Lecture 22:
Machine Learning and Causal Inference

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

May 2, 2022

Machine Learning vs. Traditional Statistics

A priori specification vs. "letting the data tell us" about

- specification of conditioning sets (which Xs to include and how to do so),
- sources of effect heterogeneity, or
- causal structure.

Machine Learning vs. Traditional Statistics

A priori specification vs. "letting the data tell us" about

- specification of conditioning sets (which Xs to include and how to do so),
- sources of effect heterogeneity, or
- causal structure.

Machine learning emphasizes regularization and predictive validity so that:

- Models *grow in complexity* but regularization creates friction in doing so,
- Models are assessed in their predictive validity, typically using *hold out samples* and cross-validation.

Machine Learning and Causal Inference

Illustrations:

- CIA with high-dimensional X.
- Effect heterogeneity and optimal treatments.

CIA with high-dimensional X

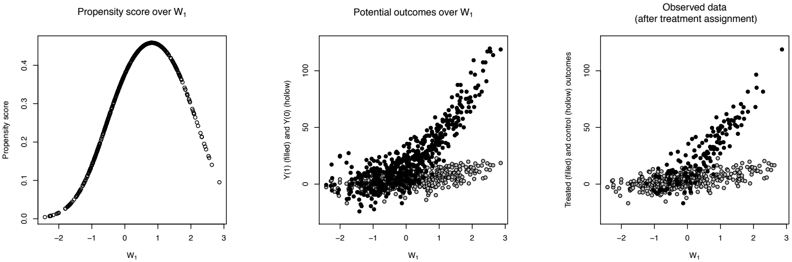
CIA with high-dimensional X

- Idea: what if we have *tons* of covariate data?
- Makes CIA more plausible.
- But implementation is challenging.
- What Xs to include? How to do it?

CIA with high-dimensional X

- Idea: what if we have *tons* of covariate data?
- Makes CIA more plausible.
- But implementation is challenging.
- What Xs to include? How to do it?
- Maybe machine learning can help?
  - By targeting the propensity score, we can let the machine learn an identifying covariate set and specification!
  - (In principle, could also target the potential outcome distributions, but one would have to do that separately for all outcomes of interest.)

Perils of Standard Practice, Promise of Machines



- Suppose we want to estimate the ATT.
- Suppose we have a set of covariates.
- But unbeknownst to us, treatment and outcome confounded by only *one* of them; the rest are just noise.
- And, it is confounded in an irregular way.
- How well do conventional methods do in these circumstances? How sensitive are they to increasing noise?

Perils of Standard Practice, Promise of Machines

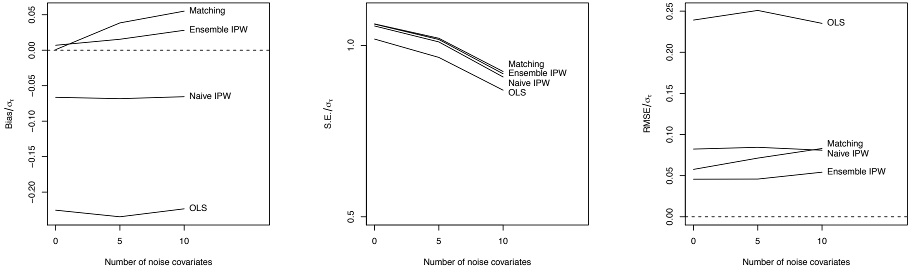


Figure 3: Simulation results. From left to right, the graphs show bias, standard error (S.E.), and root mean square error (RMSE) for the different estimators from 250 simulation runs as the number of noise covariates increases from 0 to 10. All results are standardized relative to the standard deviation of the true sample ATT across the simulation runs.

(Samii et al. 2017)

Regularization to the Rescue

- Regularization penalizes model complexity.
- Linear regression examples:
  - Recall OLS loss function: $\hat{\beta}_{OLS} = \min_{\beta} \frac{1}{N} \sum_{i=1}^{N} (Y_i - X_i' \beta)^2$
  - No constraints on $\beta \Rightarrow$ "overfit" with high dimensional X.

