



Working Paper Series

Lorenz regression analysis of inequality of opportunity and explained inequality

Alexandre Jacquemain

Cédric Heuchenne

Philippe Van Kerm

ECINEQ 2026 699

Lorenz regression analysis of inequality of opportunity and explained inequality

Alexandre Jacquemain

UCLouvain

Cédric Heuchenne

UCLouvain, University of Liège

Philippe Van Kerm

University of Luxembourg

Abstract

Inequality of opportunity (IOp) is typically operationalized as explained inequality: inequality in the conditional expectation of an advantage given circumstances. Existing methods work in two stages, first estimating a model for the outcome, and then applying an inequality index to the fitted values. Any bias in the first stage can therefore propagate to the second stage. By contrast, this paper builds instead on Lorenz regressions which optimize the object of interest directly -- explained inequality -- under a flexible single-index model. We show that this provides a robust and unified framework adapted to both continuous and binary outcomes and illustrate the method for income- and education-based IOp, where it integrates naturally into existing work while improving estimation stability, interpretability, and robustness.

Keywords: Explained inequality; Gini coefficient; Single-index model; Variable selection; Shapley value decomposition; Inequality of opportunity

JEL Classification: C14; C52; D31; D63

Lorenz regression analysis of inequality of opportunity and explained inequality

Alexandre Jacquemain^{*1}, Cédric Heuchenne^{†1,2} and Philippe Van Kerm^{‡3}

¹CAPE, UCLouvain Saint-Louis, Brussels, Belgium

²HEC Liège, University of Liège, Liège, Belgium

³University of Luxembourg, Esch-sur-Alzette, Luxembourg

Abstract

Inequality of opportunity (IOp) is typically operationalized as explained inequality: inequality in the conditional expectation of an advantage given circumstances. Existing methods work in two stages, first estimating a model for the outcome, and then applying an inequality index to the fitted values. Any bias in the first stage can therefore propagate to the second stage. By contrast, this paper builds instead on Lorenz regressions which optimize the object of interest directly – explained inequality – under a flexible single-index model. We show that this provides a robust and unified framework adapted to both continuous and binary outcomes and illustrate the method for income- and education-based IOp, where it integrates naturally into existing work while improving estimation stability, interpretability, and robustness.

Keywords: Explained inequality; Gini coefficient; Single-index model; Variable selection; Shapley value decomposition; Inequality of opportunity

JEL classification: C14; C52; D31; D63

^{*}Corresponding author. UCLouvain Saint-Louis, Boulevard du Jardin Botanique 43, 1000 Brussels, Belgium. Email: alexandre.jacquemain@uclouvain.be. Alexandre Jacquemain acknowledges financial support from the Fédération Wallonie-Bruxelles through the *Actions de Recherche Concertées* program, grant number 23/28-001.

[†]Email: cedric.heuchenne@uclouvain.be

[‡]Email: philippe.vankerm@uni.lu

1 Introduction

The literature on the measurement of inequality of opportunity aims to capture the extent to which differences in an economic advantage – earnings, income, wealth – are driven by personal *circumstances*: factors beyond a person’s responsibility, such as parental background, sex, or ethnicity. This is rooted in the idea that not all inequalities are unfair. Advantage arising from a person’s own effort should be preserved. Advantage arising from circumstances beyond a person’s control is not – or less – morally admissible, and may call for compensation (Dworkin, 1981; Roemer, 1998; Roemer and Trannoy, 2015, 2016).

A large empirical literature has developed to operationalize these principles into measures of inequality of opportunity – or of “unfair inequality”. Under the most common, *ex ante utilitarian* approach, individuals facing the same circumstances are each attributed the conditional expectation of advantage given those circumstances and inequality of opportunity is then the inequality across these conditional expectations (Van de gaer, 1993; Roemer and Trannoy, 2015; Ramos and Van de gaer, 2016; Brunori et al., 2018). The objective, in other words, is to measure inequality in $m(X) := \mathbb{E}[Y|X]$ – the conditional expectation of the advantage variable Y given circumstances X – rather than inequality in Y itself. One seeks to capture the unfairness in the distribution of economic advantage as inequality explained by individual circumstances.

Implementation thus requires the choice of an inequality index $I(\cdot)$ and an estimate of the conditional expectation $m(\cdot)$. When the vector of circumstances is small and discrete, $m(x)$ can be estimated non-parametrically as the sample average of Y given $X = x$, with the population partitioned according to the distinct values taken by X . The resulting $\hat{m}(x)$ are then passed to any standard inequality functional – the Gini coefficient, the mean log deviation, or the Theil index – preferably one satisfying desirable properties such as the principle of transfer. This returns a simple measure of *explained inequality*, as in, e.g., Cowell and Jenkins (1995). When the inequality functional is additively decomposable into a between-group and a within-group component – as the Theil index and the mean log deviation are – the resulting IOp index is also exactly equal to the between-group term of the index decomposition.

Difficulties arise when the dimension of the set of circumstances grows relative to the available sample size. Partitioning the population along multiple dimensions produces small cells, which precludes reliable estimation of the conditional expectations and introduces an upward bias in the estimated level of explained inequality (Ramos and Van de gaer, 2016; Brunori et al., 2019). The most immediate response is to replace direct, nonparametric estimation of $m(x)$ with a model-based prediction. One first fits a conditional outcome model – e.g., a log-linear model with circumstances entered additively, possibly with a limited number of higher-order and interaction terms – and uses the resulting model predictions as the $\hat{m}(x)$ fed into the chosen inequality functional (Bourguignon et al., 2007; Ferreira and Gignoux, 2011). A key issue with this approach is the trade-off between mis-specification bias, if the chosen model fails to approximate $m(x)$ everywhere, and an upward, overfitting bias, if $m(x)$ is modelled “too flexibly” to the data. The recent response has been to adopt machine learning estimators of $m(\cdot)$, such as regression trees and random forests (Carranza, 2022; Brunori et al., 2018; Bernardo et al., 2025; Salas-Rojo and Rodríguez, 2022) or high-dimensional, penalized (such as LASSO) regression (Hufe et al., 2022). These methods take a large number of circumstances and their interactions as input and produce predictions of $m(x)$ that balance bias and variance.

One drawback remains however. The first-stage model is optimized to predict $m(\cdot)$ – by cross-validation, or a similar criterion – not to estimate the inequality index $I(m(X))$ well. These are different objectives, and nothing in the fitting criterion ties the two together. Escanciano and Terschuur (2022) show that this mismatch is not innocuous: biases in the first-stage prediction, even when small relative to the prediction error they were designed to minimize, can contaminate the resulting estimate of $I(m(X))$, and so distort the measured level of inequality of opportunity.

This paper presents an empirical approach to the measurement of IOp based on recently developed penalized Lorenz regression that addresses these concerns (Heuchenne and Jacquemain, 2022; Jacquemain et al., 2024). It preserves the possibility of considering high-dimensional circumstances and high-order interactions in the estimation of inequality of opportunity, with a penalization criterion selecting the best-fitting, parsimonious specification. Unlike in the two-stage approach, however, the parameters of the conditional

expectation model are estimated to maximize explained inequality in the outcome variable directly – not to predict the intermediate $m(X)$ well. The conditional expectation is specified as a generic single-index model, $m(x) = H(x^\top \theta)$, where the link function $H(\cdot)$ does not itself need to be estimated and is therefore very flexible. This strategy preserves the adaptability of machine learning approaches while resolving the bias-propagation problem described above. It also delivers asymptotic normality and valid bootstrap confidence intervals.¹

Built into the model is the choice of the Gini coefficient as the inequality metric. What is estimated is, in effect, the *explained Gini coefficient*. This choice of metric is what makes it possible to bypass estimation of $H(\cdot)$. It also brings the well-known robustness properties of the Gini coefficient against extreme observations (Cowell and Victoria-Feser, 1996). Unlike the variance of logarithms, sometimes used as an inequality-of-opportunity metric alongside log-linear models in two-stage approaches, the Gini coefficient satisfies the standard desirable properties of inequality measures, notably the principle of transfer (Cowell, 2011; Jenkins and Van Kerm, 2009).

The model also provides a unified framework to study both a continuous advantage variable, such as income or earnings, and a binary outcome, such as educational attainment or poverty status. Most studies of inequality of opportunity for a binary outcome (e.g., Paes de Barros et al., 2009; Krafft and Alawode, 2018; Palmisano et al., 2022) adopt ad hoc adaptations of the two-stage approach: a first-stage predicts the probability of a positive outcome given circumstances, and inequality in the predicted values is then measured in a second step. We show here that the Gini coefficient of predicted outcomes provides a natural inequality metric for binary outcomes as well and that maximizing it is equivalent to optimizing the area under the ROC curve (AUC) of a linear classifier, connecting the measurement of inequality of opportunity in binary advantage variables to a widely used classification criterion. This extends the Lorenz regression beyond the continuous outcomes

¹Lorenz regression is different from “Gini regression” (Olkin and Yitzhaki, 1992; Yitzhaki and Schechtman, 2013; Mussard and Souissi-Benrejeb, 2019). The Gini regression builds robust alternatives to ordinary least squares using Gini-based variability measures – for instance by minimizing the Gini mean difference of the residuals, or by using covariate ranks as instruments. It targets robust estimation of a standard regression model whereas Lorenz regression targets direct estimation of the explained Gini coefficient itself.

considered in Heuchenne and Jacquemain (2022) and Jacquemain et al. (2024).

Using data from the European Union Statistics on Income and Living Conditions (EU-SILC), we illustrate application of the methods to inequality of opportunity in three European countries – Belgium, Spain and Norway – that exhibit very different socio-economic structures. We examine two different outcomes: household income and tertiary education attainment. The application further illustrates (i) the use of a Shapley value decomposition in the context of Lorenz regression to attribute explained inequality to individual circumstances, and (ii) adaptation of the penalized Lorenz regression estimation algorithm to incorporate survey weights as typically required for use with survey data. Lorenz regression provides a coherent analytical framework and turns out to be more parsimonious than a classic two-stage approach – selecting a smaller set of circumstance combinations as relevant to IOp. Results show, for example, that in Norway the Lorenz regression retains only 8 opportunity sets for the older cohort of 2019, against 20 for the penalized log-linear competitor, while estimating a *higher* level of income IOp (7.0% against 5.7%). Across the three countries, the Shapley decomposition points to parental education, occupation, and origin as the dominant circumstances behind unequal opportunities, with the contribution of parental education rising over time.

Note that while we develop and illustrate the Lorenz regression approach for inequality of opportunity, such a model has much broader applicability and can be used wherever the object of interest is explained inequality in some expected outcome. This is the case for many empirical questions in economics. How much of the earnings gap between workers reflects differences in education and experience (Oaxaca, 1973; Blinder, 1973)? How much of the dispersion in firms’ stock returns reflects exposure to systematic risk factors (Fama and French, 1992)? How much of the variation in market concentration across industries reflects differences in technology or entry costs (Bresnahan and Reiss, 1991; Sutton, 1991)? Any setting in which a researcher wants an estimator whose objective function *is* the inequality measure itself, rather than a byproduct of a separately optimized model, stands to benefit from the approach presented here.

The remainder of the paper is organized as follows. Section 2 formalizes inequality of opportunity and the explained Gini coefficient, introduces the Lorenz regression, and

presents its extensions to binary outcomes, penalization, and Shapley value decomposition. Section 3 presents an application to inequality of opportunity in income and education in Belgium, Spain, and Norway: it describes the data and implementation, benchmarks the Lorenz regression against a penalized parametric two-stage approach, and documents the substantive role of circumstances, including their Shapley value contributions, in each country. Section 4 concludes.

2 Inequality of opportunity, explained inequality and Lorenz regression

Let $(X^\top, Y)^\top \in \mathbb{R}^p \times \mathbb{R}$, with $0 < \mathbb{E}[Y] < \infty$, where $\mathbb{E}[\cdot]$ denotes the expected value. In an inequality of opportunity context, the response Y is the advantage variable and covariates $X = (X_1, \dots, X_p)^\top$ are circumstances. Although we focus on the inequality of opportunity interpretation, the tools introduced in this section apply to any setting concerned with explained inequality in expected outcomes.

2.1 Inequality of opportunity: circumstances, effort and the *ex ante* utilitarian view

The central idea behind equality of opportunity is to distinguish, among the determinants of an economic advantage, those for which individuals should not be held responsible from those for which they should. Roemer (1998) calls the first set of determinants *circumstances* – characteristics such as parental background or sex, over which an individual has no control – and the second *effort*. A specific realization of the vector of circumstances defines a *type* and its corresponding *opportunity set*: the potential outcomes available to the set of individuals sharing the same circumstances as a function of the effort they exert.

Equality of opportunity requires that individuals of different types – facing different circumstances – have the same prospects, irrespective of effort. Under the popular *ex ante* approach we follow here, this is operationalized by attaching a summary measure of advantage to each opportunity set; equality of opportunity then requires equality between these

summary measures across types of individuals. In the *utilitarian* version of this approach, the summary measure is the conditional expectation of advantage given circumstances, $m(X) = \mathbb{E}[Y|X]$, so that inequality of opportunity is measured as inequality in $m(X)$. See Roemer and Trannoy (2015, 2016), Ramos and Van de gaer (2016), or Brunori et al. (2018) for more detailed reviews.

Income or earnings are the most commonly used measures of advantage in this literature (e.g., Bourguignon et al., 2007; Ferreira and Gignoux, 2011; Mitnik et al., 2020), but other outcomes have been studied, including wealth (e.g., Graeber, König, and Hilbert, Graeber et al.; Salas-Rajo and Rodríguez, 2022), health status (e.g., Rosa Dias, 2009; Bricard et al., 2020), educational performance or attainment (e.g., Ferreira and Gignoux, 2014; Palmisano et al., 2022; Marrero et al., 2024) or access to education (e.g., Hoyos and Narayan, 2011).

