Heterogeneous treatment effects

(A high-level overview)

Pablo Geraldo

STAT 256

Today's plan

- Quick recap
- Old school
    - Stratification and subgroup analysis
    - Regression with interactions
- New wave
    - Naively import from ML
    - Import but tailoring to CI
- Extensions and limitations

Quick recap

Conditional estimands

In general, we can target causal estimands at different levels of aggregation:

Individual treatment effect (ITE)

$$\tau_i = Y_{1i} - Y_{0i}$$

Conditional average treatment effect (CATE)

$$\tau(x) = E(Y_{1i} - Y_{0i} | X = x)$$

Average treatment effect (ATE)

$$\tau = E(Y_{1i} - Y_{0i})$$

And other variants such as ATT, ATC, LATE, are *also* examples of conditional effects (of a different type, for sure)

What heterogeneity?

Two related, but different concepts, are those of causal interaction and essential heterogeneity.

Causal interaction refers to the situation in which we have two interventions, and their effect is not additive. For example, if we have

$$E(Y|do(x,z)) \neq E(Y|do(x,z^*))$$

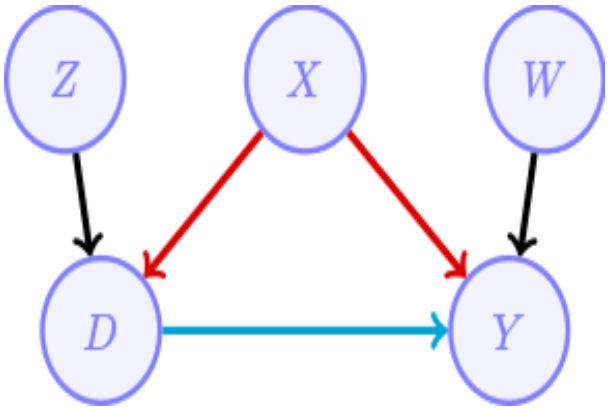
for some z^*, we say that x and z causally interact.

Essential heterogeneity, usually appearing in econometrics, refers to the situation in which subjects *choose* to participate on certain program based on their (unobserved) expected returns. More on this at the end.

Graphical models for effect modification

Recall DAGs are non-parametric, so by default they assume that every variable can *interact* with each other.

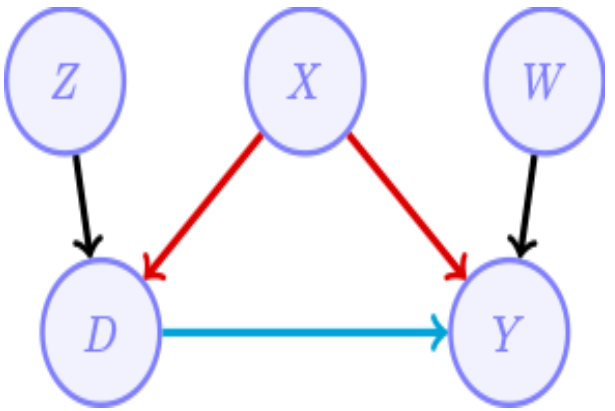
Which variables can be sources of treatment effect heterogeneity in the following DAG?



Graphical models for effect modification

Recall DAGs are non-parametric, so by default they assume that every variable can *interact* with each other.

Which variables can be sources of treatment effect heterogeneity in the following DAG?



Thinking in terms of the structural equations usually helps

The "machine learning DAG" (?)

Sometimes, you would find oversimplifications of the data generating process to explain or justify conditional ignorability.

But trying to think carefully about your data generating process cannot be replaced by convenience-based assumptions!

The "machine learning DAG" (?)

Sometimes, you would find oversimplifications of the data generating process to explain or justify conditional ignorability.

But trying to think carefully about your data generating process cannot be replaced by convenience-based assumptions!

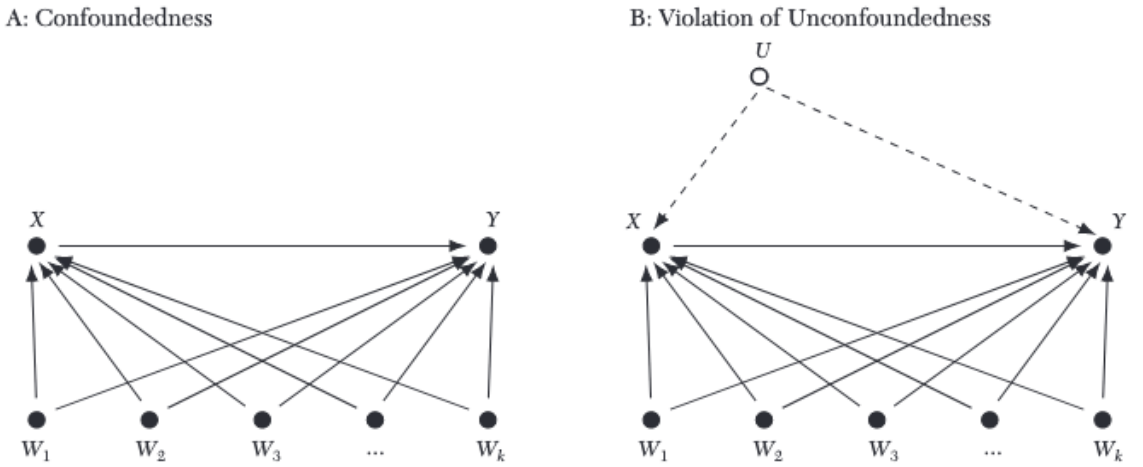


Figure 8. Unconfoundedness with Multiple Observed Confounders

Source: Imbens (2020)