Regularization to the Rescue

- Regularization penalizes model complexity.
- Linear regression examples:
  - Recall OLS loss function: $\hat{\beta}_{OLS} = \min_\beta \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2$
  - No constraints on $\beta \Rightarrow$ "overfit" with high dimensional X.
  - Least absolute shrinkage selection operator (LASSO):

$$\begin{aligned}
\hat{\beta}_{LASSO} &= \min_\beta \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2 \text{ s.t. } \sum_{k=1}^{K}|\beta_k| \leq c \\
&= \min_\beta \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2 + \lambda \sum_{k=1}^{K}|\beta_k|.
\end{aligned}$$

Regularization to the Rescue

- Regularization penalizes model complexity.
- Linear regression examples:
  - Recall OLS loss function: $\hat{\beta}_{OLS} = \min_\beta \frac{1}{N}\sum_{i=1}^N (Y_i - X_i'\beta)^2$
  - No constraints on $\beta \Rightarrow$ "overfit" with high dimensional X.
  - Least absolute shrinkage selection operator (LASSO):

$$\begin{aligned}\hat{\beta}_{LASSO} &= \min_\beta \frac{1}{N}\sum_{i=1}^N (Y_i - X_i'\beta)^2 \text{ s.t. } \sum_{k=1}^K |\beta_k| \leq c \\ &= \min_\beta \frac{1}{N}\sum_{i=1}^N (Y_i - X_i'\beta)^2 + \lambda \sum_{k=1}^K |\beta_k|.\end{aligned}$$

  - Ridge (a.k.a. Tikhanov) regularization:

$$\hat{\beta}_{Ridge} = \min_\beta \frac{1}{N}\sum_{i=1}^N (Y_i - X_i'\beta)^2 + \lambda \sum_{k=1}^K \beta_k^2.$$

Regularization to the Rescue

- Regularization penalizes model complexity.
- Linear regression examples:
  - Recall OLS loss function: $\hat{\beta}_{OLS} = \min_\beta \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2$
  - No constraints on $\beta \Rightarrow$ "overfit" with high dimensional X.
  - Least absolute shrinkage selection operator (LASSO):

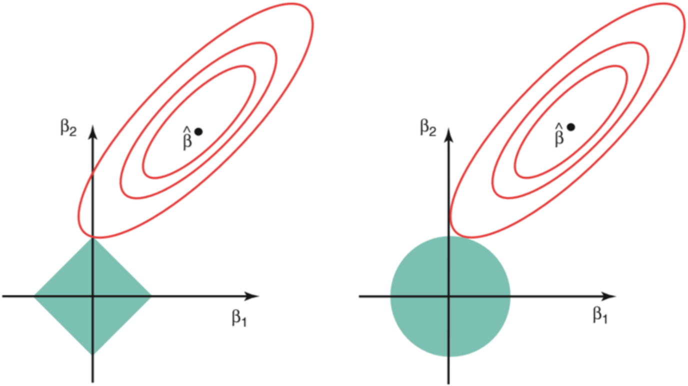
$$\begin{aligned}\hat{\beta}_{LASSO} &= \min_\beta \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2 \text{ s.t. } \sum_{k=1}^{K}|\beta_k| \le c \\ &= \min_\beta \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2 + \lambda \sum_{k=1}^{K}|\beta_k|.\end{aligned}$$

  - Ridge (a.k.a. Tikhanov) regularization:

$$\hat{\beta}_{Ridge} = \min_\beta \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2 + \lambda \sum_{k=1}^{K}\beta_k^2.$$

  - "Elastic net" combines LASSO and Ridge.
  - Identified even when $K > N$.