2.2 The explained Gini coefficient

The Gini coefficient for discrete and continuous outcomes The Gini coefficient of Y is formally given by

$$G_Y := \frac{2 \mathbb{C}[Y, F_Y(Y)]}{\mathbb{E}[Y]}, \quad (1)$$

where $F_Y(\cdot)$ is the cumulative distribution function (CDF) of Y and $\mathbb{C}[\cdot, \cdot]$ is the covariance operator. The situation of perfect equality occurs when the advantage is the same across the population, yielding $G_Y = 0$. When $Y \in \mathbb{R}^+$, the situation of perfect inequality occurs when one individual receives all advantages, yielding $G_Y = 1$.

The Gini coefficient is a normalized measure of association between a random variable and the rank it induces. For discrete variables, the raw CDF is not an appropriate ranking function. Because it jumps at each mass point, it systematically assigns excessively high ranks. A mean-rank correction addresses this issue. Whenever the ranked variable is discrete, $\tilde{F}_Y(y) = P(Y < y) + \frac{1}{2} P(Y = y)$. In the binary case, this simplifies the Gini coefficient to

$$G_Y = \frac{2 \mathbb{C}[Y, \tilde{F}_Y(Y)]}{\mathbb{E}[Y]} = P(Y = 0).$$

In a tertiary education context, for example, the Gini coefficient is simply the proportion of individuals without a tertiary education degree.

The concentration index The concentration index of Y with respect to X_k is defined as

$$\text{CI}_{Y,X_k} := \frac{2 \mathbb{C}[Y, F_{X_k}(X_k)]}{\mathbb{E}[Y]},$$

and measures the association between the ranking of an individual in the distribution of circumstance X_k and the value of the advantage he or she receives.

The explained Gini measure The explained Gini coefficient is defined as

$$G_{Y,X} := G_{m(X)} \tag{2}$$

$$= \frac{2 \mathbb{C}[Y, F_m(m(X))]}{\mathbb{E}[Y]}, \tag{3}$$

where $F_m(\cdot)$ is the CDF of $m(X)$. As pointed out in Heuchenne and Jacquemain (2022), $G_{m(X)} \leq G_Y$, allowing us to express the explained Gini coefficient as a share of total inequality. Equation (2) indicates that the explained Gini coefficient is the Gini coefficient of the conditional expectation. The equivalent formulation in (3) is a direct consequence of the law of iterated expectation and indicates that the explained Gini coefficient can be spelled out as the concentration index of Y with respect to $m(X)$.

A common approach to estimate the explained Gini coefficient is to first fit a model for $m(X)$, yielding fitted values $\hat{m}(X)$, and plug those in an estimator for the Gini coefficient. However, Escanciano and Terschuur (2022) show that this estimator suffers from bias propagation and recommend exploiting Equation (3) so that fitted values are only used to rank observations, leading to a more stable estimator. As we show in the next section, the Lorenz regression goes one step further. By assuming a single-index model, fitted values do not need to be computed.

2.3 Lorenz regression for continuous outcomes

This section introduces the Lorenz regression as an inferential tool for inequality of opportunity in the continuous case. For a more complete description, see Heuchenne and Jacquemain (2022) for the original, unpenalized, procedure and Jacquemain et al. (2024) for the penalized procedure.

We assume the following single-index model

$$m(x) = H(x^\top \theta_0), \quad (4)$$

where $H(\cdot)$ is a strictly increasing link function and θ_0 is a vector of unknown weights, which satisfies $\|\theta_0\| = 1$, where $\|\cdot\|$ denotes the L2-norm. The constraint on the parameter space is imposed in order to ensure the identifiability of the model. To simplify the exposition, we assume throughout Sections 2.3 and 2.4 that $X^\top \theta_0$ is a continuous random variable.²

The key assumption behind the monotone single-index model is that an individual's rank in the distribution of expected advantages depends linearly on the circumstances, via the index $x^\top \theta_0$. The nonparametric link function $H(\cdot)$ transforms ranks into advantage values. This semi-parametric framework is more general and flexible than the standard log-linear model $\mathbb{E}[\log Y|X = x] = x^\top \theta_0$ usually assumed in the applied literature.

Given the monotonicity of the link function, one can rewrite the explained Gini coefficient as the concentration index of the advantage Y with respect to the linear index $X^\top \theta_0$. Formally,

$$G_{Y,X} = \frac{2 \mathbb{C}[Y, F_{\theta_0}(X^\top \theta_0)]}{\mathbb{E}[Y]}, \quad (5)$$

where $F_{\theta_0}(\cdot)$ is the CDF of $X^\top \theta_0$. This formulation is coherent with the perspective developed in Escanciano and Terschuur (2022). It has the extra advantage of not requiring full computation of the conditional expectation since Equation (5) depends on the linear index only.

The weight vector θ_0 is the outcome of the following maximization program (Proposition 1 in Heuchenne and Jacquemain (2022))

$$\theta_0 := \arg \max_{\theta, \|\theta\|=1} \frac{2 \mathbb{C}[Y, F_\theta(X^\top \theta)]}{\mathbb{E}[Y]}.$$

This provides the weight vector with the intuitive appeal of maximizing explained inequality, in the sense of yielding the linear index of circumstances that maximizes the concentration index between the advantage and the index. This identification result directly translates into an estimation procedure in the empirical setting.

²In practice, however – for instance in our empirical application where circumstances are all categorical – this assumption may not hold. Adaptations and theoretical results for the Lorenz regression in the fully discrete case are discussed in Section 3 of Heuchenne and Jacquemain (2022).

Let $(X_1^\top, Y_1)^\top, \dots, (X_n^\top, Y_n)^\top$ be an i.i.d. sample of size n with the same distribution as $(X^\top, Y)^\top$. Denote by $\hat{F}_i(\theta)$ the following estimator of $F_\theta(X_i^\top \theta)$

$$\hat{F}_i(\theta) := \frac{1}{n} \sum_{j=1}^n \left[\mathbb{1}\{X_i^\top \theta \geq X_j^\top \theta\} - \frac{1}{2} \mathbb{1}\{X_i^\top \theta = X_j^\top \theta\} \right], \quad (6)$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. This estimator complies with the mean-rank convention introduced earlier and has the advantage of always having an empirical mean of 0.5. The Lorenz regression (LR) estimates the vector of weights θ_0 and the explained Gini coefficient $G_{Y,X}$ with

$$\hat{\theta} := \arg \max_{\theta, \|\theta\|=1} \frac{1}{n} \sum_{i=1}^n Y_i \hat{F}_i(\theta), \quad (7)$$

$$\hat{G}_{Y,X} := \frac{2}{\bar{Y}} \frac{1}{n} \sum_{i=1}^n Y_i \hat{F}_i(\hat{\theta}) - 1, \quad (8)$$

where $\bar{Y}$ is the sample mean advantage. As discussed in Heuchenne and Jacquemain (2022), $\hat{G}_{Y,X}$ is a consistent and asymptotically normal estimator of the explained Gini coefficient for any single-index model satisfying Equation (4).

The explained Gini coefficient $\hat{G}_{Y,X}$ measures the inequality in advantage that would be observed if individuals were assigned their expected advantage given their circumstances. To convey how large the explained Gini is relative to total inequality, we also define the Lorenz- R^2 as:

$$\text{LoR}^2 := \frac{\hat{G}_{Y,X}}{\hat{G}_Y}.$$

This statistic reflects the share of total inequality attributable to unequal opportunities. By construction, $\text{LoR}^2 \in [0, 1]$. Values close to one indicate that most observed inequality is due to circumstances, whereas values close to zero suggest that inequality is primarily driven by effort.

High-dimensional estimation In many applications, a large number of circumstances are available, and this number grows rapidly when interactions are allowed. Since some of them may be irrelevant for explaining inequality of opportunity, the application would benefit from an automatic selection procedure. The maximization program in Equation (7) is designed to find, for each sample, the weight vector that explains the most inequality. As a

result, even irrelevant circumstances receive non-zero weights, simply because they help fit noise in the sample. This lack of sparsity may lead to an overestimation of the explained Gini coefficient $\hat{G}_{Y,X}$. The penalized Lorenz regression (PLR) proposed by Jacquemain et al. (2024) adds a penalty term to automatically guide the selection of relevant circumstances and to avoid overfitting explained inequality. Also, the non-differentiable ranks displayed in Equation (6) are replaced with a smooth approximation. Formally, the weight vector is estimated with

$$\hat{\theta} = \arg \max_{\theta, \|\theta\|=1} \left\{ \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n Y_i K\left(\frac{X_i^\top \theta - X_j^\top \theta}{h}\right) - \sum_{k=1}^p p_\lambda(|\theta_k|) \right\}, \quad (9)$$

where $K(\cdot)$ is the integral of a kernel function with bandwidth h . The penalty function $p_\lambda(x)$ is either the LASSO or the SCAD penalty, each depending on a regularization parameter λ that controls the severity of the penalty. In both cases, the algorithm produces a regularization path by gradually decreasing λ from a large value that shrinks all coefficients to zero. The optimal λ value is selected using a bootstrap-based out-of-sample measure, tailored to explained inequality. Specifically, for each bootstrap resample, θ_0 is estimated using PLR on the resampled data, and then used to compute the explained Gini coefficient on the out-of-bag observations – those not included in the resample. This quantity is then averaged across bootstrap draws, providing a model selection criterion that aligns directly with IOp measurement.

2.4 Lorenz regression for binary outcomes

The explained Gini coefficient remains a meaningful measure of IOp in the case where the outcome variable is binary. A formal equivalence between the explained Gini and the area under the ROC curve connects inequality explanation to classification performance. This equivalence justifies applying the Lorenz regression framework to binary outcomes, allowing us to leverage its full toolbox – including penalization and Shapley value decomposition – while remaining grounded in a principled measure of IOp.

With a binary outcome, the single-index model becomes

$$m(x) = P(Y = 1|X = x) = H(x^\top \theta_0). \quad (10)$$

The function $m(X)$ is a random variable taking values in $[0, 1]$. It represents the conditional probability of obtaining the advantage $Y = 1$, given an individual's circumstances X . The explained Gini coefficient measures the inequality in these conditional probabilities, quantifying the extent to which individuals' chances of obtaining the advantage depend on their circumstances. Using Equation (2.1) in Yitzhaki and Schechtman (2013), the explained Gini coefficient can be rewritten as

$$G_{Y,X} = \frac{\mathbb{E}[|m(X_1) - m(X_2)|]}{2\mathbb{P}(Y = 1)}, \quad (11)$$

where $(X_1^\top, Y_1)^\top$ and $(X_2^\top, Y_2)^\top$ are i.i.d. copies of $(X^\top, Y)^\top$. The numerator in Equation (11) is the expected absolute difference in conditional probabilities between two individuals randomly drawn from the population.

This contrasts with the *dissimilarity index*, defined as

$$DI_{Y,X} = \frac{\mathbb{E}[|m(X) - \mathbb{P}(Y = 1)|]}{2\mathbb{P}(Y = 1)},$$

for which the numerator is the mean absolute deviation of the conditional probability of obtaining the advantage. While both statistics measure the dependency of chances on circumstances, the explained Gini coefficient focuses on pairwise disparities, whereas the dissimilarity index centers on deviation from the mean. The dissimilarity index has been used in, e.g., Paes de Barros et al. (2009), Hoyos and Narayan (2011), Krafft and Alawode (2018), Palmisano et al. (2022) to measure educational inequality of opportunity, with the advantage defined as having obtained tertiary education. In their approach, $m(X)$ is estimated using a probit model of the form $\mathbb{P}(Y = 1|X = x) = \Phi(x^\top\theta_0)$, where $\Phi(\cdot)$ is the CDF of the standard Normal distribution. In comparison, the advantage of using the Lorenz regression is twofold. First, the assumed single-index model is more flexible than, say, a probit or logit model. Second, it avoids estimating the full model, requiring only the estimation of the ranking of individuals in terms of their conditional probabilities.

With a binary advantage variable, the measurement of IOp can be naturally framed as a classification problem, where the goal is to assess how well circumstances discriminate between advantaged ($Y = 1$) and disadvantaged ($Y = 0$) individuals. Formally, the aim is to find a decision rule $\delta(X)$, with $\delta : \mathbb{R}^p \rightarrow \{0, 1\}$, that determines whether the individual is

advantaged $\delta(X) = 1$, or disadvantaged $\delta(X) = 0$. The performance of the rule is assessed by the true-positive rate (TPR) and false-positive rate (FPR), defined as

$$\text{TPR}(c) := \mathbb{P}(\delta_c(X) = 1 | Y = 1)$$

$$\text{FPR}(c) := \mathbb{P}(\delta_c(X) = 1 | Y = 0),$$

where we highlight that the decision rule depends on a threshold c such that $\delta_c(X) = 1$ if c is exceeded and $\delta_c(X) = 0$ otherwise. Choosing a value of c entails a specific trade-off between TPR and FPR. This is materialized with the receiver operating characteristic (ROC) curve, defined as

$$\text{ROC}(p) := \text{TPR}(c : \text{FPR}(c) = p),$$

where $p \in [0, 1]$. The higher the ROC curve, the better is the classification since it means a higher true-positive rate for a given false-positive rate. For a more complete review of the ROC curve, see Chapter 4 of Pepe (2004).

