Lorenz regressions

A statistical contribution to the quantification of explained inequality

Alexandre Jacquemain

A thesis submitted to the
UCLouvain
in partial fulfillment of the
requirements for the degree of
Doctor of Sciences

Thesis Committee:

Prof. **Cédric Heuchenne**	*ULiège*	Advisor
Prof. **Eugen Pircalabelu**	*UCLouvain*	Co-advisor
Prof. **Philippe van Kerm**	*LISER*	
Prof. **Juan Carlos Escanciano**	*Universidad Carlos III de Madrid*	
Prof. **Christian Hafner**	*UCLouvain*	Secretary
Prof. **Rainer von Sachs**	*UCLouvain*	Chairman

June 2023

You don't start out writing good stuff. You start out writing crap and thinking it's good stuff, and then gradually you get better at it. That's why I say one of the most valuable traits is persistence.

Octavia Butler

I believe that virtually all the problems in the world come from inequality of one kind or another.

Amartya Sen

Acknowledgements

Si c'est la persistence qui permet d'aller au bout d'une thèse, cette persistence serait vaine sans l'ensemble des rencontres qui rythment la vie du doctorant. De nombreuses personnes ont ainsi contribué directement ou indirectement à ce projet, je leur en suis extrêmement reconnaissant. Tout d'abord, je souhaiterais remercier mon promoteur Cédric Heuchenne pour m'avoir guidé durant ces six dernières années. Tu m'as permis de travailler sur un sujet combinant parfaitement mes différents centres d'intérêt. Sans ta polyvalence et ton investissement énorme, je n'aurais pas pu construire cette thèse. Nous avons toujours été sur la même longueur d'onde en termes de vision de la recherche, avec la volonté de proposer une solution méthodologique innovante et pertinente, et pas celle d'un impératif de publication. Finalement, j'ai toujours pu ressentir ta confiance envers mon travail. Merci pour tout cela.

Je souhaiterais remercier les membres du jury, à commencer par Eugen Pircalabelu, mon co-promoteur. Ton apport dans ce manuscrit est indéniable. Tu m'as permis de progresser en termes de rigueur et de clarté. Merci beaucoup pour ta disponibilité et tes nombreux conseils. Je remercie le Président de jury, Rainer von Sachs pour sa bienveillance et le Secrétaire, Christian Hafner, pour sa présence dans le comité depuis le début de la thèse. Merci à Philippe van Kerm pour cette collaboration. Tu as de suite vu le potentiel de la régression Lorenz dans la mesure des inégalités d'opportunité et m'a permis d'orienter l'application dans cette direction. Merci pour les multiples discussions et pour ton aide. Finally, I would like to thank Juan Carlos Escanciano for the interesting discussion following the ISNPS2022 conference as well as the very relevant comments formulated at the occasion of the private defence. Thank you for taking the time to participate in this jury.

Après plus de six années passées à travailler au sein de l'ISBA, l'institut est devenu comme une deuxième maison pour moi. Si ça ne tient très probablement pas au bâtiment à proprement parler, c'est plutôt grâce à son personnel chaleureux. Merci à tous. Je tiens d'abord à remercier l'équipe administrative : Nadja, Nancy, Maguy, Sophie et Tatiana.

Alors que les journées moroses où la thèse n'avance pas arrivent trop souvent, une extirpation salutaire m'a parfois été permise grâce aux consultances SMCS. Grâce à elles, j'ai affiné mon sens de la pédagogie et développé une sorte de pragmatisme dans les analyses statistiques. Je souhaiterais remercier Alain, Aurélie, Benjamin C., Catherine, Céline, Nathalie Le., Lieven, Séverine et Vincent pour tout ce que vous m'avez apporté. J'en profite aussi pour remercier Nora pour cette consultance qui a débouché sur une chouette collaboration.

Je souhaiterais aussi remercier les différents professeurs de la LSBA avec lesquels j'ai pu collaborer d'une façon ou d'une autre. J'ai toujours senti que je pouvais compter sur votre soutien, merci à vous. Merci en particulier à Marie-Paule, qui m'a offert l'opportunité de donner cours magistral ce dernier quadrimestre.

Je souhaiterais ensuite remercier tous les collègues doctorants et post-doctorants qui sont passés par l'ISBA. Je ne compte plus les histoires plus farfelues les unes que les autres qui s'accumulent chaque année. Loin d'être une distraction, cette convivialité permet de combattre la solitude de la thèse et de tenir jusqu'au bout. J'espère n'oublier personne, parmi les nouveaux comme les anciens, en remerciant : Aigerim, Anas, Anna, Antoine, Aurèle, Baptiste, Benjamin D., Charles, Charlotte, Chikéola, Edouard, Ensyieh, Fanny, Florian, Gabriel, Gilles, Hortense, Hugues, Jean-Loup, John-John, Keivan, Lise, Mengxue, Mickaël, Morine, Nathalie Lu., Oswaldo, Oussama, Patricia, Quentin, Rebecca, Sophie, Stefka, Stéphane, Vanessa. Merci pour tous ces moments formels et informels (surtout les informels), ces discussions et anecdotes (surtout les anecdotes), ces événements scientifiques et sociaux (surtout les sociaux).

Je n'aurais probablement jamais pu commencer cette thèse sans le soutien de ma famille, et en particulier de mes parents. Avec le recul, je me rends compte de l'importance de tous les sacrifices que vous avez faits et du temps que vous avez pris pour me soutenir et me permettre de m'épanouir. Je me souviens encore des corrections liées à mon travail de fin d'études secondaires. Je n'ai jamais eu de "review" avec plus d'éléments à changer. C'est là que j'ai appris que la rédaction n'était pas une question d'inspiration mais d'habitude et de persistence. Le mot de la fin est évidemment pour Manu, avec qui je partage ma vie depuis déjà 8 ans. Merci de m'avoir aidé à gérer émotionnellement cette thèse, avec ses aspects cycliques, ses périodes de travail intensif suivies par des moments de bloquage complet. Un des risques de la thèse est de s'y enfermer. Sans relâche, tu m'as aidé à ce que ça n'arrive pas. Je te remercie énormément pour tout ce que tu m'apportes au quotidien.

Alexandre Jacquemain, Juin 2023

Preface

In the eighth of his *Ten Short Reflections on the Future of Income Inequality*, Branko Milanovic (2016) expresses the idea that economics has methodologically shifted from a concern with representative agents and averages to a concern with heterogeneity and inequality. For an economic outcome, it means that the whole distribution is of interest, and not only the average result. To perform proper analyses requires adapting the statistical procedures to these specificities. One example is the quantile regression of Koenker and Bassett (1978), which determines explanatory factors for each quantile of the distribution, in contrast with a classical linear regression which focuses on the conditional average. This thesis proposes a new methodology in the emergent toolbox: the Lorenz regression. It is a flexible and theoretically grounded procedure used to measure the extent of inequality in a variable that can be attributed to a set of explanatory factors.

A leading application of this thesis is to be found in the context of equality of opportunity. Loosely speaking, this criterion requires that individuals' chances of economic success do not depend on circumstances beyond their control. It is an essential feature of modern discourse. Proclaimed achieved by some, it is used to justify the unequal rewards flowing from the market economy. Others consider it imperfect and use it as a justification for more reforms. Determining the extent to which opportunities are equalized is therefore crucial. Disputes concerning its definition and place are mainly the realm of philosophers and economists. How should equality of opportunity be defined? More specifically, how can one transform the notions of *chance of economic success* or *circumstances* into observable quantities? What does is it mean that chances of economic success *do not depend* on circumstances?

Once these questions are settled, inequality of opportunity needs to be measured. After all, one wants to be able to determine whether opportunities are more equalized now than before, whether it is better achieved in a country or another, what circumstances bear the most impact in the individual's fate, etc. In this respect, the perspective offered by statistics proves useful because it helps formalize and carry out the following considerations.

(i) While one is interested in a given population, say the Belgian population

over the age of eighteen, surveys are conducted on a subset of it, a representative sample. The passage from a fixed but unobserved population to an observed but random sample yields uncertainty. This uncertainty needs to be internalized and controlled in a form or another. For a given measure of inequality of opportunity, statistics determines an inferential procedure, able to characterize the value of the chosen metric in the population using the information obtained from the sample.

(ii) Using theoretically sound techniques, statistics may help select the relevant circumstances, i.e. those that most determine the chance of economic success, and in determining their respective contributions.

(iii) Finally, these statistical methods can be made readily accessible through open-source software implementations. This allows the practitioner to perform a statistical analysis of inequality of opportunity on his or her preferred data.

The characterization of these issues entails a series of choices. In this thesis, inequality is measured with the Gini coefficient, a decision justified by its interesting properties as well as its widespread use. The information contained in the explanatory factors is exploited by turning the attention to the explained Gini coefficient. This quantity is the Gini coefficient of a conditional expectation and has a natural resonance in the context of equality of opportunity. Finally, an assumption on the precise form of this conditional expectation is made with the single-index model. This choice is relevant as it brings flexibility and simplifies greatly the estimation process. In this work, we build the *Lorenz regression*, an estimation and inferential procedure for the explained Gini coefficient. The manuscript is divided into several chapters.

Chapter 1. Preliminaries. This chapter presents a selected survey of the literature relevant to this thesis. It opens with issues raised by economics, i.e. the formalization of equality of opportunity and the measurement of inequality. Then, the statistical techniques used throughout this work are presented at a general level. More specifically, the idea of semiparametric regression, with the special case of the single-index model, as well as the concept of penalized regression are introduced.

Chapter 2. The Lorenz regression. This chapter is devoted to the unpenalized Lorenz regression. In the context of the single-index model, we show that the explained Gini coefficient is the outcome of a maximization programme, where the objective function is the covariance between the response and the cumulative distribution function of a linear combination of the explanatory factors. The estimation procedure based on this result is a particular case of the monotone rank estimator of Cavanagh and Sherman (1998). We make use of this connection to provide inference for the explained Gini coefficient. A continuity correction is proposed when all covariates are discrete. The content of this chapter is joint work

with C. Heuchenne and is published in *Computational Statistics and Data Analysis*.

Chapter 3. The penalized Lorenz regression. The unpenalized Lorenz regression is prone to overestimate the explained Gini coefficient when many explanatory factors are included. This chapter addresses this issue by proposing a penalized bootstrap procedure which selects the relevant explanatory factors and produces valid inference for the explained Gini coefficient. We show the so-called oracle property of the estimator and develop a coordinate-descent algorithm to obtain it numerically. The content of this chapter is joint work with C. Heuchenne and E. Pircalabelu. A first procedure based on the LASSO penalty is published in the *Proceedings of the 22th European Young Statisticians Meeting*. A more general procedure was submitted for publication.

Chapter 4. An application to equality of opportunity. In this chapter, we use the explained Gini coefficient as a measure of inequality of opportunity, with the penalized Lorenz regression as estimation procedure. The metric thus obtained is more stable than a classical measure of inequality, which is routinely based on the Gini coefficient of fitted values computed from a log-linear regression. We also build a procedure based on the Shapley Value decomposition to break down the explained Gini coefficient into contributions of each selected circumstance. The procedure is applied on data obtained from the European Union Statistics on Income and Living Conditions. The content of this chapter is joint work with C. Heuchenne and P. Van Kerm.

Chapter 5. The LorenzRegression R package. This chapter introduces the **LorenzRegression** package, an implementation in the R language of the Lorenz and penalized Lorenz regressions. Its use is similar to customary packages performing regression analysis, making it as user-friendly as possible. It also contains methods to display and visualize the results of the analysis. Computationally intensive parts of the code are written in C++. The main functions are carefully described and illustrated on income data. The content of this chapter is joint work with C. Heuchenne and was submitted for publication.

Chapter 6. Discussion and extensions. The Lorenz regression is characterized by several constitutive features. It produces inference for a measure of explained inequality based on the Gini coefficient and assuming a statistical single-index model. Informed either by statistical or economical concerns, one may wish to challenge these bedrock elements in order to extend the reach of the procedure. This chapter proposes concrete answers to the following questions: may the Lorenz regression be applied outside the context of inequality measurement? How may one generalize the procedure to other inequality measures? Can one do without the assumption of the single-index model?

Contents

CHAPTER 1

Preliminaries

The methodology developed in this thesis, summarized under the term *Lorenz regression*, proposes an inferential procedure for the *explained Gini coefficient*. This object is central to this work. It connects semi-parametric regression techniques with inequality measurement to produce a measure of explained inequality with natural applications, e.g. in the context of equality of opportunity. Formally, it is the Gini coefficient of the conditional expectation of a response given certain covariates, assuming a single-index model.

This work is therefore indebted to several methodological directions. In this chapter, we review them and set out the necessary vocabulary and notations. Before turning to this task, we introduce the `Ilocos` dataset, provided in the R package **ineq** (Zeileis, 2014), as it will be used to provide applied examples. The literature review opens in Section 1.2 with the notion of inequality of opportunity, which gives a motivation for the use of the Lorenz and penalized Lorenz regression. More specifically, this introduction provides background material for Chapter 4. The literature on inequality measurement is discussed in Section 1.3. The main notions introduced are the Lorenz curve, Gini coefficient and concentration index. Section 1.4 presents the broad idea of the semi-parametric regression, an interesting compromise between the efficiency of parametric techniques and the flexibility of non-parametric ones. The particular case of the single-index model is tackled in Section 1.5. These sections provide sufficient material to understand the Lorenz regression procedure developed in Chapter 2. Section 1.6 introduces the concept of penalized regression. It provides supportive material to the penalized Lorenz regression set out in Chapter 3.