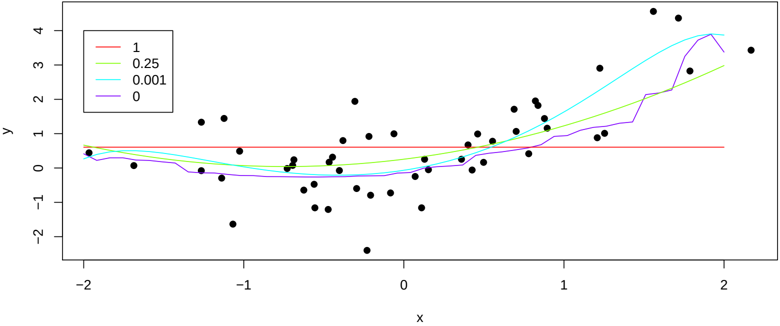
Regularization to the Rescue



(James et al. 2013, *An Intro. to Stat. Learning*)

Regularization to the Rescue

LASSO with 8th degree polynomial:



	lambda=1	lambda=0.25	lambda=0.001	lambda=0
x0	0.61	0.26	-0.07	-0.08
x1	0.00	0.58	0.71	0.93
x2	0.00	0.39	0.92	1.02
x3	0.00	0.00	0.09	-0.48
x4	0.00	0.00	0.00	-0.06
x5	0.00	0.00	0.01	0.33
x6	0.00	0.00	-0.01	-0.03
x7	0.00	0.00	-0.00	-0.05
x8	0.00	0.00	-0.00	0.00

Regularization to the Rescue

- Regularization penalities can be applied to other methods as well:
  - Trees/forests: penalize wrt tree depth.
  - Classifers: penalize wrt to curviness of classification frontier.
  - Etc.
- Regularization “tuning parameters” (e.g., λ) selected to minimize cross-validated error.

Example: Samii et al. 2017

For outcome Y_k, define the retrospective intervention effect (RIE) for A_j as,

$$\psi_j = \underbrace{\mathrm{E}\left[Y(\underline{a}_j, A_{-j})\right]}_{\text{counterfactual mean}} - \underbrace{\mathrm{E}\left[Y\right]}_{\text{observed mean}},$$

where A_{-j} refers to elements of A other than A_j. The RIE differs slightly from the average

We use this identification result to construct an inverse-propensity score weighted (IPW) estimator of the RIE:

$$\hat{\psi}_j^{IPW} = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{I(A_{ji} = \underline{a}_j)}{\hat{g}_j(\underline{a}_j|W_i, A_{-ji})}Y_i\right) - \bar{Y} \tag{2}$$

where N is the sample size and $\hat{g}_j(\underline{a}_j|W_i, A_{-ji})$ is a consistent estimator for $\Pr[A_j = \underline{a}_j|W_i, A_{-ji}]$. In

- Need to estimate the propensity score ($\hat{g}_j(.)$).
- Have 114 covariates reduced to 23 indices on war, political, economic, and social characteristics at demobilization; 9 demographic traits; and 47 municipality fixed effects.
- Conventional approaches would have a hard time.
- Try a machine learning ensemble instead.

Propensity Score Ensemble

Ensemble: why pick one approach when you can try them all?

- Vanilla and t-regularized logistic regression (benchmarks).
- Kernel regularized least squares.
- Bayesian additive regression trees (a version of random forest).
- ν-regularized support vector machine.

Ensemble

- Kernel regularized least squares:
  - "Duality" of regression as basis expansion and regression as kernel weighted average ("kernel trick").
  - For each unit, solve for c based on

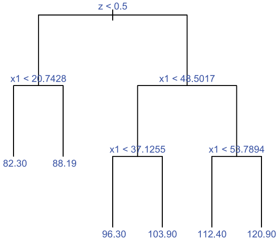
$$f(x^\star) = c_1 k(x^\star, x_1) + c_2 k(x^\star, x_2) + \ldots + c_N k(x^\star, x_N)$$
$$= c_1(\text{similarity of } x^\star \text{ to } x_1) + c_2(\text{sim. of } x^\star \text{ to } x_2) + \ldots + c_N(\text{sim. of } x^\star \text{ to } x_N).$$

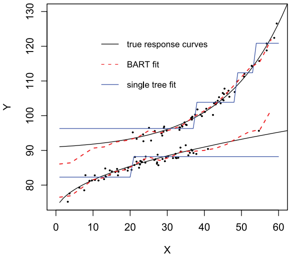
    (Hainmueller & Hazlett, 2013).
  - Regularized to penalize complexity in the c vectors.
  - Generate pscores from KRLS fits.
  - Effective for characterizing local nonlinearities and interactions.