Assuming the same single-index model as in Equation (10), Pepe et al. (2006) showed that the linear decision rule

$$\delta_c(X) = 1 \quad \text{if } X^\top \theta_0 \geq c$$

$$\delta_c(X) = 0 \quad \text{if } X^\top \theta_0 < c$$

has a dominating ROC curve, meaning that for each value of the false-positive rate, it has the highest rate of true-positive. This result implies that the optimal classifier depends only on the weight vector θ_0 , not on the link function. Building on this, Pepe et al. (2006) propose an estimator for θ_0 based on maximizing the empirical area under the ROC curve (AUC).

The connection between classification performance and inequality explanation can be formalized by two propositions. Proposition 1 shows that the explained Gini coefficient is directly proportional to the AUC of the optimal classifier, recasting the measurement of explained inequality as a problem of classifying individuals based on covariates. Proposition 2 shows that maximizing the empirical AUC yields the same estimator as the (non-penalized) Lorenz regression. The Lorenz regression framework can therefore be leveraged

in this setting. The penalized version enables automatic selection of relevant circumstances while mitigating overfitting of their classification power. Moreover, the Shapley value decomposition introduced in Section 2.5 offers an additive decomposition of the AUC across circumstances.

Proposition 1. *Let $(X^\top, Y)^\top \in \mathbb{R}^p \times \{0, 1\}$ follow the single-index model assumed in Equation (10). Then, it holds that*

$$G_{Y,X} = P(Y = 0)(2 \text{AUC}_{\theta_0} - 1)$$

where AUC_{θ_0} denotes the AUC associated to the optimal classifier $X^\top \theta_0$.

Proof. Let $(X_1^\top, Y_1)^\top$ and $(X_2^\top, Y_2)^\top$ be i.i.d. copies of $(X^\top, Y)^\top$. We start by rewriting the area under the curve and the explained Gini coefficient. First, using Result (4.6) in Pepe (2004), the AUC can be written as

$$\text{AUC}_{\theta_0} = P(X_1^\top \theta_0 > X_2^\top \theta_0 | Y_1 = 1, Y_2 = 0).$$

Starting from Equation (5), the explained Gini coefficient can be rewritten as

$$G_{Y,X} = \frac{2 \mathbb{E}[Y_1 \mathbb{1}\{X_1^\top \theta_0 > X_2^\top \theta_0\}]}{\mathbb{E}[Y_1]} - 1.$$

Furthermore, it holds

$$\begin{aligned} \mathbb{E}[Y_1 \mathbb{1}\{X_1^\top \theta_0 > X_2^\top \theta_0\}] &= P(Y_1 = 1, X_1^\top \theta_0 > X_2^\top \theta_0) \\ &= P(Y_1 = 1, Y_2 = 0, X_1^\top \theta_0 > X_2^\top \theta_0) + P(Y_1 = 1, Y_2 = 1, X_1^\top \theta_0 > X_2^\top \theta_0) \\ &= \text{AUC}_{\theta_0} P(Y = 0) P(Y = 1) + \frac{1}{2} (P(Y = 1))^2, \end{aligned}$$

which closes the proof. □

Proposition 2. *Let $(X_1^\top, Y_1)^\top, \dots, (X_n^\top, Y_n)^\top$ be an i.i.d. sample of size n with the same distribution as $(X^\top, Y)^\top$. Let $\mathcal{I}_1 := \{i = 1, \dots, n : Y_i = 1\}$ and $\mathcal{I}_0 := \{i = 1, \dots, n : Y_i = 0\}$. Define the estimator proposed by Pepe et al. (2006) as*

$$\hat{\theta}_{\text{AUC}} := \arg \max_{\theta, \|\theta\|=1} \sum_{i \in \mathcal{I}_1} \sum_{j \in \mathcal{I}_0} \mathbb{1}\{X_i^\top \theta \geq X_j^\top \theta\}.$$

Then, it holds, $\hat{\theta} = \hat{\theta}_{\text{AUC}}$, where $\hat{\theta}$ is obtained using a non-penalized Lorenz regression.

Proof. Assuming that the index is continuous and ties do not occur, the estimator $\hat{\theta}$ of the Lorenz regression can be expressed as the maximizer of $G_n(\theta)$, where

$$\begin{aligned} G_n(\theta) &= \sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{X_i^\top \theta \geq X_j^\top \theta\} \\ &= \sum_{i \in \mathcal{I}_1} \sum_{j=1}^n \mathbb{1}\{X_i^\top \theta \geq X_j^\top \theta\} \\ &= \sum_{i \in \mathcal{I}_1} \sum_{j \in \mathcal{I}_0} \mathbb{1}\{X_i^\top \theta \geq X_j^\top \theta\} + \sum_{i \in \mathcal{I}_1} \sum_{j \in \mathcal{I}_1} \mathbb{1}\{X_i^\top \theta \geq X_j^\top \theta\}. \end{aligned}$$

Notice that the second sum is a constant since $\sum_{j \in \mathcal{I}_1} \mathbb{1}\{X_i^\top \theta \geq X_j^\top \theta\}$ evaluated at each $i \in \mathcal{I}_1$ yields the rank vector $\{1, \dots, |\mathcal{I}_1|\}$. Hence, $\hat{\theta}_{\text{AUC}} = \hat{\theta}$. $\square$

2.5 Shapley value contributions to explained inequality

The Shapley value decomposition is a well-established tool in linear regression for assessing the contribution of each covariate to model fit. It operates by fitting all combinations of sub-models that omit one or more variables, and recording the associated loss in R^2 . Shorrocks (2013) extended this idea to distributional analysis, while Hoyos and Narayan (2011) applied it to the case of IOp. Building on this literature, we can apply the Shapley value decomposition to the explained Gini coefficient.

To handle high-dimensional settings, we first use a penalized Lorenz regression to select a relevant subset of circumstances. The Shapley value decomposition is then computed using non-penalized Lorenz regressions restricted to this selected set. This two-step approach balances interpretability and parsimony. The resulting decomposition is additive, meaning that the individual contributions sum to the total explained Gini coefficient. Each contribution can thus be interpreted as the share of IOp attributable to a given covariate. Each categorical variable is treated as a single unit. Omitting a categorical variable from a sub-model means removing all associated dummies. While Shapley value decompositions are computationally demanding due to the number of sub-models required, this burden is reduced by focusing on the selected covariates. Even when starting from a large initial set of circumstances, the number of sub-models remains tractable if the model selected by penalization is sparse.

Formally, let K denote the number of covariates selected by the penalized procedure, where a categorical covariate counts as one, regardless of its number of categories. Each sub-model is indexed by a pair (κ, m) , where $\kappa \in \{0, \dots, K\}$ denotes the number of covariates excluded by this sub-model and $m \in \{1, \dots, M_\kappa = \binom{K}{\kappa}\}$ indexes the particular subset of variables excluded. Let $\widehat{G}_m^\kappa$ denote the explained Gini coefficient estimated from a non-penalized Lorenz regression fitted on the covariates included in sub-model (κ, m) . In particular, $\widehat{G}_1^0$ corresponds to the total explained Gini coefficient, obtained from the full set of selected covariates. Also, we set $\widehat{G}_m^K = 0$.

The contribution of the k th covariate is defined as the average marginal effect of including that covariate across all sub-models in which it appears. It is obtained as

$$c_k := \frac{1}{K} \sum_{l=0}^{K-1} \frac{\sum_{m \in N_k^l} \left(\widehat{G}_m^l - \widehat{G}_{m_{-k}}^{l+1} \right)}{\binom{K-1}{l}},$$

where N_k^l is the set of all sub-models that exclude l covariates and include the k th. The notation $(m_{-k}, l+1)$ is used for the corresponding sub-model where the k th covariate has been removed. By construction, the decomposition is additive

$$\widehat{G}_{Y,X} = \widehat{G}_1^0 = c_1 + \dots + c_K.$$

3 Application: IOp in education and income in Belgium, Spain and Norway

We apply here Lorenz regression analysis of inequality of opportunity in income and in education in three countries – Belgium, Spain and Norway – that differ widely in socio-economic structures. In international comparisons, Norway is often found to be a low IOp country and Spain a high IOp country with Belgium standing in-between the two; see Ferreira et al. (2026).

3.1 Data

We use data from the European Union Statistics on Income and Living Conditions (EU-SILC). EU-SILC is an official micro-data source for calculation of social indicators in Eu-

ropean countries. It is available annually since 2003 and provides detailed household- and individual-level annual income information for representative samples of the population in all EU Member States, either extracted from administrative sources, or elicited directly from respondents. We use micro-data from the September 2024 release of EU-SILC, which includes cross-sectional data up to 2023 (doi.org/10.2907/EUSILC2004-2023). Our analysis draws on three cross-sections of the data, collected in 2011, 2019 and 2023. In those three years, EU-SILC collected retrospective information about childhood circumstances of respondents aged from 25 to 59, referring mostly to the period when the respondent was around 14 years old (Andreoli and Fusco, 2019).

We focus on three countries (Belgium, Spain and Norway) and, within each year, distinguish two age cohorts: young individuals (age in $[25, 45]$) and old individuals (age in $(45, 59]$).

Our two applications differ in the definition of the advantage variable. For I-IOP, we measure advantage using single-adult equivalent household annual disposable income (as in, e.g., Brzezinski, 2020; Ramos and Van de gaer, 2021), denoted by Y_I . For E-IOP, the advantage variable, denoted by Y_E , is binary and indicates whether the individual has attained a tertiary education degree, formally defined as reaching an ISCED level of 500 or above (as in Palmisano et al., 2022).

Table 1 lists the variables included in the analysis. The first two rows describe the advantage variables. The remaining ones present the included circumstances as well as their categories. Variables observed at the level of the parents are available both for the mother and for the father. Data about parental occupation (MI and FI) are not available for Belgium and Spain in 2023. Besides parental information, we use the individual's sex (S), a dummy determining whether the individual lived with both parents (H), the tenancy status (O) and the financial situation (F) of the household when the individual was aged 14. We restrict our analytical sample to observations without any missing value and for which both parents are known.

The performance of the penalized Lorenz regression can be harmed by categories with extremely low frequencies.³ We therefore exclude categories with fewer than 10 individuals

³One simple illustration appears in the bootstrap procedure. If only a very few individuals share

Table 1: Variables considered in the analysis

Variable	Name	Categories
Income	Y_I	/
Tertiary education	Y_E	ISCED $\geq$ 500 (1), ISCED $<$ 500 (0)
Sex	S	female (1), male (2)
Parent education	ME, FE	low (1), medium (2), high (3)
Parent activity	MA, FA	not active (1), self-employed (2), employed (3)
Parent ISCO3	MI, FI	elementary (1), semi-skilled (2), top-ranked (3)
Parent native	MN, FN	not native (1), native (2)
Household composition	H	both parents (1), else (2)
Ownership	O	home owned (1), rented (2)
Financial situation	F	bad (1), moderate (2), good (3)

Variables are coded with two letters. The first indicates whether it provides information about the mother (M) or the father (F). The second relates to the covariate. Categories are identified with one digit, indicated in the third column of the table. For example, MA2 indicates that the mother is self-employed.

as well as the corresponding observations. This rules out self-employed mothers (MA2) in Belgium 2011 and Norway 2019, as well as self-employed fathers (FA2) in Norway 2019.

Categorical covariates are introduced as dummies. Several papers have shown that the fit of penalized regressions depends on the chosen reference category (see, for example, Chiquet et al., 2016). We alleviate this problem by including all dummies in the regression. To the main effects, we add all two-way interactions with, again, no fewer than 10 individuals. Descriptive statistics of the final datasets are given in the Online Supplement.

The I-IOP and E-IOP applications rely on different sample restrictions, reflecting the properties of their respective advantage variables. Concerning I-IOP, the method accommodates non-positive responses, so we retain observations with negative income. However, we exclude outliers based on their empirical influence on the Gini coefficient, keeping only those with an influence value below 2.5. This threshold, determined empirically, ensures the same category for a covariate, it might happen that the bootstrap resample contains none of these individuals. In that case, the regression cannot be performed on the proper set of covariates.

the stability of the analyses without removing too much information. Across the samples, this operation deletes at most 0.8% of the observations. For E-IOp, we restrict the sample by excluding individuals who are currently in education. Table 2 reports the final sample sizes for both applications. For each I-IOp sample, the number of excluded outliers and the number of observations with a non-positive income are reported in Table 7 in the Online Supplement. Observations with a non-positive income are retained for the Lorenz regression analysis but excluded for the parametric competitor, see Section 3.3.

Table 2: Final sample sizes

	I-IOp						E-IOp					
	Old			Young			Old			Young		
	2011	2019	2023	2011	2019	2023	2011	2019	2023	2011	2019	2023
Belgium	1683	2369	2096	2557	3331	3017	1614	2284	2080	2362	3152	2878
Spain	5861	7101	12538	8379	8133	11615	5594	6991	12188	7662	7708	10664
Norway	1010	489	1215	1400	1122	1506	995	493	1172	1246	1122	1323

3.2 Implementation

The Lorenz regression procedure developed in Jacquemain et al. (2024) needs to be adapted to account for the EU-SILC sample design. First, the bootstrap procedure needs to preserve the intra-household dependence by resampling at the household level rather than at the individual level. Second, EU-SILC includes personal weights that correct for unequal probabilities of selection and non-response. Let π denote the vector of normalized sampling weights, satisfying $\pi_i \geq 0 \forall i = 1, \dots, n$ and $\sum_{i=1}^n \pi_i = 1$. We first adjust the estimation of the distribution function of the index. Equation (6) becomes

$$\hat{F}_i(\theta) := \sum_{j=1}^n \pi_j \left[\mathbb{1}\{X_i^\top \theta \geq X_j^\top \theta\} - \frac{1}{2} \mathbb{1}\{X_i^\top \theta = X_j^\top \theta\} \right],$$

In the non-penalized procedure, the objective function (7) and the estimator of the explained Gini coefficient (8) are replaced with

$$\hat{\theta} = \arg \max_{\theta, \|\theta\|=1} \sum_{i=1}^n \pi_i Y_i \hat{F}_i(\theta),$$

$$\hat{G}_{Y,X} = \frac{2}{\bar{Y}} \sum_{i=1}^n \pi_i Y_i \hat{F}_i(\hat{\theta}) - 1.$$

Likewise, the objective function of the penalized procedure (9) becomes

$$\hat{\theta} = \arg \max_{\theta, \|\theta\|=1} \left\{ \sum_{i=1}^n \sum_{j=1}^n \pi_i \pi_j Y_i K \left(\frac{X_i^\top \theta - X_j^\top \theta}{h} \right) - \sum_{k=1}^p p_\lambda(|\theta_k|) \right\}.$$

The solution is similar to the adaptation of the computation of the Gini coefficient proposed by Lerman and Yitzhaki (1989).