Throughout this thesis, we aim to explain the inequality of a random variable $Y \in \mathbb{R}$, using the information contained in a random vector $X = (X^1, \ldots, X^p)^\intercal \in \mathbb{R}^p$. The variable Y is therefore called the *response*. In the applications it is an

economic outcome, e.g. a measure of income. In the particular case of equality of opportunity, we call it the *advantage variable*. The vector X gathers the explanatory variables, called *covariates*. In the case of equality of opportunity these correspond to *circumstances*, i.e. variables over which the individual has no control. The estimation is based on a sample $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$ of n independent and identically distributed copies of $(X^\intercal, Y)^\intercal$.

We conclude this section with some notations and abbreviations. The notations $\mathbb{E}[\cdot]$, $\mathbb{V}[\cdot]$ and $\mathbb{C}[\cdot,\cdot]$ correspond to the expected value, variance and covariance operators. For an event A, the indicator function $\mathbb{1}\{A\} \in \{0,1\}$ takes value 1 if A occurs and 0 otherwise. For a generic function $f(\cdot)$, we denote the first and second derivatives by $f'(\cdot)$ and $f''(\cdot)$. For a vector b, we denote its k-th element by b_k. The notation $||.||$ refers to the L2-norm of a vector. For a matrix B, we denote by B_{kl} the element at row k and column l. We abbreviate the probability density function by PDF, and the cumulative distribution function by CDF. Also, the abbreviation DGP is used for the data generating process. Finally, the notation i.i.d stands for independent and identically distributed.

1.1 The Ilocos dataset

The `Ilocos` dataset contains economic data on 632 households in the Philippines. We use total household income (`income`) as response variable and the following covariates :

- `sex`: a factor with two levels, `"male"` (518 observations) and `"female"` (114 observations), indicating the biological sex of the household head.
- `family.size`: a numerical variable indicating the family size. Results range from 1 to 13, with a mean of 5.19 and a median of 5.
- `urbanity`: a factor with two levels, `"rural"` (301 observations) and `"urban"` (331 observations).
- `province`: a factor with four levels: `"Ilocos Norte"` (65 observations), `"Ilocos Sur"` (68 observations), `"La Union"` (116 observations) and `"Pangasinan"` (383 observations).

1.2 Inequality of opportunity

In economics and political philosophy, there exist many ways to materialize the notion of equality of opportunity. This thesis follows the route traced by John Roemer in his classical book *Equality of Opportunity* (1998). Roemer starts with a distinction between two types of conditions of success. *Circumstances* correspond to variables over which individuals have no control, such as the education of

their parents, or their country of birth. In contrast, conditions of success relating to the individuals' own choices are labelled as *effort*. In essence, equality of opportunity (EOP) requires that relevant economic outcomes, called *advantages*, do not depend on circumstances. Similarly, inequality of opportunity (IOP) emerges whenever differentials in economic outcomes are explained by individuals having different circumstances.

In what follows, we formalize the notion of EOP. We call *opportunity set* one possible value of the vector of circumstances. Then, we partition the population into *types* such that two individuals are of the same type if and only if they have the same opportunity set. On that basis, there exist two approaches to characterize EOP: the ex-ante and the ex-post perspectives.

The *ex-post* perspective is a direct translation of the idea that EOP occurs when individuals exerting the same *degree* of effort are rewarded equivalently. Formally, two individuals in different types are in the same *tranche* if they correspond to the same quantile in their respective distributions of effort. In other words, a tranche gathers individuals exerting the same degree of effort. EOP then requires that, for each tranche, the advantage is the same across the types. A difficulty in this procedure is that one needs to be able to rank observations based on effort.

In the *ex-ante* approach, one first computes a value for the opportunity set in each type. Then, EOP requires that this value is the same across the types. Here, the challenge consists in finding a relevant valuation procedure for the opportunity set. One solution is proposed by the *ex-ante utilitarian* approach, which values the opportunity set by the conditional expectation of the advantage given the opportunity set.

A new distinction arises as we move to measuring IOP. A first approach entails to compare the observed distribution of the advantage with a counterfactual distribution that assumes EOP. For a given inequality index, an *indirect measure* is constructed as the difference between the value of the index obtained on the observed distribution and the one computed on the counterfactual. Alternatively, *direct measures* of IOP examine the inequality that is left when, for each individual, we only take into account the information provided by his or her circumstances. In this work, we provide an ex-ante utilitarian and direct measure of IOP. Formally, this implies to find a relevant measure of the inequality of the conditional expectation of the advantage given the circumstances.

The next section introduces the measure of inequality that we use throughout this thesis, i.e. the Gini coefficient. The sections that follow address the question of how to model the conditional expectation.

1.3 Inequality measurement

We view *inequality* as a relative measure of dispersion of a random variable, relative in the sense that it is independent of the scale of the variable. Throughout this thesis, we are interested in analyzing the inequality of the response variable Y. We divide this section in two parts. The first introduces tools available to describe the inequality of Y. The second addresses the question of how to include information provided by the covariates X.

Let $Y \in \mathbb{R}$ with $\mathbb{E}[Y] > 0$ be a continuous random variable with CDF $F_Y(\cdot)$. We therefore do not require that all incomes be strictly positive.[1] We will make use of two main tools to describe the inequality of Y. The first is the Lorenz curve, a graphical representation of inequality where the cumulative share of the response is displayed for each cumulative share of the population. The second is the Gini coefficient. It summarizes inequality in an index ranging from 0 (perfect equality) to 1 (perfect inequality). For a thorough exposition of the Lorenz curve and Gini coefficient, we refer the interested reader to Chapter 2 of Yitzhaki and Schechtman (2013).

The *Lorenz curve* (LC) of Y at p, where p is a fixed probability, is defined as

$$\mathrm{LC}_Y(p) \coloneqq \frac{\mathbb{E}[Y\mathbb{1}\{F_Y(Y) \leq p\}]}{\mathbb{E}[Y]}.$$

Notice that $\mathrm{LC}_Y(0) = 0$ and $\mathrm{LC}_Y(1) = 1$. A graph of the curve obtained on the `Ilocos` data is displayed in red and dotted in Figure 1.1. The Lorenz curve is always convex and, when $Y \in \mathbb{R}^+$, strictly increasing. Two extreme scenarios are displayed in black. In the case of perfect equality, the curve coincides with the 45° line and all individuals receive the same share of income. If $Y \in \mathbb{R}^+$, perfect inequality traces a curve that is equal to 0 for $p < 1$ and jumps to 1 for $p = 1$. This corresponds to the situation where one individual gathers the whole mass of income. The *Gini coefficient* is defined as twice the area between the 45° line and the Lorenz curve. It is also equivalent to the following definition

$$\mathrm{Gi}_Y \coloneqq \frac{2\mathbb{C}[Y, F_Y(Y)]}{\mathbb{E}[Y]}.$$

In case of perfect equality, the coefficient reaches its lower bound, i.e. $\mathrm{Gi}_Y = 0$. In case of perfect inequality and if $Y \in \mathbb{R}^+$, it reaches the upper bound $\mathrm{Gi}_Y = 1$. In the `Ilocos` dataset, the Gini coefficient of income is of 42%.

In a Lorenz curve, the vertical and horizontal axes are constructed using the same

[1] In practice, the only restriction is $\mathbb{E}[Y] \neq 0$. If $E[Y] < 0$, the analysis would be performed on $-Y$.

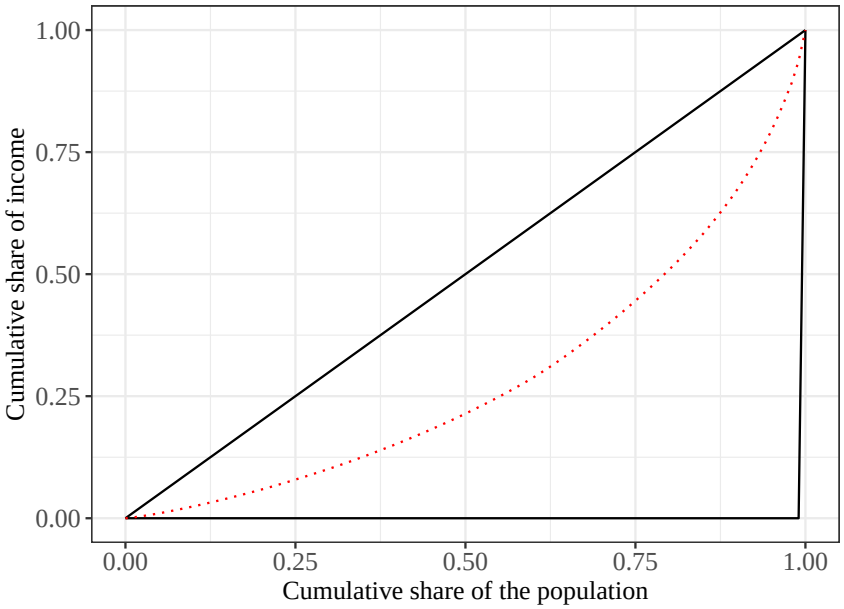

Figure 1.1: Lorenz curve of `income` on the `Ilocos` data

variable. Indeed, the curve traces the cumulative share of Y owned by each cumulative share of the population, ordered in terms of variable Y again. In a concentration curve, the two axes may be based on different variables. More precisely, the *concentration curve* (CC) of Y with respect to a new variable W at p is defined as

$$\mathrm{CC}_{Y,W}(p) \coloneqq \frac{\mathbb{E}[Y\mathbb{1}\{F_W(W) \leq p\}]}{\mathbb{E}[Y]}.$$

where $F_W(\cdot)$ is the CDF of W. Similarly to the Lorenz curve, $\mathrm{CC}_{Y,W}(0) = 0$ and $\mathrm{CC}_{Y,W}(1) = 1$. It coincides with the 45° line in two instances, either in the case of perfect equality, when all individuals receive the same value of Y, or in the case of independence between Y and W. If the cumulative ranks of W correlate positively (negatively) with the response, the concentration curve lies below (above) the 45° line. Two extreme scenarios of perfect inequality arise. The first is similar to the Lorenz curve: the concentration curve is 0 for $p < 1$ and jumps to 1 for $p = 1$. This happens when the individual with the highest value of W holds all the income in the population. In the second, the concentration curve is 0 for $p = 0$ but attains 1 as soon as $p > 0$. This occurs when the individual with the lowest value of W holds all the income. Figure 1.2 displays the Lorenz curve of `income` (red and dotted) and the concentration curve of `income` with respect to `family.size` (green and dashed) in the `Ilocos` dataset. We observe that the curve is below the 45° line.

This means that bigger families tend to gather more income. The *concentration*

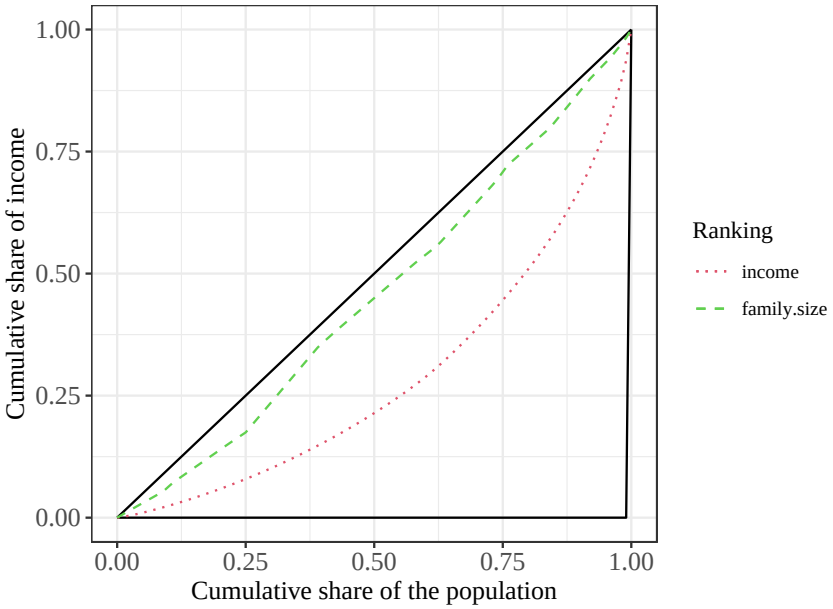

Figure 1.2: Concentration curve of `income` with respect to `family.size` on the `Ilocos` data

index is the equivalent of the Gini coefficient when the ordering is based on W. Formally, it is defined as

$$\text{CI}_{Y,W} = \frac{2\mathbb{C}[Y, F_W(W)]}{\mathbb{E}[Y]}.$$

In case of independence or perfect equality, one obtains $\text{CI}_{Y,W} = 0$. In case of perfect inequality, one has $\text{CI}_{Y,W} = 1$ if the association between Y and W is positive and $\text{CI}_{Y,W} = -1$ if the association is negative. In the Ilocos data, the concentration index of `income` with respect to `family.size` is of 9%. The concentration curve and concentration index are heavily used in health economics (Wagstaff et al., 1991; Kakwani et al., 1997). In this context, Y is a health outcome and W is a measure of socioeconomic status. The concentration index therefore measures health inequality when individuals are ordered based on their socioeconomic status.