Ensemble

- Bayesian additive regression trees:
  - Predict pscores with:

$$Y = g(z, x; T_1, M_1) + g(z, x; T_2, M_2) + \cdots + g(z, x; T_m, M_m) + \epsilon,$$





(Hill, 2011).

  - Bayesian regularization to penalize tree complexity.
  - Effective for characterizing nonlinearities, interactions, and non-smooth relationships.

Ensemble

- ν-support vector classification and regression:
  - Also works off the "kernel trick."
  - Fit a classifier for pscores:

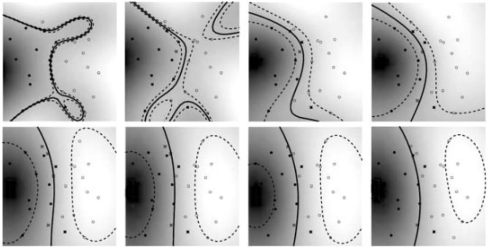


Figure 4. Toy problem (task: to separate circles from disks) solved using ν-SV classification, with parameter values ranging from $\nu = 0.1$ (top left) to 0.8 (bottom right). The larger we make ν, the more points are allowed to lie inside the margin (depicted by dotted lines). Results are shown for a Gaussian kernel, $k(x, x') = \exp(-\|x - x'\|^2)$ (from [1]).

(Chen et al., 2005).

  - For binary outcomes, minimize classification error.
  - *nu*-regularization to penalize complexity in the support vectors.
  - Effective for characterizing nonlinearities and interactions.

Ensemble

- SuperLearner prediction takes mse-minimizing combination:

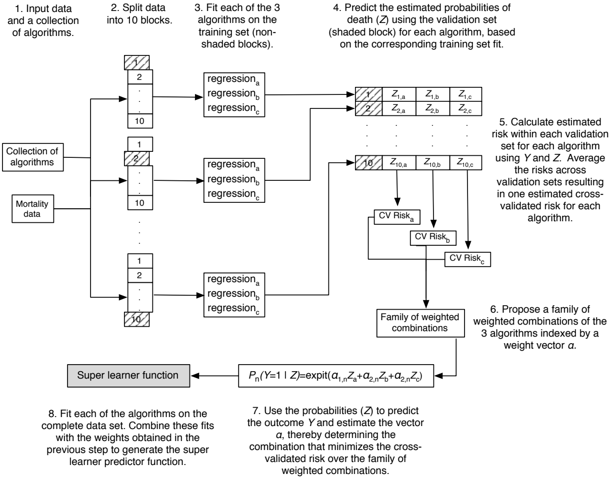


Fig. 3.2 Super learner algorithm for the mortality study example

(Van Der Laan & Rose, 2011).

Results

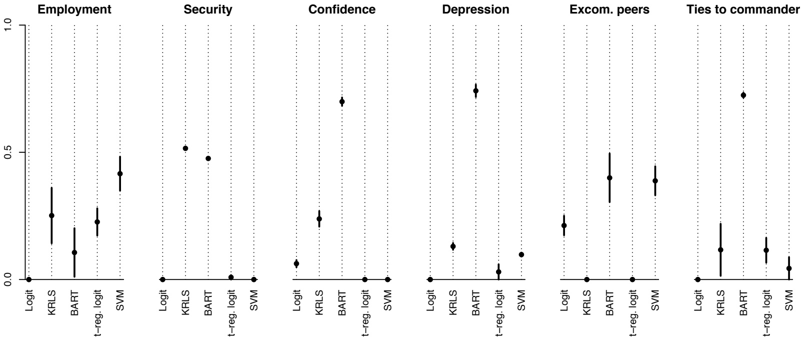
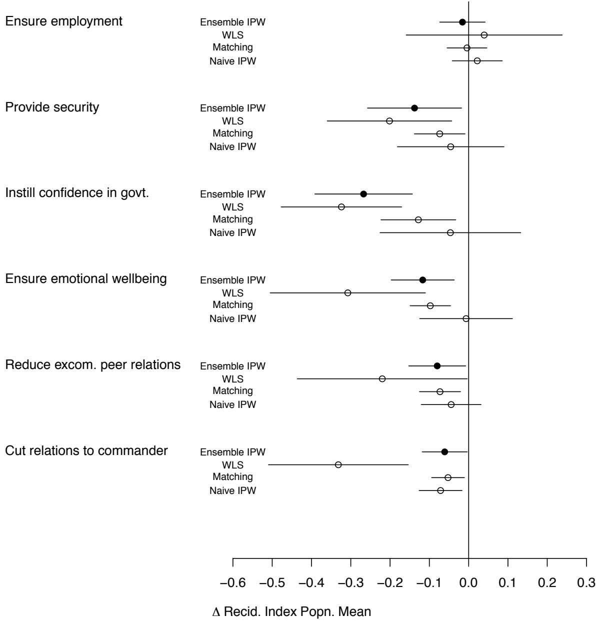