All computations are performed in R 4.5.0. Lorenz regressions are estimated with the `LorenzRegression` package (Jacquemain and Heuchenne, 2026). In each dataset, we first fit a penalized Lorenz regression to select relevant covariates. The choice of penalty (LASSO or SCAD) and of the tuning parameter λ relies on a bootstrap selection score. As documented in Jacquemain and Heuchenne (2026), this score may plateau, suggesting that a slightly less optimal but sparser model is available. To balance fit and parsimony, we normalize the score to one at its maximum and retain the sparsest model with a value above 0.99. Finally, a non-penalized Lorenz regression is refitted on the selected covariates, and its results form the basis of our analysis. Inequality of opportunity is then measured by the explained Gini coefficient, $\hat{G}_{Y,X}$.

3.3 Results

3.3.1 Lorenz regression vs. parametric linear models

We first compare estimates of inequality of opportunity and the number of distinct types and opportunity sets obtained from Lorenz regression in comparison to a parametric benchmark. We benchmark the Lorenz regression against penalized generalized linear models (GLM), estimated with the `glmnet` package (Tay et al., 2023). Following Bourguignon et al. (2007) and Palmisano et al. (2022), we use a log-linear regression for I-IOP and a

logistic regression for E-IOp. The penalty parameter is chosen by cross-validation. We transform the score so that higher is better and the optimum is attained at 1. To ensure a fair comparison with the Lorenz regression, we retain the sparsest model whose score remains above 0.99. These models provide two estimates of inequality of opportunity. The first uses the Gini coefficient of fitted values, $\widehat{G}_{\hat{Y}}$, which corresponds to the conventional approach. The second, motivated by the critique of Escanciano and Terschuur (2022), relies on the concentration index of the outcome with respect to the fitted values, $\widehat{CI}_{Y,\hat{Y}}$, which is closer in spirit to the explained Gini coefficient $\widehat{G}_{Y,X}$.

IOp estimates and Lorenz- R^2 for income IOp are presented in Figure 1. Estimates of IOp and Lorenz- R^2 for educational IOp are in Figure 2. Both figures display IOp estimates for the Lorenz regression ($\widehat{G}_{Y,X}$) and parametric alternatives ($\widehat{CI}_{Y,\hat{Y}}$ and $\widehat{G}_{\hat{Y}}$), together with total inequality ($\widehat{G}_Y$). Countries appear as separate facets, years on the horizontal axis, methods as colors, and age cohorts as line types, allowing a direct comparison across dimensions. We present an analogous figure for the Lorenz- R^2 . Tables 3 and 5 also report indicators of model size, using two complementary perspectives. The first is the number of selected circumstances, which reflects how many variables are retained as relevant for explaining inequality. The second is the number of estimated opportunity sets, defined as the distinct combinations of selected circumstances that yield different fitted values.

Inequality of opportunity in income (I-IOp). Overall, the three approaches deliver estimates of I-IOp of comparable magnitude. For Belgium and Spain, explained Gini coefficients from the Lorenz regression ($\widehat{G}_{Y,X}$) are very close to the parametric estimates obtained with the concentration index ($\widehat{CI}_{Y,\hat{Y}}$), whereas those based on the Gini coefficient of fitted values ($\widehat{G}_{\hat{Y}}$) are consistently higher. Given the instability issues discussed in Escanciano and Terschuur (2022), this pattern may reflect upward bias in the latter. In Norway, by contrast, the Lorenz regression yields systematically higher IOp estimates than both alternative models.

Lorenz regression tends to select fewer variables and generate substantially fewer opportunity sets (Table 3). In Belgium and Spain, this translates into sparser models with similar levels of estimated IOp, though in the specific case of the old cohort in 2011, the

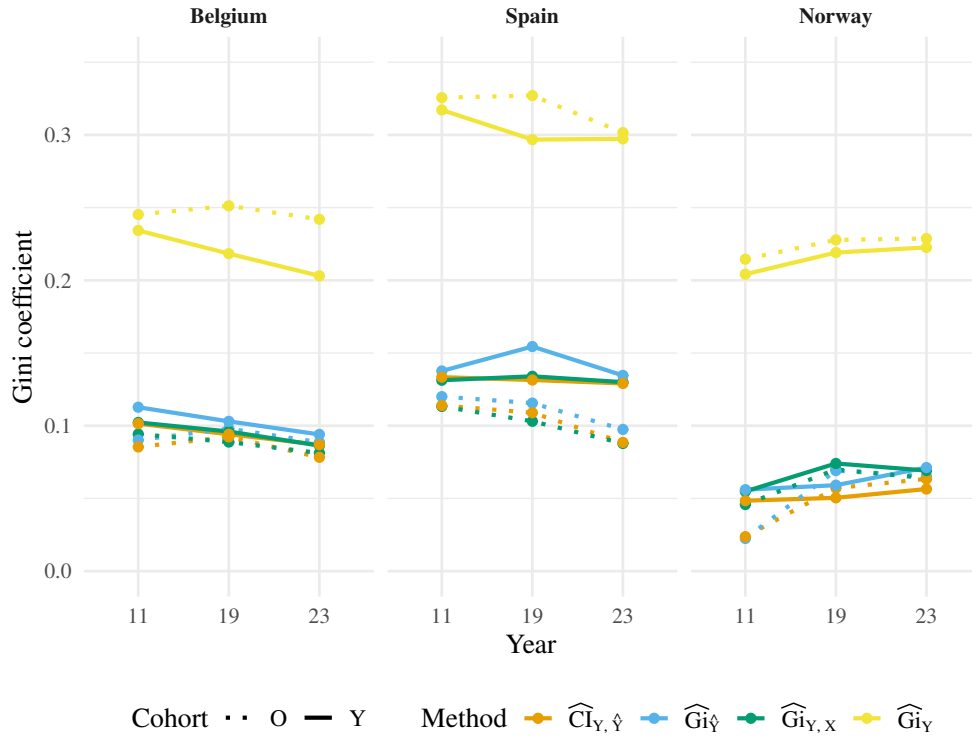
Lorenz regression delivers a 0.9 percentage point higher IOp despite comparable model size. In Norway, the contrast is sharper: the Lorenz regression consistently yields higher IOp while generally relying on fewer opportunity sets than the penalized log-linear regression. For example, for the old cohort in 2019, it retained 8 opportunity sets versus 20 for the penalized GLM but estimated a higher level of IOp (7.0% against 5.7%). These patterns highlight the main advantage of the Lorenz regression procedure. By placing the inequality explanation at the core of its optimization, the procedure is able to identify the structural relationships that are the most constitutive of unequal opportunities. In practice, this leads to similar or higher levels of explained inequality with more parsimonious models.

Table 3: Selected circumstances and generated opportunity sets – I-IOp

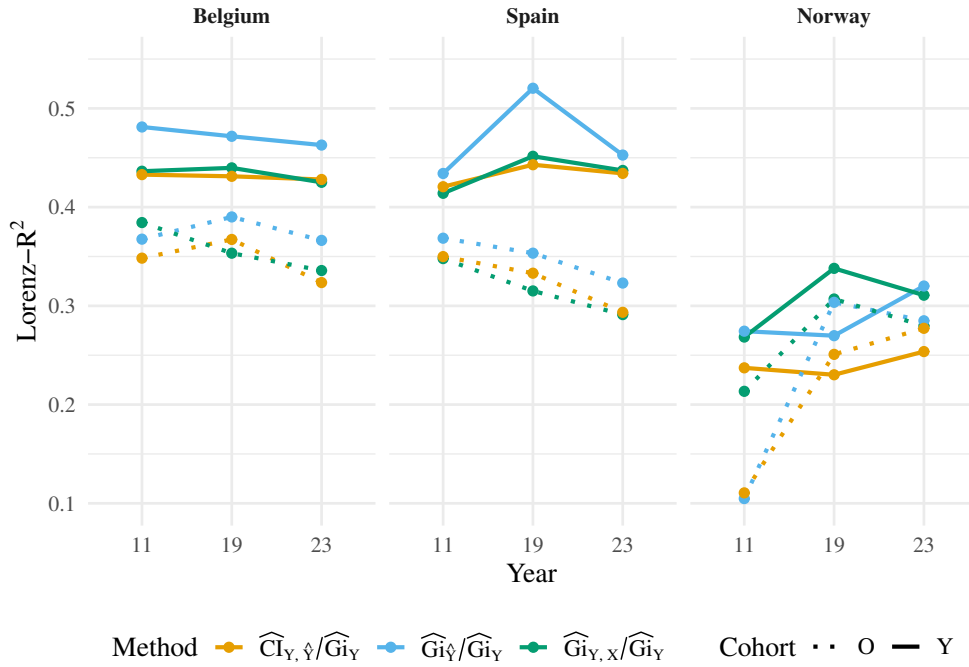
	Number of circumstances						Number of opportunity sets		
	included			selected: GLM / PLR			GLM / PLR		
	2011	2019	2023	2011	2019	2023	2011	2019	2023
<i>Belgium</i>									
Old	297	333	231	20 / 19	29 / 10	12 / 9	144 / 194	496 / 63	101 / 46
Young	384	429	291	20 / 11	29 / 17	24 / 11	435 / 139	785 / 587	413 / 75
<i>Spain</i>									
Old	321	318	291	31 / 13	33 / 7	46 / 14	400 / 78	568 / 29	825 / 183
Young	429	447	300	39 / 12	46 / 14	31 / 13	888 / 122	1203 / 228	852 / 78
<i>Norway</i>									
Old	149	160	249	2 / 9	6 / 4	9 / 6	3 / 46	20 / 8	43 / 17
Young	197	315	378	10 / 9	8 / 9	15 / 12	87 / 41	30 / 53	101 / 84

Table 4 reports confidence intervals for the Lorenz regression estimates of explained Gini coefficients and Lorenz- R^2 . For the explained Gini coefficient, the lengths of most intervals are around 2 to 3 percentage points. The intervals can therefore be sizable relative to the magnitude of IOp, which translates into wide intervals for the Lorenz- R^2 , especially in

Figure 1: Estimates of absolute and relative I-IOp



(a) Absolute I-IOp



(b) Relative I-IOp

smaller samples such as Norway.

Table 4: Confidence intervals for the Lorenz regression – I-IOp

	Explained Gini coefficient			Lorenz- R^2		
	2011	2019	2023	2011	2019	2023
<i>Belgium</i>						
Old	[8.0, 10.8]	[7.6, 10.2]	[6.8, 9.4]	[33.6, 43.3]	[30.8, 39.9]	[28.5, 38.7]
Young	[9.1, 11.3]	[8.7, 10.5]	[7.6, 9.7]	[39.7, 47.6]	[40.6, 47.3]	[38.2, 46.8]
<i>Spain</i>						
Old	[10.2, 12.5]	[9.1, 11.5]	[8.0, 9.6]	[31.6, 38.0]	[28.2, 34.8]	[26.7, 31.6]
Young	[12.1, 14.2]	[12.3, 14.5]	[12.2, 13.8]	[38.6, 44.2]	[42.0, 48.3]	[41.3, 46.2]
<i>Norway</i>						
Old	[3.1, 6.0]	[4.7, 9.3]	[4.8, 8.0]	[15.1, 27.5]	[22.0, 39.3]	[21.7, 34.3]
Young	[4.3, 6.7]	[6.1, 8.7]	[5.6, 8.2]	[21.2, 32.5]	[28.5, 39.1]	[25.7, 36.4]

Inequality of opportunity in education (E-IOp). Across all three countries, the different approaches yield remarkably consistent estimates of E-IOp. However, as visible from Table 5, the Lorenz regression always selects fewer circumstances and generates fewer opportunity sets than the penalized GLM, echoing the pattern observed for income. The educational results thus reinforce the advantage of the Lorenz regression of being able to provide stable estimates, while relying on more parsimonious, interpretable models.

Table 6 reports confidence intervals for the Lorenz regression estimates. As with income, intervals are relatively wide, particularly for smaller samples such as in Norway. Compared to I-IOp, where the most striking feature was the cross-country ranking, E-IOp is marked by a robust temporal decline: in all cases except the young cohort of Norway, the 2023 intervals lie entirely below those of 2011. Examining Lorenz- R^2 highlights further patterns. For the old cohort in 2023, the country ranking in relative inequality is confirmed by non-

overlapping intervals. Among young cohorts, the only significant difference is Spain versus Belgium. Cohort wedges in the Lorenz- R^2 are confirmed: for Spain in 2019 and for both Spain and Norway in 2023.

