Machine Learning for the Social Sciences

III. Machine Learning For Causal Effect Estimation

Andrew Peterson

Université de Genève

Table of contents

Review: Potential Outcomes and Selection Bias

Causal inference with observational data: Intuition

- Claim: UN Peacekeepers lead to more conflict deaths ("let them fight it out")
- Evidence: Pr(death|UNPKO) = large!
- Big data: if existed, lots of data says Pr(death| poor, UNPKO) = larger!, etc.

Observed & Potential Outcomes

Imagine UN peacekeepers were sent to countries 1 and 3 but not to country 2, and the number of subsequent conflict deaths was observed. What does it mean to calculate the difference-in-means of the observed outcomes? (Unobserved potential outcomes in grey)

Case	No UN	UNPKO	treatment effect
	$Y_i(0)$	$Y_i(1)$	τ_i
1	1,539	$Y_1(1)$	$Y_1(1)$ - 1,539
2	$Y_2(0)$	2,394	2,394 - $Y_2(0)$
3	239	$Y_3(1)$	$Y_3(1)$ - 239
Average observed outcomes			Diff-in-means
	14,186	3,083	$\hat{\bar{\tau}} = 1,505$

But obviously we cannot conclude from this (alone) that the presence of UN peacekeepers causes an additional 1,505 deaths on average.

Observed & Potential Outcomes

But obviously we cannot conclude from this (alone) that the presence of UN peacekeepers causes an additional 1,505 deaths on average. It may be that if we were able to observe all potential outcomes, it would look like this:

Case	No UN	UNPKO	treatment effect
	$Y_i(0)$	$Y_i(1)$	τ_i
1	1,539	587	-952
2	12,408	2,394	-10,014
3	239	102	-137
Average (all) outcomes			Diff-in-means estim
	889	2,394	$\bar{\tau} = -11,103$

Selection bias, regression style

- Consider again the (na ive) regression approach
- Perhaps for our hospital effect we estimate:

$$deaths_i = \beta_0 UNPKO + \beta_1 X_i + \epsilon_i$$

- If the X_i omits some confounder (e.g. conflict intensity), then we would expect $\beta_0 > 0$
- Adding this such a control we would see that $\beta_0 < 0$ (unless actors are more uncertain with UNPKO, etc...)

Propensity score weighting

- Intuition: healthy people shouldn't bias up the condition of those who do not visit hospitals
- And: if you took a person of a given sickness and then assigned them to be hospitalized or not, we would see the true effect
- So weight observations based on whether they are a 'good test': those who would never be hospitalized, or would be hospitalized no matter what, do not provide much information

Selection bias

- That is, it's at least possible that there is an unobserved confounder, such as: troubled
  - conflict $\rightarrow$ UNPKO
  - troubled conflict $\rightarrow$ deaths
- Similarly with US campaign contributions...anywhere people select into treatment...

Selection bias: Approaches

- Many possible solutions, depending on data, etc.
- If confounders *observed*, simple (naive) regression: control for symptoms, age, etc.
- but high-dimensional: which variables? functional form? interactions? ... *ad hoc*, researcher-driven
- Instead: machine learning driven model selection

Potential outcomes framework

- Under what conditions could we just compare the observed outcomes for groups that do and don't receive the treatment?
- Under what conditions could we do so while controling for variables in a regression setup?

Average Treatment Effect

$$ATE = \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]$$

Selection Bias

To see how selection bias enters in, add and subtract the same value ($\mathbb{E}[Y_i(0)|D_i = 1]$) in our formula for the ATE:

$$\begin{aligned}
&= \mathbb{E}[Y_i(1)|D_i = 1] && - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] && + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] && + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]
\end{aligned}$$

- In blue, the ATE among treated individuals
- In red, selection bias

Selection Bias: Random Assignment

$$= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \qquad + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]$$

- ATE among treated individuals, selection bias
- If random assignment, $Y_i(1), Y_i(0), X \perp D_i$

  $\implies$ selection bias $= 0$

  $\implies \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0)] = \text{ATE}$

  (Assumptions: SUTVA & exclusion restriction, i.e. outcome affected by treatment assignment only through the treatment)

Potential outcomes and causality (from before)

- random draw from population, non-interference (SUTVA)
- unconfoundedness / CIA
- linearity in parameters (X_i)
- constant effects (treatment effect same for all individuals)

Selection Bias: Under CIA

Roughly speaking, if we had a good model for predicting treatment assignment from X, we could eliminate the selection bias problem.