Several methods have been proposed to make use of the information provided by covariates X. Mostly, the literature has focused on decomposing the Gini coefficient of Y or the concentration index of Y with respect to W into contributions of the covariates. The simplest case arises when the response Y neatly decom-

poses into a linear combination of X. For example, Lerman and Yitzhaki (1985) introduced a decomposition of the Gini coefficient of income in the contributions of various income sources. Using this methodology, Garner (1993) assessed the role of different household budget components in total expenditure inequality. Yitzhaki (1994b) analyzed the distributional effects of commodity taxation. Assuming a linear regression model between Y and X, Wagstaff et al. (2003) introduced a decomposition of the concentration index in the context of health economics, see also Jones and Nicolás (2004) and van Doorslaer and Koolman (2004). Interestingly, these tools are also used outside the realm of inequality measurement. Shalit and Yitzhaki (2003) and Denuit et al. (2014) used a decomposition of the concentration curve to check the efficiency of a financial portfolio.

When the covariates X are discrete, they induce a partition of the population into a number of groups. Yitzhaki (1994a) proposed a decomposition of the Gini coefficient into three components: within-group inequality, between-groups inequality and an overlapping term. Therefore, contrary to the decomposition of the variance, the Gini coefficient does not decompose neatly into within-group and between-groups terms. Within-group inequality boils down to computing the Gini coefficient inside each group. Between-groups inequality is obtained by the Gini coefficient applied on the vector of group means. The final term measures the extent to which the distributions of the different groups overlap.

In the context of IOP measurement, a different route is generally taken. First, *fitted values* are obtained by estimating the conditional expectation of the advantage given the circumstances. Then, IOP is measured with the estimated Gini coefficient of the fitted values. The information contained in the covariates is therefore exploited at the level of the conditional expectation and not to perform a decomposition of the inequality measure. The conditional expectation can be estimated with the nonparametric estimator proposed by Van de Gaer (1993). In this case, the fitted value of an individual is the empirical mean of the advantage of the individuals with the same opportunity set. However, the number of types grows exponentially with the number of circumstances. In order to avoid unreliable estimation, the researcher should restrict the number of circumstances considered. At the other extreme lies the parametric approach of Bourguignon et al. (2007), where fitted values are obtained from a linear model on log-advantages. The main drawback with this method lies in its lack of flexibility. For a more thorough review of the literature on IOP measurement, we refer the reader to Ramos and Van de gaer (2016) and Brunori et al. (2018).

1.4 Semi-parametric regression

Regression analysis aims to estimate the relationship between a response variable and a vector of covariates. It is made of three steps. First, the researcher specifies what the relationship measures. Here, we consider that the quantity of interest is the conditional expectation of Y given X. Other quantities could be used, for

example the conditional quantile function. Second, a statistical model is proposed. Third, the model is estimated with a suitable procedure. The first two steps are illustrated with Equation (1.4.1).

$$\mathbb{E}[Y|X = x] = m(x). \tag{1.4.1}$$

The flexibility of the model is determined by the amount of information injected in the function $m(\cdot)$. At one end of the spectrum lies the fully *nonparametric* approach, where only mild regularity conditions are imposed, e.g. differentiability. At the other extreme lies the *parametric* approach, in which case $m(\cdot)$ is fully specified using a finite number of parameters. The simplest example of the latter is the linear regression model, where

$$\mathbb{E}[Y|X = x] = x^\intercal \beta_0,$$

and $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^\intercal$ is a vector of regression coefficients. Parametric models benefit from two main advantages. They allow for precise estimation and are easy to interpret. In economics, the coefficients of a linear regression are interpreted as marginal effects since

$$\frac{\partial \mathbb{E}[Y|X = x]}{\partial x^k} = \beta_{0,k}.$$

Hence, $\beta_{0,k}$ is the answer to the following question: with how many units does Y increase or decrease when x^k is increased of one unit ? More generally, imposing a parametric form to $m(\cdot)$ is convenient for estimation purposes, as well as to materialize restrictions imposed by economic theories. However, it also results in a lack of flexibility. The proposed model may not account for the phenomenon that generated the observed data. In that case, the model is misspecified and the conclusions drawn from the analysis are flawed.

In nonparametric regression, no structural form is imposed on $m(\cdot)$, which makes it less subject to misspecification. In practice, the resulting fit adapts more flexibly to the data. This is illustrated in Figure 1.3, which displays parametric and nonparametric regression fits for the `Ilocos` data, where the response is log-income and the covariate is family size. The parametric fit is obtained by a simple linear regression, while the nonparametric fit is obtained with a Nadaraya-Watson estimator (Nadaraya, 1965; Watson, 1964), implemented in the R package **np** (Hayfield and Racine, 2008).

However, because the model structure needs to be estimated from the data, nonparametric regression suffers from a lack of precision. Since the domain of $m(\cdot)$ is in $\mathbb{R}^p$, the amount of data required to estimate the conditional expectation at a satisfactory rate increases dramatically as p becomes large. This issue is known as the *curse of dimensionality*, see Stone (1980).

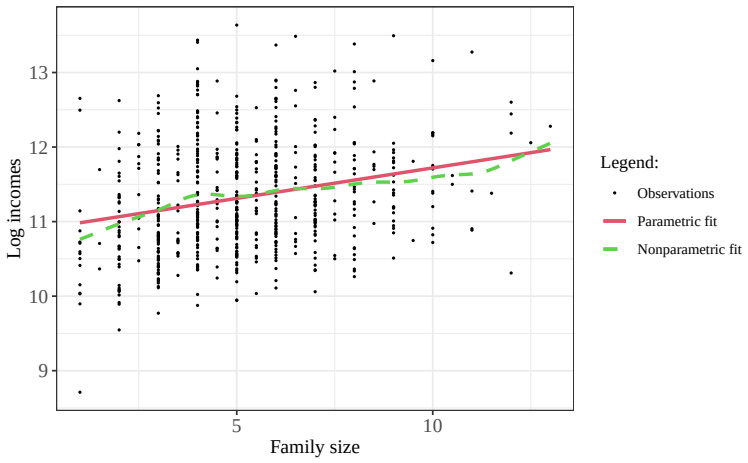

Figure 1.3: Regression of log-income on family size on the `Ilocos` data

Semiparametric regression strives to find a middle-ground between parametric and nonparametric techniques. Its model structure involves both an unknown function and a finite number of parameters. For example, the single-index model assumes that

$$\mathbb{E}[Y|X = x] = H(x^\intercal \theta_0),$$

where $H : \mathbb{R} \to \mathbb{R}$ is an unknown function and $\theta_0 \in \mathbb{R}^p$ is a vector of parameters. The random variable $X^\intercal \theta_0$ is called the *index* and the function $H(\cdot)$ is called the *link function*. This model structure does not suffer from the curse of dimensionality since $H(\cdot)$ is unidimensional. One contribution of the parametric part is therefore to achieve *dimension reduction*. Another is to retain interpretability. Consider two covariates X^k and X^l observed on the same scale. Let the marginal rate of substitution of X^k for X^l be defined as

$$\text{MRS}_{k,l} := \frac{\partial \mathbb{E}[Y|X = x]/\partial x^k}{\partial \mathbb{E}[Y|X = x]/\partial x^l}.$$

It represents the rate at which X^l must increase if X^k is reduced of one unit and one wishes to keep the expected response at the same level. In a single-index model, this marginal rate of substitution is obtained as

$$\text{MRS}_{k,l} = \frac{\theta_{0,k}}{\theta_{0,l}}.$$

The nonparametric part brings more flexibility since $H(\cdot)$ is left unspecified. We refer the interested reader to Horowitz (2009) for a more complete exposition of nonparametric and semiparametric techniques with applications to economics. The next section focuses on the characterization and estimation of a single-index model.

1.5 The single-index model

We recall that the single-index model assumes

$$\mathbb{E}[Y|X = x] = H(x^\intercal \theta_0). \tag{1.5.1}$$

Throughout this thesis, we also assume that $H(\cdot)$ is strictly increasing. This allows us to frame the division of labour between the parametric and nonparametric parts in a particular way. The index identifies the ranks of the conditional expectation, while the link function transforms them into actual predictions. Contrary to the linear regression case, we only assume that the ordering structure of the conditional expectation is obtained as a linear combination of the covariates.

Without further restrictions, the model is not identifiable. This issue is well explained in Chapter 2 of Horowitz (2009). We illustrate it here in the case of a strictly increasing link function. First, the parameter vector θ_0 needs to be normalized. Let $k > 0$ and $\theta_0^* = k\theta_0$, then $X^\intercal\theta_0$ and $X^\intercal\theta_0^*$ produce the same ranks. Hence, the model is not identifiable. In the literature, the constraint $\theta_{0,1} = 1$ is often used. In our case, this choice is not reasonable since we want to produce feature selection in Chapter 3. Instead, we require $||\theta_0||= 1$, where we recall that $||\cdot||$ denotes the L2-norm.

A second problem emerges with the presence of discrete covariates. For illustration, consider that X is made of two binary variables coded as $\{0, 1\}$. In this case, $X^\intercal\theta_0 \in \{z_1 \coloneqq 0, z_2 \coloneqq \theta_{0,1}, z_3 \coloneqq \theta_{0,2}, z_4 \coloneqq \theta_{0,1} + \theta_{0,2}\}$. Remark that the vectors $\theta_0 = (1/\sqrt{3}, \sqrt{2/3})^\intercal$ and $\theta_0^* = (1/2, \sqrt{3}/2)^\intercal$ produce the same ranking, i.e. $\{z_1 < z_2 < z_3 < z_4\}$ and are therefore indistinguishable. Even though this model is not identifiable, it is still informative. For example, $\theta_0^{**} = (-1/2, \sqrt{3}/2)^\intercal$ is distinguishable from the other two vectors since it produces the ranking $\{z_2 < z_1 < z_4 < z_3\}$. As we will argue and formalize in Chapter 2, θ_0 is identifiable up to a region. Identifiability is restored by the following conditions, based on Theorem 2.1 in Horowitz (2009).

(SI1) $H(\cdot)$ is a strictly increasing and differentiable function.

(SI2) X contains at least one continuous variable. If it contains discrete variables, they must not divide the support of $X^\intercal\theta_0$ into disjoint subsets.

(SI3) The support of X is not contained in any proper linear subspace of $\mathbb{R}^p$.

(SI4) The norm $||\theta_0||= 1$.

In the remaining of this section, we review two estimation procedures of a single-index model: the weighted nonlinear least squares (WNLS) estimator of Ichimura (1993), and the monotone rank (MR) estimator of Cavanagh and Sherman (1998).

For a more complete exposition, we refer again the interested reader to Chapter 2 of Horowitz (2009).

At first glance, the estimation of the link function and parameter vector are intertwined. If the link function were known, the model would be fully parametric and an estimator for θ_0 could be obtained as the solution of the weighted least-squares problem

$$\min_{\theta} \sum_{i=1}^{n} \omega(X_i)[Y_i - H(X_i^\intercal \theta)]^2, \tag{1.5.2}$$

where $\omega(\cdot)$ is a weight function. For a fixed θ, the link function can be estimated with the kernel regression estimator

$$\hat{H}(z, \theta) \coloneqq \frac{1}{nh\hat{g}(z, \theta)} \sum_{i=1}^{n} Y_i \kappa\left(\frac{z - X_i^\intercal \theta}{h}\right),$$

where $\hat{g}(z, \theta)$ is the kernel density estimator

$$\hat{g}(z, \theta) \coloneqq \frac{1}{nh} \sum_{i=1}^{n} \kappa\left(\frac{z - X_i^\intercal \theta}{h}\right),$$

with bandwidth h and kernel function $\kappa(\cdot)$. For an unknown link function, an estimation procedure for θ_0 is obtained by plugging $\hat{H}(X_i^\intercal \theta, \theta)$ into Equation (1.5.2). Ichimura (1993) adapts this estimation procedure in three ways. First, it avoids that $\hat{g}(X_i^\intercal \theta, \theta)$ gets too close to 0. Second, observation i is excluded from the computation of $\hat{H}(X_i^\intercal \theta, \theta)$. Third, it includes the weight function also in the estimation of the link function. More precisely, let $g(\cdot, \theta)$ be the PDF of $X^\intercal \theta$. Also, define $A_x \coloneqq \{x : g(x^\intercal \theta, \theta) \geq \eta, \forall \theta \in \Theta\}$ and $A_{nx} \coloneqq \{x : ||x - x^*|| \leq 2h, \text{for some } x^* \in A_x\}$, where η is a small positive constant and Θ is the parameter space. The WNLS estimator solves

$$\min_{\theta} \sum_{i=1}^{n} \mathbb{1}\{X_i \in A_x\} \omega(X_i)[Y_i - \hat{H}_i(X_i^\intercal \theta, \theta)]^2, \tag{1.5.3}$$

where

$$\hat{H}_i(z, \theta) \coloneqq \frac{1}{nh\hat{g}_i(z, \theta)} \sum_{j \neq i} \mathbb{1}\{X_j \in A_{nx}\} \omega(X_j) Y_j \kappa\left(\frac{z - X_j^\intercal \theta}{h}\right)$$

$$\hat{g}_i(z, \theta) \coloneqq \frac{1}{nh} \sum_{j \neq i} \mathbb{1}\{X_j \in A_{nx}\} \omega(X_j) \kappa\left(\frac{z - X_j^\intercal \theta}{h}\right).$$

Call $\hat{\theta}_{\text{WNLS}}$ the solution to (1.5.3). The consistency and asymptotic normality of $\hat{\theta}_{\text{WNLS}}$ are established in Theorems 5.1 and 5.2 in Ichimura (1993).