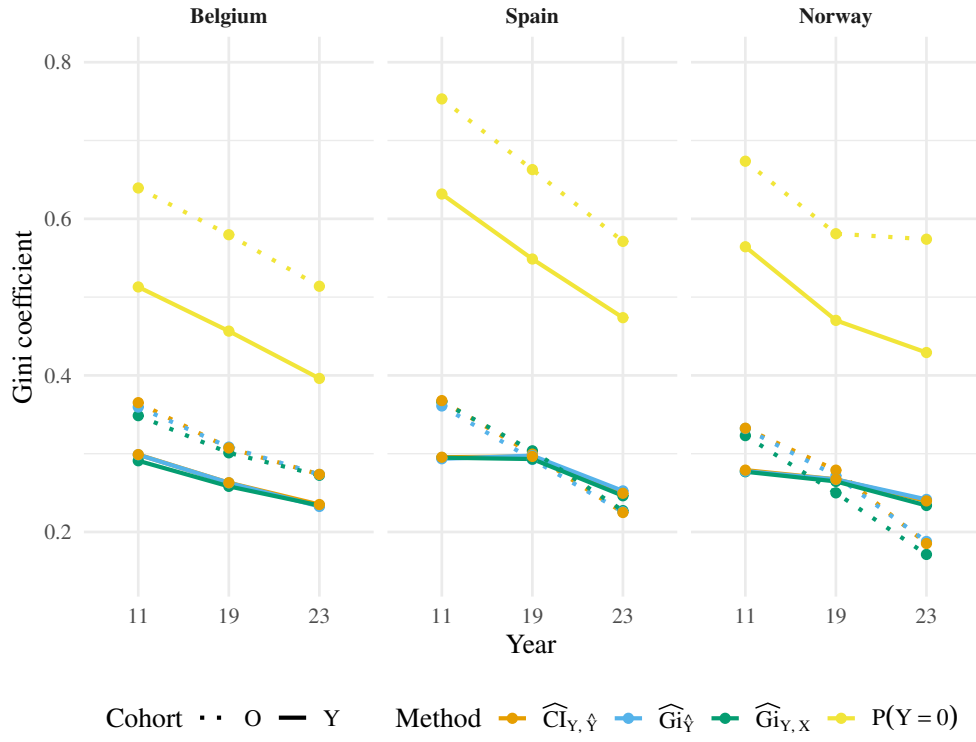
Table 5: Selected circumstances and generated opportunity sets – E-IOp

	Number of circumstances						Number of opportunity sets		
	included			selected: GLM / PLR			GLM / PLR		
	2011	2019	2023	2011	2019	2023	2011	2019	2023
<i>Belgium</i>									
Old	241	275	185	14 / 6	22 / 12	19 / 14	103 / 19	504 / 156	324 / 182
Young	319	371	245	29 / 14	24 / 14	20 / 12	701 / 354	820 / 336	405 / 226
<i>Spain</i>									
Old	263	260	245	26 / 17	24 / 20	27 / 23	270 / 140	398 / 260	631 / 437
Young	353	371	254	26 / 17	26 / 14	27 / 13	962 / 456	762 / 250	744 / 204
<i>Norway</i>									
Old	107	116	182	17 / 9	7 / 4	7 / 4	179 / 77	39 / 11	19 / 9
Young	155	261	296	21 / 16	22 / 16	16 / 11	309 / 226	302 / 224	142 / 102

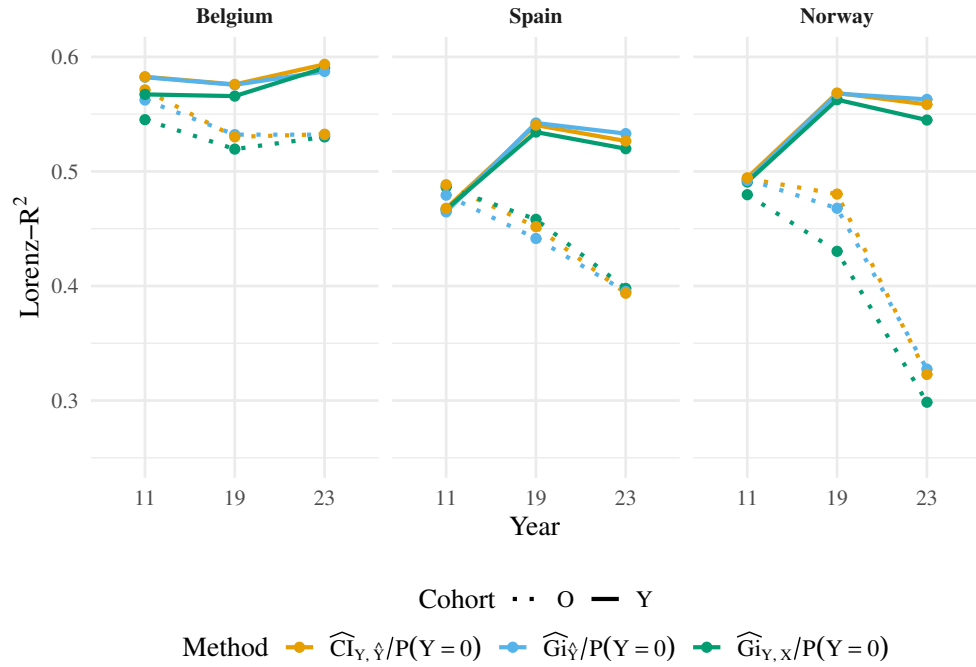
3.3.2 Substantive results and the role of circumstances in Lorenz regression-based models

Cross-country patterns. Cross-country rankings differ between the two outcomes. For income (Figure 1), Spain records both the highest total inequality and the highest IOp; Belgium sits at an intermediate level of inequality but produces a Lorenz- R^2 strikingly close to Spain's, implying a comparable relative role for circumstances; Norway combines total inequality close to Belgian levels with a markedly lower IOp, and thus the smallest Lorenz- R^2 of the three countries. This ranking is stable across cohorts and years. For education (Figure 2), the ranking differs: Spain has the highest non-attainment, followed by Norway,

Figure 2: Estimates of absolute and relative E-IOp



(a) Absolute E-IOp



(b) Relative E-IOp

Table 6: Confidence intervals for the Lorenz regression – E-IOp

	Explained Gini coefficient			Lorenz- R^2		
	2011	2019	2023	2011	2019	2023
<i>Belgium</i>						
Old	[31.2, 38.5]	[27.4, 32.8]	[24.2, 30.2]	[49.2, 59.8]	[47.7, 56.2]	[48.0, 58.0]
Young	[26.4, 31.8]	[23.7, 28.0]	[21.2, 25.6]	[52.4, 61.0]	[52.9, 60.2]	[55.3, 62.8]
<i>Spain</i>						
Old	[33.8, 39.4]	[28.1, 32.7]	[21.2, 24.3]	[44.9, 52.3]	[42.5, 49.2]	[37.3, 42.3]
Young	[27.2, 31.7]	[27.3, 31.4]	[23.0, 26.3]	[43.4, 50.0]	[50.2, 56.6]	[49.3, 54.7]
<i>Norway</i>						
Old	[27.8, 36.8]	[19.4, 30.6]	[13.4, 20.9]	[41.5, 54.4]	[34.0, 52.0]	[23.6, 36.1]
Young	[24.4, 30.9]	[23.3, 29.6]	[20.7, 26.1]	[43.9, 54.3]	[50.8, 61.7]	[49.3, 59.7]

with Belgium lowest, yet absolute IOp levels are similar across countries, especially among the young – strikingly, Belgium records the highest IOp for the old cohort despite having the highest tertiary attainment. Relative to non-attainment, Belgium stands out further: it has systematically the highest Lorenz- R^2 , so that despite the smallest attainment gaps, family background explains the largest share of the (smaller) remaining disparities.

Trends over time. Time patterns also diverge across outcomes. Between 2011 and 2023, income IOp followed no single trajectory: it declined in Belgium (more markedly among the young), declined steadily among Spain’s older workers while remaining essentially flat for Spain’s young workers, and rose modestly in Norway for both cohorts. The Spanish contrast is instructive: because total inequality fell for both cohorts, the Lorenz- R^2 moved in opposite directions – up 2.3 percentage points for the young, down 5.7 for the old – showing how the relative measure can diverge from the level. Educational IOp, in contrast, declined almost everywhere alongside rising tertiary attainment, but asymmetrically across

cohorts: the decline was much steeper for the old cohort in Spain (share of disparities explained by background: 39.8% old vs. 52.0% young by 2023) and Norway (29.9% vs. 54.5%), whereas Belgium’s decline was roughly parallel across cohorts, preserving the cohort ranking.

Age cohorts. A consistent pattern emerges across both outcomes when comparing age cohorts. For income, total inequality is higher among the old, yet the Lorenz- R^2 is systematically higher among the young (e.g., in Spain 2023: total inequality 29.7% young vs. 30.2% old, but explained Gini 13.0% vs. 8.8%). As differences accumulate over the life cycle, a shrinking share is attributable to circumstances. Education tells a related story from a different angle: in 2011, IOp was higher among the old cohort everywhere, but by 2023 this gap had narrowed (Spain), reversed (Norway), or persisted only in Belgium – and the Lorenz- R^2 shows the same widening wedge favouring the young that is observed for income. Together, these results point to a common life-cycle regularity, visible in both outcomes: circumstances weigh more heavily on advantages measured earlier in life.

Leading circumstances: education, origin, and occupation. Turning to the Shapley value decomposition (Figures 3 and 6), parental education, origin, and occupation dominate both applications, jointly explaining about 75% of I-IOp and between 49% and 99% of E-IOp. Parental education is the leading circumstance in both cases (median contribution of 29% for income, 48% for education) and rises over time in Belgium and Spain for both outcomes. The heatmaps (Figures 4–5 and 7–8) reveal a further, more subtle regularity: over-representation of low parental education among the least advantaged has been *rising* over time in all three countries (e.g., from 1.2 to 2.8 times more frequent in Belgium), even as its prevalence in the population fell sharply, whereas over-representation of high parental education among the most advantaged has been *falling* (e.g., from 7.6 to 5.1 times in Spain). Disadvantage is becoming more concentrated even as advantage becomes more diffuse. The remaining two circumstances diverge sharply between outcomes. Parental origin, the second pillar for income (median 26%, driven by strong over-representation of foreign-born parents among the least advantaged, especially in Spain), is a minor circum-

stance for education (median 3.6%). Parental occupation, conversely, contributes similarly for income (15–20% on average) but more for education (median 31%), with fathers in elementary occupations more over-represented among the least advantaged on the education ladder than on the income ladder, particularly in Belgium (2.0–4.3 vs. 0.7–1.9 times).

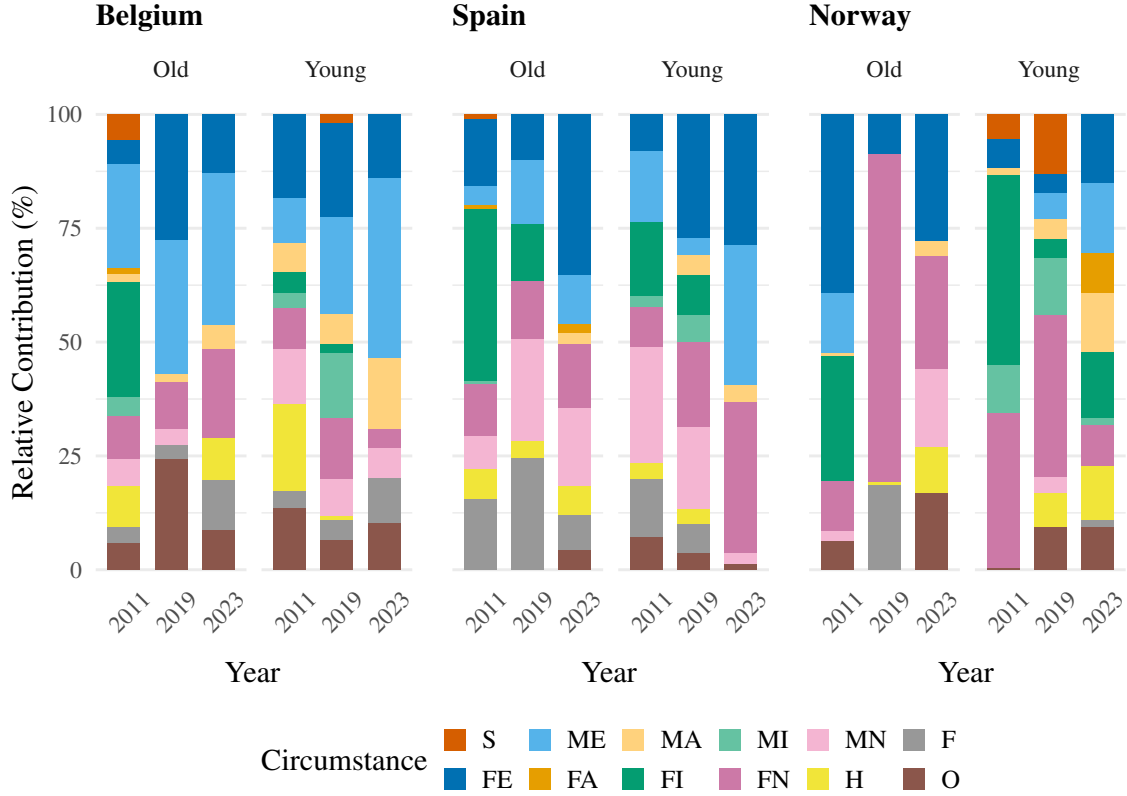
Paternal characteristics tend to dominate over maternal ones in both applications, especially in Spain and Norway – where gaps reach up to 80 percentage points for income and 95 for education – while Belgium is comparatively the most balanced country. Parental activity is the one clear exception, running almost entirely through the mother in both applications.

