Improving Precision in Survey Experiments — A New Method to Use Ordinal Variables for Blocking

Simon Heuberger
6 March 2019

American University
Washington, DC

Outline

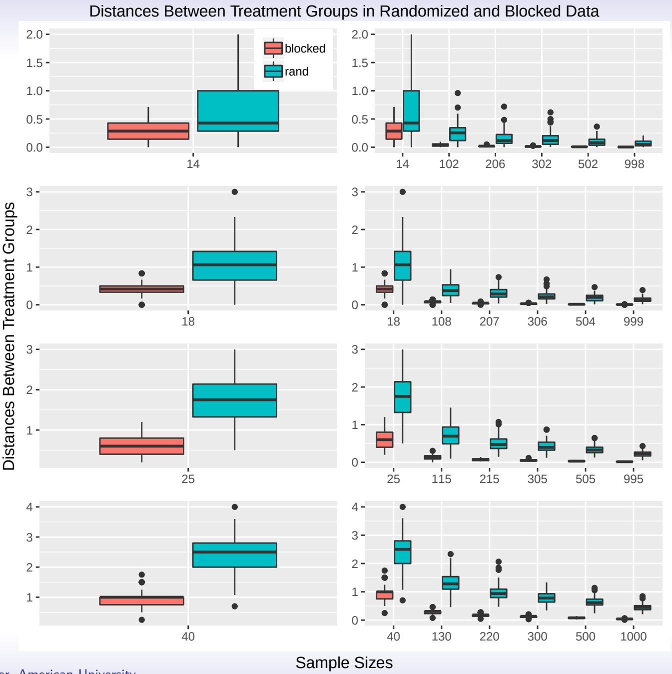
- Basics of survey experiments
- Basics of blocking
- Basics of ordinal variables
- Method:
    - Machine learning through Ordered Probit Model (OPM)
    - Blocking
    - Ordinal variable: Education
- Setup of eventual R package

Survey Experiments

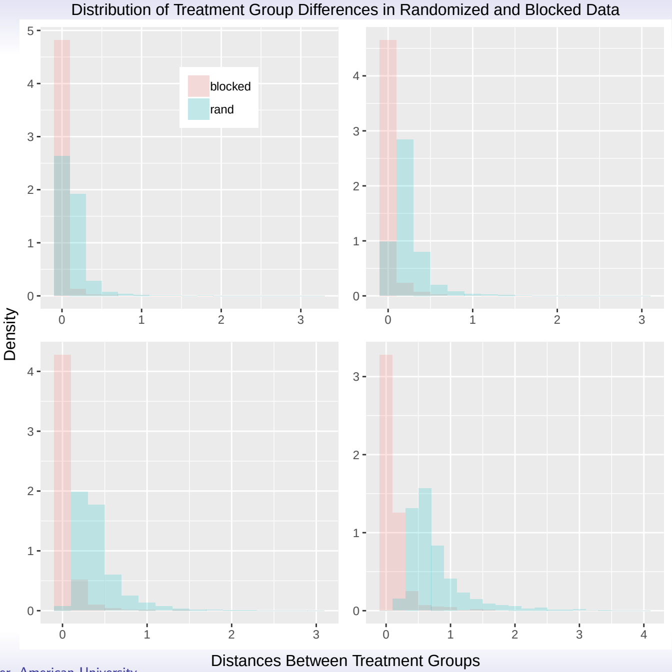
- Collect demographics and contain questions with some type of treatment
- Created to uncover effect of treatment on public opinion and/or behavior
- Example: "A disease breaks out in the US."
  - Treatment group 1: "With Program A, 200 out of 600 people will live."
  - Treatment group 2: "With Program A, 400 out of 600 people will die."
  - "Do you support or oppose program A?"
- Results usually estimated with OLS regression
- DV is respondents' answer to treatment question; EVs are demographics
- Crucial for (internally) valid results: Balance
  - Randomization, i.e. flip a coin
  - More advanced: Blocking

Blocking

- Using covariates to create pre-assignment groups ("blocks") of similar units
  - Unit similarity estimated with Mahalanobis or Euclidian distance
- Randomization takes place within blocks
- Ensures an equal proportion of treated and untreated units
- Can improve precision of causal estimates over randomization
- Most crucially: Guarantees balance
  - Especially relevant with small samples
  - Especially relevant with high number of treatment groups



Distances Between Treatment Groups in Randomized and Blocked Data
blocked
rand
Distances Between Treatment Groups
14
102
206
302
502
998
18
108
207
306
504
999
25
115
215
305
505
995
40
130
220
300
500
1000
Sample Sizes



