An overview of (some) statistical methods to estimate inequality of opportunity

Winter School on Inequality and Collective Welfare Theory (IT17)

Paolo Brunori (LSE & University of Florence)

LSE International Inequalities Institute

Nicholas Kaldor's Stylized facts

> *"facts as recorded by statisticians are always subject to numerous snags and qualifications, and for that reason are incapable of being summarized"*

According to Kaldor economists should work from

> *"a stylized view of the facts [and] concentrate on broad tendencies, ignoring individual detail"*

Kaldor, 1961 p. 2

Definitions

- *EOP* = equality of opportunity (free to choose from same opportunity set);
- *IOP* = a summary index measuring to what extent *EOP* is violated;
- C = potential sources of *IOP*.

$I\hat{O}P$: 'This is not a causal identification'

- Can we estimate the effect of circumstances?
- Attempts: sibling correlation, experiments, quasi-experiments;
- Partial and limited external validity;
- I am not sure (even theoretically) possible for the cumulative effect of all circumstances.

Understanding the role of circumstances and 'choices'

This project turned out to be like peeling away layers of an onion. [...] There is no way to separate a person from the accumulated effects of her interactions with her circumstances, including her opportunities, because the product of those accumulated interactions is the person.

Fishkin, 2014 p. 64

Model assumptions (Roemer, 1998)

- Outcome (y) produces same welfare for all individuals;
- Agreement about a list of circumstances that should not affect the outcome (C);
- Roemer suggests 'any variable outside individual control';
- In practice: any observable exogenous variable.

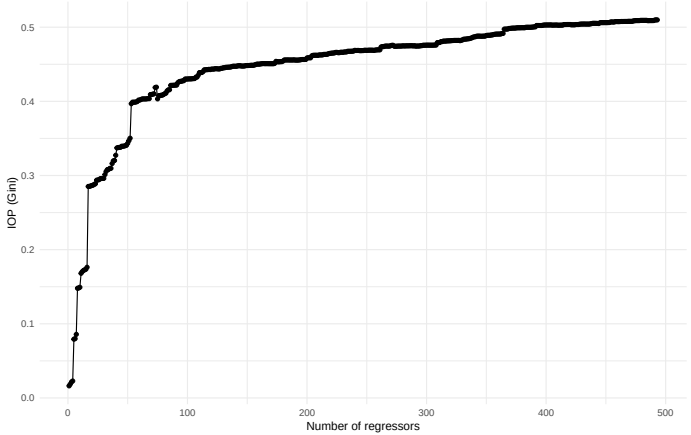
What does 'affect' mean?

- Degree of statistical association (Checchi & Peragine, 2010; Ferreira & Gignoux, 2011);
- find a sufficiently reach data source;
- $I\hat{O}P = I(\hat{y})$, where $\hat{y} = \hat{f}(\mathrm{C})$;
- interpreted as lower bound.

A lower bound of what?

- If IOP is not causally defined $I\hat{O}P$ cannot be a lower bound;
- But even if assumptions for causal interpretation hold;
- Still $I\hat{O}P > IOP$ if the model is sufficiently overfitted.
- $I\hat{O}P$ is always interpretable conditional on data used and model specified.

OLS-based $I\hat{O}P$ in South Africa



International Inequalities Institute

Source: Brunori, Ferreira, Salas-Rojo (2023), Data: NIDS, 2017.

Robust $I\hat{O}P$

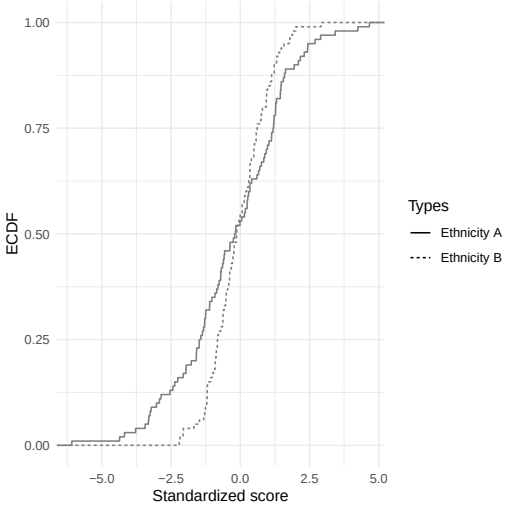
$\rightarrow$ find an approach to make comparisons across time and space meaningful;

- A candidate: "to what extent C covary with y?" $\rightarrow$ "to what extent C can predict y?"
- $I\hat{O}P$ is still dependent on observable C;
- But we have a criterion to select $f()$.

