Inequality of Opportunity in the United Kingdom

A Supervised Learning Approach

Bruno Fagnola

bruno.fagnola@glasgow.ac.uk

Overview

1. Motivation
2. Theoretical framework
3. Methodology
4. Empirical application
   - Data
   - Implementation
   - Opportunity structure in the UK
   - Model performance
5. Final remarks

Motivation

- Inequality might be one of the most important issues of the 21st century
- Trends on global inequality, *between*-country and *within*-country income inequality
- Dynamics of net household income and individual gross income in the United Kingdom

Motivation

- To what extent is the idea of *meritocracy* valid?
- Do circumstances matter for different attainments?
- Research questions:
  - Are individuals in the UK constrained by circumstances?
  - If so, which are the most relevant?
  - Does inequality of opportunity translate into income inequality?

Equality of Opportunity as a concept

- John E. Roemer's approach to Equality of Opportunity
- Effort and circumstances
- Principle of 'leveling the playing field'
- *Level* of effort vs. *Degree* of effort
- Limitations:
  - Not a theory of distributive justice
  - Largely normative approach

Regression Trees

- Predict an outcome $y \in Y$ as a function of inputs $X = (X^1, ..., X^j, ..., X^k)$
- Take a sample of the data $S = \{(x_i, y_i)\}_{i=1}^{S}$ and divide the population into non-overlapping groups $G = \{g_1, ..., g_m, ...g_M\}$
- Predict value of observation i as conditional expectation of the corresponding group: $\hat{f}(x_i) = \hat{u}_{m(i)} = \frac{1}{N_m} \cdot \sum_{j \in g_m} y_j$
- Construction by *recursive binary splitting*
- Conditional inference trees: test partial hypothesis for distribution independence $H_0^j : D(Y|X^j) = D(Y)$ with global null hypothesis $H_0 = \bigcap_{j=1}^{k} H_0^j$

Empirical application: Data

- Quarterly Labour Force Survey (QFLS) by ONS
- Quarter corresponding to July-September, 2019
- Individuals in working age (30 to 59 year old's)
- Target variable: average gross hourly pay
- Dismiss lower 0.5th and upper 99.5th percentiles of the distribution
- Sample size: 6,853 observations