Four types of effect modification

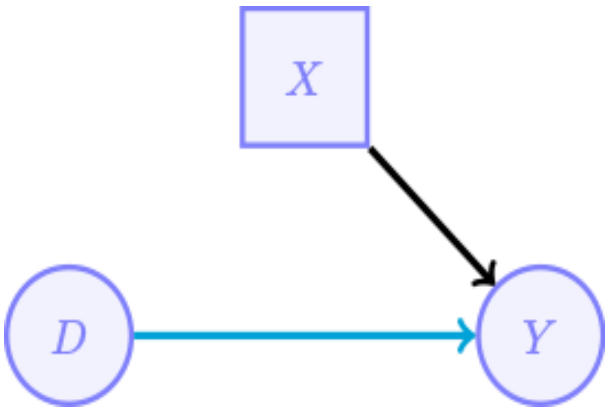
Using graphical models, VanderWeele and Robins proposed four types of effect modification.

- Direct effect modification
- Indirect effect modification
- Effect modification by proxy
- Effect modification by common cause

Four types of effect modification

Using graphical models, VanderWeele and Robins proposed four types of effect modification.

- **Direct effect modification**

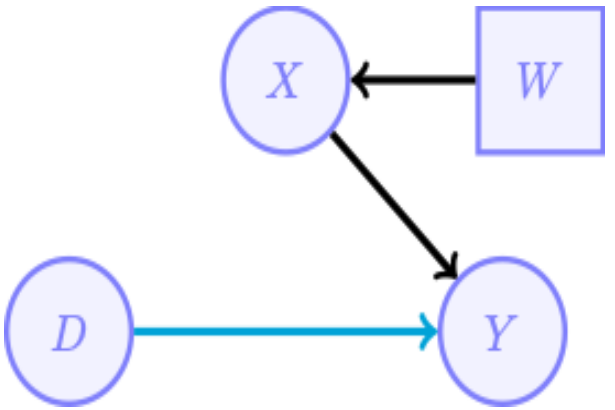


- Indirect effect modification
- Effect modification by proxy
- Effect modification by common cause

Four types of effect modification

Using graphical models, VanderWeele and Robins proposed four types of effect modification.

- Direct effect modification
- **Indirect effect modification**

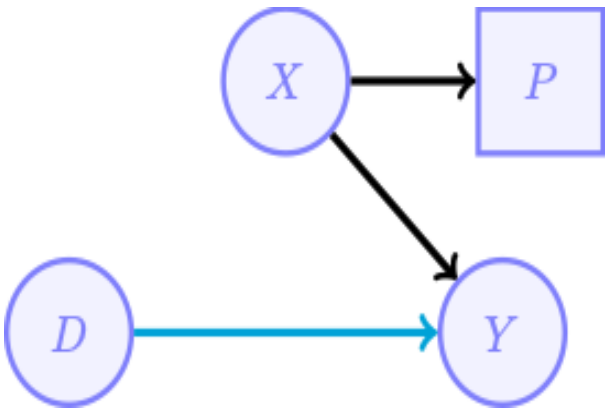


- Effect modification by proxy
- Effect modification by common cause

Four types of effect modification

Using graphical models, VanderWeele and Robins proposed four types of effect modification.

- Direct effect modification
- Indirect effect modification
- **Effect modification by proxy**



- Effect modification by common cause

Four types of effect modification

Using graphical models, VanderWeele and Robins proposed four types of effect modification.

- Direct effect modification
- Indirect effect modification
- Effect modification by proxy
- **Effect modification by common cause**

Old school

Estimating heterogeneity

Traditionally, there have been many ways to estimate effect heterogeneity.

- In regression models, including interaction terms.
  - Need to specify which variables one is interacting
  - Need to specify which the functional form of such interactions

Recent "best practice" proposals:

Hainmueller, Mummolo, and Xu (2018)

Mize (2019)

- In stratification approaches, one can either use balanced propensity score strata, or do subgroup analysis for a set of previously defined comparison groups.

Limitations of old school approaches

- Stratification approach only feasible with low-dimensional and discrete X.
    - It won't produce very nuanced estimates in general
    - Good for testing, not necessarily for discovering
- Regression with interaction approach too restrictive with respect to the heterogeneity
    - Need to select variables to "interact"
    - Include or not main terms?
    - No attention to overlap/distributional issues
    - Usually presented as hypothesis-based, but mostly fishing
- Two opposite risks: fishing and extreme rigidity.
    - A new wave of estimators has been developed trying to address those issues.
    - The goal: flexibility without arbitrariness.

New wave

Source: This section is based on Wager's class notes

Using ML to estimate CATE: two approaches

- **Transformed Outcome Regression**
  - A transformation of the observed outcome whose expectation equals to our quantity of interest (CATE)
  - **Pros**: valid of any off-the-shelf ML algorithm
  - **Cons**: high variance!

- **Response Surface Modeling**
  - Using the observed outcomes conditional on the treatment status to model $E(Y|X, D = 1)$ and $E(Y|X, D = 0)$
  - **Pros**: more familiar (just a supervised problem?)
  - **Cons**: algorithms and CV need adaptation (cross-fitting, honesty)

In both cases we need experimental data, or

$$(Y_1, Y_0) \perp\!\!\!\perp D|X$$

Transformed Outcome

Let's define the following transformation of the observed outcome

$$Z \equiv D\frac{Y}{e(X)} - \frac{(1-D)Y}{1-e(X)}$$

Where

Y is the observed outcome

D is the treatment status (1 if treated, 0 if control)

$e(X)$ is the propensity score $P(D=1|X)$

Transformed Outcome

It can be shown that

$$E(Z|X) = \frac{E(Y_1|X)}{e(X)} P(D = 1|X) - \frac{E(Y_0|X)}{1 - e(X)} P(D = 0|X)$$

Where

$E(Y|X, D = 1) = E(Y_1|X)$ by potential outcome consistency, and

$E(D|X) = P(D = 1|X)$ by expectation of a binary variable.