Do not rush

- If *IOP* measurement is a prediction problem $\rightarrow$ use supervised ML!
- But depending on the data, accuracy-interpretability trade-off cab be an issue;
- Your definition of *EOP* may be not equality in $\mathbb{E}[y|\mathrm{C}]$.

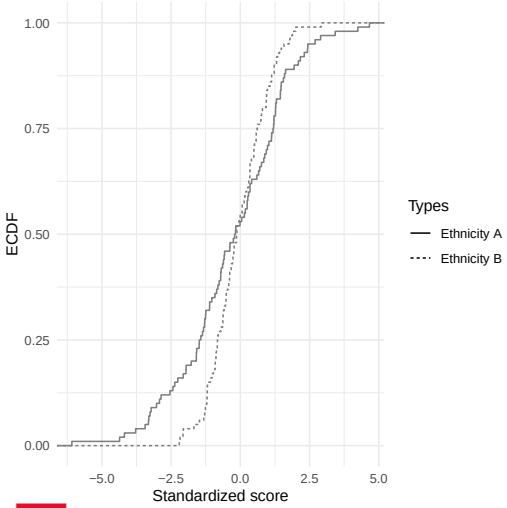
Roemer 'ex-post' EOP (1998)

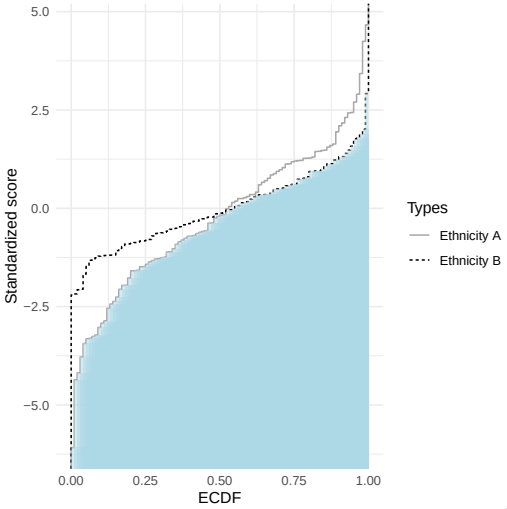


International Inequalities Institute

Source:

Roemer 'ex-post' EOP (1998)

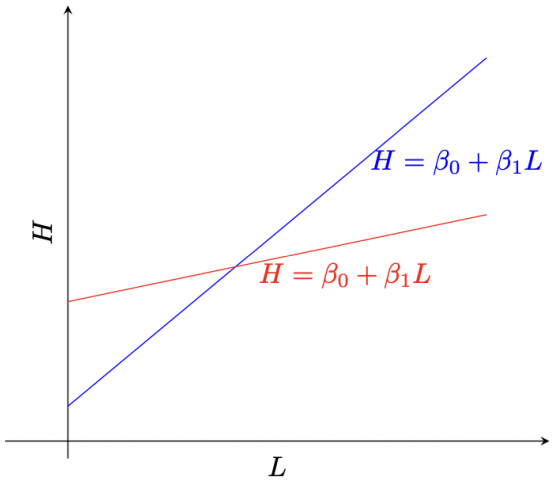




LSE International Inequalities Institute

Source:

Fleurbaey and Shokkaert health 'fairness gap' (2009)



LSE International Inequalities Institute

Roemerian types

- Adopting ex-post *IOP* and fairness gaps make natural to define roemerian types;
- Then type-specific distributions can be estimated;
- We consider two approaches from unsupervised and supervised ML;
  1. latent class model (Li Donni et al., 2015)
  2. tree-based methods (Zeileis, Hothorn, Hornik, 2023)

LCA $\in$ Latent variable models

	Manifest variables	
Latent variables	**Continuous**	**Categorical**
Continuous	Factor analysis	Item response theory
Categorical	Latent profile analysis	Latent class analysis

Source: Wikipedia

LCA assumptions

1. Individuals belonging to a given class have same probability to have a particular response to all manifest variables;

2. **Local independence**: manifest variables are independently distributed conditional on class membership.

