# The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach

ZEW Research Seminar

17.01.2019

Paolo Brunori

*University of Florence*

Guido Neidhöfer

*ZEW*

# Motivation

- equality of opportunity: a very successful political ideal

# Margaret Thatcher

*First, that the pursuit of equality itself is a mirage. What's more desirable and more practicable than the pursuit of equality is the pursuit of equality of opportunity.*

Speech to the Institute of SocioEconomic Studies
New York, September 15, 1975

# Raul Castro

*Socialismo significa justicia social e igualdad, pero igualdad de derechos, de oportunidades, no de ingresos.*

Speech at the Asamblea Nacional del Poder Popular
La Habana, July 11, 2008

# Motivation, cnt.

- equality of opportunity (EOp): a very successful political ideal

- two reasons:

  1. EOp = equality + freedom

  2. EOp is sufficiently vague

# Literature

- a "third generation" paper on inequality of opportunity:
- first generation (theory): moral philosophers and welfare economists Rawls (1971), Dworkin (1981), Arneson (1989) and Cohen (1989), Roemer (1998);
- second generation (measurement): Lefranc et al. (2009), Checchi and Peragine (2010), Bourguignon et al. (2007), Ferreira and Gignoux (2011);
- third generation (econometric specification): Li Donni et al. (2015), Brunori, Hufe and Mahler (2018).

# Roemer's Model

$$y_i = g(C_i, e_i)$$

- $y_i$: individual's $i$ outcome;
- $C_i$: circumstances beyond individual control;
- $e_i$: effort.

# Types and effort tranches

- Romerian type: set of individuals sharing exactly the same circumstances;

- effort tranche: set of individuals exerting the same effort;

- no random component:
  same type and same tranche $\rightarrow$ same outcome;

- there is equality of opportunity if and only if:

$$e_i = e_j \iff y_i = y_j \ , \ \forall i, j \in 1, ..., n$$

$\Rightarrow$ IOP = within-tranche inequality.

# Equality of opportunity: weaker definition

- Van de Gaer (1993): a weaker principle of equal opportunity;
- type outcome distributions = opportunity sets;
- IOP = inequality between opportunity set values;
- utilitarian approach: IOP = between-type inequality.

# Equality of opportunity: weaker definition, cnt

- Van de gaer's approach is the most popular in empirical analysis;
- World Bank Human Opportunity Index (Barros et al, 2008);
- measures obtained with the two approaches differ conceptually and empirically;
- between-type approach: no need to measure effort.

# Effort identification

- effort: observable and not observable choices;
- Roemer's identification strategy, two assumptions:
  1. monotonicity: $\frac{\partial g}{\partial e} \geq 0$
  2. orthogonality: $e \perp\!\!\!\perp C$
- degree of effort = quantile of the type-specific outcome distribution;

# 3-step estimation

1. identification of Romerian types;

   $->$ (weaker) IOP = between-type inequality

2. measurement of degree of effort exerted;

3. (Roemer) IOP = within-tranche inequality

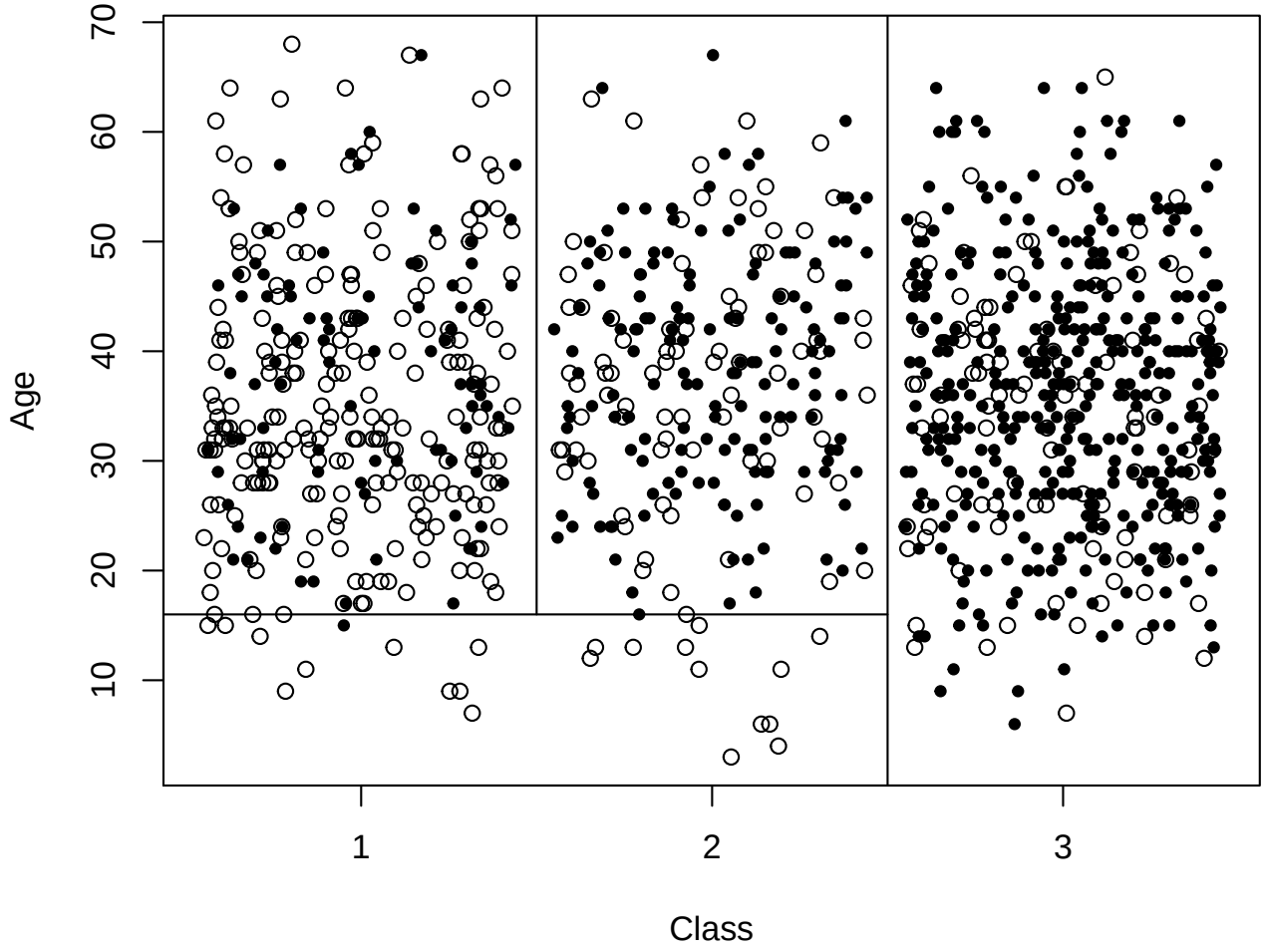
# Roemerian types

- first generation papers tried a direct implementation of Roemer's theory;
- unobservable circumstances (downward bias);
- sparsely populated types (upward bias);
- the trade-off is now solved maximizing out-of-sample IOP.