Since we assume that the link function is strictly increasing, it is possible to derive an estimator of θ_0 irrespective of $H(\cdot)$ as $X^\intercal\theta_0$ only affects the ordering structure of the conditional expectation. A first estimator in this category is the maximum rank correlation (MRC) estimator of Han (1987), obtained as

$$\hat{\theta}_{\text{MRC}} \coloneqq \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{1}\{Y_i > Y_j\} \mathbb{1}\{X_i^\intercal \theta > X_j^\intercal \theta\}.$$

Another candidate is the monotone rank (MR) estimator, introduced by Cavanagh and Sherman (1998) and defined as

$$\hat{\theta}_{\text{MR}} \coloneqq \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j \neq i} M(Y_i) \mathbb{1}\{X_i^\intercal \theta > X_j^\intercal \theta\}, \tag{1.5.4}$$

where $M(\cdot)$ is an increasing function, determined by the user. Theorems 1 and 2 in Cavanagh and Sherman (1998) determine the consistency and asymptotic normality of $\hat{\theta}_{\text{MR}}$. In this thesis, the MR estimator is of special interest since the Lorenz regression developed in Chapter 2 involves a special case of this estimator.

1.6 Penalized regression

Regression analysis often involves variable selection. The purpose of this step consists in finding the subset of the variables which accounts best for the phenomenon under study. To achieve this, the chosen model should strike a balance between goodness of fit and parsimony. The most straightforward way to proceed, called *subset selection*, consists in comparing models using a criterion which encompasses these two aspects such as, for example the AIC criterion, see Akaike (1998). In most cases, it is not feasible to perform this comparison across all possible submodels, as it would involve fitting 2^p models, and stepwise searches are implemented. This approach has the disadvantage of being unstable. Indeed, small changes in the data may result in very different models being selected, see Breiman (1996).

An alternative consists in the use of penalized regression. This procedure adds to the objective function underlying the estimation process a penalty on the amplitude of each coefficient. If the penalty is sparsity-inducing, it will lead to a selection of the covariates. We divide the remaining part of this section in three parts. We start by focusing on linear regression. The first part introduces the least absolute shrinkage and selection operator (LASSO) developed by Tibshirani (1996). The second part presents the smoothly clipped absolute deviation (SCAD) proposed by Fan and Li (2001). Then, we move to the single-index model and introduce the penalized maximum smoothing rank correlation (PMSRC) estimator of Lin and Peng (2013).

In the context of linear regression, a penalized estimator of β_0 is found by solving

$$\min_{\beta} \left(\sum_{i=1}^{n} [Y_i - X_i^\intercal \beta]^2 + \sum_{k=1}^{p} p_\lambda(|\beta_k|) \right), \tag{1.6.1}$$

where $p_\lambda(\cdot)$ is the *penalty function* and $\lambda > 0$ is the *penalty parameter*, sometimes also called *regularization parameter*. The LASSO estimator $\hat{\beta}_{\text{LASSO}}(\lambda)$ of β_0 solves (1.6.1) with $p_\lambda(x) = \lambda x$. The LASSO is therefore a constrained optimization problem, where the constraint is imposed on the L1-norm of the coefficient vector. It is important to notice that, in general, no analytic solution is available. Several algorithms have been proposed to solve the LASSO efficiently, for example least-angle regression (Efron et al., 2004) or the coordinate-descent algorithm of Friedman et al. (2007). We focus our explanation on the latter because it bears some similarity with the SCAD-FABS algorithm introduced in Chapter 3. Assume that each covariate is standardized such that it has a mean of 0 and a variance of 1. In the very specific case of a single covariate, the LASSO has a simple analytic expression, given by

$$\hat{\beta}_{\text{LASSO}}(\lambda) = S(\hat{\beta}, \lambda) \coloneqq \begin{cases} \hat{\beta} - \lambda & \text{if } \hat{\beta} > 0 \text{ and } |\hat{\beta}| > \lambda \\ \hat{\beta} + \lambda & \text{if } \hat{\beta} < 0 \text{ and } |\hat{\beta}| > \lambda \\ 0 & \text{if } |\hat{\beta}| \leq \lambda \end{cases} \tag{1.6.2}$$

where $S(\cdot, \lambda)$ is called the *soft-thresholding* function and $\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i$ is the simple least-squares estimator. The coordinate-descent algorithm exploits this idea in the case of multiple covariates. At each step, it updates only one coefficient by solving a univariate problem, where the response variable is given as the partial residuals obtained with all the other covariates. For coefficient k, the update rule is the following. Consider an estimator $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^\intercal$ and define the partial residual for observation i as $Y_i^{(k)} = \sum_{l \neq k} \tilde{\beta}_l X_i^l$, where X_i^l is the i-th observation of variable X^l. The update rule is

$$\tilde{\beta}_k(\lambda) = S\left(\frac{1}{n} \sum_{i=1}^{n} X_i^k (Y_i - \tilde{Y}_i^{(k)}), \lambda \right).$$

This step is repeated for $k = 1, \ldots, p, 1, \ldots, p, \ldots$ until convergence is met. In practice, this procedure is applied to a decreasing grid of values for λ and each solution is used as starting value for the next problem. The coordinate-descent is the method of choice in the widely used R package **glmnet** (Friedman et al., 2010). We apply it on the `Ilocos` data, with model `log(income) ~ sex*family.size + sex*urbanity + province`, i.e. we regress log-incomes on family size, urbanity, province and sex, and allow for an interaction between sex and family size and sex and urbanity. Figure 1.4 is a *traceplot*, showing the evolution of each coefficient value as the L1-norm increases.

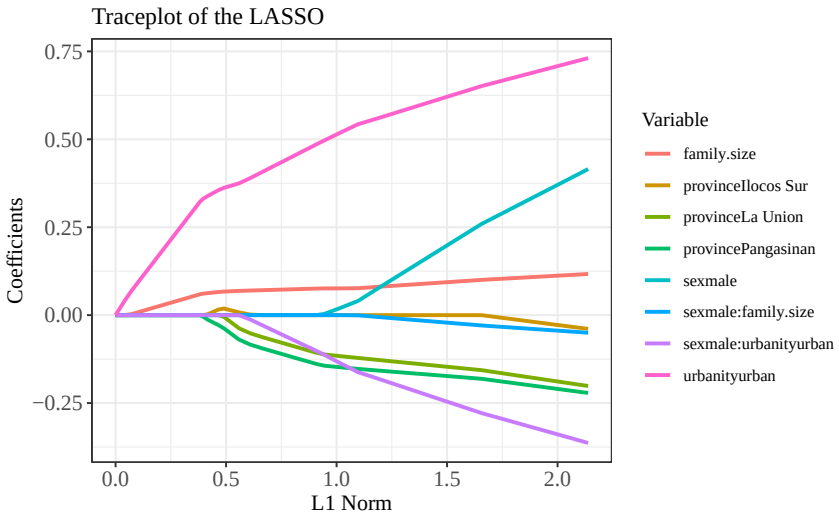

Figure 1.4: Traceplot of the LASSO algorithm on the `Ilocos` data

As a motivation for the SCAD, Fan and Li (2001) introduce three criteria that an appropriate penalty function should satisfy.

1. *Sparsity.* The penalty should ensure that irrelevant covariates are discarded. To do so, the procedure should be able to set coefficients exactly equal to zero. The LASSO satisfies this property. Also, subset selection can be seen as a sparse procedure.

2. *Continuity.* The penalty is a continuous function. This ensures that the penalized estimator does not move too much when the original coefficient changes slightly. Again, the LASSO satisfies this property. However, subset selection does not. If the coefficient is too small, it is set to zero. But as soon as it passes a given threshold it simply gets unpenalized.

3. *Unbiasedness.* The penalized estimator is unaffected by the penalization when the original estimator is large enough. This property is satisfied by subset selection for the same reason as before. It is not satisfied by the LASSO as its first derivative is constant.

Short of satisfying the unbiasedness requirement, the LASSO is not suitable to perform proper inference on the parameter vector. One solution is the debiased LASSO, see Zhang and Zhang (2014). Another lies in adapting the penalty to ensure that large enough coefficients are unpenalized. This is the idea underlying

the SCAD penalty, whose first derivative satisfies

$$p'_\lambda(x) = \begin{cases} \lambda & \text{if } x \leq \lambda \\ \frac{a\lambda - x}{a-1} & \text{if } \lambda < x \leq a\lambda \\ 0 & \text{if } x > a\lambda \end{cases} \tag{1.6.3}$$

where $a > 2$ is an arbitrary constant. A plot of the SCAD is displayed in Figure 1.5 for the usual value $a = 3.7$ and different choices of λ. The case $\lambda = 1$ here corresponds to the LASSO.

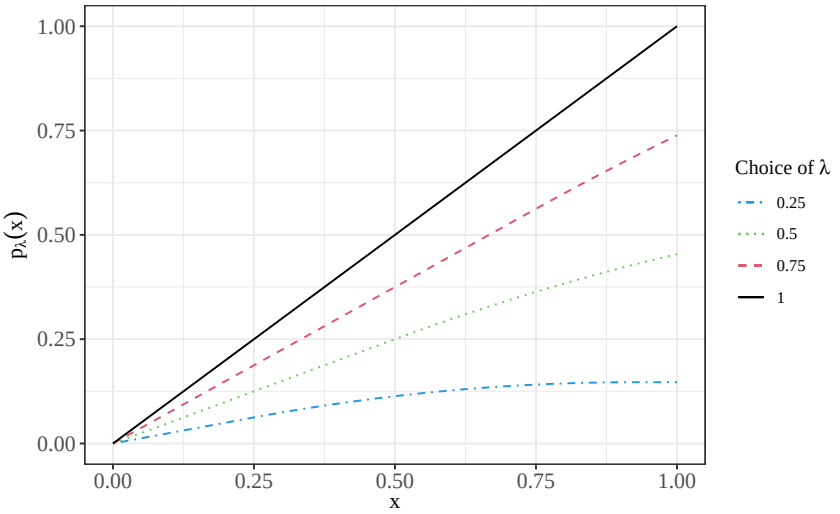

Figure 1.5: Plot of the SCAD penalty for different choices of λ

By construction, the SCAD penalty satisfies the requirements presented above. Because of this, the resulting estimator enjoys the so-called *oracle property*, which ensures two specific guarantees. First, the variable selection process is consistent, in that it sets to zero the estimated coefficients of the irrelevant covariates with a probability tending to one, see Theorem 1 in Fan and Li (2001). The second concerns inference and states that the estimated coefficients of the active covariates are asymptotically normal, with a rate of convergence unaffected by the selection process, see Theorem 2 in Fan and Li (2001).

We now turn to penalized procedures for single-index models. In Section 1.5, we drew a distinction between techniques where the link function is estimated and rank-based procedures avoiding this issue by exploiting the monotonicity of the link function. A penalized example of the former is the profile least-square estimator of Liang et al. (2010), proposed in the slightly more general context of a partially

linear single-index model.[2] The link function is replaced by a local linear estimator. Then, the parameter vectors are estimated by solving a SCAD-penalized least-squares programme. The oracle property of the estimator is established in Theorem 2 in Liang et al. (2010). We now turn to rank-based penalized procedures and focus on the penalized maximum smoother rank correlation (PMSRC) estimator of Lin and Peng (2013). They start from the MRC estimator of Han (1987) and adapt it in two ways. First, they replace the discrete objective function by the following smooth approximation

$$S_n(\theta) \coloneqq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{1}\{Y_i > Y_j\} \Phi\left(\frac{(X_i - X_j)^\intercal \theta}{h} \right)$$

where $\Phi(\cdot)$ is the CDF of a standard normal random variable and $h > 0$ is a bandwidth. Second a SCAD penalty is added to the optimization programme. Formally, the PMSRC estimator is obtained as

$$\hat{\theta}_{\text{PMSRC}} \coloneqq \arg\max_{\theta} \left(S_n(\theta) - \sum_{k=1}^{p} p_\lambda(|\theta_k|) \right), \tag{1.6.4}$$

where $p_\lambda(\cdot)$ is the SCAD penalty. Theorems 1 and 2 in Lin and Peng (2013) establish the oracle property of the estimator. The numerical solution can be obtained by the Newton-Raphson algorithm developed by Lin and Peng (2013) or by the forward and backward stagewise (FABS) algorithm, a coordinate-descent algorithm proposed by Shi et al. (2018) and reviewed in Chapter 3.