Secondary circumstances. Among the remaining circumstances, patterns again contrast across outcomes. Parental activity status is a minor, mother-driven determinant in both applications (rarely above 5%). Sex plays almost no role in income IOp (below 10% nearly everywhere), consistent with household income pooling across partners, but contributes more to educational IOp (5–15%, peaking at 20% in Norway), with males over-represented among the least advantaged, especially in younger cohorts. Household composition is modest for income (up to 19% in one case, below 12% elsewhere) and even more limited for education (rarely above 5%, exceptionally up to 10% in Norway). Financial situation contributes similarly to both outcomes (mainly in 2023, and more so in Spain), but home ownership matters only for income – particularly in Belgium and Norway, where it is consistently associated with the most advantaged group – explaining less than 10% of educational IOp.

4 Conclusion

This paper presents a new tool to analyze inequality of opportunity. It revolves around the explained Gini coefficient, an inequality measure for the conditional expectation of an advantage variable given a set of circumstances. This quantity is estimated with the penalized Lorenz regression, which offers several key advantages. It is flexible because it assumes a semi-parametric single-index model. It is parsimonious in the estimation of its

Figure 3: Relative contributions of circumstances to I-IOp (Shapley value decomposition)

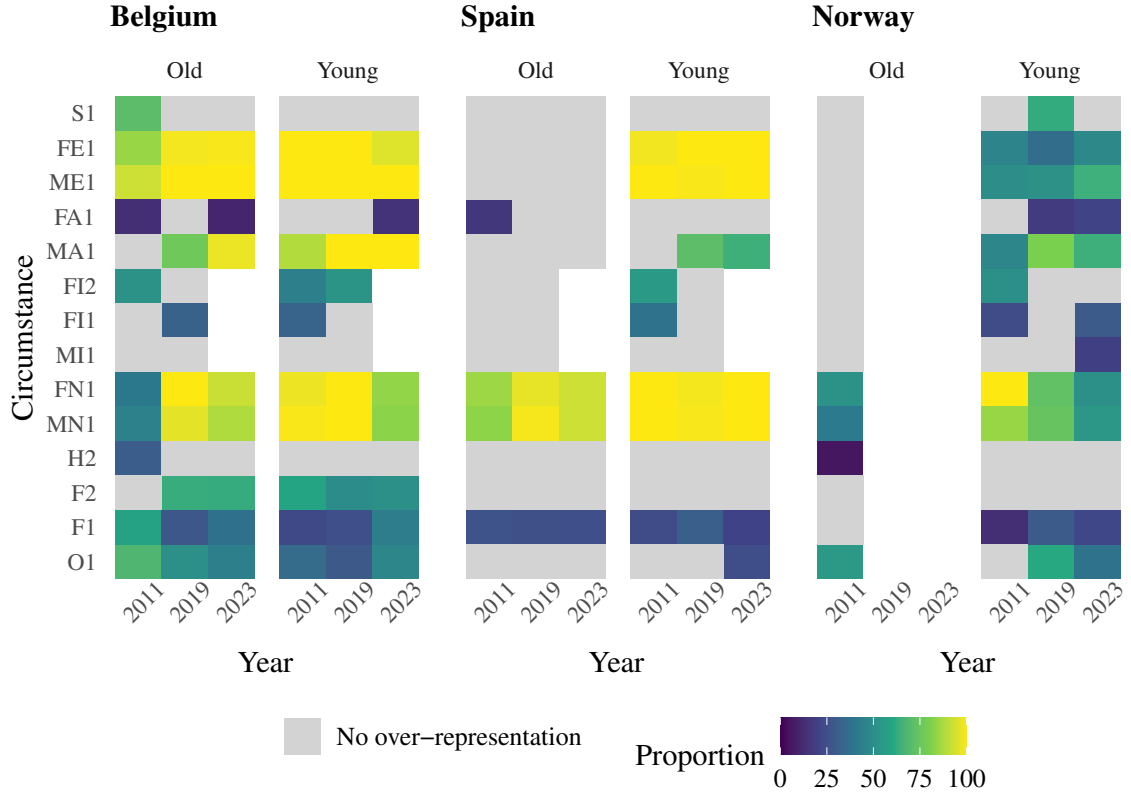


parameters since the link function needs not be estimated. It is robust since it relies on a concentration index. It also comes with a proper inferential theory and selects the relevant circumstances with a probability tending to one. It extends naturally to binary outcomes, where maximizing the explained Gini coefficient is equivalent to maximizing the AUC of a linear classifier, and it accommodates the survey weights required when working with data such as EU-SILC.

In the three countries analyzed, our procedure performs favorably compared to penalized parametric benchmarks – a log-linear model for income and a logistic model for education. Typically, it proposes a sparser model while providing comparable, and in several cases higher, estimates of IOp. Across countries, years and cohorts, estimated income IOp (Table 4) ranges from about 3% to 15%, or 15% to 48% once expressed as a share of total income inequality; estimated educational IOp (Table 6) ranges from about 13% to

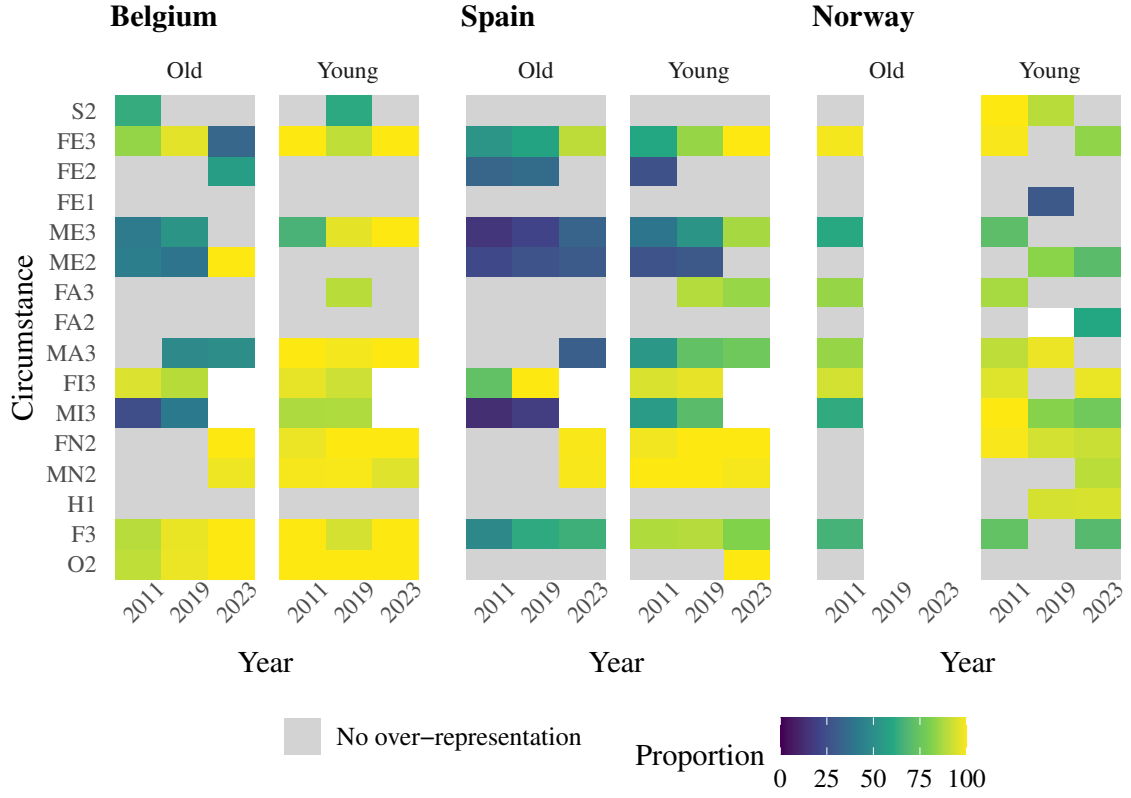
39%, or 24% to 63% as a share of total disparities in attainment. Norway tends to record the lowest income IOp, both in absolute and relative terms, while cross-country rankings are closer for education. In relative terms, opportunities are generally more unequal for the young-age cohorts in both outcomes, even though total inequality is typically higher in the old-age cohorts. Using the Shapley value decomposition, we further find that the most important circumstances throughout are the parents' levels of education, types of occupation and nationalities.

Figure 4: Heatmap - I-IOp- low advantage



Note: A neutral color is used whenever the frequency is not sufficiently higher than in the overall sample – less than three times higher or with an absolute difference below 10 percentage points. Categories never meeting this criterion in any dataset are excluded. Among the retained categories, color intensity reflects the frequency within the group. The linear index is a discrete variable, whose distinct values characterize different opportunity sets, see Tables 3 and 5. As a consequence, the proportion of individuals within the first and last deciles may be different from 10%. We exclude from the analysis situations where the least advantaged set or the most advantaged set includes less than 5% or more than 15% of individuals.

Figure 5: Heatmap - I-IOp- high advantage



Note: A neutral color is used whenever the frequency is not sufficiently higher than in the overall sample – less than three times higher or with an absolute difference below 10 percentage points. Categories never meeting this criterion in any dataset are excluded. Among the retained categories, color intensity reflects the frequency within the group. The linear index is a discrete variable, whose distinct values characterize different opportunity sets, see Tables 3 and 5. As a consequence, the proportion of individuals within the first and last deciles may be different from 10%. We exclude from the analysis situations where the least advantaged set or the most advantaged set includes less than 5% or more than 15% of individuals.

Figure 6: Relative contributions of circumstances to E-IOp (Shapley value decomposition)