Therefore

$$E(Z|X) = E(Y_1|X) - E(Y_0|X) = \tau(X)$$

Why has this estimator a high variance?

The two-regression approach

One immediately intuitive strategy to estimate effect heterogeneity is apparent when rewriting the CATE as

$$\tau(x) = \mu_1(x) - \mu_0(x),$$

where

$$\mu_d(x) = E(Y_i | X_i = x, D_i = d)$$

We can see that $\mu_d(x)$ is simply a conditional expectation and we can just fit a flexible model to each of them.

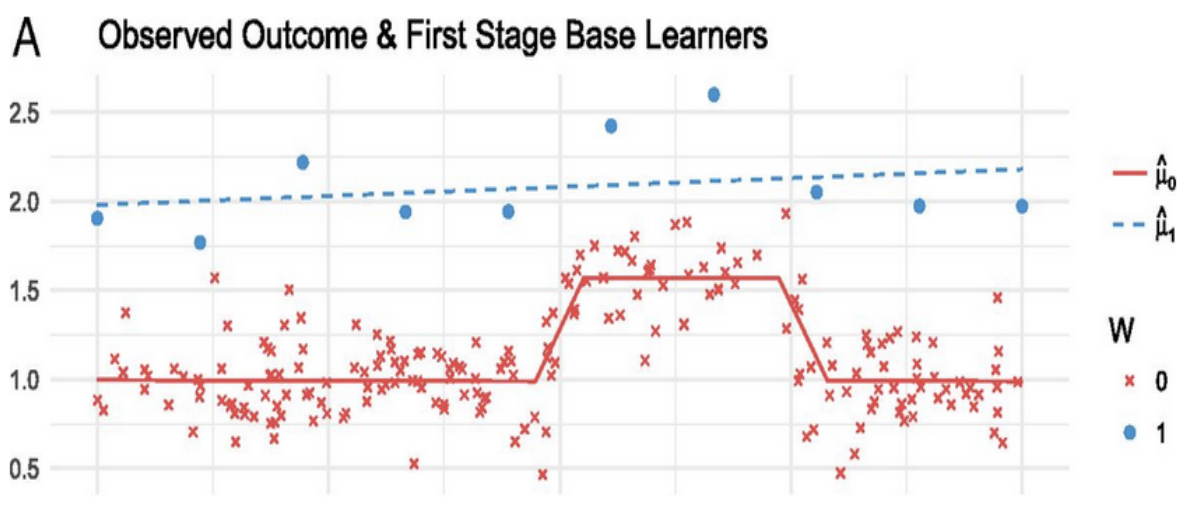
Our CATE estimator then becomes

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

Problems with the two-regression approach

Despite being intuitive and sometimes sufficient, modeling μ_0 and μ_1 independently has some drawbacks:

- **Regularization bias**: when the sample size of treated and control units is too different, and we are flexibly modeling both response surfaces, $\hat{\mu}_0$ and $\hat{\mu}_1$ would be subject to different regularizations



Künzel et al (2019)

Problems with the two-regression approach

Despite being intuitive and sometimes sufficient, modeling μ_0 and μ_1 independently has some drawbacks:

- **Regularization bias**: when the sample size of treated and control units is too different, and we are flexibly modeling both response surfaces, $\hat{\mu}_0$ and $\hat{\mu}_1$ would be subject to different regularizations.
- **Covariate shift**: when the propensity score $e(x)$ vary considerably across the support of X, then the fit of $\hat{\mu}_0$ and $\hat{\mu}_1$ would be influenced by different regions of X.

¿Are these cases likely to happen in practice?

Semi-parametric approaches

One possible solution to the previous problems is to estimate the CATE under a semi-parametric framework. Consider the following model:

$$Y_{di} = f(X_i) + d\tau(X_i) + \epsilon_i(d)$$

where $\tau(x) = \psi(x)\beta$ for some basis functions ψ.

Under unconfoundedness, it can be shown that we can rewrite this model

$$Y_i - m(X_i) = (D_i - e(X_i))\psi(X_i)\beta + \epsilon_i$$

where $m(x) = E(Y_i|X_i = x) = f(X_i) + e(X_i)\tau(X_i)$

Here we would model the $X \to Y$ and $X \to D$ relationships using flexible approaches (like ML).

Then we fit a regression for the residualized Y on the residualized treatment D (aka, residuals-on-residuals regression), using cross-fitting.

Robinson (1988)

Loss function for the CATE

What about cases in which limiting ourselves to the semi-parametric regression model is still too restrictive?

For example, with complex and high-dimension X, we might not want to stick to functions $\psi(x)\beta$, and instead try to "discover" a good approximation to the CATE.

Loss function for the CATE

Recall that $\mu_d(x) = E(Y_{di}|X = x)$.

It is easy to see that, under unconfoundedness, $E(\epsilon_i(D_i)|X_i, D_i) = 0$

where $\epsilon_i(d) = Y_{di} - (\mu_0(X_i) + d\tau(X_i))$.

