Causal Inference II:

Methods

Vincent Dorie

Columbia University

January 10, 2019

Nonparametric Tests

If we have no confounders (randomized experiment), simple nonparametric tests can be used

- The only sources of randomness are noise and treatment assignment
- Treatment assignment mechanism is known
- $P(\text{test statistic}(Z, y) \geq \text{observed}) =$
$$\#\{\text{random assignments} : t(z, y) \geq t\}/\#\text{total assignments}$$

Ideal Scenario

Unit i	**Female** x_{1i}	**Age** x_{2i}	**Treat-ment** z_i	**if** $z_i = 0$, $y_i(0)$	**if** $z_i = 1$, $y_i(1)$	**Observed outcome** y_i
Audrey	1	40	0	**140**		140
Anna	1	40	1		**135**	135
Bob	0	50	0	**150**		150
Bill	0	50	1		**140**	140
Caitlin	1	60	0	**160**		160
Cara	1	60	1		**155**	155
Dave	0	70	0	**170**		170
Doug	0	70	1		**160**	160

- Every treated subject is *matched* to a control based on age and female
- The average treatment effect can be estimated by averaging over differences within matches

Ideal Scenario

Unit i	**Female** x_{1i}	**Age** x_{2i}	**Treat-ment** z_i	**if** $z_i = 0$, $y_i(0)$	**if** $z_i = 1$, $y_i(1)$	**Observed outcome** y_i
Audrey	1	40	0	**140**		140
Anna	1	40	1		**135**	135
Bob	0	50	0	**150**		150
Bill	0	50	1		**140**	140
Caitlin	1	60	0	**160**		160
Cara	1	60	1		**155**	155
Dave	0	70	0	**170**		170
Doug	0	70	1		**160**	160

- Every treated subject is *matched* to a control based on age and female
- The average treatment effect can be estimated by averaging over differences within matches -7.5

Exact Matching

For every unique set of covariates

1. Find all treated and control observations with those values
2. If either set is empty abort unless the empty group is the inferential one
3. Take average of treated, average of controls
4. Take difference between averages
5. Take a weighted average of all matched differences, weights coming from how many inferential observations are present

$$\begin{aligned} E[Y(1) - Y(0)] =& E_X \left[E_{Y(1),Y(0)|X}[Y(1) - Y(0) \mid X] \right] \\ =& \sum_x P(X = x) \left[\bar{Y}(1)_{x_i = x} - \bar{Y}(0)_{x_i = x} \right] \end{aligned}$$

Exact Matching Continued

Exact matching fails:

- Covariate space is large, sparse
- Any continuous covariate

Use approximate matches with any concept of distance between observations and a threshold

- Exact: $d(X_i, X_j) = 0$ if $X_i = X_j$, ∞ otherwise
- Mahalanobis: $d(X_i, X_j) = \sqrt{(X_i - X_j)^\top S^{-1}(X_i - X_j)}$, S an estimate of the covariance between variables

Balancing

Within each exact matched set observations are *comparable* and *balanced* on their covariates

With approximate matches, we want similar

- A *balancing score*, $b(X)$, is a summary of the observations such that for any value of that score, the distribution of covariates will be similar between treated and control, $Z \perp X \mid b(X)$

The *propensity score*, $e(X) := P(Z = 1 \mid X)$, is an optimal balancing score

Propensity Score

Observations with the same probability of being treated are directly comparable, as being treated or not is a coin flip

We also don't ever know the true propensity score, but we can estimate it

- Usually a logistic regression: $\operatorname{logit} P(Z = 1 \mid X) = \alpha + X\beta$
- Assess balance of covariate distributions across matches
- Refine the model if balance is poor

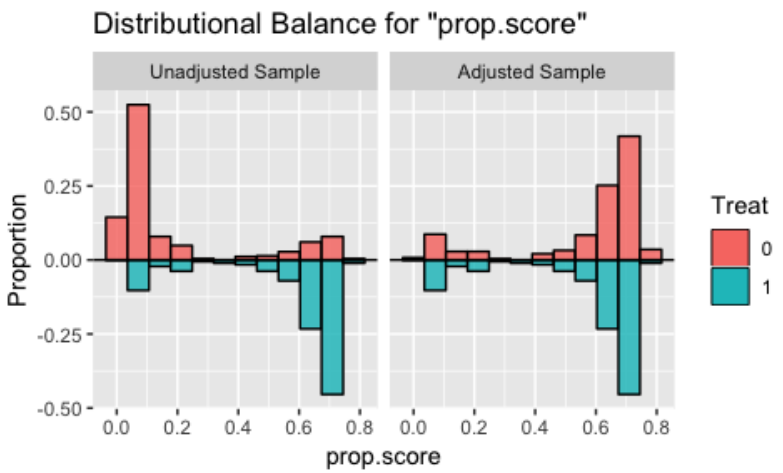
Matching on the propensity score uses the distance in that space and a rule to assign control observations to treated

Assessing Balance

Matching essentially reduces to *weighting*, in that we take averages across all observations weighting each unit

Many matching rules give 0/1 weights

One can look at the distributions of covariates when weighted by, e.g. standardized differences in means or any distributional test