Figure 8: Weights applied to propensity score predictions from each prediction method. The values of weights run along the *y*-axis, and prediction methods run along the *x*-axis. Results are grouped by intervention. The weights are constrained to be no less than zero and to sum to one for each intervention. The black bars show the range of the weights over the 10 imputation runs, and the dots show the means.

Results


Ensure employment
Provide security
Instill confidence in govt.
Ensure emotional wellbeing
Reduce excom. peer relations
Cut relations to commander
Ensemble IPW
WLS
Matching
Naive IPW
–0.6 –0.5 –0.4 –0.3 –0.2 –0.1 0.0 0.1 0.2 0.3
Δ Recid. Index Popn. Mean

Principles for Regularization

(Belloni, Chernozhukov, and Hansen, 2014; Chernozhukov et al. 2017)

Principles for Regularization

- Previous example illustrated potential benefits of regularization.
- Estimator is was an IPW estimator and so consistency depended on *consistent pscore estimation*. Ensemble and regularization were targeted toward this.[1]

[1] Nice discussion is here: `http://www.unofficialgoogledatascience.com/2016/06/to-balance-or-not-to-balance.html`

Principles for Regularization

- Previous example illustrated potential benefits of regularization.
- Estimator is was an IPW estimator and so consistency depended on *consistent pscore estimation*. Ensemble and regularization were targeted toward this.[1]
- IPW has some attractive features, but may be less efficient than methods that incorporate outcome modeling.
- Belloni et al. (2014) develop some principles for regularization and covariate selection with methods rooted in outcome modeling.

[1] Nice discussion is here: `http://www.unofficialgoogledatascience.com/2016/06/to-balance-or-not-to-balance.html`

Principles for Regularization

- Following Belloni et al. (2014), suppose CIA and linearity.
- Outcome and treatment equation:

$$Y_i = \alpha D_i + X_i'\theta_Y + \zeta_i \quad \text{and} \quad D_i = X_i'\theta_D + \nu_i.$$

  with $\mathrm{E}[\zeta_i|D,X] = 0$, $\mathrm{E}[\nu_i|X] = 0$, $\mathrm{E}[\zeta_i\nu_i|X] = 0$.
- Implies a "reduced form" equation in terms of X:

$$\begin{aligned} Y_i &= \alpha(X_i'\theta_D + \nu_i) + X_i'\theta_Y + \zeta_i \\ &= X_i'(\alpha\theta_D + \theta_Y) + (\alpha\nu_i + \zeta_i) \\ &= X_i'\pi + \varepsilon_i \end{aligned}$$

Principles for Regularization

- Following Belloni et al. (2014), suppose CIA and linearity.
- Outcome and treatment equation:

$$Y_i = \alpha D_i + X_i'\theta_Y + \zeta_i \quad \text{and} \quad D_i = X_i'\theta_D + \nu_i.$$

  with $\mathrm{E}[\zeta_i | D, X] = 0$, $\mathrm{E}[\nu_i | X] = 0$, $\mathrm{E}[\zeta_i \nu_i | X] = 0$.
- Implies a "reduced form" equation in terms of X:

$$\begin{aligned} Y_i &= \alpha(X_i'\theta_D + \nu_i) + X_i'\theta_Y + \zeta_i \\ &= X_i'(\alpha\theta_D + \theta_Y) + (\alpha\nu_i + \zeta_i) \\ &= X_i'\pi + \varepsilon_i \end{aligned}$$

- Parameter of interest is α. By partial regression we know:

$$\mathrm{Cov}[\nu, \varepsilon] = \mathrm{Cov}[\nu, \alpha\nu + \zeta] = \alpha \mathrm{Var}[\nu] \Leftrightarrow \alpha = \frac{\mathrm{Cov}[\nu, \varepsilon]}{\mathrm{Var}[\nu]}.$$

- Suppose we have lots of potential Xs. This motivates a machine learning approach to fit the treatment and reduced form equations.