# Romerian types, cnt

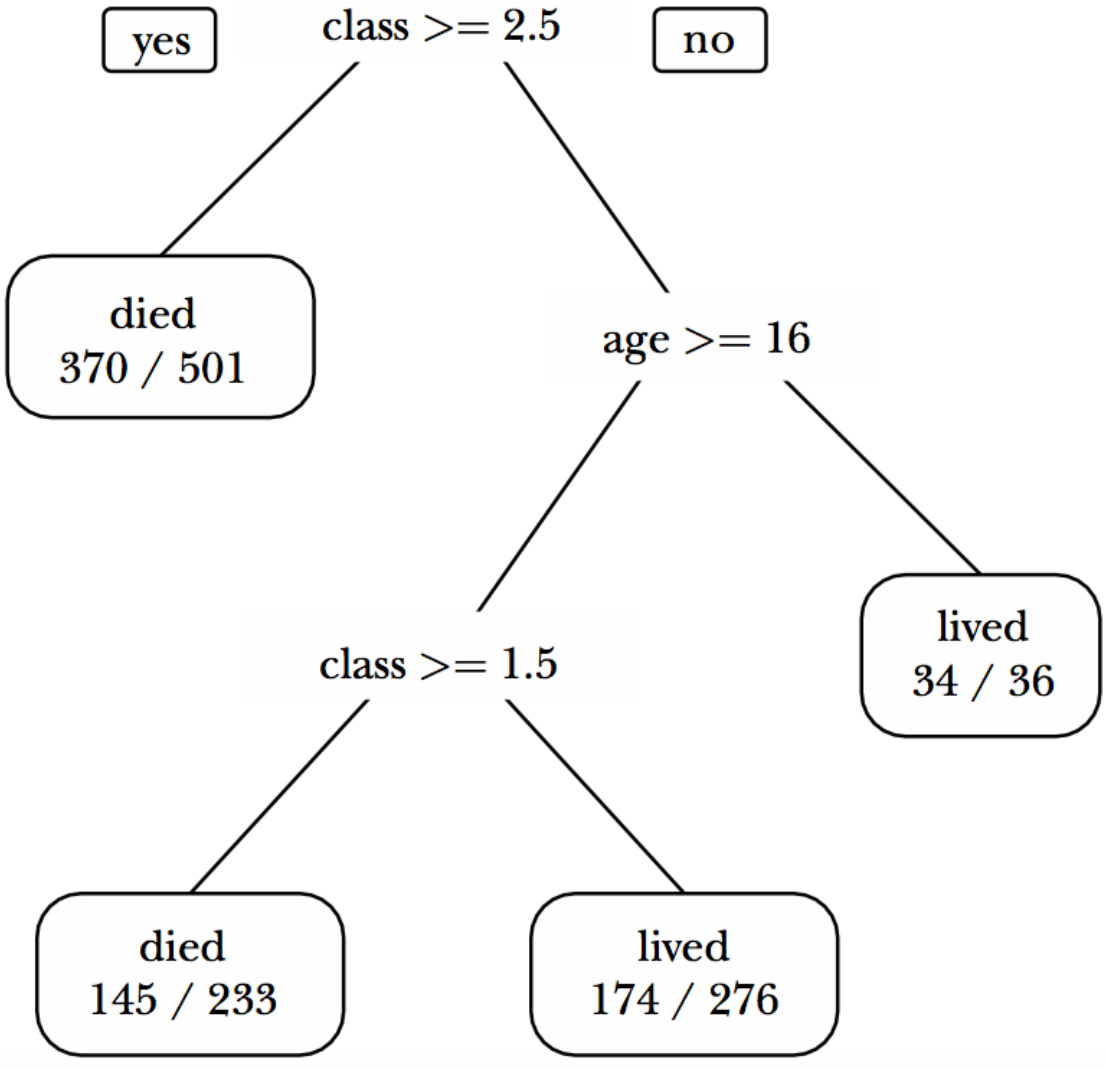
- we use regression tree to identify types;
- a tree is an algorithm to predict a dependent variable based on observable predictors (Morgan and Sonquist,1963; Breiman et al.,1984)
- the population is divided into non-overlapping subgroups
- prediction of each observation is the the mean value of the dependent variable in the group

# What is a tree? cnt.




*Source: Varian, 2014*

# What is a tree? cnt.




*Source: Varian, 2014*

# What is a tree? cnt.

- overfitted models explain perfectly in-sample
- but perform poorly out-of-sample (low out-of-sample IOP)
- different solutions lead to different type of trees

# Conditional inference trees

- we use *conditional inference trees* (Hothorn et al., 2006);
- splitting are based on a sequence of statistical test;
- Brunori, Hufe, Mahler (2018): highly interpretable and outperform standard methods to identify types.

# The algorithm

- choose $\alpha$
- $\forall p$ test the null hypothesis of independence:
  $H^{C_p} = D(Y|Cp) = D(Y)$, $\forall C_p \in \mathbf{C}$
- if no (adjusted) p-value $< \alpha \rightarrow$ exit the algorithm
- select the variable, $C^\star$, with the lowest p-value
- test the discrepancy between the 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

# Effort

- recall: IOP quantifies to what extent individuals exerting the same degree of effort obtain the same outcome;

- standard approach: choose an arbitrary number of quantiles;

- low efficiency and limited comparability across studies.

# Bernstein polynomials

- violation of the EOP principle: how far is income of individual at the j-th quantile of his type income distribution from what expected?

- approximate the ECDF with a polynomial;

- for any quantile $\pi \in [0, 1]$ we can predict the expected outcome in all types;

- we use Bernstein polynomials.

# Bernstein polynomials

- introduced in 1912 by Sergei Bernstein
- today: mathematical basis for curves' approximation in computer graphics
- outperform competitors (kernel estimators) in approximating distribution functions (Leblanc, 2012)