From cobalt package vignette by Noah Greifer

Stratification

Matching often discards observations, instead we can stratify:

- Divide observations into strata based on any convenient criteria (e.g. quantiles of propensity score)
- Estimate treatment effects within strata
- Take a weighted average

Matching is stratification with as many categories as units of interest

Inverse Weighting

We can use all of the observations and have less bias by directly weighting observations

- $w_{ATE} = z/e + (1-z)/(1-e)$
- $w_{ATT} = z + (1-z) \cdot e/(1-e)$
- Re-weighting observations to create a psuedo-population that is balanced

Derives from rewriting $E[Y(1)] = E_X \left[\frac{E_{Y(1)|X}[Y(1) \cdot Z|X]}{P(Z=1|X)} \right]$

Small/large propensity scores give large weights, can use *stabilized* weights instead

Propensity Score Methods

Proponents highlight:

- Better than what came before/more principled than matching directly on covariates
- Using the response variable tempts researchers to get result they want to see, checking balance is the only target
- Alternatives often also require parametric assumptions

Criticisms of propensity score methods:

- Require accurate specification of propensity score model
- Ignore information about the response variable, may focus too much on imbalance caused by a variable that is not a confounder
- Can't target individual effects, only population ones

Regression

One can instead directly model the response variable:

$$E[Y \mid Z, X] = \alpha + \beta X + \tau Z$$

- τ is the treatment effect
- Can incorporate propensity score weights (population weights)
- Can complexify the model to handle nonlinearities, interactions, and non-constant treatment effects

Also requires assumptions about the model

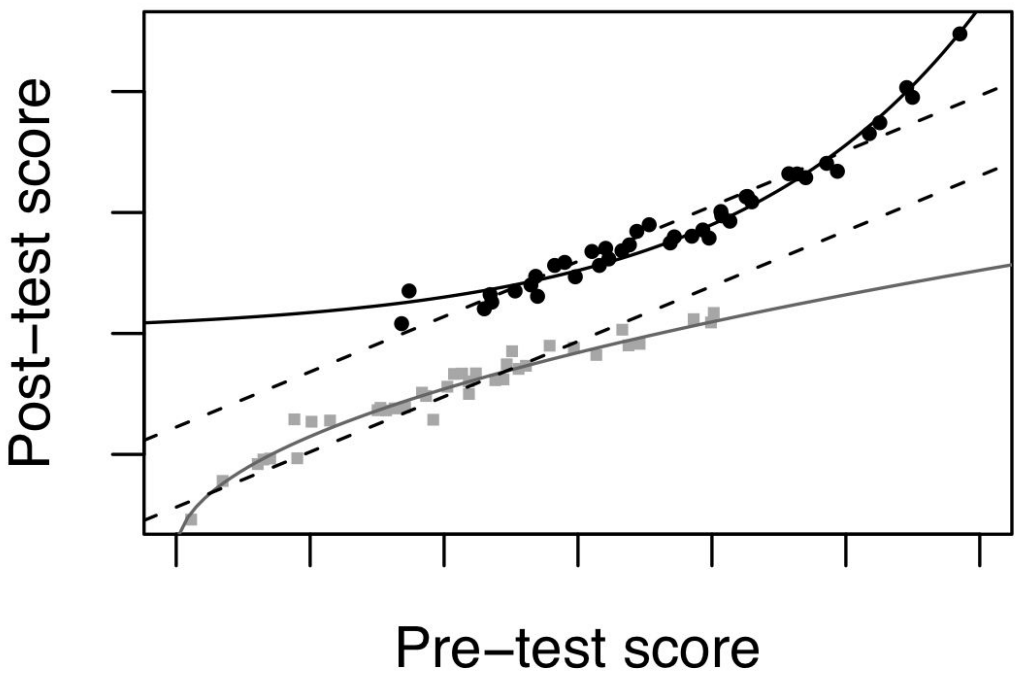
Doubly robust methods model both treatment assignment and response surface, and are accurate if only one of them is correctly specified

Non-parametric Regression

For an arbitrary f (Gaussian process, random forest, NN, etc), fit:

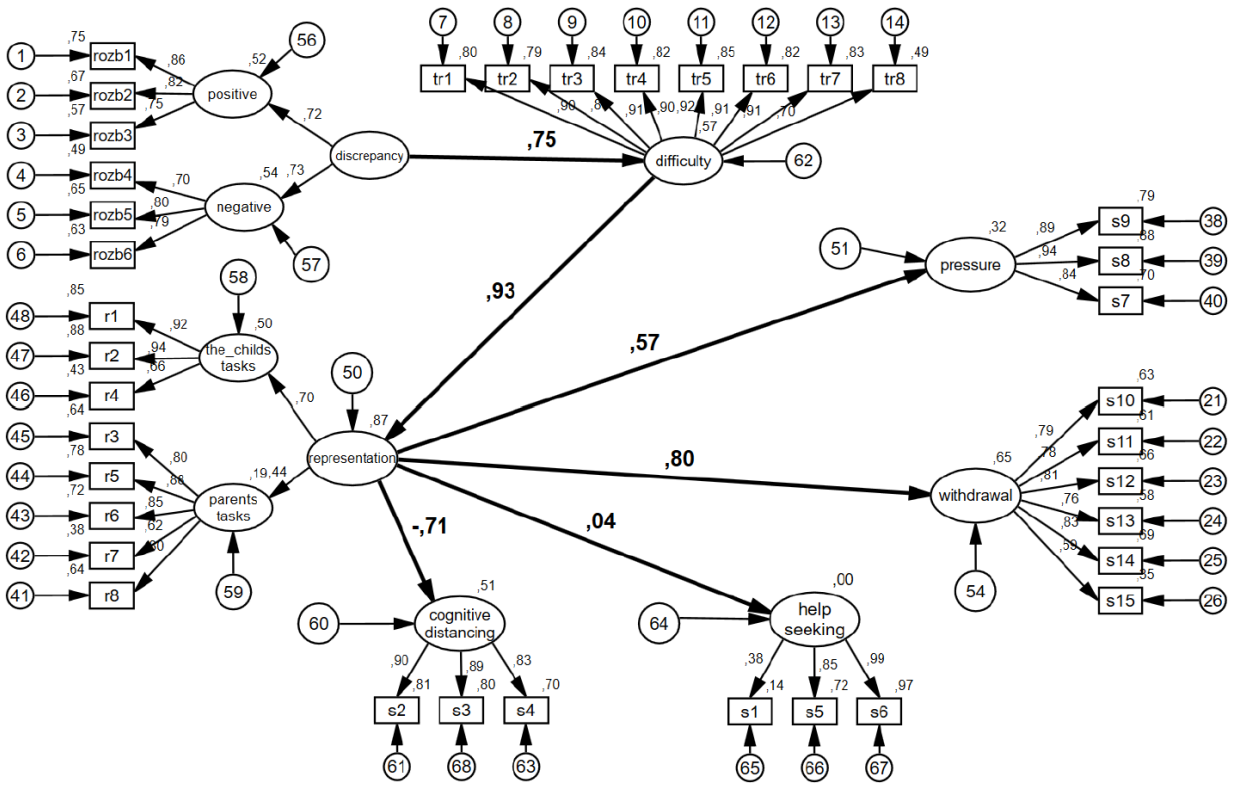
$$E[Y \mid Z = z, X = x] = f(x, z)$$

and use model to predict individual counterfactuals



Structural Equation Models

Structural equation models (Wright, Pearl) are a way of fitting an arbitrary causal DAG

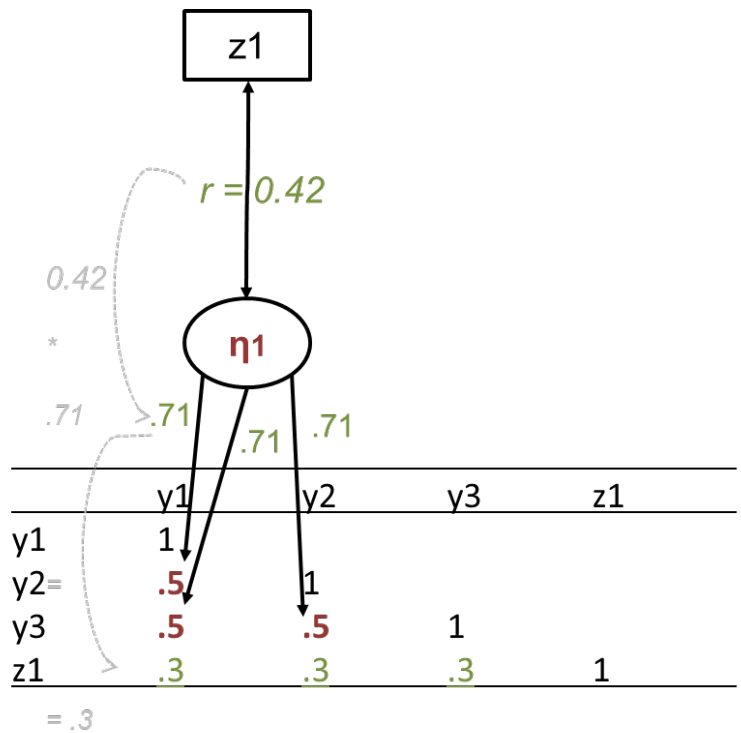


Szymańska and Dobrenko

Structural Equation Models Continued

SEMS interject a latent, structural model in between the inputs and outputs

E.g.: measure "openness" on a survey together with counting how many novel uses they can come up with for a sequence of every day objects. Assume there is an underlying/latent "creativity" that drives the process



Edelsbrunner & Thurn

Structural Equation Models Continued

Advantages:

- The structural model can have arbitrary complexity and connectedness
- Allows one to conceive of moderation at any point in the DAG

Disadvantages:

- Strong (conditional) independence assumptions
- Often requires specifying a lot of parametric models

How to Pick a Method

General advice:

1. Use whatever method is most appropriate to your field
2. Relax and test as many assumptions as possible

Models that flexibly model treatment assignment and the response surface tend to be superior in most settings, not yet clear on how to best combine the two