[2] In this case, one assumes that $\mathbb{E}[Y|X = x, W = w] = H(x^\intercal \theta_0) + w^\intercal \beta_0$, where W is a vector of covariates for which the impact on the response is known to be linear.

CHAPTER 2

The Lorenz regression

The Lorenz regression is an estimation procedure for the Gini coefficient of the conditional expectation of a response given a vector of covariates, assuming a single-index model. The measure of inequality thus obtained is called the explained Gini coefficient and it boils down to a concentration index between the response and a linear combination of the covariates. Unlike methods based on decompositions of inequality measures, the procedure does not assume a linear relationship between the response and the covariates. Inference can be performed by noticing a similarity with the monotone rank estimator, introduced by Cavanagh and Sherman (1998). A continuity correction is presented in the presence of discrete covariates. The Lorenz-R^2 is a goodness-of-fit measure evaluating the proportion of explained inequality and is used to build a test of joint significance of several covariates. Monte-Carlo simulations and a real-data example are presented. The content of this chapter is based on the manuscript

> Cédric Heuchenne and Alexandre Jacquemain. Inference for monotone single-index conditional means: A Lorenz regression approach. *Computational Statistics & Data Analysis*, 167(C), 2022. ISSN 0167-9473. doi: 10.1016/j.csda.2021.107347.

2.1 Introduction

The Lorenz regression attempts to quantify the explanatory power of a set of covariates concerning the inequality of a response. As we reviewed in Section 1.3, the literature on this subject tends to be clustered around either parametric approaches, lacking flexibility, and nonparametric ones, yielding unreliable estimates when too many covariates are included. An example of the former is the decomposition method of Wagstaff et al. (2003), which assumes a linear model between the response and the covariates. In the IOP context, Bourguignon et al. (2007),

estimate the inequality in the conditional expectation of the advantage given the circumstances, using a log-linear model. In the nonparametric framework, Yitzhaki (1994a) derives a decomposition of the Gini coefficient when the covariates partition the population into a number of groups. Using the same idea of a partition, Van de Gaer (1993) proposes a nonparametric measure of IOP. However, both methods are unreliable when the number of covariates gets too large. They also presuppose that the covariates are discrete.

In this chapter, we propose to measure the inequality of the response reproduced by some covariates with the explained Gini coefficient. We argue that it strikes a balance between flexibility and interpretability. It corresponds to the Gini coefficient of the conditional expectation of the response given the covariates and is therefore a natural ex-ante utilitarian measure of IOP. Still, it is based on a semiparametric single-index model, which is more flexible than usual parametric approaches. We refer the reader to Chapter 1 for an exposition of the relevant literature. The context of IOP is explained in Section 1.2. The measurement of inequality is tackled in Section 1.3. The literature concerning semiparametric regression is exposed in Section 1.4. More specifically, the single-index model is thoroughly discussed in Section 1.5. In what follows, we recall the main features of the single-index model and the optimization programme underlying the monotone rank estimator. We assume

$$\mathbb{E}[Y|X = x] = H(x^\intercal \theta_0), \tag{2.1.1}$$

where $H(\cdot)$ is a strictly increasing link function and θ_0 is a parameter vector satisfying $||\theta_0||= 1$. Further conditions are required to ensure the identifiability of the model, see conditions (SI1)-(SI4) in Section 1.5. The estimator of θ_0 proposed by the Lorenz regression is a special case of the MR estimator introduced by Cavanagh and Sherman (1998) and defined as

$$\hat{\theta}_{\text{MR}} \coloneqq \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j \neq i} M(Y_i) \mathbb{1}\{X_i^\intercal \theta > X_j^\intercal \theta\}.$$

where $M(\cdot)$ is an increasing function chosen by the user. In the context of the Lorenz regression, it is the identity function.

This chapter is divided into several sections. In Section 2.2, we introduce the main notions of this chapter, i.e. the explained Gini coefficient and the proportion of explained inequality, and illustrate their use with a simple example. The explained Gini coefficient is the Gini coefficient of a conditional expectation. However, thanks to the structure of the single-index model, it can be rewritten as the solution of a maximization programme. This opens the door to the inference procedure established in Section 2.3. The consistency and asymptotic normality of the estimators are obtained by noticing a similarity with the MR estimator. As the

covariance matrix of the estimator of θ_0 is difficult to estimate, we advocate for the use of bootstrap. Beyond constructing usual confidence intervals, we propose a test of joint significance for several covariates. Finally, a genetic algorithm is proposed to obtain a numerical solution of the estimator. Thus far, it is assumed that X contains at least one continuous covariate. In Section 2.4, we tackle the identifiability of the model and the estimation of θ_0 and $\mathrm{Gi}_{Y,X}$ when all covariates are discrete. The vector θ_0 is identifiable up to a region, which is determined by the ordering of the index that it induces. In terms of estimation, we propose to break the ties in the ordering of the index with a random assignment. We compare our solution with the mean-rank assignment and present some arguments in favor of our proposal. Section 2.5 illustrates the performance of the estimation procedure via Monte-Carlo simulations. A real data application is presented in Section 2.6. Section 2.7 provides a discussion over the advantages of the procedure as well as its limitations. The proofs of the results presented throughout the chapter are relegated to Section 2.8.

2.2 The Lorenz regression methodology

We call *explained Gini coefficient* the Gini coefficient of the conditional expectation of Y given X in the assumed single-index model, and denote this quantity by $\mathrm{Gi}_{Y,X}$. Throughout this chapter, we also use the notation F_θ for the CDF of $X^\intercal\theta$ and $F_H(\cdot)$ for the CDF of $H(X^\intercal\theta_0)$. The logic underlying the Lorenz regression can be understood from the chain of equations presented in Proposition 2.1.

Proposition 2.1. *Let $(X^\intercal, Y)^\intercal \in \mathbb{R}^p \times \mathbb{R}$, with $0 < \mathbb{E}[Y] < \infty$, be continuous random variables satisfying Equation (2.1.1) with $H(\cdot)$ strictly increasing. Assume conditions (SI1) to (SI4) presented in Section 1.5 are satisfied. Then, it holds that*

$$\mathrm{Gi}_{Y,X} \coloneqq \frac{2\mathbb{C}[H(X^\intercal\theta_0), F_H(H(X^\intercal\theta_0))]}{\mathbb{E}[H(X^\intercal\theta_0)]} \tag{2.2.1}$$

$$= \frac{2\mathbb{C}[Y, F_H(H(X^\intercal\theta_0))]}{\mathbb{E}[Y]} \tag{2.2.2}$$

$$= \frac{2\mathbb{C}[Y, F_{\theta_0}(X^\intercal\theta_0)]}{\mathbb{E}[Y]} \tag{2.2.3}$$

$$= \max_{\theta, ||\theta||=1} \frac{2\mathbb{C}[Y, F_\theta(X^\intercal\theta)]}{\mathbb{E}[Y]}. \tag{2.2.4}$$

Denoting by θ^* the value of θ which solves (2.2.4), it also holds that $\theta^* = \theta_0$ is unique.

In what follows, we sketch the proof of each of these results and detail their implications.

- Equation (2.2.2) is a direct consequence of the law of iterated expectations. In this formulation, the conditional expectation only intervenes through the ordering structure it generates, i.e. through $F_H(H(X^\intercal\theta_0))$. In the context of estimation, this tells us that one does not need a good estimator of the conditional expectation per se. One only needs an estimator that reproduces well its CDF. Notice that this result holds irrespective of whether the single-index model is assumed.
- Equation (2.2.3) exploits the monotonicity of the link function $H(\cdot)$. It indicates that, in the context of the assumed single-index model, an estimator for the explained Gini coefficient can be obtained without having to estimate the nonparametric part, i.e. the link function $H(\cdot)$. Only the parametric vector θ_0 needs to be estimated.
- Equation (2.2.4) opens the door to a simple estimation procedure for θ_0 and for $\mathrm{Gi}_{Y,X}$. This procedure would rely on maximizing the empirical covariance between Y and the empirical CDF of the index $X^\intercal\theta$. We now turn to the proof of this final result. Notice that, in the continuous case, $F_\theta(X^\intercal\theta) \sim U[0,1]$, where $U[0,1]$ denotes the continuous uniform distribution defined on the unit interval. Therefore, it holds that

$$\underset{\theta,||\theta||=1}{\arg\max} \frac{2\mathbb{C}[Y, F_\theta(X^\intercal\theta)]}{\mathbb{E}[Y]} = \underset{\theta,||\theta||=1}{\arg\max}\, \mathbb{E}[Y F_\theta(X^\intercal\theta)]. \tag{2.2.5}$$

  Consider $(X_1^\intercal, Y_1)^\intercal$ and $(X_2^\intercal, Y_2)^\intercal$, two i.i.d copies of $(X^\intercal, Y)^\intercal$. It holds that

$$\begin{aligned}\mathbb{E}[Y F_\theta(X^\intercal\theta)] &= \mathbb{E}[Y_1 \mathbb{1}\{X_1^\intercal\theta \geq X_2^\intercal\theta\}] \\ &= \frac{1}{2}\mathbb{E}[Y_1 \mathbb{1}\{X_1^\intercal\theta \geq X_2^\intercal\theta\}] + \frac{1}{2}\mathbb{E}[Y_2 \mathbb{1}\{X_2^\intercal\theta \geq X_1^\intercal\theta\}] \\ &= \frac{1}{2}\mathbb{E}[H(X_1^\intercal\theta_0) \mathbb{1}\{X_1^\intercal\theta \geq X_2^\intercal\theta\}] + \frac{1}{2}\mathbb{E}[H(X_2^\intercal\theta_0) \mathbb{1}\{X_2^\intercal\theta \geq X_1^\intercal\theta\}],\end{aligned}$$

  where the last line exploits the law of iterated expectation. Evaluated at θ_0, and using the monotonicity of $H(\cdot)$, the objective function becomes

$$\mathbb{E}[Y F_{\theta_0}(X^\intercal\theta_0)] = \frac{1}{2}\mathbb{E}[\max\{H(X_1^\intercal\theta_0), H(X_2^\intercal\theta_0)\}],$$

  which establishes that the maximum is indeed attained at θ_0. The proof of its uniqueness is more technical and is obtained from the proof of Theorem 1 in Cavanagh and Sherman (1998).

A second object of interest is the *proportion of explained inequality* (PEI), denoted by $\mathrm{PEI}_{Y,X}$ and defined as

$$\mathrm{PEI}_{Y,X} \coloneqq \frac{\mathrm{Gi}_{Y,X}}{\mathrm{Gi}_Y}.$$

assuming perfect equality cannot occur, i.e. $\mathrm{Gi}_Y > 0$. In this thesis, we will often refer to $\mathrm{Gi}_{Y,X}$ as measuring *explained inequality* and to Gi_Y as measuring *total inequality*. The PEI therefore measures the proportion of explained inequality over total inequality. This terminology is justified by the following proposition.

Proposition 2.2. *Let $(X^\intercal, Y)^\intercal \in \mathbb{R}^p \times \mathbb{R}$ be continuous random variables, with $0 < \mathbb{E}[Y] < \infty$. Then, it holds that*

$$\mathrm{PEI}_{Y,X} \in [0, 1].$$

The lower bound is a direct consequence of the fact that both the numerator and denominator are Gini coefficients and therefore bounded from below by 0. The upper bound is a corollary of Theorem 3.1 and Remark 6.2 in Das Gupta (1999), which shows that, at any point, the concentration curve of Y with respect to any variable W lies above the Lorenz curve of Y. Interestingly, Proposition 2.2 does not rely on the assumption of the single-index model. The ratio between the Gini coefficient of the conditional expectation of the response given the covariates and the Gini coefficient of the response is always between 0 and 1.

We now illustrate the notions of explained Gini coefficient and proportion of explained inequality in the IOP context by means of a simple example. Consider three datasets A, B and C. For a given measure of economic advantage Y and a set of circumstances X, we estimate the Gini coefficient of Y, the explained Gini coefficient of Y given X and the PEI. We obtain the estimates $\widehat{\mathrm{Gi}}_Y$, $\widehat{\mathrm{Gi}}_{Y,X}$ and $\widehat{\mathrm{PEI}}_{Y,X}$ displayed in Table 2.1. Total inequality, i.e. the inequality of the economic

Table 2.1: (Explained) Gini coefficients and PEI in datasets A, B and C (all results are expressed in percentage points)

	A	B	C
$\widehat{\mathrm{Gi}}_Y$	23.4	24.5	33.1
$\widehat{\mathrm{Gi}}_{Y,X}$	10.2	9	11.6
$\widehat{\mathrm{PEI}}_{Y,X}$	43.6	36.7	35

advantage, is the lowest in dataset A. However, there is more inequality of opportunity in dataset A than in dataset B since the explained Gini coefficient is larger in the former case. It is also in dataset A that unequal opportunities explain the largest share of total inequality since it displays the largest PEI. Finally, dataset C suffers both from the largest total inequality and the largest IOP. Still, it is in that dataset that unequal opportunities explain the lowest share of total inequality. In practice, datasets A and B could correspond to two different age-cohorts of the same country. Namely, A would correspond to young individuals and B to older

ones. In the beginning of the work-life, earnings inequality is lower but is more heavily related to circumstances. As seniority increases, the burden of circumstances becomes lighter even though total inequality is larger. This would explain the results observed in Table 2.1 for A and B. Dataset C could correspond to another country displaying both more total inequality and more IOP, even though the latter explains a lower share of the former. This fictitious illustration is not arbitrary as the estimates are taken from the real data application analyzed in Chapter 4.[1]

Further interpretations can be drawn from the Lorenz regression by analyzing the coefficient vector θ_0. First, the sign of each element $\theta_{0,k}$ is easy to interpret. A value of zero indicates that the circumstance does not contribute to the inequality of the advantage, i.e. it does not create inequality of opportunity. A positive (negative) coefficient rather refers to a notion of concordance (discordance). In order to understand what it entails, consider that the associated covariate represents the number of siblings. A positive coefficient corresponds to a situation where a larger share of the advantage is amassed by the individuals with larger families. If, instead, more advantage is gathered in the hands of individuals with less siblings, the coefficient would then exhibit a negative value.