We can further rewrite the residualized outcome model as

$$Y_i - m(X_i) = (D_i - e(X_i))\tau(X_i) + \epsilon_i$$

where $m(x) = E(Y_i|X_i = x) = \mu_0(X_i) + e(X_i)\tau(X_i)$

This is equivalent to

$$\tau(\cdot) = \text{argmin}_{\tau'}\Big\{E\Big(\Big[(Y_i - m(X_i)) - (D_i - e(X_i))\tau'(X_i)\Big]^2\Big)\Big\}$$

Loss function for the CATE

This is equivalent to

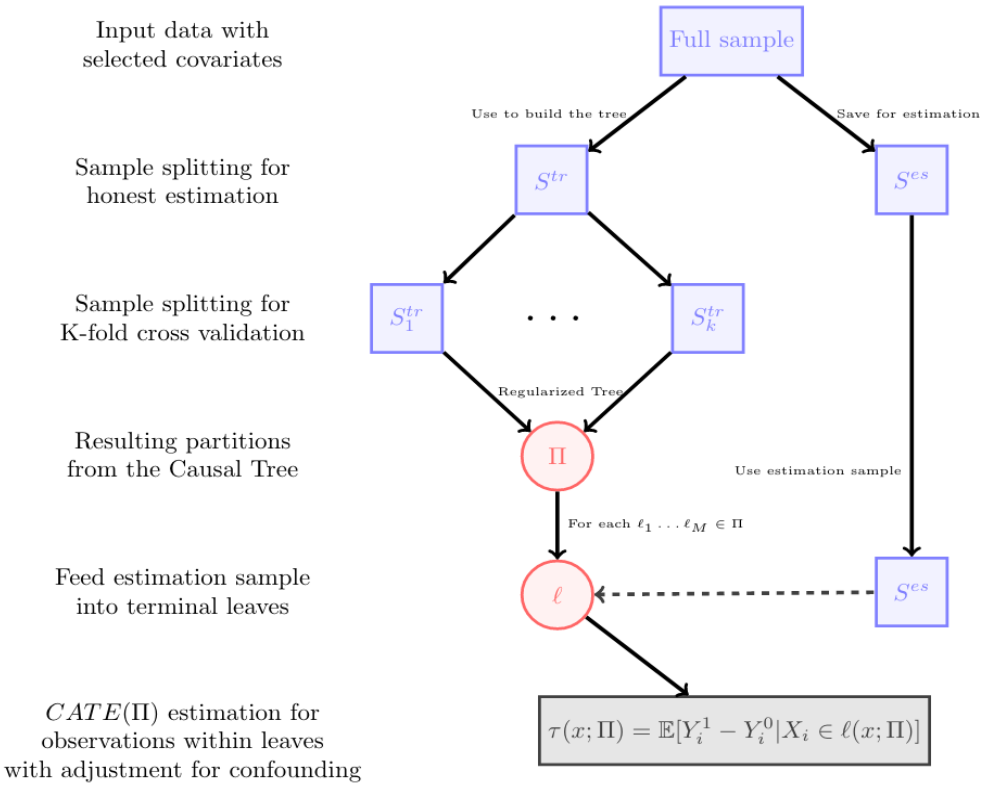
$$\tau(\cdot) = \text{argmin}_{\tau'} \left\{ E\left(\left[(Y_i - m(X_i)) - (D_i - e(X_i))\tau'(X_i) \right]^2 \right) \right\}$$

But recall that we don't have μ_d nor $e(X_i)$, both need to be estimated.

So one would implement this approach:

- Replacing the loss function with the plug-in version;
- Potentially adding a regularizing term;
- Estimation using cross-fitting.

Example: Causal Trees



Source: Brand et al. (2021)

Validating treatment heterogeneity

Finally, we have the problem of how to validate treatment heterogeneity found in these flexible ways just described.

Recall that, most of the times, the aim of such analysis is targeting policies or designing tailored treatment regimes, so we need estimates that are precise enough but not all over the place.

Some suggestions:

- Use cross-validation to select the best performing model
- Use ensemble methods when competitors show similar performance
- Use these methods as an *honest* group finding strategy, and then do subgroup analysis.
- Go back to the semi-parametric strategy, using the more flexible approaches as a benchmark for the ψ functions.

Extensions and limitations

HTE under unconfoundedness

So far we have assumed either randomization or conditional ignorability in order to identify HTE. Under this assumption, we will see that ML methods could help estimation by restricting our search for HTE.

An intuitive way of limiting the search for heterogeneity is to use an uni-dimensional index that summarizes all covariate's information: the Propensity Score (PS).

This is a natural extension in the use of Propensity Scores: if we can identify conditional effects in each strata defined by $X = x$, then we can do the same using $P(D = 1|X = x)$ instead.

HTE and selection bias

Recall that we can decompose the difference-in-means estimator as the true ATE plus two bias terms:

$$
\begin{aligned}
&E(Y_1 - Y_0) - && \text{True ATE} \\
&\pi[E(Y_0|D=0,X) - E(Y_0|D=1,X)] + && \text{baseline bias} \\
&(1-\pi)[E(Y_1|D=1,X) - E(Y_1|D=0,X)] && \text{response bias}
\end{aligned}
$$

Under conditional ignorability, both baseline and response to the treatment biases are ruled out.

HTE and propensity scores

If we are able to assume selection on observables, then all the bias-removing information in the covariates can be summarized by $e(X_i)$

This means that including a $Treatment \times PS$ terms (however flexible) is *sufficient* to capture treatment heterogeneity based on observables.