Correlated C (local dependence)

	Occupation "High"	Occupation "Low"
Education "High"	260	140
Education "Low"	240	360

$$\text{P(occupation=high} \mid \text{education=high)} \neq \text{P(occupation} = \text{High} \mid \text{education} = \text{low)}$$
$$260/400 \neq 240/500$$

Local independence

Type A	Occupation "High"	Occupation "Low"
Education "High"	240	60
Education "Low"	160	40

Type B	Occupation "High"	Occupation "Low"
Education "High"	20	80
Education "Low"	80	320

LCA models

- LCA assigns a probability to type membership maximizing local independence;
- Fixing the number of types probabilities can be estimated by maximum likelihood;
- Individuals are assignment to type based on max probability;
- Number fo type selected by panalized goodness of fit (e.g. BIC).

Latent types pros

- We all know that Roemerian types do not exist;
- Provide a criterion to select $f()$;
- Latent types are interesting to study.

LCA item response probabilities

Table 3: Latent type membership by mother education (Portugal -- 3 latent types)

Mother education	Type 1	Type 2	Type 3
Illiterate	11.50%	4.10%	84.40%
Low	74.90%	6.70%	18.40%
Medium	25.80%	71.10%	3.10%
High	0.00%	100.00%	0.00%

Source: EU-SILC, 2011

Source: Brunori, Trannoy, Guidi (2021)

Latent types cons

- LCA minimize covariance of C, does not maximize $COV(y, \mathrm{C})$ (conservative $I\hat{O}P$);
- All categories of all C are used;
- LCA are data-espensive (the number of parameters (N) is growing with number of latent types (L), number of circumstances (C), and number of categories of variable c (R_c)):

$$N = \sum_{c=1}^{\mathrm{C}} (R_c - 1)(L - 1)$$

- Penalized likelihood criteria will favour parsimonious models (conservative $I\hat{O}P$);

Possible developments

- Find a method to pre-select 'useful' C;
- Find a method to trade-off local independence and need to explain $COV(y, \mathrm{C})$;
- Explore the use of other latent variable models when some C is continuous.

Using Latent types

- **Interesting application**: Carrieri, Davillas, Jones (2020) 'A latent class approach to inequity in health using biomarker data', Health Economics;

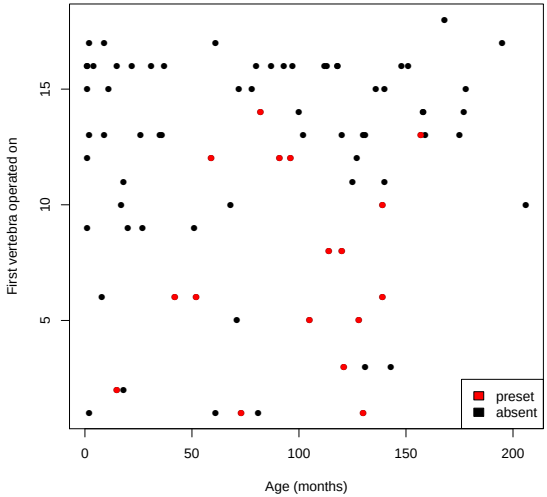
- **Understanding LCA**: Collins and Lanza (2009) 'Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences';

- **Implementation in R**: Linzer and Lewis (2011) 'poLCA: An R Package for Polytomous Variable Latent Class Analysis', Journal of Statistical Software.