Average gross hourly pay

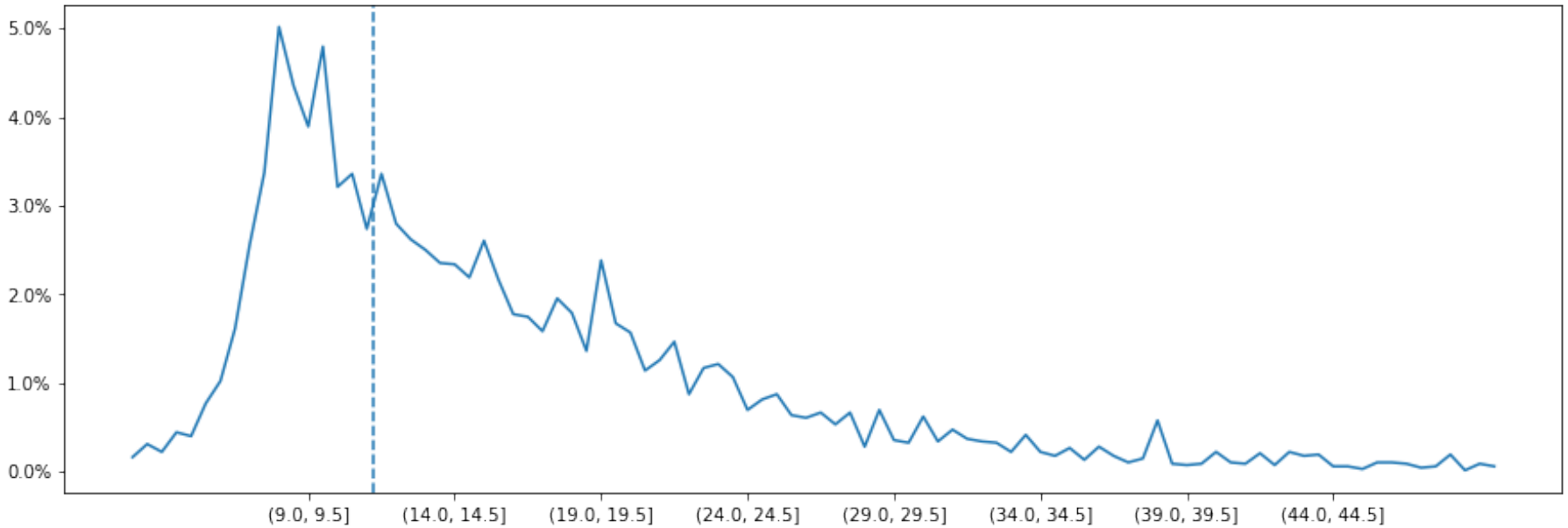

Figure 1: Distribution of hourly wages corresponding to the period Jul-Sep 2019 (own creation). Figures are all in GBP. Values are rounded to two decimal places and binned in intervals of 50 pence - Mean is represented by the dashed line.

Covariates

Variable	Values	Code QLFS
Gender	0. Male 1. Female	SEX
Nationality	0. United Kingdom 1. Other European country 2. Rest of the world 3. Asia 4. European Union	NATOX7 _EUL_Main
Ethnicity	0. White 1. Indian 2. Other ethnic group 3. Bangladeshi 4. Mixed 5. Pakistani 6. Black 7. Other Asian 8. Chinese	ETHUKEUL
Religion	0. Non-religious 1. Christian 2. Muslim 3. Hindu 4. Other 5. Sikh 6. Buddhist 7. Jewish	RELIG11
Number of GCSE's	0. More than five 1. Fewer than five	NUMOL5

Variable	Values	Code QLFS
Health condition	0. No 1. Yes	LIMITA
Household composition	0. Living with one or both parents 1. Living with other family members 2. Not living with family	SMHCOMP
Main earner in the household	0. Father 1. Mother 2. Joint 3. Other family member 4. No earners	SMEARNER
Occupation of parent	0. Managers and directors 1. Professionals 2. Technicians 3. Administrative occupations 4. Sales and customer service 5. Skilled workers 6. Other service occupations 7. Plant and machine operatives 8. Elementary occupations	SMSOC101