- If, further, the observed units are representative, then we can estimate the ATE
- If, alternatively, we had the true outcome model (for predicting y from X, then we could control for any potential confounders (but this contradicts the premise of this investigation...)

ML for Average Treatment Effects

Overview: Estimating ATE, high-dimensional

- semi-parametric methods (1990s-)
- targeted maximum likelihood (early 2000s)
- ML (Lasso)
- ML for double selection
- Double ML

(We will focus on the last 3 ML approaches).

High-dimensional sparse methods & LASSO

Considering many covariates, possibly many more than obs ($p >> n$)
Rate of convergence in estimating the (non-parametric) mean regression function:

- oracle rate $\sqrt{s/n}$
- Lasso $\sqrt{s/n}\sqrt{\log p}$
- Use Lasso to chose vars, OLS to estimate parameters: similar convergence and smaller bias (Belloni & Chernozhukov 2013)

Double selection: Belloni, et al (2013)

As mentioned above, want to be good at (1) predicting treatment (propensity model) and (2) predicting the outcome (outcome model)

- Use Lasso to select variables for outcome model
- Use Lasso to select variables for propensity model
- OLS regression of Y on all selected variables

(with tuning based on cross-validation)

Frisch Waugh Lovell / Partial Regression

- Say we want some particular coefficient for the kth regressor.
  e.g. $k = 2$ for the model $y = c + x_1 + x_2 + x_3 + \epsilon$
- If $\tilde{y}$ is from residualizing y on xs other than k, and similarly $\tilde{x_k}$
- We can estimate the regression $\tilde{y} = \tilde{x}_k + \nu$ or calculate directly:

$$\hat{\beta}_2 = \frac{\text{cov}(\tilde{y}, \tilde{x}_k)}{\text{var}\, \tilde{x}_k}$$

- Or in R as as follows. (pf: Hansen 2018, §3.16)

Frisch Waugh Lovell (R code)

1. regress y on x1 to get residuals e_1

```
lm1 <- lm(y~x1+x3, data=df)
e_1 <- lm1$residuals
```

2. regress x2 on x1 to get resids $\tilde{x}_2$:

```
lm2 <- lm(x2~x1+x3, data=df)
xtilda_2 <- lm2$residuals
```

3. regress the residualized y on the residualized x_2:

```
lm3 <- lm(e_1~xtilda_2)

print(lm3$coefficients['xtilda_2'])
```

Neyman Orthogonality Condition

› Roughly: Evaluating the nuisance function at a particular point is valid even under 'local' mistakes

$$D = \partial_\nu \mathbb{E}\phi(W, \theta_0, \nu)|_{\nu=\nu_0} = 0$$

Double machine learning: Chernozhukov, et al. (2017)

- Intuition: Combine ideas of double robustness, ML, and orthogonality
- In particular: naive ML prediction is biased.
- Semi-parametric model: for treatment D, controls Z,
  $Y = D\theta_0 + g_0(Z) + U$
- $\mathbb{E}[U|Z, D] = 0$

Double machine learning: Chernozhukov, et al. (2017)

1. (like double selection) Use ML to predict Y and D from Z o get $\mathbb{E}[\hat{Y}|Z]$, $\mathbb{E}[\hat{D}|Z]$

2. FWL to cleanly estimate the effect (and de-bias):
Residualize Y and D of the control variables:
$\tilde{Y} = Y - \mathbb{E}[\hat{Y}|Z]$;
$\tilde{D} = D - \mathbb{E}[\hat{D}|Z]$

3. Regression of $\tilde{Y}$ on $\tilde{D}$ to get treatment effect θ_0

Questions?