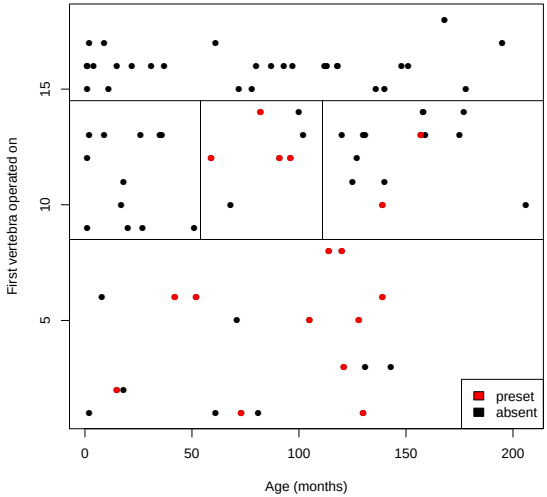
Tree-based types

- Supervised ML will directly learn about $COV(y, \mathrm{C})$ from data;
- Need to identify types $\rightarrow$ tree-based methods.

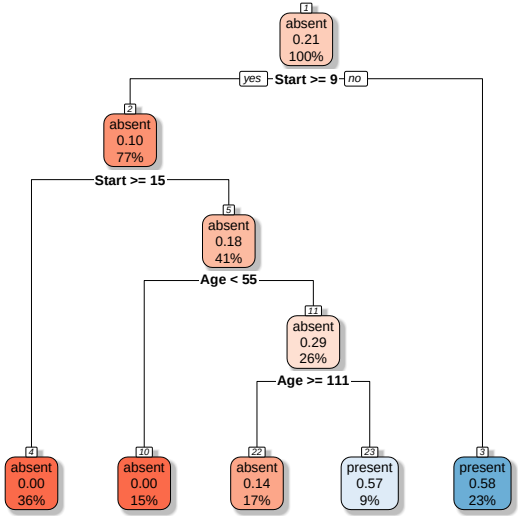
Kyphosis after pediatric spinal surgery



Kyphosis after pediatric spinal surgery



Kyphosis after pediatric spinal surgery

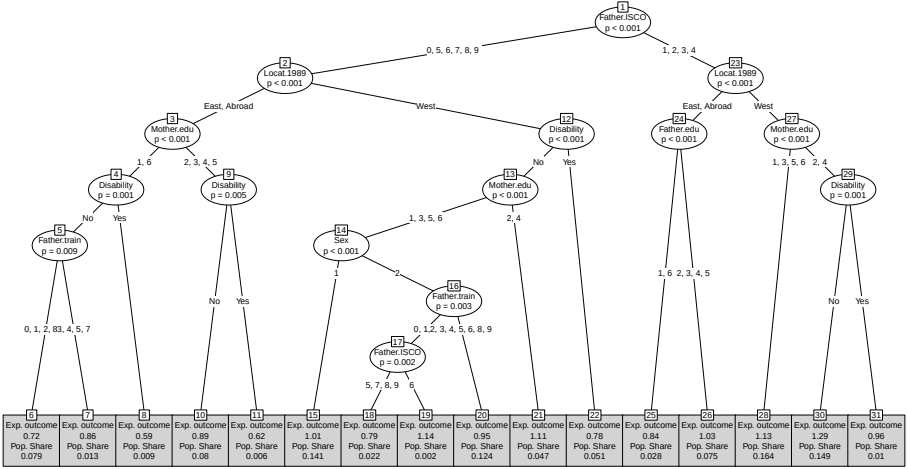


International Inequalities Institute

Conditional inference trees (Hothorn et al., 2006)

- choose a confidence level (1-α);
- $\forall c$ test the null hypothesis of independence: $H^c = CORR(y, c) = 0$, $\forall c \in$ C;
- if no (adjusted) p-value $< \alpha \rightarrow$ exit the algorithm;
- select the variable, $c^\star$, with the lowest p-value;
- test the discrepancy between subsamples for each possible binary partition based on $c^\star$;
- split the sample by selecting the splitting point that yields the lowest p-value;
- repeat the algorithm for each of the resulting subsample.