Figure 2: List of circumstances chosen for the study

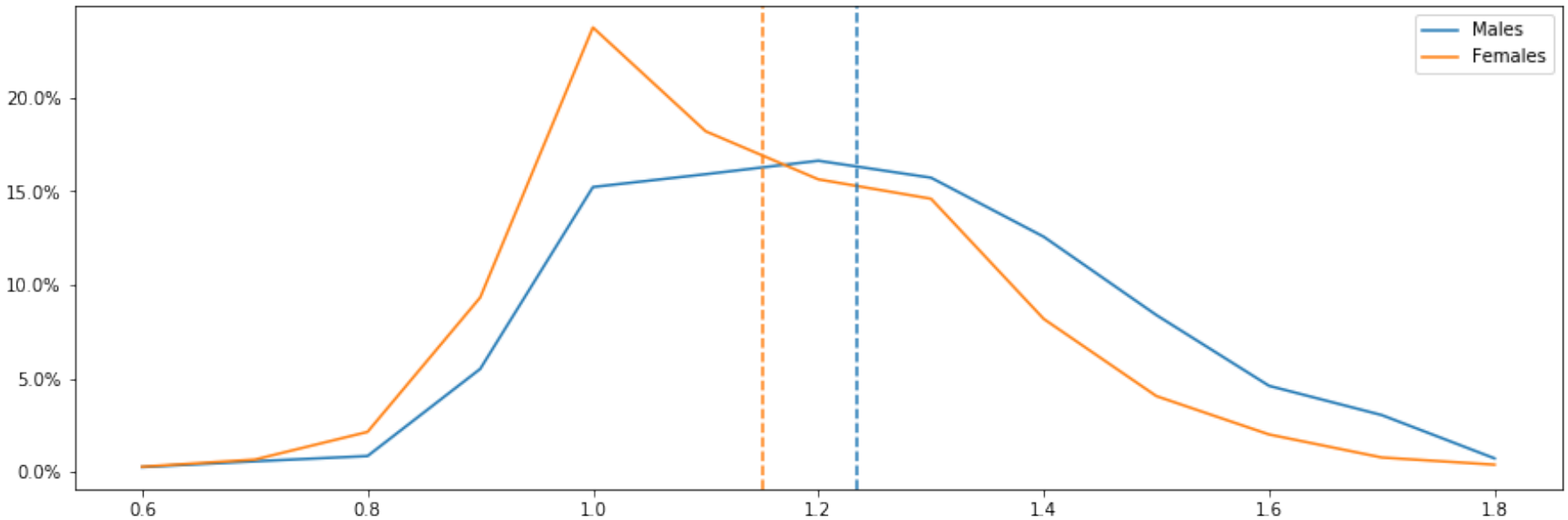

Figure 3: Distribution of hourly wages by gender, expressed in logs, for the period Jul-Sep 2019 (own creation). Means are represented by the dashed lines.

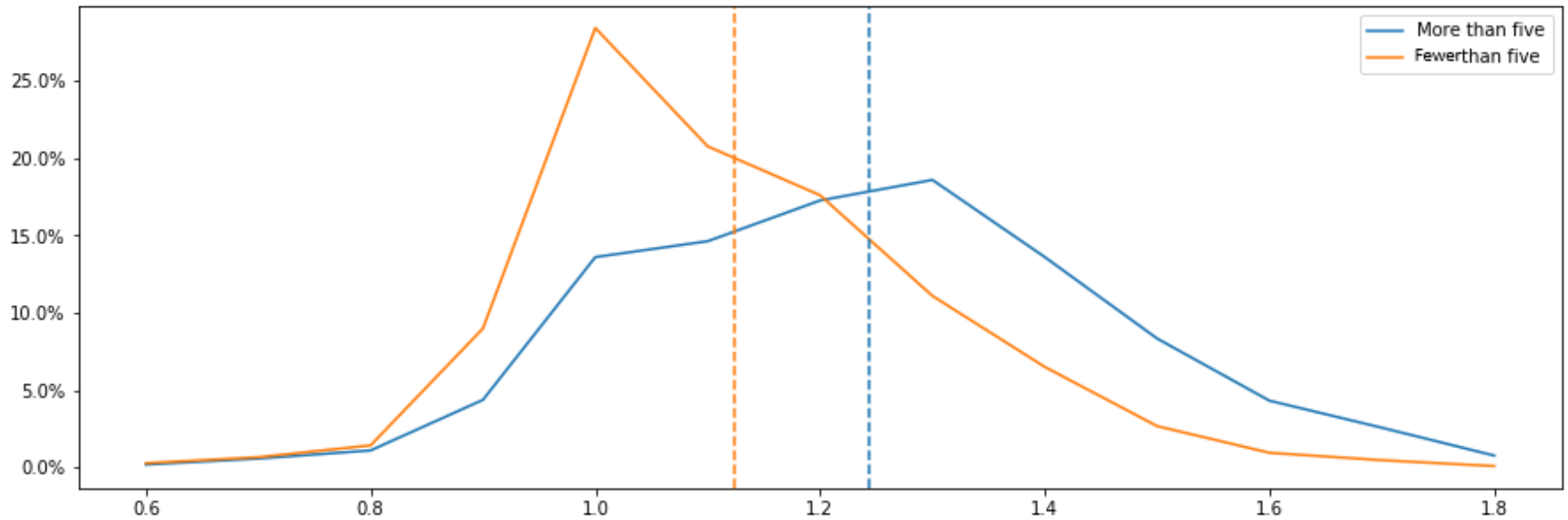

Figure 4: Distribution of hourly wages by number of GCSE's, expressed in logs, for the period Jul-Sep 2019 (own creation). Means are represented by the dashed lines.