Distribution of Treatment Group Differences in Randomized and Blocked Data
blocked
rand
Density
Distances Between Treatment Groups

Categorical Variables

- Data that can be divided into groups
- Nominal: No intrinsic ordering
  - Gender, Race, Occupation, Party ID
  - Often used as binary variables
- Interval: Ordered, evenly spaced
  - Income
  - Often made numeric
- Ordinal: Ordered, not evenly spaced
  - Education
  - Often made numeric

Ordinal Variables

- Example for an important ordinal variable: Education
- Education levels: “Elementary school”, “Some high school”, “HS grad”, “Some college”, “College grad”
- Previous plots simulated by assigning interval: 1, 2, 3, 4, 5
  - Arbitrary
  - Not based on any data-driven reasons
- Much better: Estimate underlying latent continuous structure
  - Machine Learning: Ordered Probit Model

Ordered Probit Model (OPM) Theory

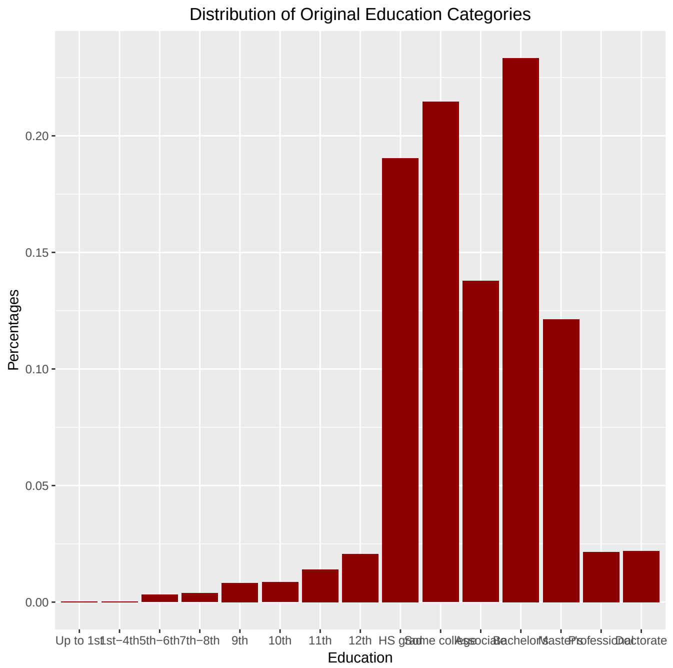
- $\exists$ $\boldsymbol{X}$, an $n \times k$ matrix of explanatory variables
- $\boldsymbol{Y}$ observed on ordered categories: $\boldsymbol{Y}_i \in [1, \ldots, k]$, for $i = 1, \ldots n$
- $\boldsymbol{Y}$ assumed to be produced by unobserved latent continuous variable $\boldsymbol{Y}^*$
- $\boldsymbol{Y}^*$ is continuous from $-\infty$ to ∞
- $\boldsymbol{Y}^* = \boldsymbol{X}_i\beta + e_i, e_i \sim N(0, 1), \forall i = 1, ..., N$
- Linear model creates numerical thresholds:
  $\boldsymbol{Y}_i^* : \ \delta_0 \underset{c=1}{\longleftrightarrow} \delta_1 \underset{c=2}{\longleftrightarrow} \delta_2 \underset{c=3}{\longleftrightarrow} \delta_3 \ldots \delta_{C-1} \underset{c=C}{\longleftrightarrow} \delta_C$
- Thresholds partition variable into regions corresponding to ordinal categories
- Linear model bins observations between thresholds according to the EVs

OPM in Plain English

- Find an externally and internally valid data set
- Train model on data with education as DV and meaningful covariates as EVs
- This model:
  - Estimates cutoff thresholds between categories
  - Bins data cases according to linear predictors
  - Binned cases determine which variable categories make sense

