) **balancing score**

Simplified recipe for propensity score analysis

- Estimation of counterfactual:
  - Compute/Model π_i so that X_i are balanced like an RCT
  - Weight outcomes: $w_i = \frac{T_i}{\pi_i} + \frac{1-T_i}{1-\pi_i}$
  - Assess effect with outcome regression on treatment indicator just like RCT
- Required ingredients:
  - Propensity score models for $\pi_i = P(T_i|X_i)$
  - Balance assesment metrics: SMD or KS statistic
  - Outcome estimators:

$$\hat{\mu}_{IPW} = \frac{\sum_i \frac{T_i}{\pi_i} Y_i}{\sum_i \frac{T_i}{\pi_i}}$$

An intial application: Airborne Delay at JFK from GDP

- How would average hourly airbone delay change if a GDP was not applied at JFK?
- Unit of analysis (i): arrival hour at JFK (7102; Sep 2013-Aug 2014)
- Covariates (X): Hourly (forecasted) weather and traffic
- Treatment (T): Hour is treated (GDP) or in control (no GDP)
- Outcome (Y): Average hourly airborne delay
- Data snapshot - (mostly) publicly available in ASPM:

wind	vis	snow	TS	rain	fog	xwind	Arr	AD	qlength	T
10	10.0	0	0	0	0	9.4	31	4.3	18	1
8	10.0	0	0	0	0	6.1	27	7.7	1	0
12	0.5	0	0	1	1	2.1	21	4.1	0	0
7	10.0	0	0	0	0	6.1	22	-8.7	0	0
8	10.0	0	0	0	0	0.0	33	5.1	102	1

Data manipulation and alignment

- Analysis requires hourly TAF, ASPM (various modules), TFMI
- ASPM has hourly values for average airborne delay
- Queue length from ASPM is 15-min based
  - Arrival Demand - Effective Arrivals
- TAF is (nominally) generated for 0,6,12,18h
- TFMI is event based
- GDP initations (root advisories) can be modified
- Must follow the sub time-series to get actual start/stop times
- End-result is ‘status’ of each hour (GDP or No-GDP)

Propensity Score for balancing imbalanced groups

- Recall definition of PS:

$$\pi_i = P(T_i = 1|X_i)$$

- Used to balance treatment and control groups, which are imbalanced across covariates

covariate	E(Y1;t=1)	E(Y0;t=0)	p
qlength	53.6	2.3	0.00
arrivals	34.8	21.9	0.00
visibility	7.7	9.3	0.00
windspeed	15.2	11.4	0.00
ceiling	228.0	481.7	0.00
crosswind	8.6	7.0	0.00

Propensity Score modeling of binary treatments

- Parametric estimation of propensity score: Linear Logistic regression

$$\texttt{logit}(\hat{\pi}_i) = \frac{P(T_i = 1|X_i)}{1 - P(T_i = 1|X_i)} = F(X_i) = \beta X_i$$

- GBM (using **t** trees) is often better (more flexible and robust) than linear logistic regression

$$F(X_i) = \sum_{k}^{t} f_k(X_i)$$

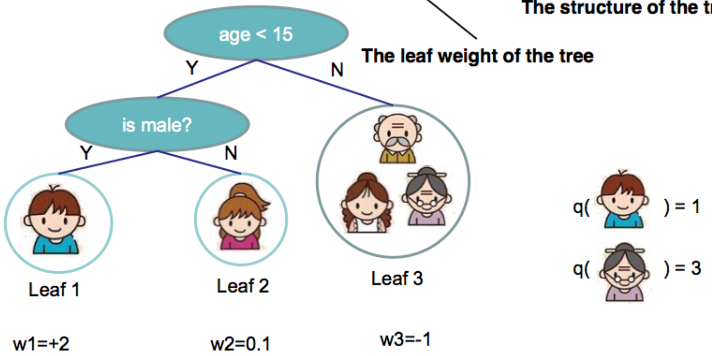
What is a GBM tree?