Principles for Regularization

$$\begin{aligned}
Y_i &= \alpha D_i + X_i'\theta_Y + \zeta_i \\
D_i &= X_i'\theta_D + \nu_i \\
Y_i &= X_i'(\alpha\theta_D + \theta_Y) + (\alpha\nu_i + \zeta_i) = X_i'\pi + \varepsilon_i.
\end{aligned}$$

- Regularizing wrt D misses Xs for which θ_Y large, θ_D small.
- Regularizing wrt Y misses Xs for which θ_D large if α small, or for which θ_Y small.

Principles for Regularization

$$
\begin{aligned}
Y_i &= \alpha D_i + X_i'\theta_Y + \zeta_i \\
D_i &= X_i'\theta_D + \nu_i \\
Y_i &= X_i'(\alpha\theta_D + \theta_Y) + (\alpha\nu_i + \zeta_i) = X_i'\pi + \varepsilon_i.
\end{aligned}
$$

- Regularizing wrt D misses Xs for which θ_Y large, θ_D small.
- Regularizing wrt Y misses Xs for which θ_D large if α small, or for which θ_Y small.
- Better to choose X wrt to both the D and Y equations.

Principles for Regularization

- Chernozhulov et al. (2017) propose "double machine learning" (DML).
- Based on FWL:
  - Regularized regression of D on X. Get residuals.
  - Regularized regression of Y on X. Get residuals.
  - DML estimator is residual-residual regression.

Skepticism

Causal inference is not just about adding Xs:

- Regularized methods work when DGP is in fact sparse (either few Xs matter or few interactions matter).
- Recall possibility of "bias amplification."
- D'Amour et al. (2019):
  - Recall that CIA has an overlap condition.
  - This limits how different covariate distributions can really be across treatment and control, either in terms of number of covariates or extent of difference for any given covariate.

Characterizing Effect Heterogeneity

Characterizing Effect Heterogeneity

- Another area of active develop is machine learning methods to characterize effect heterogeneity.
- Various uses:
  - Optimal treatment regimes (e.g., Imai & Strauss 2011).
  - Extrapolation (Hotz et al. 2005; Dehejia et al. 2017; Gechter et al. 2019).
  - Exploratory analyses (e.g., Angrist et al. 2013; Athey & Imbens 2016).

Characterizing Effect Heterogeneity

Wager & Athey (2018) "causal forest":

- Model $\hat{\tau}(X) = \hat{\mathrm{E}}[Y_1 - Y_0 | X]$ using random forest.
- Regularization tuning parameters selected via cross validation.
- "Black box" conditional treatment effect estimator.

Characterizing Effect Heterogeneity

Compare to Athey & Imbens (2016) “causal tree”:

- Want to identify effect heterogeneity in an exploratory (not confirmatory) way and be able to interpret result.
- $\Rightarrow$ Tree approach: “partition of the population according to treatment effect heterogeneity”—a “Causal Tree.”
- Sacrifices some predictive accuracy for the sake of interpretability.
- Similar idea in Imai & Ratkovic (2013) using LASSO-penalized SVM.

Characterizing Effect Heterogeneity: Policy Targeting

Targeting impact versus deprivation*

Johannes Haushofer[1], Paul Niehaus[2], Carlos Paramo[3], Edward Miguel[3], and Michael Walker[3]

[1]Stockholm University
[2]University of California, San Diego
[3]University of California, Berkeley

April 21, 2022

Characterizing Effect Heterogeneity: Policy Targeting

Suppose social welfare problem is allocate treatments to maximize

$$\sum_i W(Y_{i0} + T_i(Y_{i1} - Y_{i0}))$$

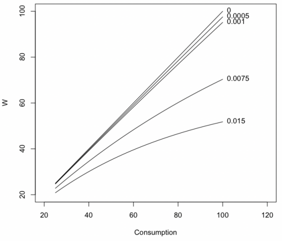
where $W' > 0$ and $W'' < 0$. How to target cash transfers to maximize?