# Bernstein polynomial of degree 4

$$B_4(x) = \sum_{v=0}^{4} \beta_v b_{v,4}$$

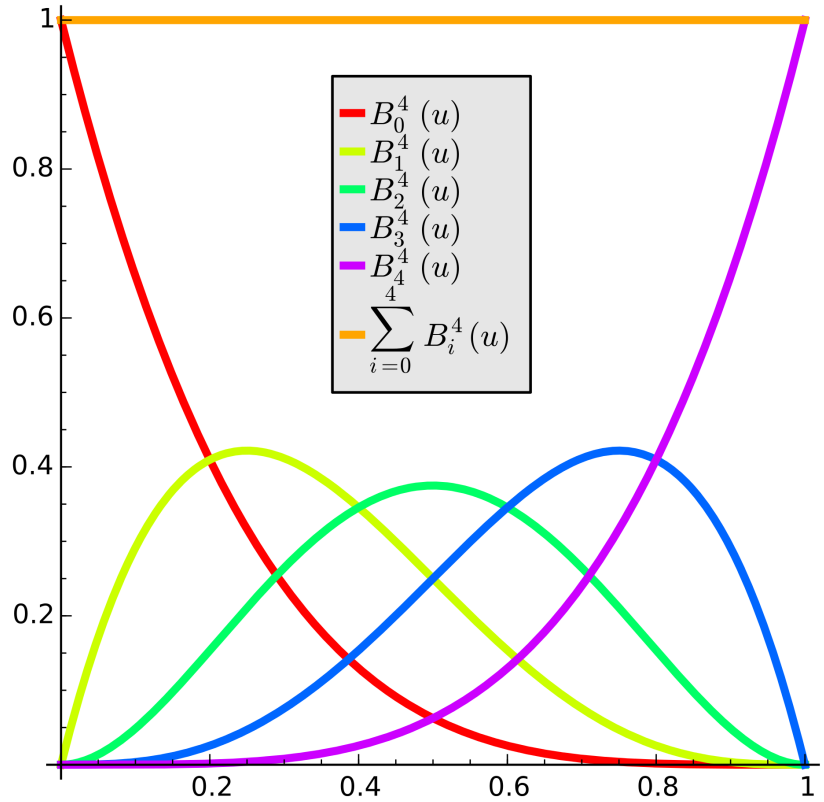
where $\beta_v b_{v,4}$ is the v-th Bernstein basis polynomial

$$b_{v,k} = \binom{k}{v} x^v (1-x)^{k-v}$$

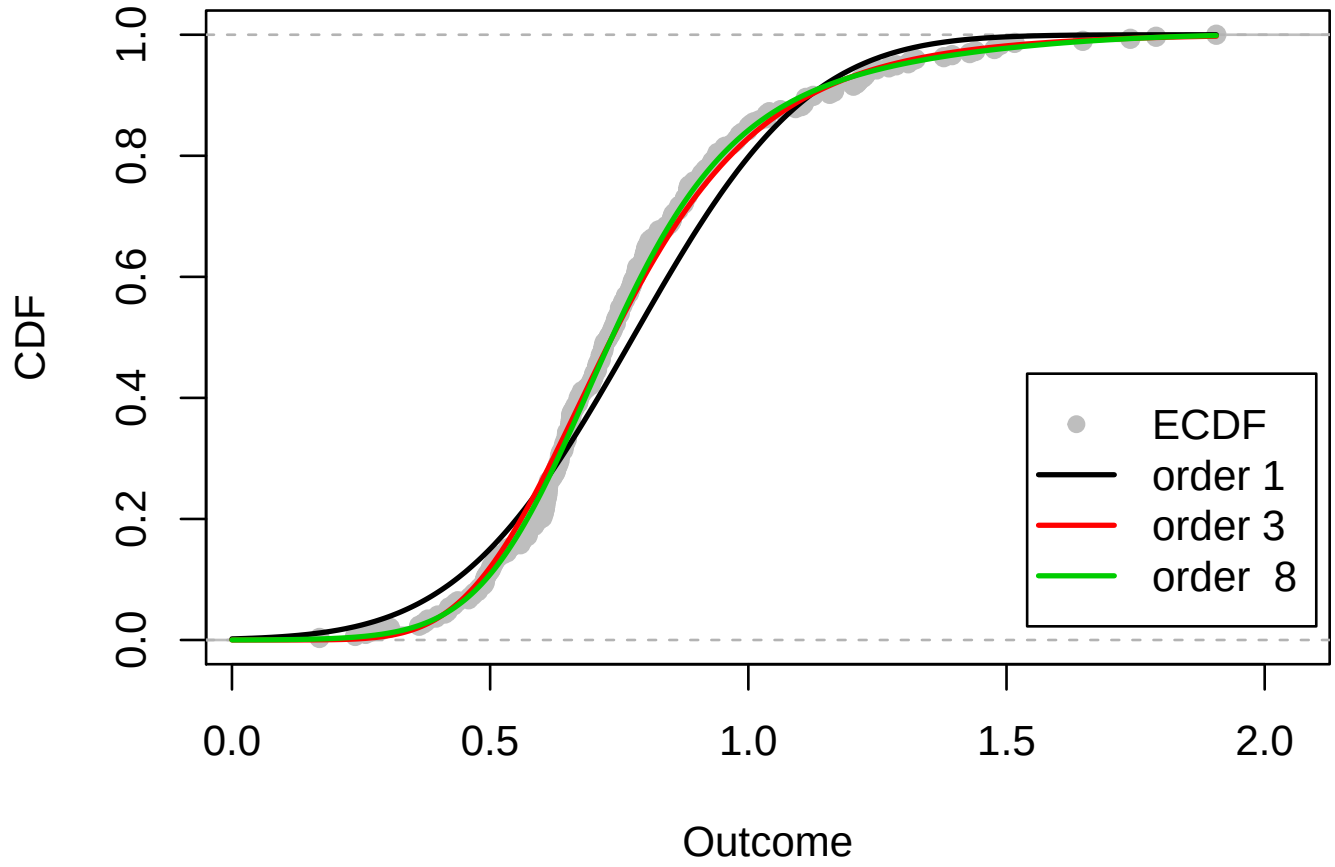
example

$$\begin{aligned} b_{0,4} &= (1-x)^4 \\ b_{1,4} &= 4x(1-x)^3 \\ b_{2,4} &= 6x^2(1-x)^2 \\ b_{3,4} &= 4x^3(1-x) \\ b_{4,4} &= x^4 \end{aligned}$$