Xie, Brand, and Jann(2012) proposed three methods for estimating treatment effect heterogeneity using propensity scores:

- Stratification-Multilevel method
- Matching-smoothing method
- Smoothing-differencing method

HTE and propensity scores

Xie, Brand, and Jann(2012) proposed three methods for estimating treatment effect heterogeneity using propensity scores:

- **Stratification-Multilevel method**
  - Construct balanced strata
  - Estimate strata-specific effects
  - Multilevel model of the second-level trend
- Matching-smoothing method
- Smoothing-differencing method

HTE and propensity scores

Xie, Brand, and Jann(2012) proposed three methods for estimating treatment effect heterogeneity using propensity scores:

- Stratification-Multilevel method
- **Matching-smoothing method**
  - Construct a balanced matched sample
  - Estimate pair (or block) specific effects
  - Non-parametric modeling of the trend
- Smoothing-differencing method

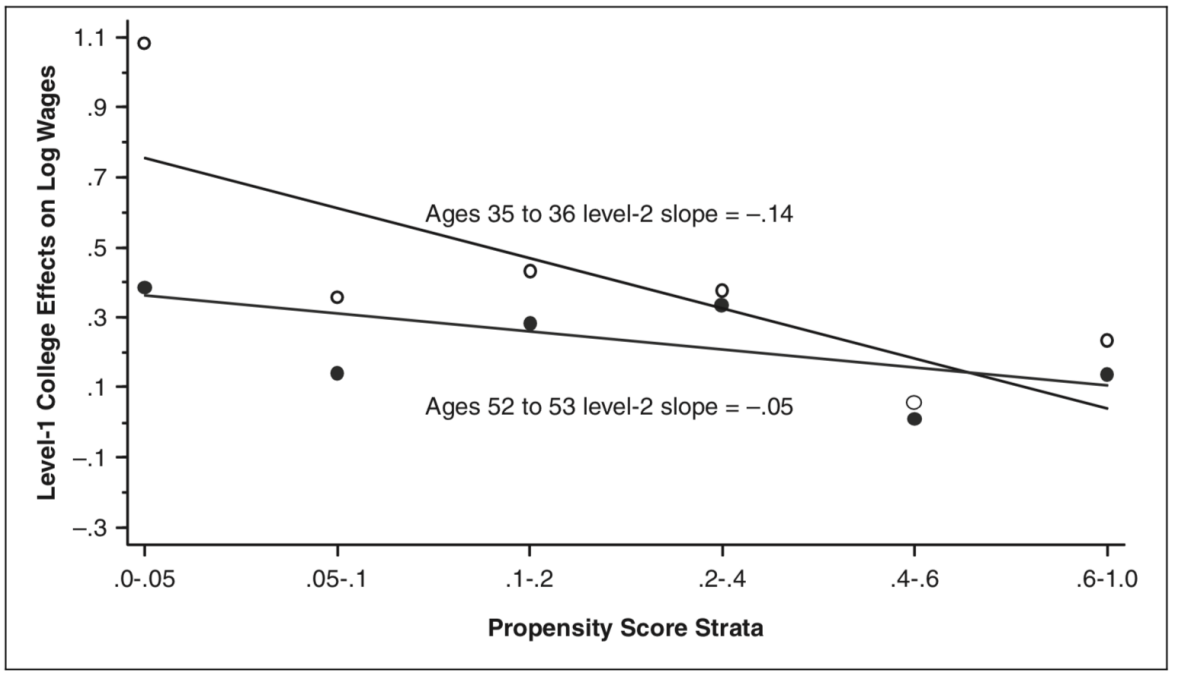
HTE and propensity scores

Xie, Brand, and Jann(2012) proposed three methods for estimating treatment effect heterogeneity using propensity scores:

- Stratification-Multilevel method
- Matching-smoothing method
- **Smoothing-differencing method**
  - Construct a balanced matched sample
  - Fit non-parametric models for control and treatment groups
  - Take the difference between the regressions

Example: Negative selection into College

Using this method, Brand and Xie (2010) find that those with lower propensity to attend College are the ones that benefit the most from going to College.



Data: Wisconsin Longitudinal Study (1957 cohort), women sample.

HTE under SOO violations

However, what happen when there is unobserved selection into the treatment?

Breen, Choi and Holm (2015) shows that, under small departures from the conditional ignorability assumption, it is possible to:

- Find heterogeneous effects where the true effect is homogeneous.
- Commit errors of magnitude, direction, or both, when the true effect is heterogeneous.

They particularly criticized the PS-based approach because baseline bias can easily be mistaken for HTE.

HTE and selection bias

For simplicity, let's assume the potential outcomes are generated by a linear function (with constant β, δ, ϵ):

$$
\begin{aligned}
Y_1 &= \beta X_i + \delta + \epsilon_i \\
Y_0 &= \beta X_i + \epsilon_i
\end{aligned}
$$

and the selection into the treatment (D_i) is defined by a latent switching function D_i^*:

$$
D_i^* = \gamma Z_i + u_i, \text{ such that } \begin{cases} D_i = 1 & \text{if } D_i^* > 0 \\ D_i = 0 & \text{if } D_i^* \leq 0 \end{cases}
$$

HTE and selection bias

We are assuming a constant effect δ with no influence on ϵ. It can be shown that under unobservable selection (when $corr(\epsilon, u) \neq 0$), the amount of bias is heterogeneous on the propensity score:

$$\begin{aligned} &E(Y|D=1, e(X)) - E(Y|D=0, e(X)) \\ =&\delta + [E(\epsilon|D=1, e(X)) - E(\epsilon|D=0, e(X))] \end{aligned}$$

By assuming the errors are bivariate normal, We can came up with an analytical formula for this bias term:

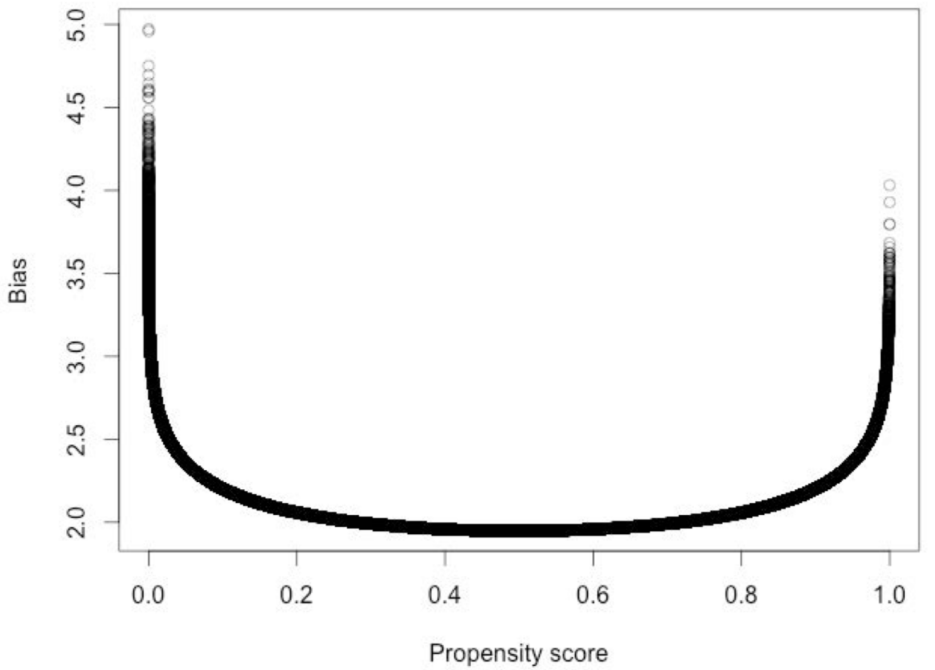
$$\rho_{\epsilon,u}\sigma_\epsilon \frac{\phi(\Phi^{-1}p(Z))}{p(Z)(1-p(Z))}$$

Complete derivation on Breen, Choi and Holm (2012)

Bias as function of propensity score

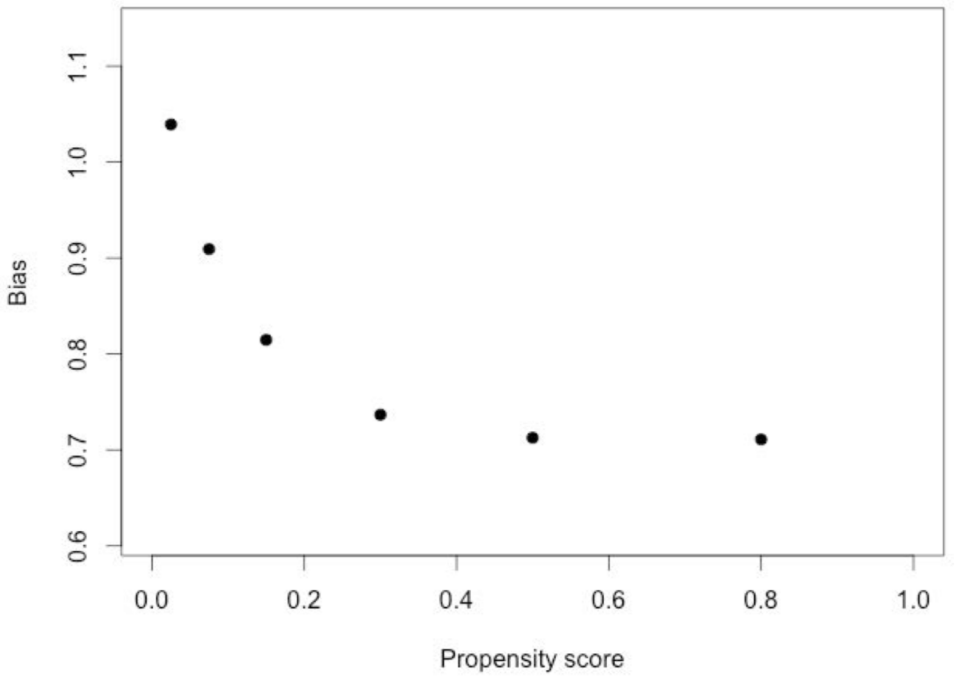
The intuition behind Breen, Choi, and Holm (2015) is that both baseline bias and the hypothesized heterogeneous effect are functions of the propensity score.

Therefore, we cannot disentable between them when conditional ignorability does not hold (or without further parametric assumptions)



Bias as function of propensity score

Constructing strata of varying width (as in the original paper), the authors are able to reproduce the negative selection pattern shown in Brand and Xie (2010)



HTE under unobserved selection

At this point, we can't make any progress unless we are willing to add some assumptions.

One approach, with a long tradition in Economics, is to directly model the selection process, allowing the agents to anticipate their potential outcomes under different treatment conditions. This is known as the Roy model.

A lot of this work has been done by James Heckman (1974, 2005; Heckman, Urzua and Vytlacil, 2006), with recent extensions by Zhou and Xie (2016, 2018)

A good general introduction to this framework can be found in Heckman (2005) The scientific model of causality

Generalized Roy model

To proceed, we have to define the potential outcomes and the selection process as follows:

$$
\begin{aligned}
Y_0 &= \mu_0(X) + \epsilon \\
Y_1 &= \mu_1(X) + \epsilon + \eta
\end{aligned}
$$

$$
\begin{aligned}
D^* &= \mu_D(Z) - V \\
D^* &= (P(Z) - U > 0) \\
D &= \mathbb{I}(D^* > 0)
\end{aligned}
$$

Where Z and V represent observed and unobserved variables respectively, X is a subset of Z, and $Z \setminus X$ are instruments. Finally, U corresponds to quantiles of V given X.