By the assumption of monotonicity of the link function, the ratio between coefficients compares the relative impact of two circumstances on the advantage variable. We recall that the marginal rate of substitution (MRS) of X^k for X^l, all other things being equal, is given by

$$\mathrm{MRS}_{k,l} := \frac{\partial \mathbb{E}[Y|X=x]}{\partial x^k} \Big/ \frac{\partial \mathbb{E}[Y|X=x]}{\partial x^l} = \frac{\theta_{0,k}}{\theta_{0,l}} \,.$$

This quantity can be consistently estimated without the need to estimate the link function $H(\cdot)$.

2.3 Inference in the Lorenz regression

This section develops as follows. We start by presenting the estimators for the quantities of interest in the Lorenz regression, i.e. the parameter vector, the explained Gini coefficient and the proportion of explained inequality. Then, we turn to inference with a twofold objective: first of constructing confidence intervals and tests for the explained Gini coefficient; second, of developing a test of joint significance of one or several covariates in explaining the inequality of the response. Finally, we present a genetic algorithm to obtain numerical solutions for the proposed estimators.

[1] See Table 4.3. Dataset A is the young-age cohort of Belgium in 2011 (BE11Y); dataset B is the old-age cohort of Belgium in 2011 (BE11O); dataset C is the old-age cohort of Spain in 2011 (ES11O).

Given n i.i.d samples $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$, a consistent and asymptotically normal estimator for the Gini coefficient of Y is given by

$$\widehat{\mathrm{Gi}}_Y \coloneqq \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Y_i}{\overline{Y}}\mathbb{1}\{Y_i \geq Y_j\} - \frac{n+1}{n}, \tag{2.3.1}$$

where $\overline{Y}$ is the empirical mean of Y. The subtraction of $(n+1)/n$ is related to the fact that the empirical CDF of Y evaluated at each observation point has expected value $(n+1)/(2n)$.

As Equation (2.2.3) indicates, the explained Gini coefficient can be expressed as the concentration index of Y with respect to $X^\intercal\theta_0$. If θ_0 were known, an estimator could therefore be obtained by replacing the indicator in Equation (2.3.1) by $\mathbb{1}\{X_i^\intercal\theta_0 \geq X_j^\intercal\theta_0\}$. In practice, θ_0 is unknown and needs to be estimated. As we already mentioned, Equation (2.2.4) opens the door to an estimation procedure that does not entail to estimate nor to assume a specific form for $H(\cdot)$. More precisely, we estimate θ_0 and $\mathrm{Gi}_{Y,X}$ with

$$\overline{\theta} \coloneqq \underset{\theta,||\theta||=1}{\arg\max}\,\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}Y_i\mathbb{1}\{X_i^\intercal\theta \geq X_j^\intercal\theta\}, \tag{2.3.2}$$

$$\overline{\mathrm{Gi}}_{Y,X} \coloneqq \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Y_i}{\overline{Y}}\mathbb{1}\{X_i^\intercal\overline{\theta} \geq X_j^\intercal\overline{\theta}\} - \frac{n+1}{n}. \tag{2.3.3}$$

For ease of representation, we will often denote the estimator of the CDF of $X^\intercal\theta$ evaluated at observation i by

$$\hat{F}_i(\theta) \coloneqq \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}\{X_i^\intercal\theta \geq X_j^\intercal\theta\}.$$

The optimization programme displayed in Equation (2.3.2) therefore becomes

$$\overline{\theta} \coloneqq \underset{\theta,||\theta||=1}{\arg\max}\,\frac{1}{n}\sum_{i=1}^{n}Y_i\hat{F}_i(\theta). \tag{2.3.4}$$

The estimator $\overline{\theta}$ is a special case of the MR estimator proposed by Cavanagh and Sherman (1998) and introduced in Equation (1.5.4). The consistency of $\overline{\theta}$ therefore follows from Theorem 1 in Cavanagh and Sherman (1998) and its asymptotic normality is established by Theorem 2 in the same source. The consistency and asymptotic normality of the estimated explained Gini coefficient are derived in the more general case of the penalized Lorenz regression in Chapter 3. An estimator

for $\text{PEI}_{Y,X}$ is obtained by plugging the estimated Gini coefficients. We call this quantity the *Lorenz-R^2* and denote it by LR^2. Formally, it is given by

$$\text{LR}^2 \coloneqq \frac{\overline{\text{Gi}}_{Y,X}}{\widehat{\text{Gi}}_Y} = \frac{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{X_i^\intercal \overline{\theta} \geq X_j^\intercal \overline{\theta}\} - \frac{n+1}{n}\frac{\overline{Y}}{2}}{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{Y_i \geq Y_j\} - \frac{n+1}{n}\frac{\overline{Y}}{2}}. \tag{2.3.5}$$

Proposition 2.3 establishes that the Lorenz-R^2 always lies between 0 and 1. A value close to 0 appears when the covariates do not help in reproducing the inequality of Y. A value close to 1 is attained when they help reproduce almost all the inequality. The Lorenz-R^2 is therefore a goodness of fit measure based on the proportion of inequality reproduced by the covariates.

Proposition 2.3. *Let* LR^2 *be defined as in Equation* (2.3.5)*, we have*

$$\text{LR}^2 \in [0, 1].$$

Proof. The proposition is a direct consequence of the following chain of equations.

$$\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{Y_i \geq Y_j\} \geq \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{X_i^\intercal \overline{\theta} \geq X_j^\intercal \overline{\theta}\} \tag{2.3.6}$$

$$\geq \frac{n+1}{n}\frac{\overline{Y}}{2}. \tag{2.3.7}$$

The inequality in (2.3.6) is trivial so we focus on proving (2.3.7). Denote by $(Y_{(1)}, \ldots, Y_{(n)})^\intercal$ the response vector permuted such that the observations are ordered in increasing order in terms of the vector $X^\intercal\overline{\theta}$. For example, $Y_{(1)}$ is the response for the observation with the lowest value of the index $X^\intercal\overline{\theta}$. Consider $\overline{\theta}_- \coloneqq -\overline{\theta}$. We can write

$$\sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{X_i^\intercal \overline{\theta} \geq X_j^\intercal \overline{\theta}\} = \sum_{i=1}^n Y_{(i)} i$$

$$\sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{X_i^\intercal \overline{\theta}_- \geq X_j^\intercal \overline{\theta}_-\} = \sum_{i=1}^n Y_{(i)}(n-i+1).$$

Now recall that $\overline{\theta}$ is the solution to Equation (2.3.2). Therefore, it holds that

$$\sum_{i=1}^n Y_{(i)} i \geq \sum_{i=1}^n Y_{(i)}(n-i+1). \tag{2.3.8}$$

The desired inequality is then easily obtained after rearranging the terms. □

The proof of Proposition 2.3 enriches the understanding of the bounds of the Lorenz-R^2. The upper bound is attained when (2.3.6) reaches the equality. This

happens when the estimated index $X^\intercal\overline{\theta}$ perfectly reproduces the ranking of the response vector $(Y_1, \ldots, Y_n)^\intercal$. The lower bound is obtained when (2.3.8) reaches the equality. This occurs when the covariates have no explanatory power, in the sense that $\overline{\theta}$ and $-\overline{\theta}$ yield the same value of the objective function displayed in (2.3.2).

We now move to the issue of constructing tests and confidence intervals. Cavanagh and Sherman (1998) discuss the estimation of the asymptotic covariance matrix of the MR estimator. More specifically, two approaches are presented. The first is based on numerical derivatives of the objective function but it yields unstable results. The second relies on a nonparametric estimation of several model quantities, e.g. the density function of the index and the link function. However, this method is numerically intensive and removes the advantage of not having to estimate the link function. Besides, both methods involve crucial choices of smoothing parameters. We direct the interested reader to Subbotin (2007) for a thorough discussion of the drawbacks of these methods. An alternative lies in the use of bootstrap. The advantage of this method lies in the absence of any smoothing parameter. Interestingly, Subbotin (2007) establishes the consistency of bootstrap procedures for the MR estimator. Confidence intervals and tests may then be constructed either by bootstrapping the asymptotic covariance matrix only (parametric bootstrap), or rather by bootstrapping the distribution of $\overline{\theta}$ directly (basic bootstrap). Theorem 3 in Subbotin (2007) implies the consistency of the basic bootstrap, while Theorem 4 from the same source guarantees the consistency of the parametric bootstrap. A third method to construct confidence intervals consists in directly plugging the quantiles of the bootstrap distribution of $\overline{\theta}$ (percentile bootstrap). More precisely, $(1-\alpha)$-level confidence intervals for $\theta_{0,k}$ are given by

$$\begin{aligned}
\mathrm{CI}_{\mathrm{Basic}} &:= \left[2\overline{\theta}_k - q_{\overline{\theta}^*_k;1-\frac{\alpha}{2}}; 2\overline{\theta}_k - q_{\overline{\theta}^*_k;\frac{\alpha}{2}}\right], \\
\mathrm{CI}_{\mathrm{Percentile}} &:= \left[q_{\overline{\theta}^*_k;\frac{\alpha}{2}}; q_{\overline{\theta}^*_k;1-\frac{\alpha}{2}}\right], \\
\mathrm{CI}_{\mathrm{Parametric}} &:= \left[\overline{\theta}_k \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{\overline{\Sigma}^*_{kk}}}{\sqrt{n}}\right],
\end{aligned}$$

where $\overline{\theta}^*$ is the estimator of θ_0 in the bootstrap sample. Moreover, $q_{\overline{\theta}^*_k;a}$ is the bootstrap estimator of the a-quantile of the distribution of $\overline{\theta}^*_k$ and $\overline{\Sigma}^*$ is the bootstrap estimator of the asymptotic variance-covariance matrix of $\overline{\theta}^*$. Finally, z_a is the a-quantile of the standard normal distribution. With this in mind, it is easy to obtain bootstrap confidence intervals and to build tests concerning $\mathrm{Gi}_{Y,X}$. For example, one could test the equality of the explained Gini coefficients between two countries.

Central to the motivation of this paper is the idea of assessing whether a set of

covariates explains significantly the inequality of the response. Accordingly, we develop a bootstrap test for the joint significance of d variables. For each $i = 1, \dots, n$, assume $Y_i = H(X_i^\intercal \theta_0) + \epsilon_i$, with $H(\cdot)$ strictly increasing and ϵ_i independent from X_i. The null and alternative hypotheses of the desired test are respectively given by

$$\begin{aligned} &H_0 : \theta_{0,k}, \dots, \theta_{0,k+d-1} = 0 \\ &H_1 : \exists j \in [0, d-1] \text{ such that } \theta_{0,k+j} \neq 0, \end{aligned}$$

with $k \leq p - d + 1$. The idea underlying the following testing procedure lies in comparing the Lorenz-R^2 under the null hypothesis with the unconstrained one. Accordingly, we reject H_0 if dropping the d variables leads to a significant decrease in explained inequality. Specifically, we use the following test statistic

$$U := \frac{\mathrm{LR}^2}{\mathrm{LR}^2_{H_0}} = \frac{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{X_i^\intercal \overline{\theta} \geq X_j^\intercal \overline{\theta}\} - \frac{n+1}{n}\frac{\overline{Y}}{2}}{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{X_i^{(0)\intercal} \overline{\theta}^{(0)} \geq X_j^{(0)\intercal} \overline{\theta}^{(0)}\} - \frac{n+1}{n}\frac{\overline{Y}}{2}},$$

where $X_i^{(0)}$ is the matrix of covariates obtained after dropping columns k to $k+d-1$ from X_i and $\overline{\theta}^{(0)}$ is the estimator of θ_0 based on $X_i^{(0)}$. More precisely,

$$\overline{\theta}^{(0)} := \underset{\theta, ||\theta||=1}{\arg\max} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n Y_i \mathbb{1}\{X_i^{(0)\intercal} \theta \geq X_j^{(0)\intercal} \theta\}.$$

Intuitively, one rejects the null hypothesis if the observed value of U is sufficiently larger than 1. In order to obtain a bootstrap sample for the test statistic under H_0, the constraint imposed by the null hypothesis must be incorporated in the resampling mechanism. To do so, we compute the residuals under H_0. These are obtained as

$$\hat{\epsilon}_i^{(0)} = Y_i - \hat{H}^{(0)}\left(X_i^{(0)\intercal} \hat{\theta}^{(0)}\right),$$

where $\hat{H}^{(0)}(\cdot)$ is an estimator of $H(\cdot)$ under H_0. Then, $\epsilon_i^{(0)*}$ is randomly drawn with replacement from $\hat{\epsilon}_1^{(0)}, \dots, \hat{\epsilon}_n^{(0)}$. The bootstrap sample is obtained as $(X_i^\intercal, Y_i^*)_{i=1,\dots,n}^\intercal$, where

$$Y_i^* = \hat{H}^{(0)}\left(X_i^{(0)\intercal} \hat{\theta}^{(0)}\right) + \epsilon_i^{(0)*}.$$

The p-value of the test is approximated with $p^* = P_n(U^* \geq U_{\text{obs}})$, where P_n indicates that the probability is conditional on the original sample $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$. Also, U_{obs} is the observed value of U in the original sample and U^* is obtained as

$$U^* = \frac{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Y_i^* \mathbb{1}\{X_i^\intercal \overline{\theta}^* \geq X_j^\intercal \overline{\theta}^*\} - \frac{n+1}{n}\frac{\overline{Y^*}}{2}}{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Y_i^* \mathbb{1}\{X_i^{(0)\intercal} \hat{\theta}^{(0)*} \geq X_j^{(0)\intercal} \hat{\theta}^{(0)*}\} - \frac{n+1}{n}\frac{\overline{Y^*}}{2}},$$

where $\overline{\theta}^*$ is the unconstrained estimator of θ_0 in the bootstrap sample and $\overline{\theta}^{(0)*}$ the constrained one. Also, $\overline{Y^*}$ is the empirical mean in the bootstrap sample. Finally, p^* can be estimated via Monte-Carlo simulations.