# Bernstein polynomials, cnt



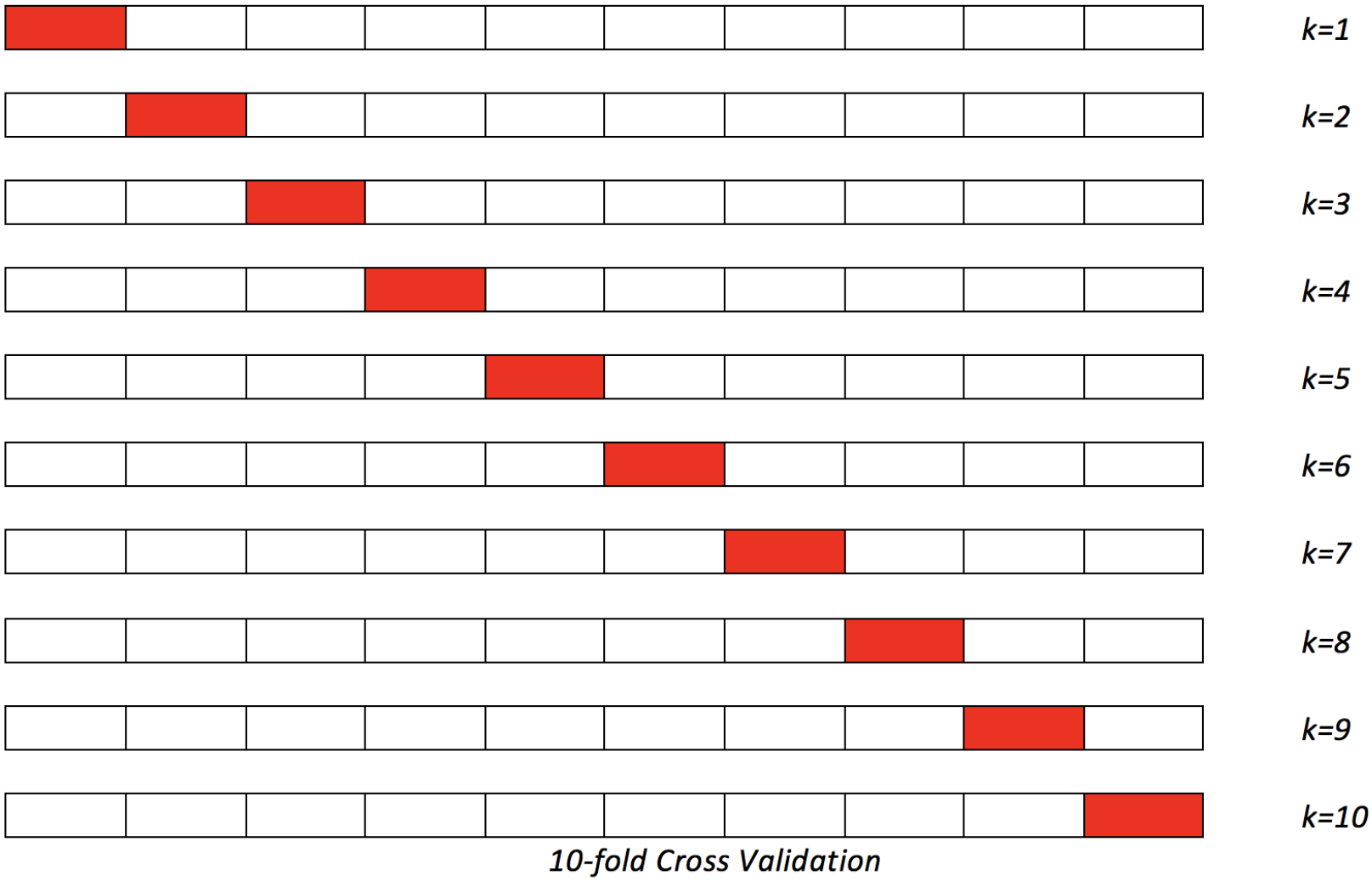
# ECDF approximation by Bernstein polynomials



# Choice of the polynomial's degree

- the polynomial is estimated with the *mlt* algorithm written by Hothorn (2018);
- out-of-sample log-likelihood to select the most appropriate order of the polynomial;
- out-of-sample log-likelihood is estimated by 5-fold cross validation;

# k-fold cross validation



# IOP estimation

- Knowing the shaper of all type-specific distribution functions we can estimate the distribution of 'unfair' inequality
- $IOP = Gini\left(\frac{y_i}{\mu_j}\right)$, $\mu_j$ expected outcome at percentile $j$;
- no longer need to choose a particular number of effort quantiles;
- number of quantiles varies to maximize estimate reliability.

# Data

- SOEP (v33) including all subsamples apart from the refugee samples;
- adult individuals (30-60);
- $y$ = age-adjusted household equivalent disposable income;

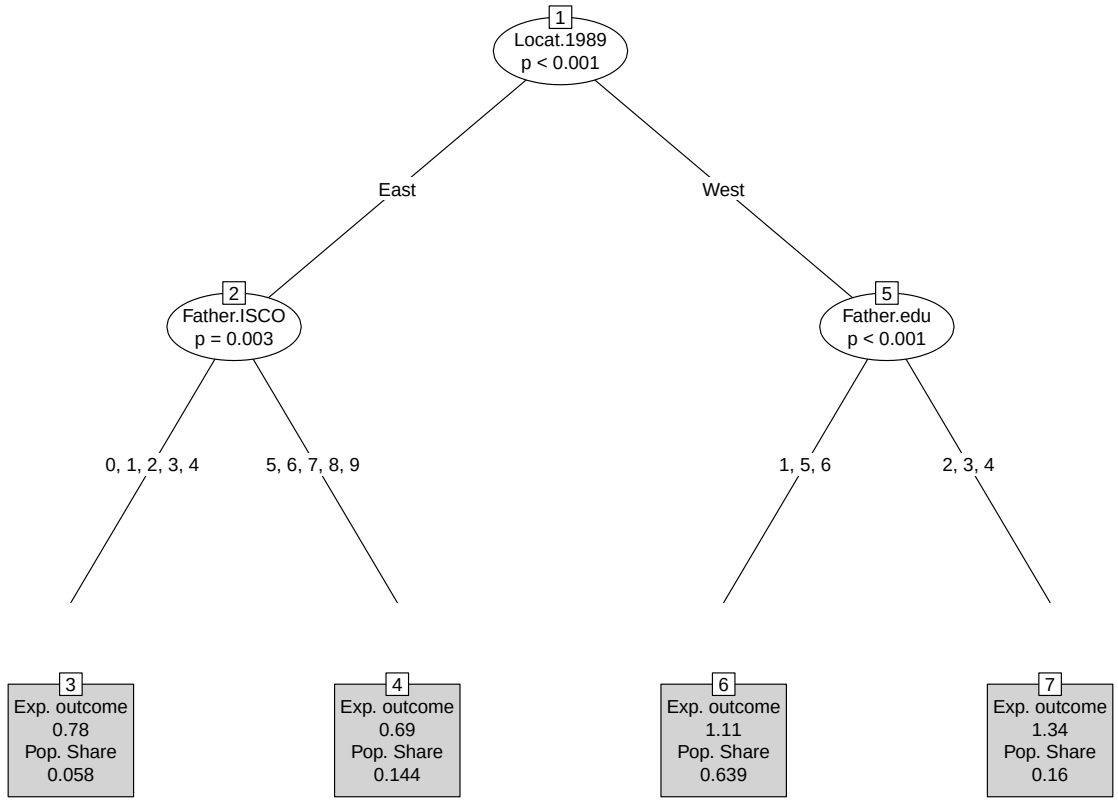
# Data, cnt.