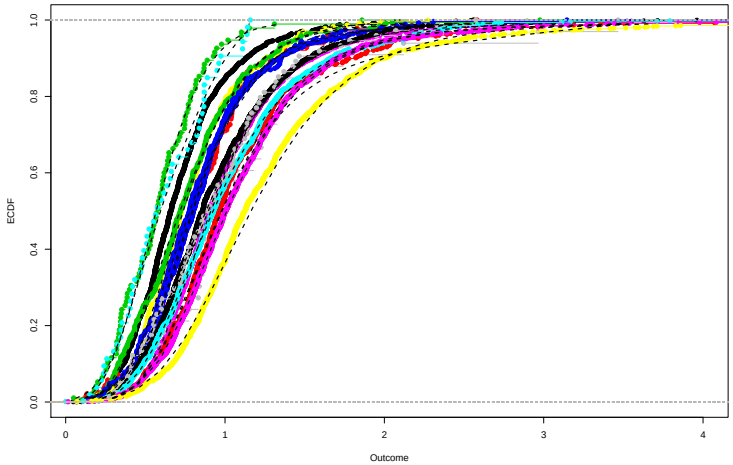
Ctree-based types in Germany



International Inequalities Institute

Source: Brunori and Nidhöfer (2020), Data: SOEP, 2016.

Ctree-based types in Germany



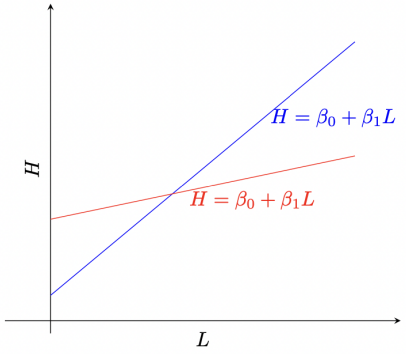
International Inequalities Institute

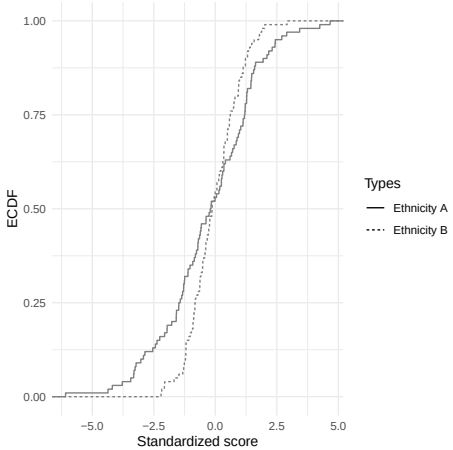
Source: Brunori and Nidhöfer (2020), Data: SOEP 2016.

Tree-based types

- Ctree splitting detects heterogeneous conditional expectations;
- The partition is consistent with EOP as nonpredictability (Brunori, Hufe, Mahler, 2023);
- May fail to detect violations of other EOP definitions (e.g. ex-post IOP or fairness gaps).

Same mean outcome, different opportunities





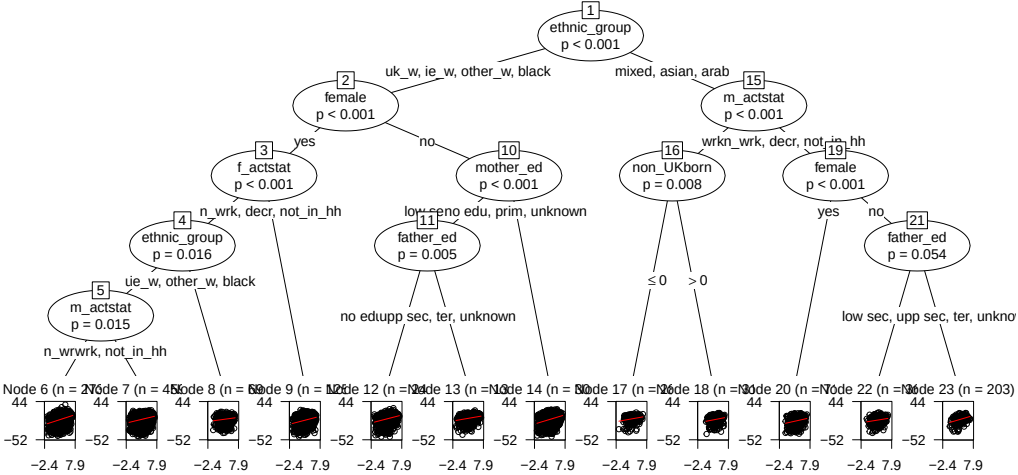
Model-based trees: general approach

- Select your definition of opportunity (e.g. $F(y)$ or (β_0, β_1));
- Define a set of parameters that approximate opportunity;
- Test for the instability of parameters across potential subgroups;
- Partition the sample when you can reject the null hypothesis of stability with sufficient confidence.