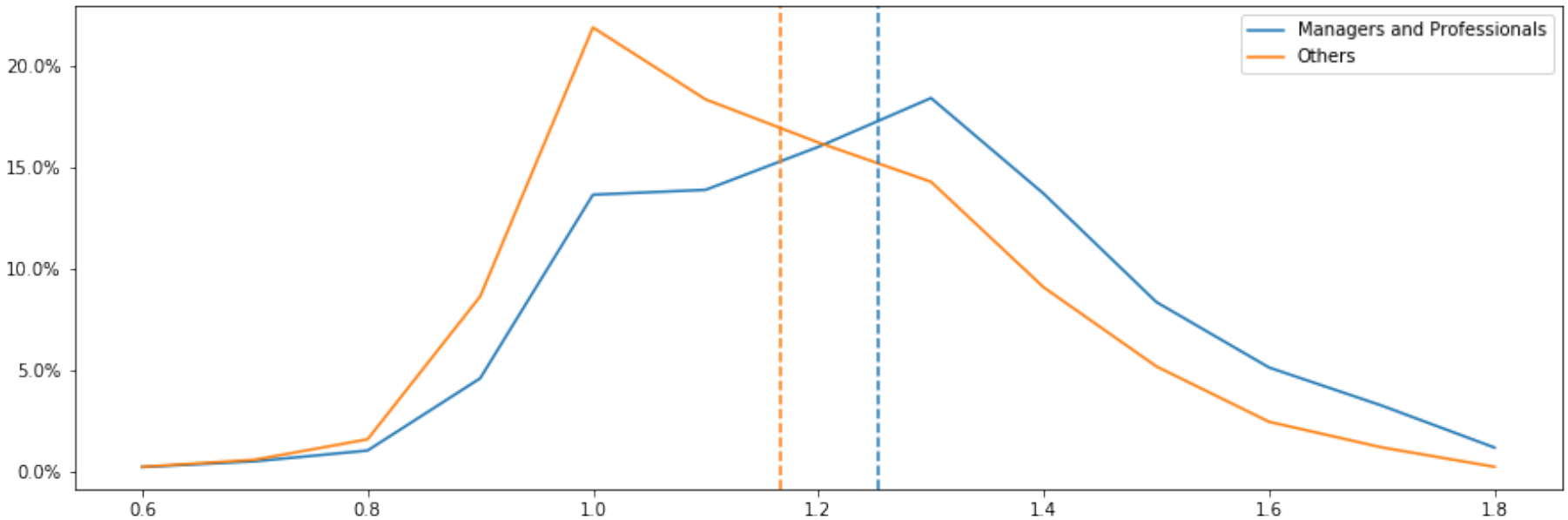

Figure 5: Distribution of hourly wages by occupation of the parent, expressed in logs, for the period Jul-Sep 2019 (own creation). Means are represented by the dashed lines.

Algorithm

1. Choose a significance level α^* and the maximum depth for the tree *max_depth*

2. Test the null hypothesis of distribution independence $H_0^{j(\omega)} : D(Y|X^{j(\omega)}) = D(Y)$ for all $X^{j(\omega)} \in \Omega$ and obtain p-value associated with each test: $p^{j(\omega)}$

3. Select the variable with the smallest p-value p^*: $X^* = \textit{argmin}\,\{p_{adj}^{j(\omega)} : X^{j(\omega)} \in \Omega\}$

   - If $p^* > \alpha^*$ or if *max_depth* has already been reached: exit the algorithm
   - If $p^* \leq \alpha^*$: select X^* as a splitting variable and continue

Algorithm (cont.)