- SOEP provides comprehensive information about circumstances beyond individual control;

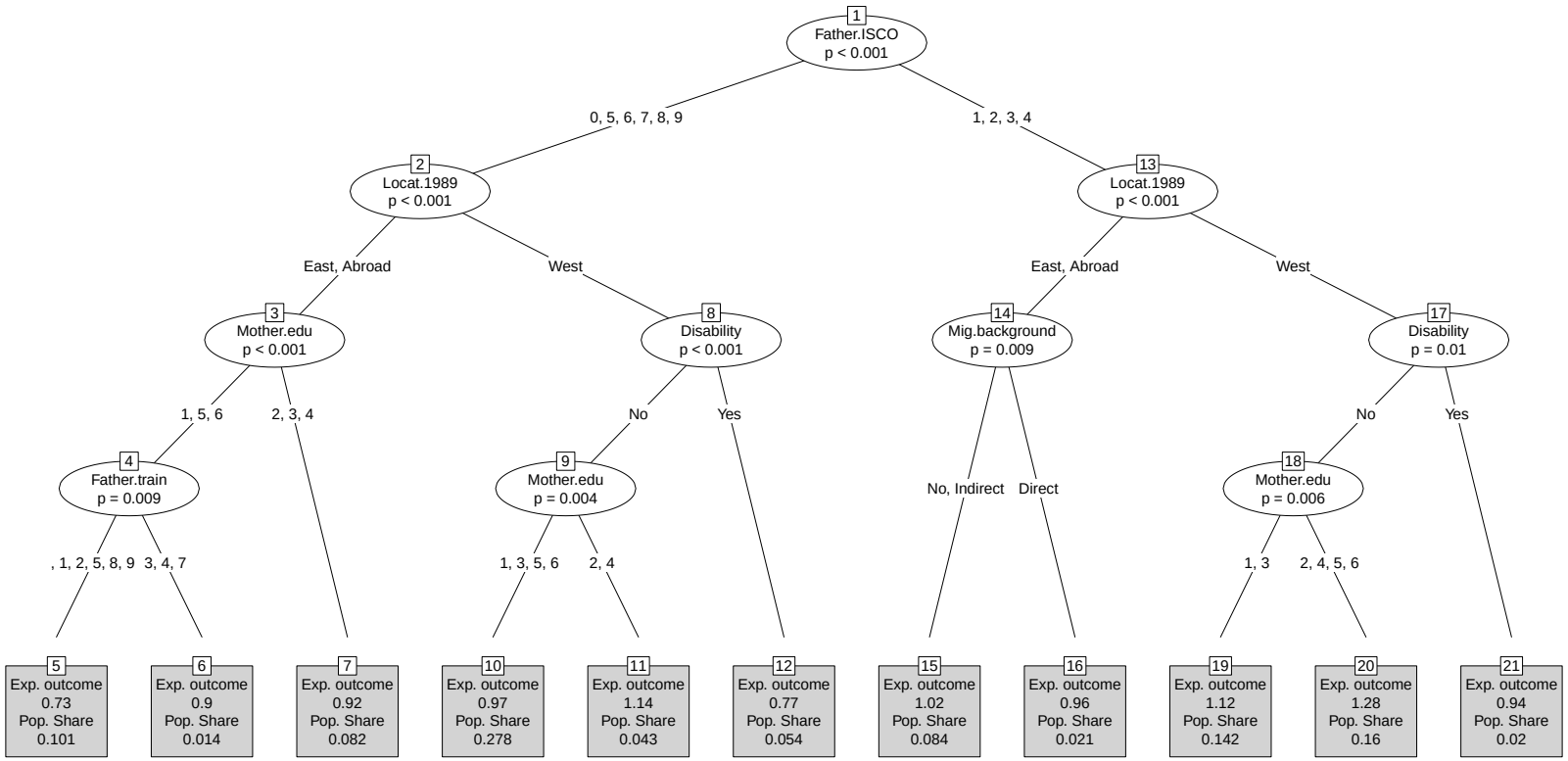
- waves considered 1992-2016;

- circumstances considered: migration background, location in 1989, mother's education, father's education, father's occupation, father's training, month of birth, disability, siblings;

# Opportunity tree in 1992



# Opportunity tree in 2016



# Mother/father raining

mtraining / ftraining

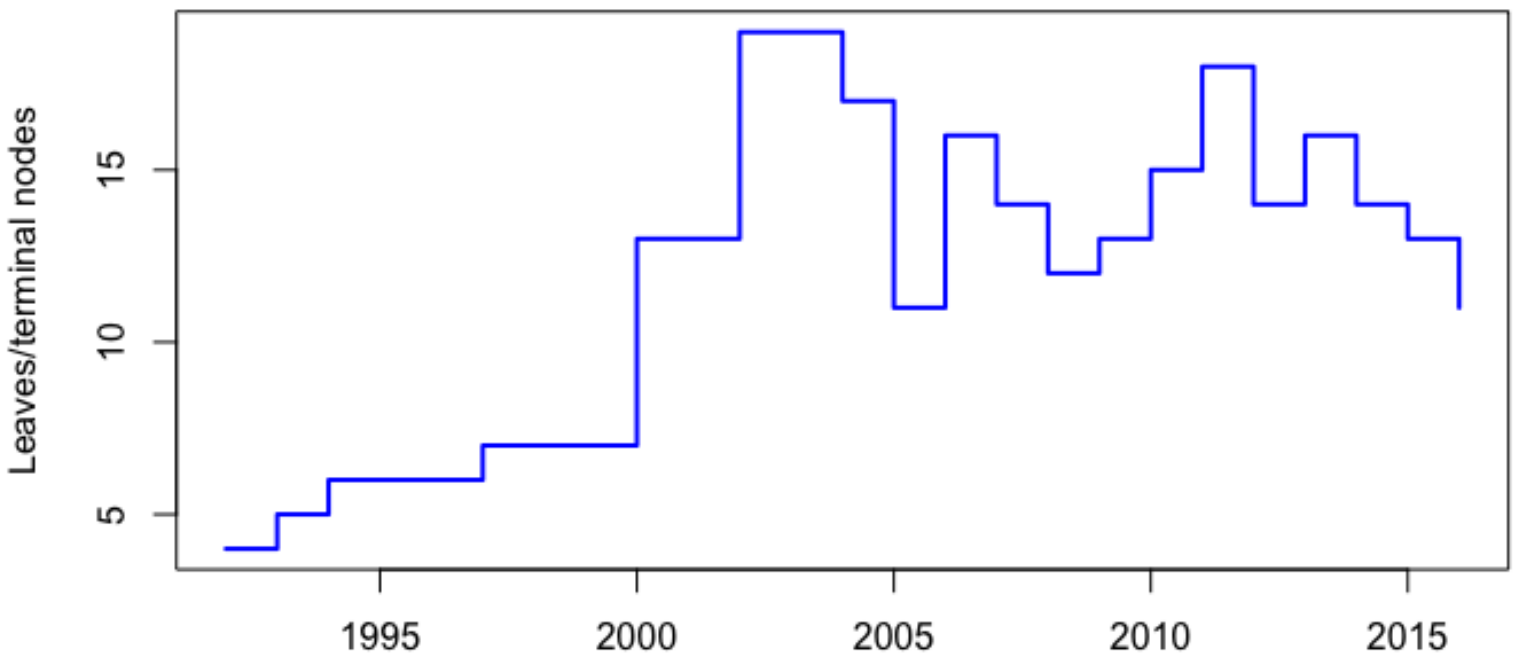
| cod. | Berufsbildung M/V | Vocational Training M/F |
|---|---|---|
| 1 | Keine Ausbildung | No vocational degree |
| 2 | Berufliche Ausbildung | Vocational Degree |
| 3 | Gewerbliche oder Landwirtschaftliche Leh | Trade or Farming Apprentice |
| 4 | Kaufm.L.,Bfs,Handel | Business |
| 5 | Gesundheitswesen, FS,Techn.-o.Meisters | Health Care or Special Technical School |
| 6 | Beamtenausbildung | Civil Service Training |
| 7 | FHS,Ingeniuerschule | Tech Engineer School |
| 8 | Hochsch.,Universit. (In- und Ausland) | College, University (in GER or Abroad) |
| 9 | Sonstige Ausbildung | Other Training |