Model-based trees (Zeileis et al., 2023)

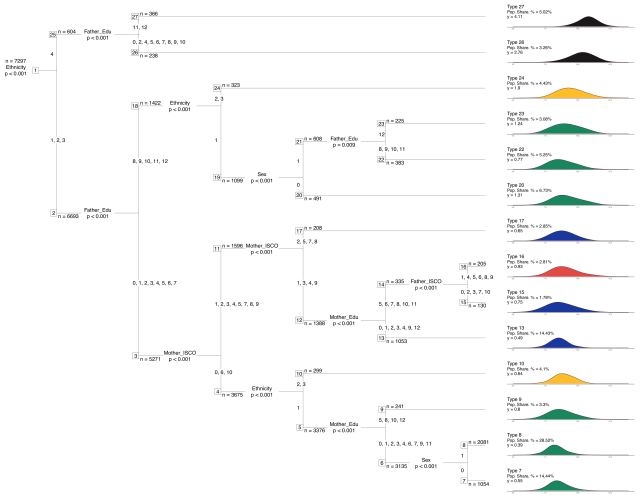
1. set a confidence level $(1 - \alpha)$;
2. fit the model in the entire sample ($h = \beta_0 + \beta_1 E + u$);
3. perform a M-fluctuation test on the stability of the parameters depending $c \in C$;
4. If H_0 is rejected a split is performed, otherwise the algorithm stops;
5. repeat 2-5 on the resulting sub-samples.

Unfair inequality in health in UK with MOB-tree



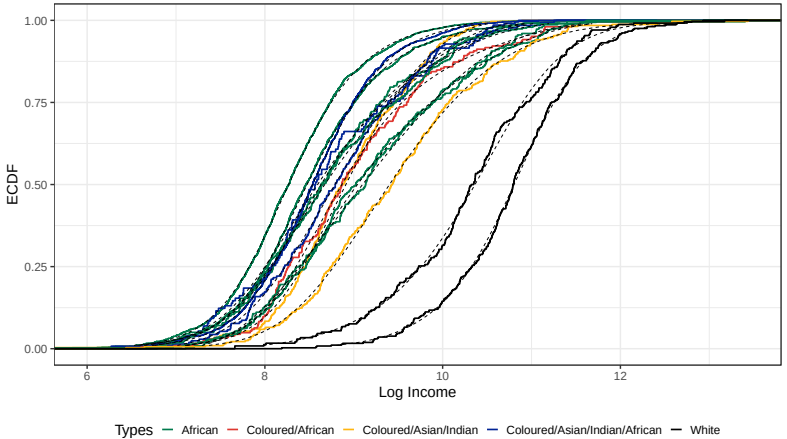
Source: *Brunori, Davillas, Jones, Scarchilli (2022), Data: UKHLS.*

Ex-post IOP in South Africa with transformation tree



Source: Brunori, Ferreira, Salas-Rojo, 2023, Data: NIDS, 2017.