Generalized Roy model

We are assuming (ϵ, η, V) are independent of Z given X, but V could be arbitrarily correlated with (η, ϵ).

Along with monotonicity, these are the same conditions to identify the LATE we have already discussed.

In short, this means that subjects can sort themselves into the treatment based on expected gains (their anticipated $Y_{1i} - Y_{0i}$).

- In the classic Roy model, $D^* = Y_1 - Y_0$ (so perfect knowledge about the effect of the treatment for oneself).
- In the generalized Roy model, the selection function could discount for the cost of changing the treatment status $(Y_1 - Y_0 - C)$, or be conditional on the information that the subjects have *ex ante* $(Y_1 - Y_0 | \mathcal{I})$, among other scenarios.

Marginal Treatment Effect (MTE)

We can then define the Marginal Treatment Effect (MTE) as the fundamental quantity in this framework:

$$\begin{aligned}\text{MTE}(x, u) &= E[Y_1 - Y_0 | X = x, U = u] \\ &= E[\mu_1(X) - \mu_0(X) + \eta | X = x, U = u] \\ &= \mu_1(X) - \mu_0(X) + E[\eta | X = x, U = u]\end{aligned}$$

Interpretation of the MTE as the effect for those indifferent between participating on not, if the instrument is set to $P(Z) = u_D$

Integrating over the distribution of X and U, we can construct other known causal quantities: ATE, ATT, ATC, and new quantities as the PRTE (Policy Relevant Treatment Effect).

Estimation of MTE

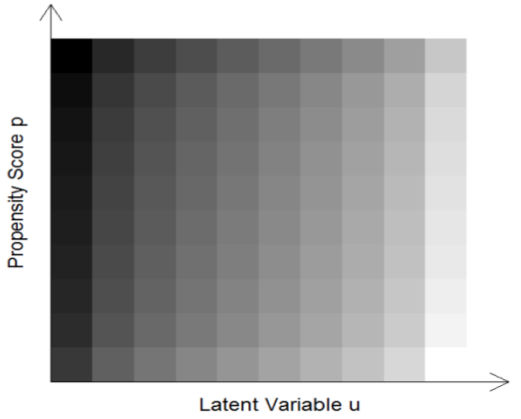
Estimation could be parametric or semi-parametric, and usually involves more stringent assumptions about the error terms (such as being additively separable).

The traditional sample selection of Heckman (1978), aka the "normal switching regression model", is a special case of the MTE under (ϵ, η, V) assumed to be jointly normal with zero means.

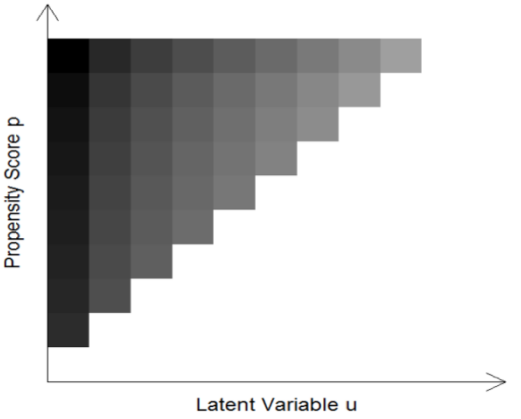
Critical conditions for estimating the MTE are continuous instruments and common support between $P(Z)$ and the IV.

Formal identification results in the control function and LIV setting given in Heckman-Vytlacil (1999, 2001, 2005) and Heckman, Urzua and Vytlacil (2006).