- A Tree is like a multi-dimensional step-function ~ decision tree

$$f_t(x) = w_{q(x)}, \quad w \in \mathbf{R}^T, q : \mathbf{R}^d \rightarrow \{1, 2, \cdots, T\}$$

The structure of the tree

The leaf weight of the tree



- graphic from xgboost: https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf

Fitting GBM: penalized loss and additive training

- GBM model based on objective: loss (Bernoulli) and penalty

$$\sum_{i}^{n} \ell(T_i, \hat{\pi}_i) + \sum_{k}^{t} \Omega(f_k)$$

- Bernoulli loss:

$$\ell_i = T_i \ln \hat{\pi}_i + (1 - T_i) \ln(1 - \hat{\pi}_i)$$

- Penalize lots of leaves and large weights:

$$\Omega_k = \gamma L_k + \lambda \sum_{j}^{L_k} w_j^2$$

- Solution: Additive Training (Boosting)
- The best GBM model for the PS is not the most accurate predictor of GDP, but one that achieves best covariate balance!

Assessing covariate imbalance *before* propensity weighting

covariate	E(Y1;t=1)	E(Y0;t=0)	p
qlength	53.6	2.3	0.00
arrivals	34.8	21.9	0.00
visibility	7.7	9.3	0.00
windspeed	15.2	11.4	0.00
ceiling	228.0	481.7	0.00
crosswind	8.6	7.0	0.00

Improved balance *after* propensity weighting:

covariate	E(Y1;t=1)	**E(Y1;t=0)**	p
qlength	53.6	44.5	0.09
arrivals	34.8	33.2	0.61
visibility	7.7	7.7	0.87
windspeed	15.2	14.3	0.72
ceiling	228.0	265.8	0.16
crosswind	8.6	8.6	0.73

Diagnositcs: propensity scores shows overlap for counterfactuals

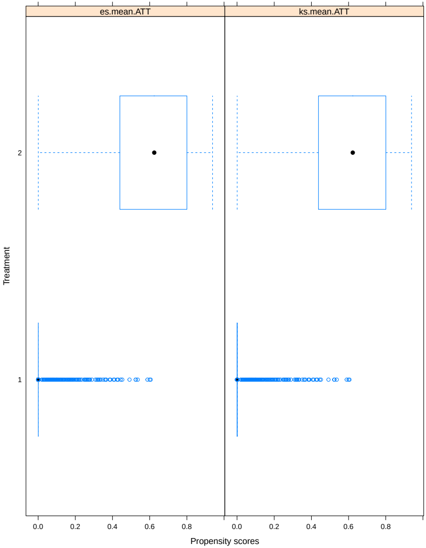


Figure 1

Outcome Analysis to estimate ATT

- Those weights which achieve best balance can now be used in weighted linear regression on treatment
- Average Treatment Effect for Treated (ATT)

$$= E(Y1|t = 1) - E(Y1|t = 0)$$

- For this analysis (unit of analysis = hour): **had a GDP not been applied to the hours at JFK when it in fact was, the hourly average airborne delay would increase on average by 1.4 minutes**
- This result was not statistically significant (std error of same magnitude)
- Let's consider alternative analysis options: unit of analysis; outcomes; binary vs continous treatment

Near term refinements of this simplified analysis

- Unit of analysis may interact: adjacent hours likely to be treated
- Hourly average airborne delay is based on ETE: schedule padding
- Alternative: use individual flight data (not public)
  - consider single origin: e.g. LAX to JFK to avoid estimating **nominal** ETE
  - outcome is now actual time enroute between treated/controls
- At individual flight level, treatment is not binary, but continuous (ground-delay)
  - requires Generalized Propensity score to estimate dose-response curve (dose: ground delay; response: airborne delay)
  - compare to simulated queueing models of airbone delay
- How to account for possible additional TFMI incurred by GDP delayed flights? (conjectered by Billimoria, 2016)

Some concluding remarks and future research options

- Casual inference/propensity score modeling is worth investigating further for ATM counterfactual estimation
- Propensity score analysis provides advantages over regression-based techniques:
  - dimensional reduction
  - grounded in rigorous statistical framework
  - robust again model mis-specification
  - avoids extrapolation
  - seperates covariate balancing from outcome analysis
- Our simplified (binary treatment) analysis focused on GDP;
  - could consider individual flights with continous treatments
  - account for pre-treatment covariates such as weather/traffic at nearby airports (NY metro)
- Consider other TFMI (e.g. Reroutes) in potential outcomes frameworks; requires examining relevant covariates (e.g. convective cloud top altitude)