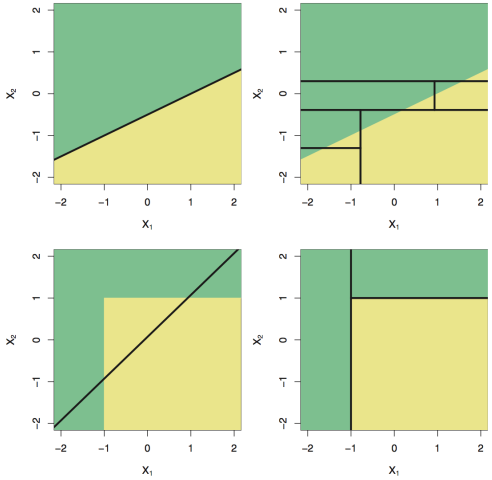
Ex-post IOP in South Africa with transformation tree



International Inequalities Institute

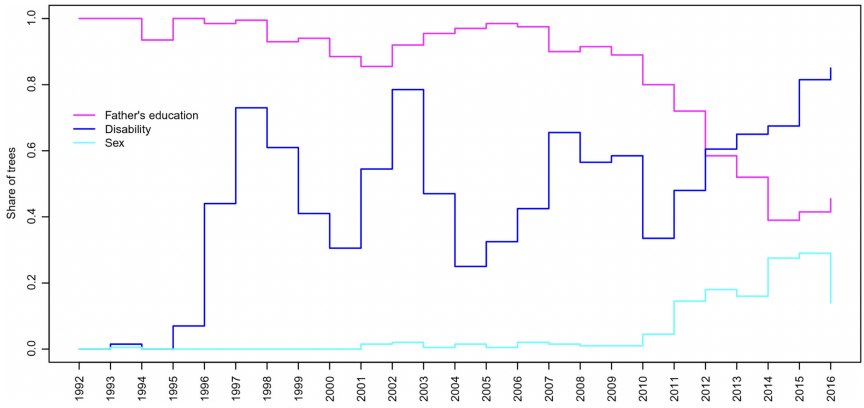
Source: Brunori, Ferreira, Salas-Rojo, 2023, Data: NIDS, 2017.

Tree-based types cons: linear DGP



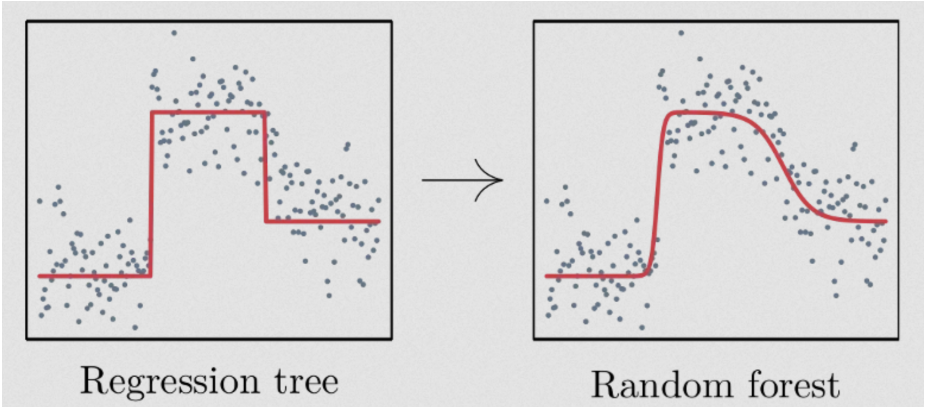
Source: James et al. (2013)

Tree-based types cons: instability



Source: Brunori and Nidhöfer (2020), Data: SOEP 1992-2016.

Forests



Source: Hothorn et al. (2018)

Bagging trees

- Bagging trees in forest is my preferred option when EOP is defined in terms of conditional expectations;

- Performs better, still interpretable, makes explicit the very essence of what we (do not know) about the DGP;

- But (open issue) it dramatically reduces $I\hat{O}P$ when EOP is defined with references to conditional distributions.

Possible future developments (ongoing)

- Modify trees to reduce their instability (Moramarco et al., 2024);
- Assess the power of the empirical exercize (ibid.);
- Practical method to adjust for sample size (Andreoli and Van Kerm, 2024);
- Use ML method that obtain prediction by both binary splitting and additive models (Annaelena Valentini today's later presentation);
- Debias $I\hat{O}P$ obtained with ML (Escanciano and Terschuur, 2023);
- Introduce some structure to a flexible $f()$ (yesterday's presentation by Francesca Subioli)!

Possible future developments

- How robust are hour estimates to missing (C)? Should we (and how) impute?
- How should we approach increasingly available administrative and genetic data?