- Target the most deprived (the "D" group)?
- Target those for whom the transfer makes the biggest impact (the "I" group)?
- Two groups may not be the same:
  - most deprived have highest marginal utility from income,
  - but not in a position to make high yielding investments.

Characterizing Effect Heterogeneity: Policy Targeting

- GRF to estimate $\widehat{Y}_{i0}$ and $\widehat{Y_{i1} - Y_{i0}}$ given large X_i.
- Classify units in terms of I and D groups.
- Solve for allocation rule that maximizes welfare
- Look at overlap with I and D groups.
- Constant absolute risk aversion specification for $W(\cdot)$:

$$W(y) = \begin{cases} \frac{1-e^{-\alpha y}}{\alpha} & \alpha \neq 0 \\ y & \alpha = 0 \end{cases}$$



Characterizing Effect Heterogeneity: Policy Targeting

Table 4: Overlap of socially optimal households to target with most deprived and most impacted

	(1)	(2)	(3)	(4)	(5)
CARA: α	CE	Most deprived	Most impacted	Choice	α_c
Panel A: Consumption					
0.0000	\$50.00	0.30	1.00	I	
0.0005	\$49.38	0.31	0.96	I	
0.0010	\$48.75	0.33	0.92	I	
0.0075	\$40.84	0.40	0.81	I	
0.0150	\$32.78	0.42	0.79	I	← 0.016
Panel B: Assets					
0.0000	\$50.00	0.31	1.00	I	
0.0005	\$49.38	0.35	0.91	I	
0.0010	\$48.75	0.39	0.83	I	← 0.007
0.0075	\$40.84	0.55	0.59	D	
0.0150	\$32.78	0.57	0.55	D	
Panel C: Income					
0.0000	\$50.00	0.47	1.00	I	
0.0005	\$49.38	0.48	0.97	I	
0.0010	\$48.75	0.50	0.93	I	
0.0075	\$40.84	0.59	0.73	I	← 0.011
0.0150	\$32.78	0.63	0.66	D	

Notes: Column 1 denotes the certainty equivalent (CE) of a 50-50 lottery over \$0 or \$100 under the specified CARA α parameter value. Column 2 (3) reports the share of households belonging to I (D) that are also "socially optimal" for a planner to treat. Socially optimal households are those in the top 50% of households ranked by potential gains from treatment using a CARA utility function for the risk aversion parameter (α) given in the row label. Reported shares are the mean of 150 5-fold GRF iterations; median ratios are similar (not shown). Column 4 reports the welfare maximizing choice between targeting the most impacted (I) and the most deprived (D) for a given α value. Column (5) reports the critical value α_c, the mean minimum value of α required to rationalize a policy targeting the most deprived instead of targeting the most impacted across the 150 estimated models. Formally, $\alpha_c = \min(\{\alpha : SW(D;\alpha) \geq SW(I;\alpha)\})$.

Remarks

- Machine learning tools enable working with many covariates in causal inference problems (operationalizing CIA, characterizing heterogeneous effects).

Remarks

- Machine learning tools enable working with many covariates in causal inference problems (operationalizing CIA, characterizing heterogeneous effects).
- Can be applied to other problems:
  - Exploring what features of complex treatments matter (e.g., "texts" as treatments Egami et al 2018).
  - Dynamic learning of optimal treatments ("causal bandits"—Lattimore et al. 2016).

Remarks

- Machine learning tools enable working with many covariates in causal inference problems (operationalizing CIA, characterizing heterogeneous effects).
- Can be applied to other problems:
  - Exploring what features of complex treatments matter (e.g., "texts" as treatments Egami et al 2018).
  - Dynamic learning of optimal treatments ("causal bandits"—Lattimore et al. 2016).
- That said, these tools *complement* or *supplement*, not replace, the randomization, discontinuities, or arguments for CIA that allow for causal identification.