4 Test the null hypothesis of distribution independence between subsamples for each partition s of X^* and obtain p-value associated with each test p^{*s}.

   Split the sample based on X^* by choosing the splitting point s that yields the smallest p-value $\hat{p}^* : X^{*\hat{\omega}} = \textit{argmin}\left\{p^{*s} : X^{*(\omega)} \in \Omega\right\}$

   - If $\hat{p}^* > \alpha^*$ dismiss the split and exit
   - If $\hat{p}^* \leq \alpha^*$ keep the split and continue

5 Repeat steps 2-4 for each resulting subsample

Opportunity structure in the UK

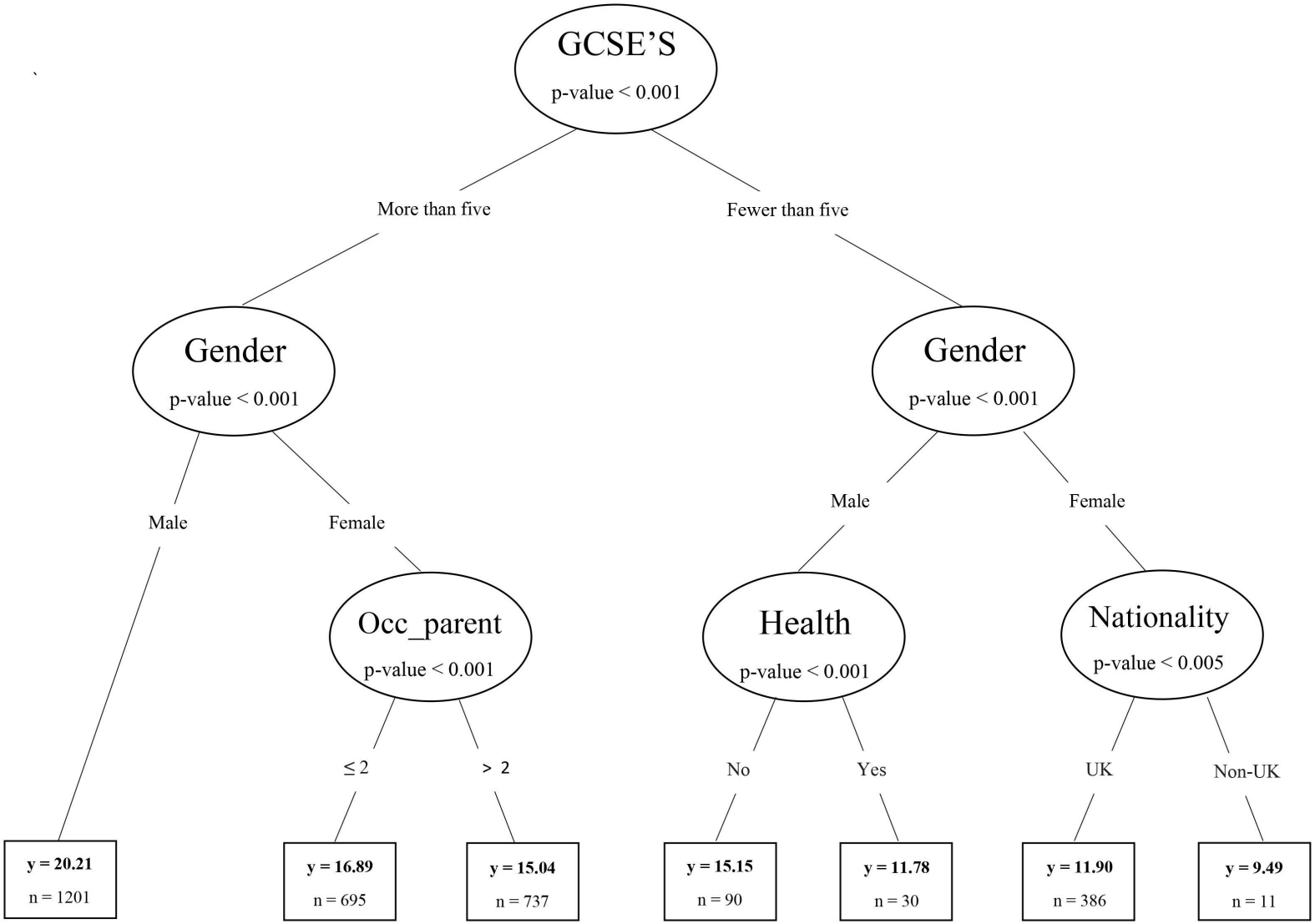

Figure 6: Conditional inference tree for the United Kingdom (own creation). T-test used for hypothesis testing