The described procedure opens the door to the estimation of the link function $H(\cdot)$. Upon estimation of θ_0, we obtain an estimated index $X^\intercal\overline{\theta}$. In a second stage, we estimate $H(\cdot)$ as a nonparametric smoother of Y given $X^\intercal\overline{\theta}$, subject to the constraint that $H(\cdot)$ is strictly increasing. To achieve this, we use the procedure laid down by Chernozhukov et al. (2009), hereafter called *rearrangement method*, which starts from an initial estimator of $H(\cdot)$ and makes it increasing by taking its quantile function. Interestingly, this rearrangement operation weakly reduces the estimation error of the initial estimator, see Proposition 1 in Chernozhukov et al. (2009).

We close this section with the issue of the numerical solution of the Lorenz regression. The estimation of θ_0 requires to solve the discrete maximization programme displayed in (2.3.2). In this respect, our contribution consists in proposing a genetic algorithm with the double advantage of being fast and reducing the risk of local minima. The *genetic algorithm* is a generic method for solving optimization problems. It iteratively proposes solution candidates and assesses their score through a fitness function. The evolution of the candidates is governed by the principles of crossover and mutation. Convergence is met when the fitness has not improved for a sufficiently large number of generations. For a quick introduction of the method and its implementation in R via the library **GA**, we refer the reader to Scrucca (2013). In our context, the only technical difficulty lies in the incorporation of the unit-norm constraint. We propose the following principles to address this issue.

(i) Candidates γ are vectors of size $p-1$ defining two possible solutions

$$\overline{\theta}_1(\gamma) = (\gamma^\intercal, -(1-||\gamma||))^\intercal$$
$$\overline{\theta}_2(\gamma) = (\gamma^\intercal, 1-||\gamma||)^\intercal.$$

(ii) We ensure that the initial population of candidates satisfies the unit-norm constraint. We adapt the mutation and crossover mechanisms to ensure that, at each iteration, the candidates satisfy $||\gamma|| \leq 1$.

(iii) The fitness function $\mathrm{Fit}(\gamma)$ associated to candidate γ is defined by

$$\mathrm{Fit}(\gamma) := \max\{\mathrm{Fit}_1(\gamma), \mathrm{Fit}_2(\gamma)\}, \tag{2.3.9}$$

where, for $k \in \{1,2\}$,

$$\mathrm{Fit}_k(\gamma) := \sum_{i=1}^{n} Y_i \hat{F}_i(\overline{\theta}_k(\gamma)).$$

2.4 Accommodating for discrete covariates

The situation where all covariates are discrete creates two distinct issues: the question of ties management in the estimation of the Gini coefficient and the problem of non-identifiability in the single-index model. We solve the first issue with a continuity correction, which amounts to break the ties at random. Concerning the second, we argue that the parameter vector θ_0 is identifiable up to a region, and this is sufficient to uniquely identify and consistently estimate the explained Gini coefficient.

The remaining of this section develops as follows. In Subsection 2.4.1, we introduce the *random* continuity correction and compare it with the alternative *mean-rank* correction. At the population level, the two corrections are equivalent, i.e. they define the same corrected explained Gini coefficient. Subsection 2.4.2 establishes the identifiability and consistency of the estimator of the corrected explained Gini coefficient, obtained with either of the two continuity correction methods. In Subsection 2.4.3, we compare the random assignment with the mean-rank solution and provide some arguments in favor of our choice. Equations (2.4.1) to (2.4.3) present the estimators of the CDF of the index in the continuous and discrete cases.

$$\hat{F}_i(\theta) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_i^\intercal\theta \geq X_j^\intercal\theta\} \qquad \text{(continuous)} \tag{2.4.1}$$

$$\dot{F}_i(\theta) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X_i^\intercal\theta > X_j^\intercal\theta] + \frac{1}{2}\mathbb{1}[X_i^\intercal\theta = X_j^\intercal\theta] \quad \text{(discrete - mean-rank)} \tag{2.4.2}$$

$$\tilde{F}_i(\theta) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X_i^\intercal\theta > X_j^\intercal\theta] + \mathbb{1}[X_i^\intercal\theta = X_j^\intercal\theta, V_i > V_j], \quad \text{(discrete - random)} \tag{2.4.3}$$

where $(V_i)_{i=1,\dots,n}$ are i.i.d copies of $V \sim U[0,1]$.

2.4.1 The random continuity correction

From Schechtman and Zitikis (2006), we know that keeping the CDF in the definition of the Gini coefficient leads to an inconsistency between its interpretation as a covariance on one side and a surface between the egalitarian line and the Lorenz curve on the other side. In order to restore the concordance between these two perspectives, it is enough to replace the CDF by a suitable alternative. One of these alternatives, presented in Lerman and Yitzhaki (1989), leads to an estimator where the average rank is given in every situation where a tie occurs. In this section, we propose another solution. It consists in a random assignment of the ranks

affected by a situation of ties. Equations (2.4.4) to (2.4.6) display the usual CDF, providing the ordering structure in the continuous case, and the two corrections used in the discrete case.

$$F_\theta(t) = P(X^\intercal\theta \leq t) \qquad \text{(continuous)} \tag{2.4.4}$$

$$\dot{F}_\theta(t) = P(X^\intercal\theta < t) + \frac{1}{2}P(X^\intercal\theta = t) \qquad \text{(discrete - mean-rank)} \tag{2.4.5}$$

$$\tilde{F}_\theta(t) = P(X^\intercal\theta < t) + VP(X^\intercal\theta = t), \qquad \text{(discrete - random)} \tag{2.4.6}$$

where $V \sim U[0,1]$.

In the full discrete case, the concordance between maximizing $\mathbb{C}[Y, F_\theta(X^\intercal\theta)]$ and $\mathbb{E}[YF_\theta(X^\intercal\theta)]$ is compromised. Essentially, this is due to the fact that $F_\theta(X^\intercal\theta) \sim U[0,1]$ no longer holds. Proposition 2.4 provides an impulse for using the random correction as it shows that $\tilde{F}_\theta(X^\intercal\theta) \sim U[0,1]$ is always guaranteed.

Proposition 2.4. *Let Z be a non-continuous random variable with CDF $F(\cdot)$ having a finite number of discontinuities, and let $\tilde{F}(\cdot)$ be defined as in Equation (2.4.6). Then, $\tilde{F}(Z) \sim U[0,1]$.*

The proof of this result is relegated to Section 2.8. The corrected objective function is $\tilde{G}(\theta) \coloneqq \mathbb{E}[Y\tilde{F}_\theta(X^\intercal\theta)]$. By Proposition 2.4, it is equivalent to maximizing the following corrected concentration index

$$\widetilde{\mathrm{CI}}_{Y,X^\intercal\theta} \coloneqq \frac{2\mathbb{C}[Y, \tilde{F}_\theta(X^\intercal\theta)]}{\mathbb{E}[Y]}.$$

The corrected explained Gini coefficient is $\widetilde{\mathrm{Gi}}_{Y,X} = \widetilde{\mathrm{CI}}_{Y,X^\intercal\theta_0}$.

In the population, the random and the mean-rank corrections are completely equivalent, in that

$$\mathbb{E}[Y\tilde{F}_\theta(X^\intercal\theta)] = \mathbb{E}[Y\dot{F}_\theta(X^\intercal\theta)] \tag{2.4.7}$$

$$\mathbb{C}[Y, \tilde{F}_\theta(X^\intercal\theta)] = \mathbb{C}[Y, \dot{F}_\theta(X^\intercal\theta)]. \tag{2.4.8}$$

Consequently, the corrected explained Gini coefficient is uniquely defined, irrespective of the choice between the two correction methods. Equation (2.4.7) emerges as a trivial consequence of the independence of V. In order to show that Equation (2.4.8) holds, it is enough to show that $\mathbb{E}[\dot{F}_\theta(X^\intercal\theta)] = \frac{1}{2}$. To do so, let $z_1, z_2, \ldots$ denote the values of $Z \coloneqq X^\intercal\theta_0$ arranged in ascending order and define

$\pi_k \coloneqq P(Z = z_k)$. We have

$$\begin{aligned}\mathbb{E}[\dot{F}_\theta(X^\intercal\theta)] &= \sum_{k\geq 1}\Big[\sum_{l=1}^{k}\pi_l - \frac{\pi_k}{2}]\Big]\pi_k \\ &= \frac{1}{2}\Big[2\sum_{k\geq 2}\sum_{l=1}^{k-1}\pi_l\pi_k + \sum_{k\geq 1}\pi_k^2\Big] \\ &= \frac{1}{2}\Big[\sum_{k\geq 1}\pi_k\Big]^2 = \frac{1}{2}.\end{aligned}$$

We now introduce the correction in the sample. We focus our exposition on the random solution. However, the estimators based on the mean-rank are similarly constructed. The corrected estimator for θ_0 is

$$\tilde{\theta} \coloneqq \underset{\theta,||\theta||=1}{\arg\max}\,\tilde{G}_n(\theta), \tag{2.4.9}$$

with the corrected objective function

$$\tilde{G}_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} Y_i\tilde{F}_i(\theta),$$

where $\tilde{F}_i(\theta)$ is defined in Equation (2.4.3). The new estimator for the explained Gini coefficient is

$$\widetilde{\mathrm{Gi}}_{Y,X} \coloneqq \frac{2}{n}\sum_{i=1}^{n}\frac{Y_i}{\overline{Y}}\tilde{F}_i(\theta) - 1. \tag{2.4.10}$$

Notice that, in contrast with the former estimator of the (explained) Gini coefficient, the subtraction is of -1. This stems from the fact that the corrected version of the empirical CDF has expected value $1/2$.

In Subsection 2.4.2, we show that both corrections provide strong statistical guarantees in terms of estimating the true corrected explained Gini coefficient. In Subsection 2.4.3, we provide some arguments in favor of using the random solution.

2.4.2 Statistical properties of the continuity corrections

As mentioned in the beginning of this section, the single-index model is not identifiable when all covariates are discrete. We accommodate this issue by adapting a construction developed in Section 5 of Cavanagh and Sherman (1998). We argue

that it best characterizes the extent of the identifiability of θ_0. Throughout this section, we assume that $(X^\intercal, Y)^\intercal$ satisfy Equation (2.1.1) with $H(\cdot)$ strictly increasing on its points of definition. Finally, we focus our exposition on the random solution but all results also hold using the mean-rank correction.

Since the formal development is quite technical, we first introduce the general idea with a simple illustration. Assume that X is made of two binary covariates. As such, X has four distinct values that we denote

$$\begin{aligned} x_1 &\coloneqq (1,0)^\intercal \\ x_2 &\coloneqq (0,1)^\intercal \\ x_3 &\coloneqq (0,0)^\intercal \\ x_4 &\coloneqq (1,1)^\intercal. \end{aligned}$$

Consider that the true parameter vector is $\theta_0 = (\frac{\sqrt{3}}{2}, \frac{1}{2})^\intercal$, such that

$$x_4^\intercal\theta_0 > x_1^\intercal\theta_0 > x_2^\intercal\theta_0 > x_3^\intercal\theta_0.$$

The parameter vector θ_0 is indistinguishable from any other value of θ that leads to the same ranking. However, it is distinguishable from values of θ leading to a different ordering. In this case, θ_0 is therefore distinguishable from, say, $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\intercal$ or $(\frac{1}{\sqrt{4}}, \frac{\sqrt{3}}{\sqrt{4}})^\intercal$, but not from $(\frac{\sqrt{2}}{\sqrt{3}}, \frac{1}{\sqrt{3}})^\intercal$. Consider now a second situation where $\theta_0 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\intercal$. This entails

$$x_4^\intercal\theta_0 > x_1^\intercal\theta_0 = x_2^\intercal\theta_0 > x_3^\intercal\theta_0.$$

Now, θ_0 is indistinguishable from any value of θ that satisfies the orderings

$$\begin{aligned} x_4^\intercal\theta > x_1^\intercal\theta > x_3^\intercal\theta \\ x_4^\intercal\theta > x_2^\intercal\theta > x_3^\intercal\theta. \end{aligned}$$

In this situation, θ_0 is indistinguishable from $(\frac{1}{2}, \frac{\sqrt{3}}{2})^\intercal$ and $(\frac{\sqrt{3}}{2}, \frac{1}{2})^\intercal$. It is however distinguishable from $(-\frac{1}{2}, \frac{\sqrt{3}}{2})^\intercal$ for example.