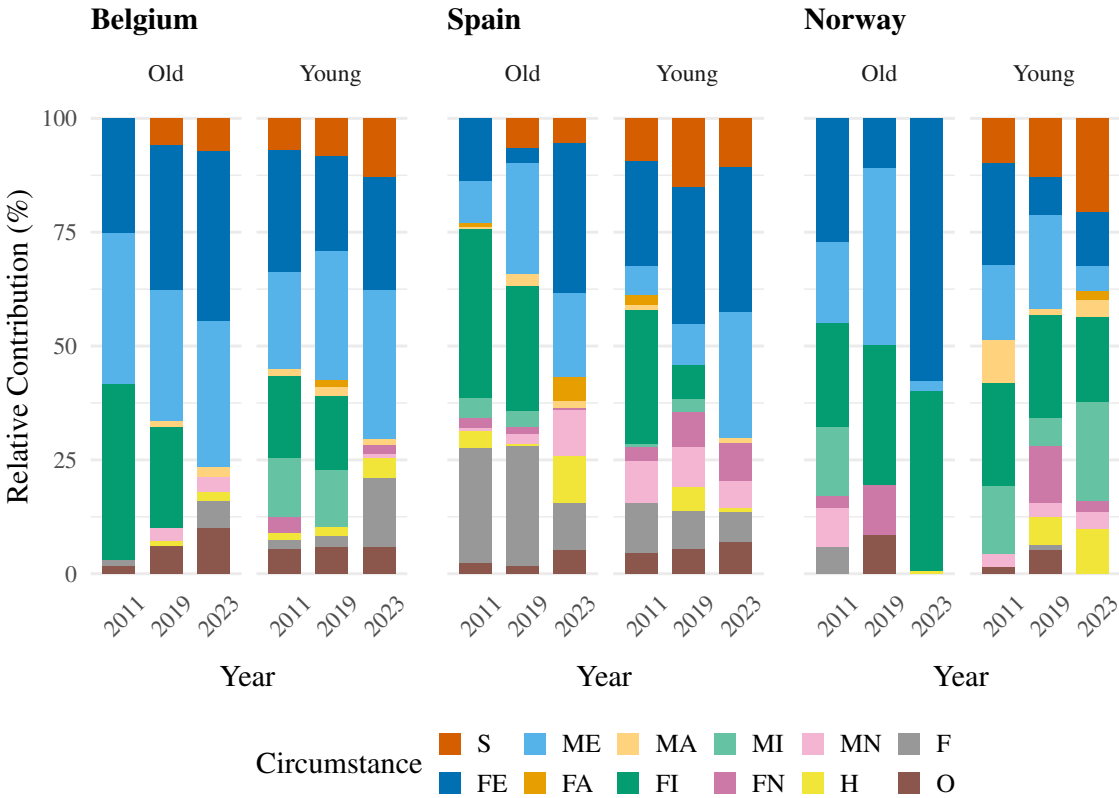
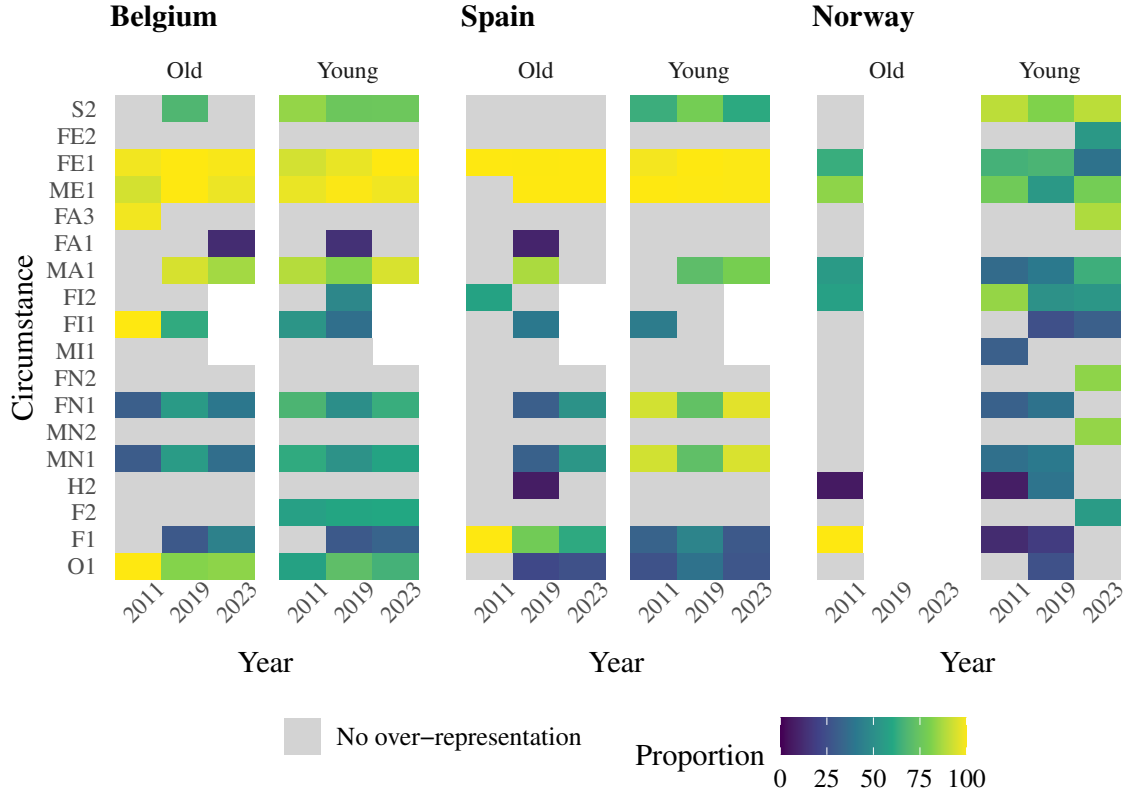
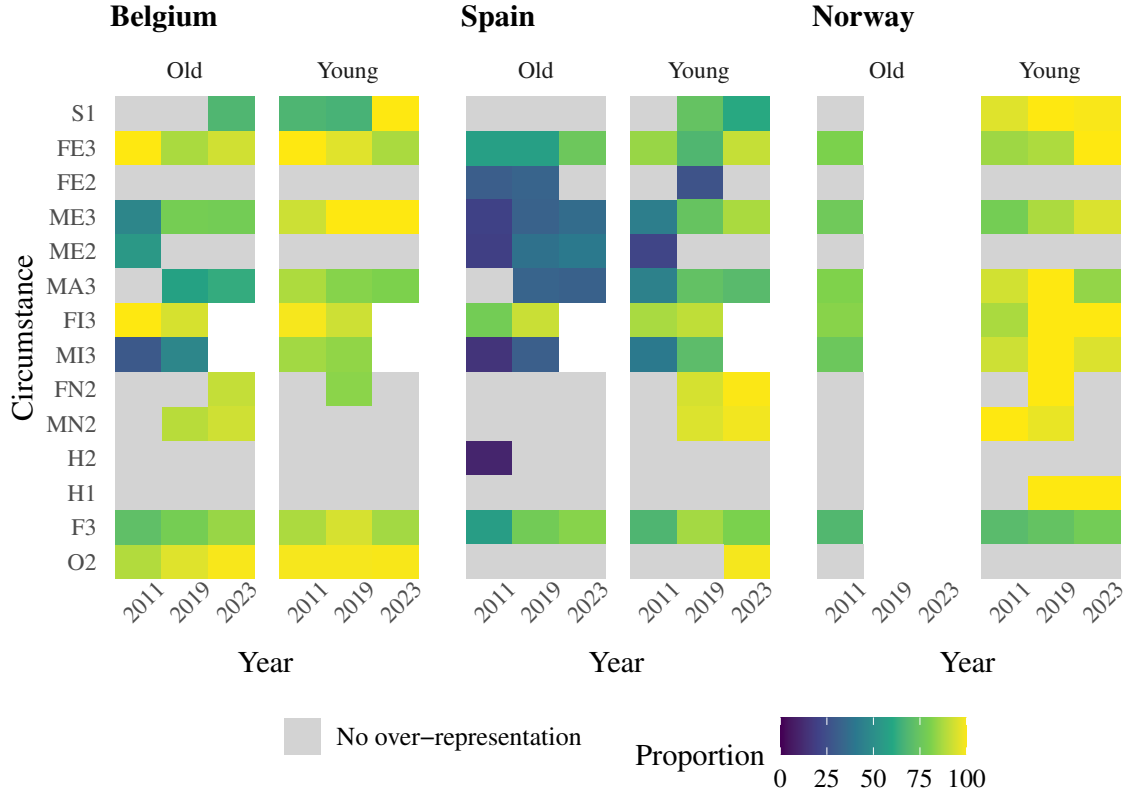


Figure 7: Heatmap - E-IOp- low advantage



Note: A neutral color is used whenever the frequency is not sufficiently higher than in the overall sample – less than three times higher or with an absolute difference below 10 percentage points. Categories never meeting this criterion in any dataset are excluded. Among the retained categories, color intensity reflects the frequency within the group. The linear index is a discrete variable, whose distinct values characterize different opportunity sets, see Tables 3 and 5. As a consequence, the proportion of individuals within the first and last deciles may be different from 10%. We exclude from the analysis situations where the least advantaged set or the most advantaged set includes less than 5% or more than 15% of individuals.

Figure 8: Heatmap - E-IOp- high advantage



Note: A neutral color is used whenever the frequency is not sufficiently higher than in the overall sample – less than three times higher or with an absolute difference below 10 percentage points. Categories never meeting this criterion in any dataset are excluded. Among the retained categories, color intensity reflects the frequency within the group. The linear index is a discrete variable, whose distinct values characterize different opportunity sets, see Tables 3 and 5. As a consequence, the proportion of individuals within the first and last deciles may be different from 10%. We exclude from the analysis situations where the least advantaged set or the most advantaged set includes less than 5% or more than 15% of individuals.

References

- Andreoli, F. and A. Fusco (2019). Application of the EU-SILC 2011 data module “intergenerational transmission of disadvantage” to robust analysis of inequality of opportunity. *Data in Brief* 25(1043016), 1–13.
- Bernardo, G., S. N. Durlauf, A. Kourtellos, and C. M. Tan (2025). Model uncertainty and measures of inequality of opportunity. Working Paper 34163, National Bureau of Economic Research.
- Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources* 8(4), 436–455.
- Bourguignon, F., F. H. G. Ferreira, and M. Menéndez (2007). Inequality of Opportunity in Brazil. *Review of Income and Wealth* 53(4), 585–618.
- Bresnahan, T. F. and P. C. Reiss (1991). Entry and competition in concentrated markets. *Journal of Political Economy* 99(5), 977–1009.
- Bricard, D., F. Jusot, A. Trannoy, and S. Tubeuf (2020). Inequality of opportunities in health and death: an investigation from birth to middle age in Great Britain. *International Journal of Epidemiology* 49(5), 1739–1748.
- Brunori, P., P. Hufe, and D. G. Mahler (2018). The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees. SSRN Scholarly Paper ID 3127234, Social Science Research Network, Rochester, NY.
- Brunori, P., V. Peragine, and L. Serlenga (2019). Upward and downward bias when measuring inequality of opportunity. *Social Choice and Welfare* 52(4), 635–661.
- Brzezinski, M. (2020). The evolution of inequality of opportunity in Europe. *Applied Economics Letters* 27(4), 262–266.
- Carranza, R. (2022). Upper and lower bound estimates of inequality of opportunity: A cross-national comparison for europe. *Review of Income and Wealth* (69), 838–860.

- Chiquet, J., Y. Grandvalet, and G. Rigaiil (2016). On coding effects in regularized categorical regression. *Statistical Modelling* 16(3), 228–237.
- Cowell, F. A. (2011). *Measuring Inequality* (3rd ed.). LSE Perspectives in Economic Analysis. Oxford: Oxford University Press.
- Cowell, F. A. and S. P. Jenkins (1995). How much inequality can we explain? A methodology and an application to the United States. *The Economic Journal* 105(429), 421–430.
- Cowell, F. A. and M.-P. Victoria-Feser (1996). Robustness Properties of Inequality Measures. *Econometrica* 64(1), 77–101.
- Dworkin, R. (1981). What Is Equality? Part 2: Equality of Resources. *Philosophy and Public Affairs* 10(4), 283–345.
- Escanciano, J. C. and J. R. Terschuur (2022). Debiased Semiparametric U-Statistics: Machine Learning Inference on Inequality of Opportunity.
- Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *The Journal of Finance* 47(2), 427–465.
- Ferreira, F. H. G. and J. Gignoux (2011). The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of Income and Wealth* 57(4), 622–657.
- Ferreira, F. H. G. and J. Gignoux (2014). The measurement of educational inequality: Achievement and opportunity. *The World Bank Economic Review* 28(2), 210–246.
- Ferreira, F. H. G., V. Peragine, P. Brunori, P. Salas-Rajo, D. Moramarco, L. Barajas, T. Barbieri, N. Daza-Báez, G. Datt, V. de Sandi, F. Farella, A. Martinez Jr., J. Nguyen, A. Park, E. Simeone, L. Sirugue, P. T. López, and G. Zotti (2026). Global estimates of opportunity and mobility: A database. Stone Center Working Paper Series 120, Graduate Center, City University of New York, USA.
- Graeber, D., J. König, and V. Hilbert. Inequality of opportunity in wealth: Levels, trends and drivers in Germany. *Oxford Bulletin of Economics and Statistics* n/a(n/a).

- Heuchenne, C. and A. Jacquemain (2022). Inference for monotone single-index conditional means: A Lorenz regression approach. *Computational Statistics & Data Analysis* 167(C).
- Hoyos, A. and A. Narayan (2011). Inequality of Opportunities Among Children: How Much Does Gender Matter? Working Paper, World Bank, Washington, DC.
- Hufe, P., A. Peichl, and D. Weishaar (2022). Lower and upper bound estimates of inequality of opportunity for emerging economies. *Social Choice and Welfare* 58(3), 395–427.
- Jacquemain, A. and C. Heuchenne (2026). LorenzRegression: An Implementation of the Lorenz and Penalized Lorenz Regressions in R. *Journal of Statistical Software* 117(6), 1–33.
- Jacquemain, A., C. Heuchenne, and E. Pircalabelu (2024). A penalised bootstrap estimation procedure for the explained Gini coefficient. *Electronic Journal of Statistics* 18(1), 247 – 300.
- Jenkins, S. P. and P. Van Kerm (2009). The measurement of economic inequality. In W. Salverda, B. Nolan, and T. M. Smeeding (Eds.), *Oxford Handbook on Economic Inequality*, Chapter 3, pp. 40–67. Oxford: Oxford University Press.
- Krafft, C. and H. Alawode (2018). Inequality of opportunity in higher education in the Middle East and North Africa. *International Journal of Educational Development* 62, 234–244.
- Lerman, R. I. and S. Yitzhaki (1989). Improving the accuracy of estimates of Gini coefficients. *Journal of Econometrics* 42(1), 43–47.
- Marrero, G. A., J. C. Palomino, and G. Sicilia (2024). Inequality of opportunity in educational achievement in Western Europe: contributors and channels. *Journal of Economic Inequality* (2), 383–410.
- Mitnik, P., A.-L. Helsø, and V. L. Bryant (2020). Inequality of opportunity for income in Denmark and the United States: A comparison based on administrative data. Working Paper 27835, National Bureau of Economic Research.

- Mussard, S. and F. Souissi-Benrejab (2019). Gini-PLS regressions. *Journal of Quantitative Economics* 17(3), 477–512.
- Oaxaca, R. L. (1973). Male-female wage differentials in urban labor markets. *International Economic Review* 14(3), 693–709.
- Olkin, I. and S. Yitzhaki (1992). Gini regression analysis. *International Statistical Review/Revue Internationale de Statistique*, 185–196.
- Paes de Barros, R., F. H. G. Ferreira, J. R. Molinas Vega, and J. S. Chanduvi (2009). *Measuring inequality of opportunities in Latin America and the Caribbean*. Palgrave Macmillan: The International Bank for Reconstruction and Development / The World Bank.
- Palmisano, F., F. Biagi, and V. Peragine (2022). Inequality of opportunity in tertiary education: Evidence from Europe. *Research in Higher Education* 63, 514–565.
- Pepe, M. S. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series. Oxford, New York: Oxford University Press.
- Pepe, M. S., T. Cai, and G. Longton (2006). Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve. *Biometrics* 62(1), 221–229.
- Ramos, X. and D. Van de gaer (2016). Approaches to inequality of opportunity: Principles, measures and evidence. *Journal of Economic Surveys* 30(5), 855–883.
- Ramos, X. and D. Van de gaer (2021). Is inequality of opportunity robust to the measurement approach? *Review of Income and Wealth* 67(1), 18–36.
- Roemer, J. E. (1998). *Equality of Opportunity*. Harvard University Press.
- Roemer, J. E. and A. Trannoy (2015). Equality of opportunity. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Volume 2, Chapter 4, pp. 217–300. Amsterdam: North-Holland.

- Roemer, J. E. and A. Trannoy (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature* 54(4), 1288–1332.
- Rosa Dias, P. (2009). Inequality of opportunity in health: evidence from a UK cohort study. *Health Economics* 18(9), 1057–1074.
- Salas-Rajo, P. and J. G. Rodríguez (2022). Inheritances and wealth inequality: A machine learning approach. *The Journal of Economic Inequality* 20(1), 27–51.
- Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *The Journal of Economic Inequality* 11(1), 99–126.
- Sutton, J. (1991). *Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration*. Cambridge, MA: MIT Press.
- Tay, J. K., B. Narasimhan, and T. Hastie (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software* 106(1), 1–31.
- Van de gaer, D. (1993). *Equality of Opportunity and Investment in Human Capital*. Ph.D. Dissertation. Leuven: KULeuven.
- Yitzhaki, S. and E. Schechtman (2013). *The Gini Methodology*, Volume 272 of *Springer Series in Statistics*. New York, NY: Springer New York.