Performance of the predictions

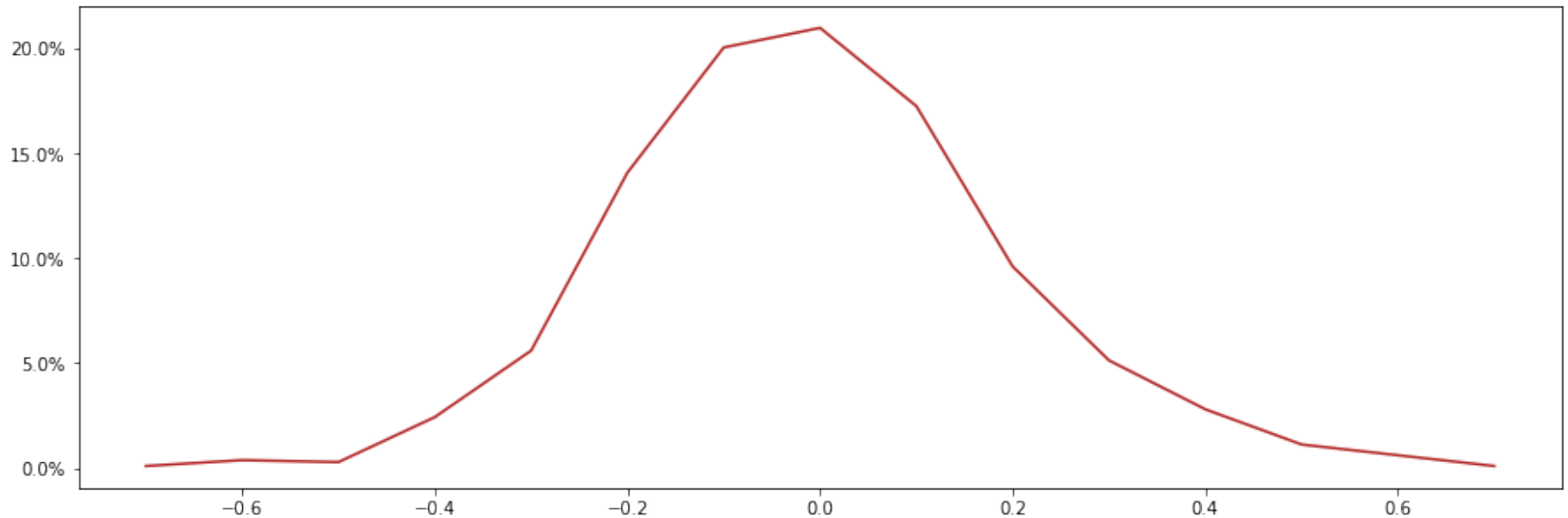

Figure 7: Distributions of errors derived from the estimations of the conditional inference tree. Differences are calculated out of sample

Final remarks

- Identification of 'Number of GCSE's' and 'Occupation of the parent' as key circumstances
- Good out-of sample accuracy of the estimations given sample size and non-comprehensive set of covariates in the data
- Encourage more studies on inequality of opportunities and methods with validation or results out of sample
- Future lines of research:
  - Construction of random forests
  - Different waves of the QLFS
  - Impacts of the COVID-19 pandemic
  - Actual measurement of inequality of opportunity