# Mother/father education

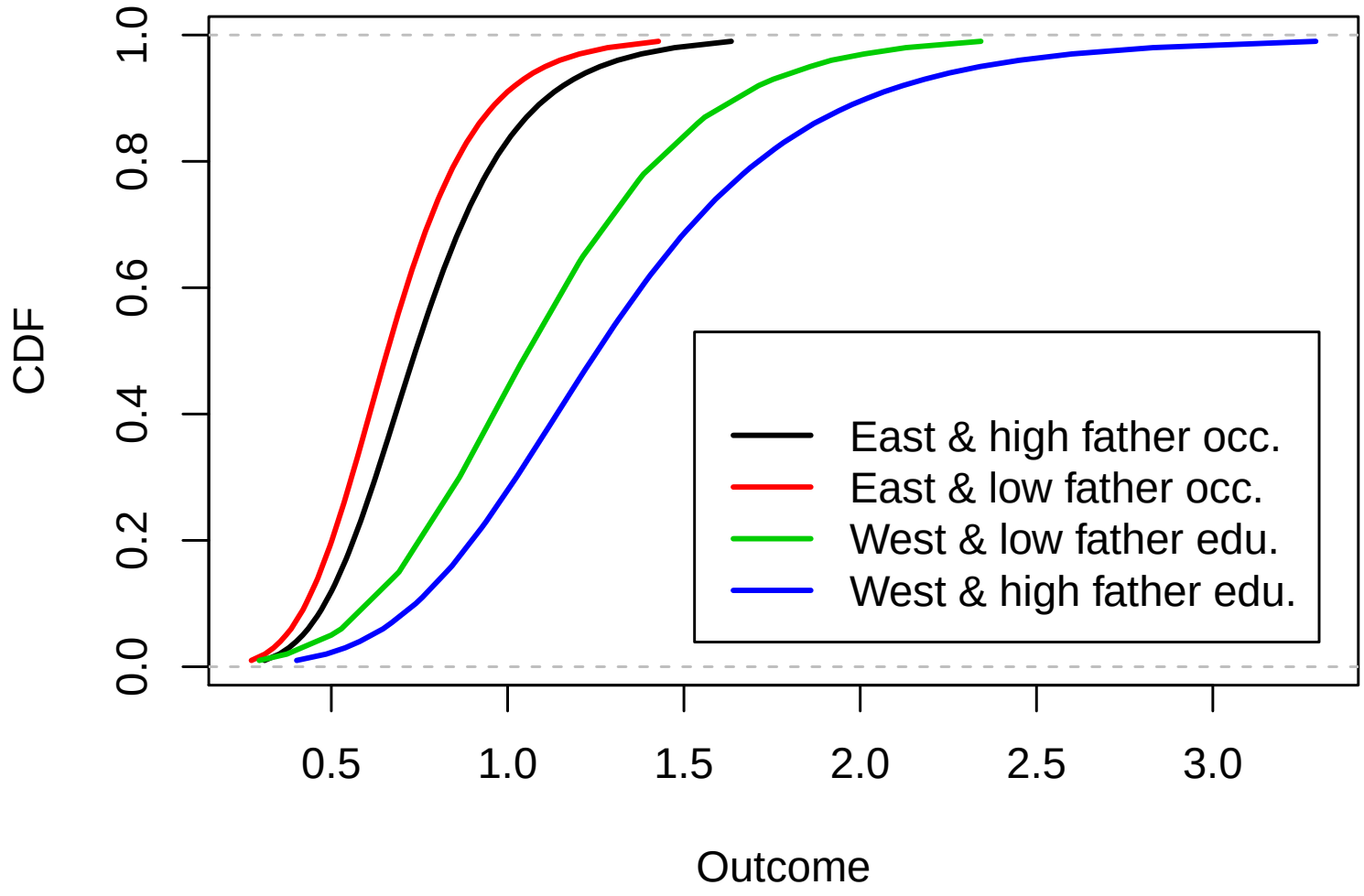
fsed / msed

| cod. | Schulbildung Vater / Mutter | Father/Mother Education |
|---|---|---|
| 1 | [1] Hauptschule | Lower Secondary |
| 2 | [2] Realschule | Intermediate Secondary |
| 3 | [3] Fachoberschule | Technical School |
| 4 | [4] Abitur | Upper Secondary |
| 5 | [5] sonstiger Abschluss | Other School Degree |
| 6 | [6] Kein Abschluss | No School Degree |
| 7 | [7] Keine Schule besucht | School not attended |