Online supplement: Descriptive analysis

This supplement presents descriptive statistics for all variables included in the study, see Table 1 for their detailed definitions. Because sample restrictions differ between the I-IOp and E-IOp analyses, both cases are reported separately. Tables 7 and 8 correspond to the I-IOp analysis, while Tables 9 and 10 pertain to E-IOp. In each case, the first column of the first table summarizes the outcome: median income for I-IOp and the percentage share of individuals with tertiary education for E-IOp. For I-IOp, the next two columns of Table 7 report the number of observations excluded as outliers and the number of observations with a non-positive income, see Section 3.1. All other columns describe the circumstances. Since all circumstances are categorical, we report the percentage share of each category.

Table 7: I-IOp- Descriptive statistics - First part

	Y	Outliers	$Y_I \leq 0$	S1	S2	ME1	ME2	ME3	FE1	FE2	FE3	MA1	MA2	MA3	FA1	FA2	FA3
BE11O	23,882	1	0	48.9	51.1	75.8	16.4	7.8	66.0	17.7	16.3	78.1		21.9	3.9	12.4	83.6
BE11Y	23,887	3	4	50.5	49.5	49.1	29.6	21.3	45.3	27.6	27.1	51.2		48.8	5.2	13.5	81.3
BE19O	25,544	2	4	48.0	52.0	69.5	21.4	9.1	63.4	21.1	15.6	58.9	12.1	29.0	4.1	17.1	78.9
BE19Y	25,382	0	3	49.8	50.2	43.7	30.1	26.3	42.9	29.4	27.7	34.0	9.8	56.2	4.4	16.9	78.7
BE23O	25,266	6	3	49.1	50.9	60.4	27.4	12.2	51.5	28.5	19.9	50.3	10.7	39.0	2.8	13.1	84.2
BE23Y	25,445	3	1	50.9	49.1	35.6	36.5	27.8	33.0	38.3	28.7	32.1	7.0	60.9	4.2	14.8	81.0
ES11O	16,087	18	43	50.0	50.0	95.4	2.6	2.0	89.8	3.9	6.3	81.4	7.8	10.8	3.0	26.4	70.6
ES11Y	15,403	27	53	48.7	51.3	86.2	7.6	6.2	80.2	9.1	10.7	69.5	6.5	24.0	3.3	22.0	74.7
ES19O	15,174	20	37	50.0	50.0	90.6	5.2	4.2	83.3	7.7	8.9	75.2	8.5	16.3	2.7	23.3	74.0
ES19Y	15,154	28	24	49.7	50.3	76.5	12.7	10.8	70.7	14.3	15.0	55.7	8.1	36.1	2.4	21.6	76.1
ES23O	16,017	30	36	50.9	49.1	87.3	7.5	5.2	79.1	9.5	11.5	71.8	8.6	19.6	4.4	26.5	69.1
ES23Y	16,203	22	40	49.4	50.6	65.3	17.7	17.0	62.9	17.4	19.7	46.5	8.3	45.1	4.0	23.1	72.9
NO11O	45,834	2	1	47.3	52.7	52.0	39.0	8.9	44.8	35.1	20.1	39.1	12.0	48.9	1.7	28.3	70.0
NO11Y	40,127	1	2	47.3	52.7	24.8	47.1	28.1	22.2	41.4	36.4	18.0	9.5	72.5	1.3	22.9	75.7
NO19O	40,968	4	0	45.9	54.1	31.8	46.3	21.9	27.2	42.8	30.1	20.2		79.8	2.7		97.3
NO19Y	36,511	0	2	50.3	49.7	22.7	39.2	38.0	18.0	42.0	40.1	16.3		83.7	4.7		95.3
NO23O	39,027	5	3	49.8	50.2	40.9	38.7	20.3	32.1	34.6	33.3	24.2	10.5	65.3	2.7	23.6	73.7
NO23Y	33,094	5	1	50.1	49.9	18.6	37.1	44.3	18.4	38.1	43.6	18.6	8.2	73.2	3.9	21.1	74.9

Table 8: I-IOp- Descriptive statistics - Second part

	MI1	MI2	MI3	FI1	FI2	FI3	MN1	MN2	FN1	FN2	H1	H2	O1	O2	F1	F2	F3
BE11O	9.1	3.7	9.0	22.7	33.6	39.9	15.6	84.4	14.9	85.1	95.2	4.8	22.5	77.5	10.1	44.3	45.6
BE11Y	9.8	7.8	31.2	17.5	29.4	48.0	21.9	78.1	22.9	77.1	90.3	9.7	18.3	81.7	8.4	39.6	52.0
BE19O	7.0	19.6	14.5	20.2	38.0	37.7	21.8	78.2	20.8	79.2	95.8	4.2	21.1	78.9	7.2	42.3	50.5
BE19Y	8.9	22.6	34.5	17.9	32.3	45.4	27.9	72.1	28.1	71.9	91.1	8.9	16.9	83.1	8.0	37.6	54.4
BE23O							25.8	74.2	26.1	73.9	95.5	4.5	20.0	80.0	7.4	40.0	52.6
BE23Y							27.6	72.4	29.1	70.9	92.4	7.6	16.7	83.3	8.2	35.2	56.6
ES11O	5.7	10.0	2.9	28.8	47.0	21.2	9.8	90.2	10.1	89.9	97.2	2.8	19.5	80.5	15.3	60.5	24.2
ES11Y	8.8	11.8	9.9	26.5	42.5	27.7	18.5	81.5	18.5	81.5	95.3	4.7	11.8	88.2	9.8	55.5	34.7
ES19O	7.9	12.3	4.6	28.6	49.8	18.9	9.6	90.4	9.9	90.1	98.7	1.3	11.7	88.3	9.6	62.1	28.3
ES19Y	11.8	17.3	15.2	23.0	48.4	26.2	17.6	82.4	17.5	82.5	97.1	2.9	9.4	90.6	7.3	52.2	40.5
ES23O							12.6	87.4	12.5	87.5	96.8	3.2	14.3	85.7	8.6	50.7	40.6
ES23Y							22.5	77.5	22.3	77.7	92.2	7.8	12.0	88.0	7.1	42.0	50.9
NO11O	13.2	26.9	20.8	17.7	44.9	35.7	7.6	92.4	6.1	93.9	98.9	1.1	8.3	91.7	5.8	50.3	43.8
NO11Y	10.6	27.5	43.9	12.8	35.9	49.9	13.7	86.3	12.4	87.6	98.1	1.9	4.9	95.1	3.7	41.5	54.8
NO19O	10.2	23.8	45.8	15.3	30.9	51.1	24.2	75.8	23.7	76.3	92.7	7.3	8.2	91.8	7.5	49.7	42.8
NO19Y	5.5	25.9	52.3	14.3	23.8	57.2	22.2	77.8	21.7	78.3	82.6	17.4	7.8	92.2	6.2	41.6	52.2
NO23O	10.0	28.8	37.0	13.6	38.0	45.8	18.4	81.6	17.8	82.2	90.6	9.4	5.9	94.1	6.4	46.6	47.0
NO23Y	6.6	25.1	49.7	9.5	31.3	55.3	29.0	71.0	29.0	71.0	82.7	17.3	8.7	91.3	5.8	41.3	52.9

Table 9: E-IOp- Descriptive statistics - First part

	Y	S1	S2	ME1	ME2	ME3	FE1	FE2	FE3	MA1	MA2	MA3	FA1	FA2	FA3
BE11O	36.1	48.7	51.3	76.3	16.1	7.6	66.5	17.3	16.1	78.1		21.9	4.0	12.4	83.6
BE11Y	48.7	49.7	50.3	50.1	28.9	21.0	46.1	28.0	25.9	51.2		48.8	5.1	13.2	81.7
BE19O	42.0	47.8	52.2	69.4	21.6	9.0	63.3	20.9	15.8	59.3	12.1	28.6	4.2	17.2	78.7
BE19Y	54.3	49.2	50.8	44.1	30.0	25.9	43.3	29.3	27.4	34.3	9.7	56.0	4.5	16.9	78.6
BE23O	48.6	48.9	51.1	60.4	27.4	12.2	51.6	28.3	20.0	50.5	10.6	38.9	2.7	13.0	84.3
BE23Y	60.4	50.2	49.8	35.9	36.7	27.4	33.1	38.4	28.4	32.0	6.9	61.1	4.3	14.7	81.0
ES11O	24.7	49.8	50.2	95.3	2.6	2.0	89.5	4.2	6.2	81.4	7.8	10.8	3.0	26.3	70.7
ES11Y	36.8	48.3	51.7	87.2	7.1	5.8	81.2	8.8	10.0	70.2	6.4	23.4	3.1	22.2	74.7
ES19O	33.7	49.8	50.2	89.8	5.7	4.5	82.8	8.3	8.9	74.9	8.4	16.7	2.6	23.5	73.9
ES19Y	45.1	49.2	50.8	77.0	12.6	10.4	71.1	14.0	14.9	56.3	7.9	35.8	2.3	21.4	76.3
ES23O	42.9	50.7	49.3	87.2	7.5	5.3	79.1	9.5	11.4	71.9	8.4	19.6	4.3	26.5	69.2
ES23Y	52.6	49.1	50.9	66.7	17.4	15.9	64.0	17.0	19.0	47.3	8.3	44.3	4.0	23.1	72.9
NO11O	32.6	47.3	52.7	52.2	39.0	8.7	44.9	35.3	19.9	38.7	12.1	49.1	1.7	28.6	69.7
NO11Y	43.6	45.8	54.2	24.6	48.8	26.7	22.3	42.5	35.1	18.2	9.5	72.3	1.4	22.9	75.8
NO19O	41.9	45.9	54.1	31.7	46.5	21.7	27.0	42.8	30.2	20.2		79.8	2.7		97.3
NO19Y	53.0	50.3	49.7	22.7	39.2	38.0	18.0	42.0	40.1	16.3		83.7	4.7		95.3
NO23O	42.6	48.8	51.2	41.2	38.6	20.2	32.5	34.1	33.3	24.5	10.5	65.0	2.7	23.6	73.7
NO23Y	57.1	48.4	51.6	18.6	39.1	42.4	18.5	38.8	42.8	17.8	8.9	73.3	4.0	21.9	74.1

Table 10: E-IOp- Descriptive statistics - Second part

	MI1	MI2	MI3	FI1	FI2	FI3	MN1	MN2	FN1	FN2	H1	H2	O1	O2	F1	F2	F3
BE11O	9.3	3.7	9.0	22.9	33.5	39.6	15.9	84.1	15.3	84.7	95.2	4.8	22.5	77.5	9.9	45.1	45.0
BE11Y	10.1	7.9	30.8	18.3	29.6	47.0	21.1	78.9	22.2	77.8	90.5	9.5	18.0	82.0	8.2	39.6	52.2
BE19O	6.7	19.5	14.4	20.2	38.1	37.5	21.9	78.1	20.8	79.2	95.8	4.2	21.3	78.7	7.2	42.6	50.2
BE19Y	9.0	22.4	34.3	18.1	32.5	45.0	27.6	72.4	27.9	72.1	91.1	8.9	17.3	82.7	8.3	37.7	54.1
BE23O							25.8	74.2	26.2	73.8	95.4	4.6	20.1	79.9	7.3	40.2	52.5
BE23Y							27.4	72.6	28.8	71.2	92.4	7.6	16.8	83.2	8.2	35.2	56.6
ES11O	5.9	9.9	2.8	28.3	47.2	21.4	9.3	90.7	9.7	90.3	97.0	3.0	19.5	80.5	14.5	60.8	24.7
ES11Y	8.9	11.5	9.4	26.8	43.2	26.9	18.8	81.2	18.7	81.3	95.6	4.4	11.9	88.1	9.6	55.9	34.4
ES19O	7.8	12.3	5.0	28.5	49.9	18.9	10.2	89.8	10.4	89.6	98.7	1.3	11.6	88.4	9.6	62.1	28.3
ES19Y	11.9	16.9	14.8	23.1	48.5	26.0	18.3	81.7	18.1	81.9	97.2	2.8	9.7	90.3	7.4	52.8	39.8
ES23O							12.7	87.3	12.7	87.3	96.8	3.2	14.3	85.7	8.6	50.6	40.8
ES23Y							23.0	77.0	22.8	77.2	92.5	7.5	12.1	87.9	7.0	42.4	50.7
NO11O	13.3	27.0	20.9	17.6	45.2	35.5	7.4	92.6	6.0	94.0	99.0	1.0	8.3	91.7	5.9	50.4	43.6
NO11Y	11.1	27.6	43.2	13.1	36.4	49.1	11.9	88.1	10.4	89.6	98.0	2.0	4.8	95.2	3.7	41.4	54.9
NO19O	10.3	23.9	45.6	15.4	30.9	51.1	24.0	76.0	23.5	76.5	92.8	7.2	8.2	91.8	7.4	49.5	43.1
NO19Y	5.5	25.9	52.3	14.3	23.8	57.2	22.2	77.8	21.7	78.3	82.6	17.4	7.8	92.2	6.2	41.6	52.2
NO23O	10.3	28.6	36.6	13.7	37.6	46.0	19.4	80.6	18.7	81.3	91.0	9.0	6.3	93.7	6.6	46.2	47.2
NO23Y	6.9	26.3	49.0	9.7	31.6	54.7	28.8	71.2	28.8	71.2	83.7	16.3	7.9	92.1	5.7	42.1	52.2