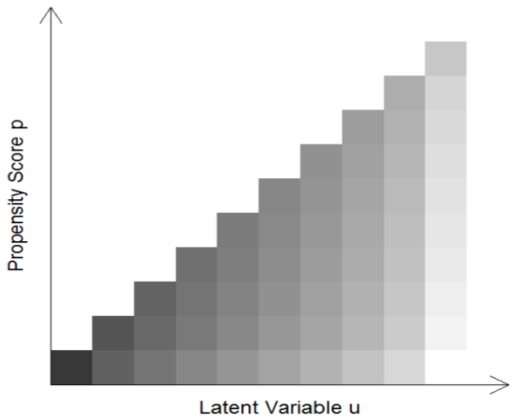
HTE as a function of MTE



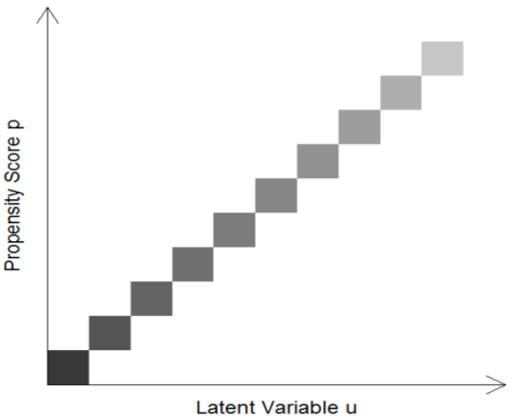
(a) Population



(b) Treated Units

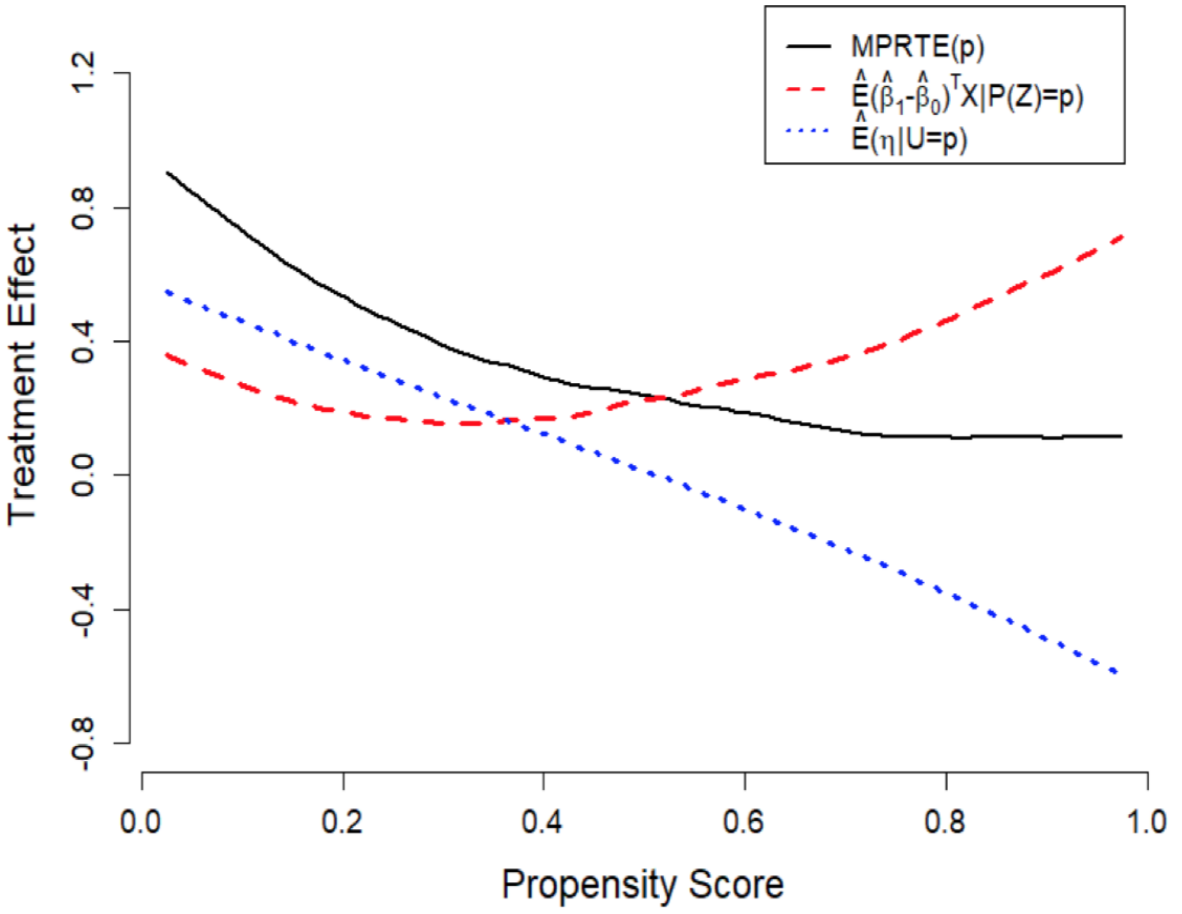


(c) Untreated Units



(d) Marginal Units

Negative selection re-visited



HTE under unobserved selection

Conceptually, what we were trying to achieve is to adjust for both observed and unobserved selection when $(Y_1, Y_0) \perp\!\!\!\perp D|X$ doesn't hold, so we would have $(Y_1, Y_0) \perp\!\!\!\perp D|X, U$

However, nothing is free in causal inference! So we have to make further (untestable) assumptions and embrace some parametrization in order to identify such models.

We can still have good reasons to do so, particularly in settings in which sorting into the treatment is plausible. Then, modeling U would likely be a better approximation than assuming conditional ignorability.

Some references

Brand, and Simon-Thomas (2013) "Causal Effect Heterogeneity." In Handbook on Causal Analysis for Social Research. Springer

Breen, Choi, and Holm (2015) "Heterogeneous Causal Effects and Sample Selection Bias." Sociological Science.

Heckman (1974) "Shadow Prices, Market Wages, and Labor Supply," Econometrica, 42, issue 4, p. 679-94

Heckman (1978) "Dummy Endogenous Variables in a Simultaneous Equation System," Econometrica Vol. 46, No. 4 (Jul., 1978), pp. 931-959

Heckman (2005) "The Scientific Model of Causality." Sociological Methodology, 35: 1-97.

Heckman, Urzua, and Vytlacil (2006) "Understanding Instrumental Variables in Models with Essential Heterogeneity," The Review of Economics and Statistics, MIT Press, vol. 88(3): 389-432

Heckman, and Vytlacil (1999) "Local instrumental variables and latent variable models for identifying and bounding treatment effects," PNAS April 13, 1999 96 (8) 4730-4734

Some references

Heckman, and Vytlacil (2001) "Policy-Relevant Treatment Effects," American Economic Review, 91 (2): 107-111.

Heckman, and Vytlacil (2005) "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," Volume73, Issue3 May 2005: 669-738

Xie, and Brand (2010) "Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." American Sociological Review, 75(2): 273-302.

Xie, Brand, and Jann (2012) "Estimating Heterogeneous Treatment Effects with Observational Data." Sociological Methodology, 42(1): 314-347.

Zhou, and Xie (2016) "Propensity Score-based Methods Versus MTE-based Methods in Causal Inference: Identification, Estimation, and Application." Sociological Methods & Research, 45(1): 3-40.

Zhou, and Xie (2018) "Heterogeneous Treatment Effects in the Presence of Self-Selection: A Propensity Score Perspective." Working Paper.