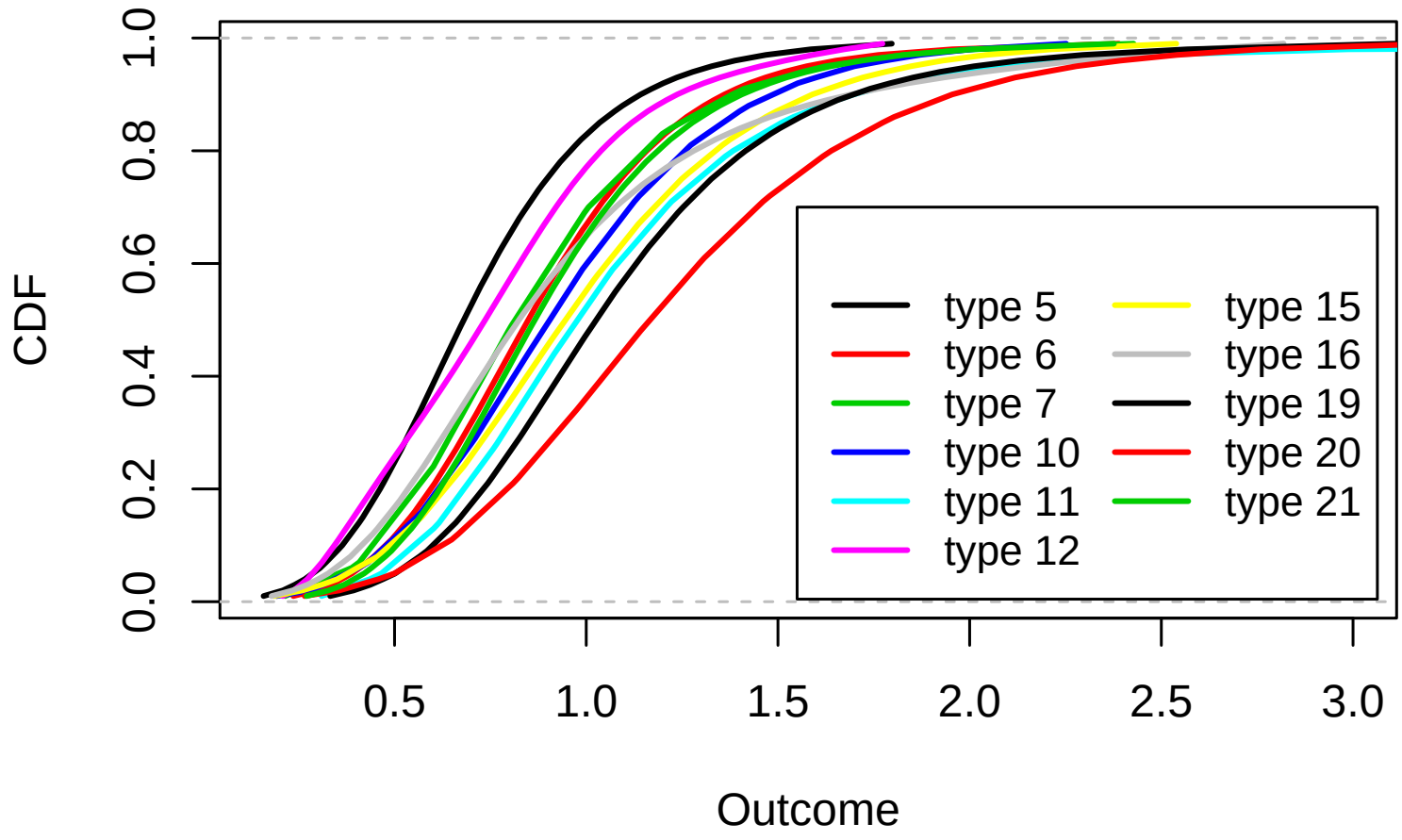
# Terminal nodes 1992-2016



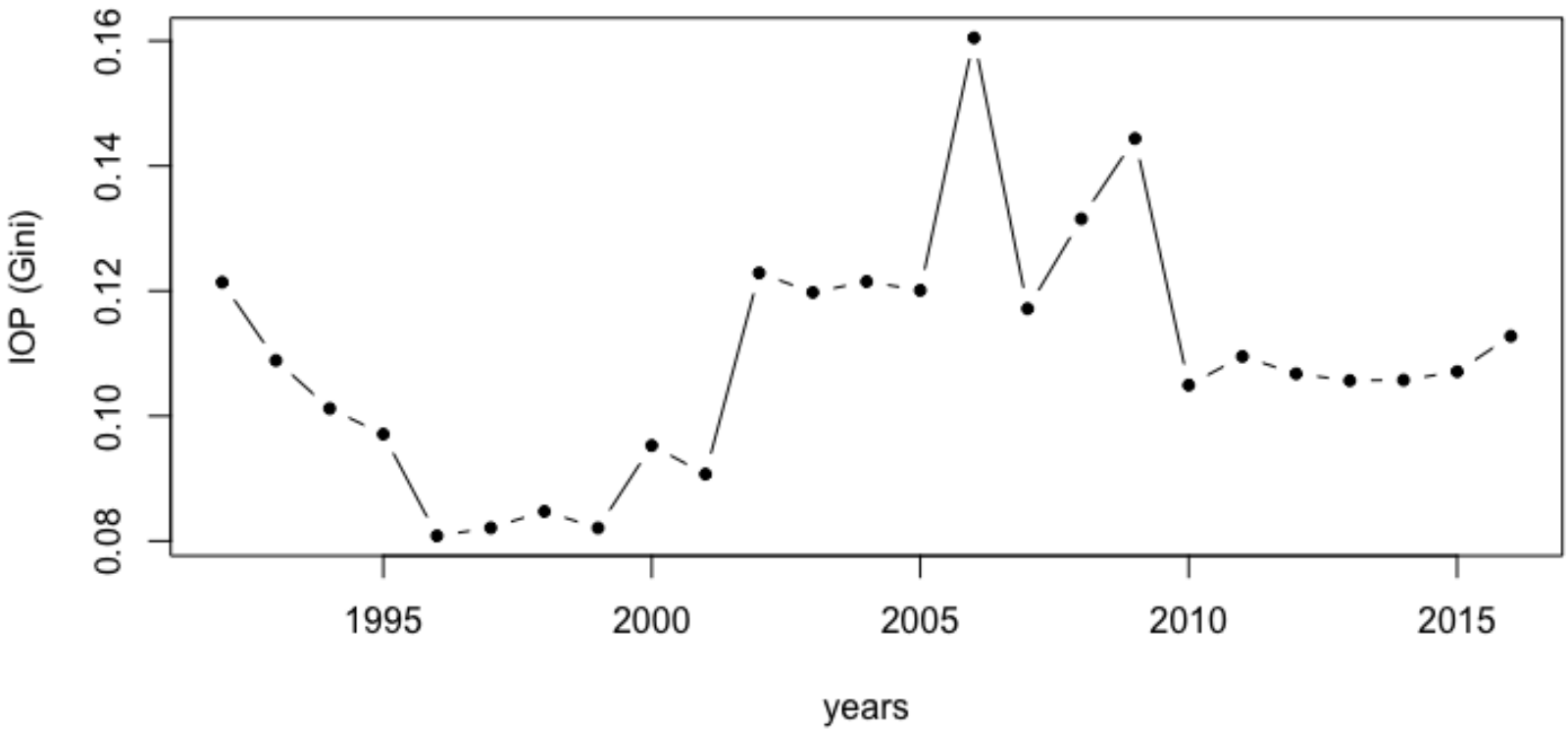
# IOP in 1992


CDF
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
2.5
3.0
Outcome
East & high father occ.
East & low father occ.
West & low father edu.
West & high father edu.

# IOP in 2016


CDF
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
2.5
3.0
Outcome
type 5
type 6
type 7
type 10
type 11
type 12
type 15
type 16
type 19
type 20
type 21