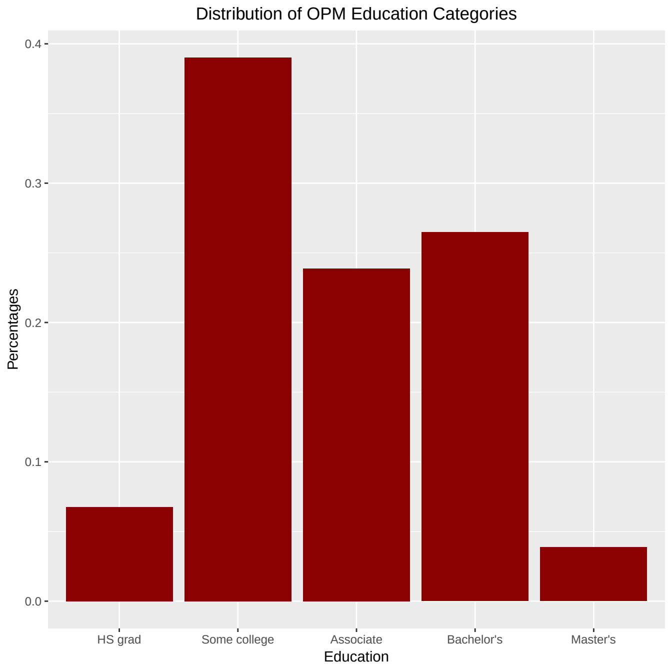
OPM for Education

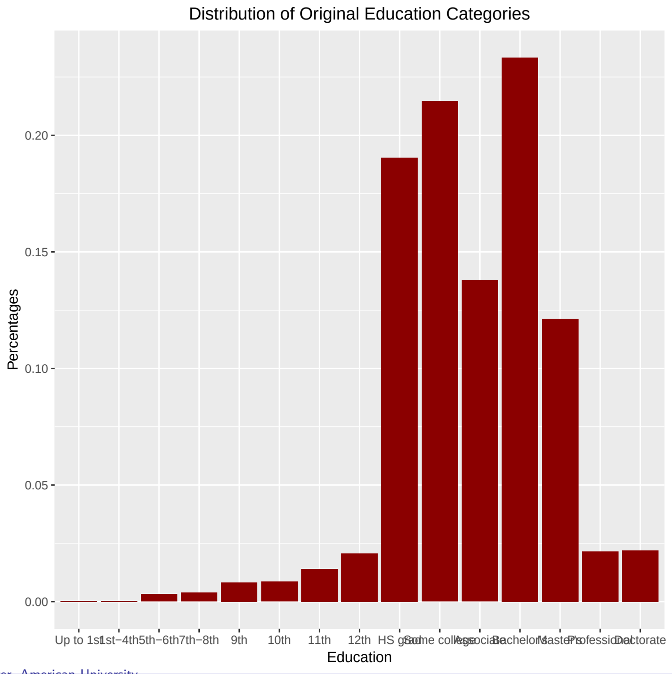
- Data: ANES 2016
- Model: $\textit{Education} \sim \textit{Gender} + \textit{Race} + \textit{Age} + \textit{Income} + \textit{Occupation} + \textit{PartyID}$


Distribution of Original Education Categories
Percentages
0.20
0.15
0.10
0.05
0.00
9th
10th
11th
12th
Doctorate
Education

Ordered Probit Threshold Estimates

Thresholds	Coefficients	SE	t.values
Up to 1st|1st-4th	−7.869	1.024	−7.681
1st-4th|5th-6th	−7.146	0.717	−9.965
5th-6th|7th-8th	−5.379	0.326	−16.515
7th-8th|9th	−4.671	0.253	−18.472
9th|10th	−3.920	0.206	−19.070
10th|11th	−3.468	0.188	−18.489
11th|12th	−2.984	0.174	−17.100
12th|HS grad	−2.511	0.166	−15.116
HS grad|Some college	−0.710	0.154	−4.607
Some college|Associate	0.384	0.154	2.500
Associate|Bachelor's	1.045	0.154	6.766
Bachelor's|Master's	2.478	0.160	15.538
Master's|Professional	4.099	0.177	23.144
Professional|Doctorate	4.838	0.197	24.589


Distribution of OPM Education Categories
0.4
0.3
0.2
0.1
0.0
Percentages
HS grad
Some college
Associate
Bachelor's
Master's
Education


Distribution of Original Education Categories
Percentages
0.20
0.15
0.10
0.05
0.00
Education

Using OPM Results to Block on Education

- OPM results:
  - Are not arbitrary
  - Are data-based
  - Use ordinal information whilst respecting uneven spaces
- Assigning numerical values to the new categories is now justifiable
- Block on numerical values the same way as before
  - `blockTools` with Mahalanobis distance
- Difference between OPM and interval method:
  - OPM gives us categories that make sense given the data
  - Interval method without any modelling has no empirical justification
- OPM uses the ordinal information to create categories that fit the data
- Through OPM, we can block whilst fully utilizing the ordinal information
- This was not possible before

Eventual R Package

- Loads trained model
- Applies trained categories to education variable
- Blocks on trained categories (`blockTools`)
- In addition: Online application

Thank you!