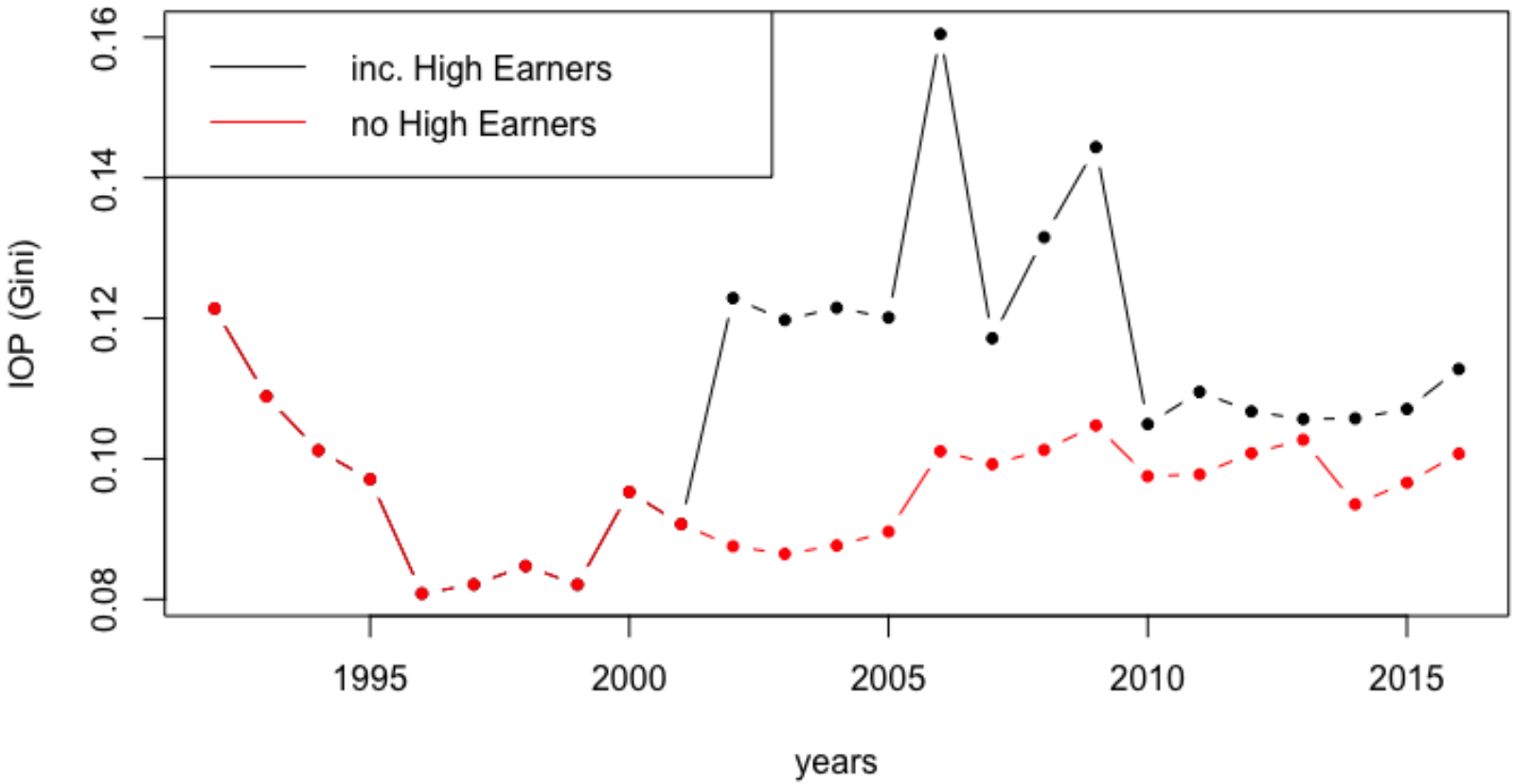
# IOP trend 1992-2016



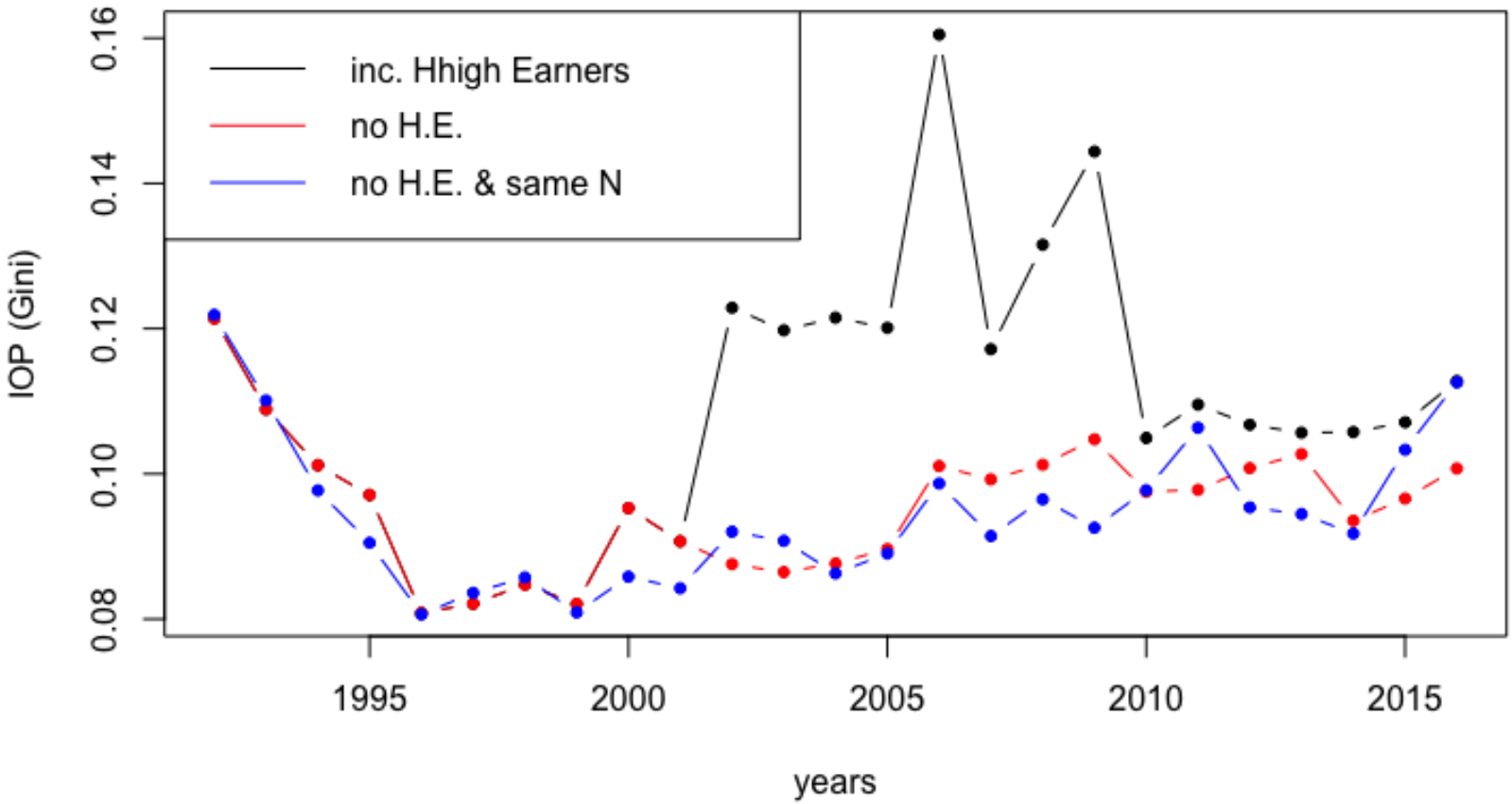
# Three open issues

- high income earners 2002 sample;
- sample size (power of the tree)
- confidence bounds

# IOP trend 1992-2016



# IOP trend 1992-2016



# Summary

- we propose an approach to estimate IOP fully consistent to Roemer's theory;
- effort identification method maximizes efficiency and comparability;
- since 1992 in Germany the opportunity structure has become more complex;
- IOP declined after reunification and increased with Hartz reforms;
- is today about 20% lower than in 1992.

# Distribution of bootstrap estimates