In conclusion, θ_0 is identified up to a region, which is characterized by the strict inequalities in the ordering of the index $x^\intercal\theta_0$. To formalize this identifiability, we define a vector of *representatives*, where each representative is an arbitrary member of its region. We will not be able to identify θ_0 but only the representative of its region. This has, however, no consequence in terms of identification of the explained Gini coefficient. We return to the example developed above in order to illustrate this point. Let $\theta_0 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\intercal$. Let $(X_1^\intercal, Y_1)^\intercal$ and $(X_2^\intercal, Y_2)^\intercal$ be two i.i.d copies of $(X^\intercal, Y)^\intercal$. We have

$$\begin{aligned} \tilde{G}(\theta_0) &= \mathbb{E}[Y_1 \mathbb{1}\{X_1^\intercal\theta_0 > X_2^\intercal\theta_0\}] + \frac{1}{2}\mathbb{E}[Y_1 \mathbb{1}\{X_1^\intercal\theta_0 = X_2^\intercal\theta_0\}] \\ &= \mathbb{E}[H(X_1^\intercal\theta_0)\mathbb{1}\{X_1^\intercal\theta_0 > X_2^\intercal\theta_0\}] + \frac{1}{2}\mathbb{E}[H(X_1^\intercal\theta_0)\mathbb{1}\{X_1^\intercal\theta_0 = X_2^\intercal\theta_0\}], \end{aligned}$$

where the first line exploits the independence of V and the second line uses the law of iterated expectation and the monotonicity assumption of the link function. Denote $p(x_k) \coloneqq P(X = X_k)$, we have that

$$\begin{aligned}\tilde{G}(\theta_0) = {} & H(x_4^\intercal\theta_0)p(x_4)\Big[p(x_1)+p(x_2)+p(x_3)+\frac{1}{2}p(x_4)\Big] \\ & + H(x_1^\intercal\theta_0)p(x_1)\Big[p(x_3)+\frac{1}{2}p(x_1)+\frac{1}{2}p(x_2)\Big] \\ & + H(x_2^\intercal\theta_0)p(x_2)\Big[p(x_3)+\frac{1}{2}p(x_1)+\frac{1}{2}p(x_2)\Big] \\ & + H(x_3^\intercal\theta_0)\frac{[p(x_3)]^2}{2}.\end{aligned}$$

Consider now $\theta = (\frac{1}{2}, \frac{\sqrt{3}}{2})^\intercal$. It holds that

$$\begin{aligned}\tilde{G}(\theta) = {} & H(x_4^\intercal\theta_0)p(x_4)\Big[p(x_1)+p(x_2)+p(x_3)+\frac{1}{2}p(x_4)\Big] \\ & + H(x_1^\intercal\theta_0)p(x_1)\Big[p(x_2)+p(x_3)+\frac{1}{2}p(x_1)\Big] \\ & + H(x_2^\intercal\theta_0)p(x_2)\Big[p(x_3)+\frac{1}{2}p(x_2)\Big] \\ & + H(x_3^\intercal\theta_0)\frac{[p(x_3)]^2}{2}.\end{aligned}$$

Noticing that $H(x_2^\intercal\theta_0) = H(x_1^\intercal\theta_0)$, we have that $\tilde{G}(\theta) = \tilde{G}(\theta_0)$. This illustrates the fact that $\widetilde{\mathrm{CI}}_{Y,X^\intercal\theta} = \widetilde{\mathrm{Gi}}_{Y,X}$ for any θ in the region of θ_0.

Alternatively, consider $\theta = (-\frac{1}{2}, \frac{\sqrt{3}}{2})^\intercal$. We have that

$$\begin{aligned}\tilde{G}(\theta) = {} & H(x_2^\intercal\theta_0)p(x_2)\Big[p(x_4)+p(x_3)+p(x_2)+\frac{1}{2}p(x_2)\Big] \\ & + H(x_4^\intercal\theta_0)p(x_4)\Big[p(x_3)+p(x_1)+\frac{1}{2}p(x_4)\Big] \\ & + H(x_3^\intercal\theta_0)p(x_3)\Big[p(x_1)+\frac{1}{2}p(x_3)\Big] \\ & + H(x_1^\intercal\theta_0)\frac{[p(x_1)]^2}{2}.\end{aligned}$$

Hence,

$$\begin{aligned}\tilde{G}(\theta) - \tilde{G}(\theta_0) = {} & [H(x_2^\intercal\theta_0) - H(x_4^\intercal\theta_0)]p(x_2)p(x_4) \\ & + \left[H(x_2^\intercal\theta_0) - \frac{H(x_1^\intercal\theta_0)+H(x_2^\intercal\theta_0)}{2}\right]p(x_1)p(x_2) \\ & + [H(x_3^\intercal\theta_0) - H(x_1^\intercal\theta_0)]p(x_1)p(x_3)\end{aligned}$$

is strictly negative by the assumption of monotonicity. This illustrates the fact that $\widetilde{\mathrm{CI}}_{Y,X^\intercal\theta} < \widetilde{\mathrm{Gi}}_{Y,X}$ for any θ outside the region of θ_0.

This construction justifies the passage from the original parameter space to a discrete vector of representatives, arbitrarily chosen from each region. We now formalize the idea in general. Suppose X has possible values $x_1, \ldots, x_N$. Let $B_1, \ldots, B_l$ be open regions where $x_j^\intercal\theta \neq x_k^\intercal\theta$ for any $k \neq j$ and such that any pair of θ share the same ordering of $\{x_1^\intercal\theta, \ldots, x_N^\intercal\theta\}$ if and only if they live in the same region. These regions are bounded by hyperplanes $H_1, \ldots, H_m$ where $x_j^\intercal\theta = x_k^\intercal\theta$ for at least one pair $k \neq j$. In our illustration, examples of these regions are given by

$$
\begin{aligned}
B_1 &= \{x_4^\intercal\theta > x_1^\intercal\theta > x_2^\intercal\theta > x_3^\intercal\theta\},\\
H_1 &= \{x_4^\intercal\theta > x_1^\intercal\theta = x_2^\intercal\theta > x_3^\intercal\theta\},\\
B_2 &= \{x_4^\intercal\theta > x_2^\intercal\theta > x_1^\intercal\theta > x_3^\intercal\theta\}.
\end{aligned}
$$

Denote by $I(\theta)$ the subspace defined by the strict inequalities in the ordering of $\{x_1^\intercal\theta, \ldots, x_N^\intercal\theta\}$ implied by θ. Our conceptualization starts from the region where θ_0 falls, whether it's an open region or a hyperplane. The important step is the following. We merge all the regions characterized by $I(\theta) \subseteq I(\theta_0)$. In the illustration, if $\theta_0 \in B_1$, or if $\theta_0 \in B_2$, the set of strict inequalities fully describes the region and no grouping needs to be made. If $\theta_0 \in H_1$, the set of strict inequalities is $I(\theta_0) = \{x_4^\intercal\theta > x_1^\intercal\theta > x_3^\intercal\theta$ and $x_4^\intercal\theta > x_2^\intercal\theta > x_3^\intercal\theta\}$. In this case, for any θ in B_1 or B_2, we have $I(\theta) \subseteq I(\theta_0)$. Hence, the three regions need to be merged. This construction formalizes the idea that θ_0 can be identified up to the region defined by the strict inequalities in the ordering of the index. The resulting regions, as well as the remaining hyperplanes and open regions, are then summarized by a vector of representatives. The new parameter space is denoted by Θ^*. Proposition 2.5 establishes the identifiability of the explained Gini coefficient. The point (i) ensures that all parameter vectors inside the region of θ_0 yield the same value of the corrected concentration index, i.e. $\widetilde{\mathrm{CI}}_{Y,X^\intercal\theta} = \widetilde{\mathrm{Gi}}_{Y,X}$. The point (ii) ensures that the concentration index is maximized only in the region of θ_0. Its proof is relegated to Section 2.8.

Proposition 2.5. *Let $(X^\intercal, Y)^\intercal$, where X is a vector of discrete covariates and Y is a continuous random variable with $\mathbb{E}[Y] < \infty$. Let $\tilde{G}(\theta)$ be defined as above and θ_0^* be the representative of the region of θ_0. Then, it holds that*

(i) $\tilde{G}(\theta) = \tilde{G}(\theta_0)$ for all θ in the same region as θ_0.

(ii) The representative θ_0^* is the unique maximizer of $\tilde{G}(\theta^*)$, for $\theta^* \in \Theta^*$.

Finally, Theorem 2.1 ensures the consistency of the estimated explained Gini coefficient.

Theorem 2.1. *Let $(X^\intercal, Y)^\intercal$, where X is a vector of discrete covariates and Y is a continuous random variable with $\mathbb{E}[Y] < \infty$. Also, let $V \sim U[0,1]$ independent of $(X^\intercal, Y)^\intercal$. Assume the single-index model with the link function $H(\cdot)$ strictly increasing on its points of definition. Besides, let $\tilde{G}(\theta)$ and $\tilde{G}_n(\theta)$ be defined as above. Then, it holds that*

(i) $\tilde{G}_n(\theta) \to \tilde{G}(\theta)$ almost surely for all $\theta^* \in \Theta^*$.

(ii) $P(\overline{\theta}^* \neq \theta_0^*) = O(1/n)$, where $\overline{\theta}^*$ is the solution to (2.4.9) maximized on Θ^* and θ_0^* denotes the representative of the region where θ_0 lies.

The strategy of the proof is exactly similar to that of Theorem 3 in Cavanagh and Sherman (1998). (i) follows from standard results on U-statistics. (ii) is guaranteed by the almost sure convergence of $\tilde{G}_n(\theta)$, the continuity of $\tilde{G}(\theta)$, and the fact that $\tilde{G}(\theta)$ is uniquely maximized in θ_0.

The findings established in this subsection prompt a more general comment. The identifiability of θ_0 is not required in order to identify and consistently estimate the suitably corrected explained Gini coefficient. What matters is that, in all instances, we are able to identify the ordering structure yielded by the index. In this respect, assumption (SI2) is not needed. However, the reasoning can be further developed. For similar reasons, assumptions (SI3), i.e. absence of multicolinearity, and assumption (SI4), i.e. the normalization of θ_0 are not needed to identify and consistently estimate the explained Gini coefficient.

2.4.3 Comparison with the mean-rank correction

As already stated, the mean-rank and random corrections are equivalent in the population and both methods provide the statistical guarantees developed in Subsection 2.4.2. In practice, however, the two methods are not equivalent. We develop two arguments in favor of the random assignment. The first consists in the idea that the mean-rank solution discriminates against solution vectors lying on the hyperplanes $\{H_1, \ldots, H_m\}$. The second is the ability of the random solution to quantify the severity of the ties issue on the value taken by the explained Gini coefficient. We elaborate these two arguments in turn.

Consider data generated by the assumed single-index model, with only two binary covariates, and using parameter vector $\theta_0 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\intercal$. In this setup, the number of bounding regions is quite low and a discrete algorithm could be used to estimate the parameter vector. Besides θ_0, we consider two alternative vectors of

parameters

$$\theta_1 \coloneqq \left(\frac{\sqrt{3}}{\sqrt{4}}, \frac{1}{\sqrt{4}}\right)^\intercal,$$
$$\theta_2 \coloneqq \left(\frac{1}{\sqrt{4}}, \frac{\sqrt{3}}{\sqrt{4}}\right)^\intercal,$$

which, without loss of generality, are the representatives of their regions. Finally, we suppose that our sample is made of $n = 20$ observations and consists in a balanced design, i.e. 5 observations in each of the 4 possible values of X. In what follows, we focus on observations for which $X = (1,0)^\intercal =: x_1$ and those for which $X = (0,1)^\intercal =: x_2$. In this way, $x_1^\intercal\theta_0 = x_2^\intercal\theta_0$, while $x_1^\intercal\theta_1 > x_2^\intercal\theta_1$ and $x_1^\intercal\theta_2 < x_2^\intercal\theta_2$. For $a \leq b$, denote by $U\{a,b\}$ the discrete uniform distribution with support $\{a, a+1, \ldots, b\}$. Table 2.2 shows the ranks that each method attaches to those observations if the index is built upon either θ_0, θ_1 or θ_2. For example, each observation with value x_1 is given a rank of 10.5 if we are considering θ_0 and the average-rank solution. They have a rank randomly drawn between 6 and 15 if we are using the random solution.

Table 2.2: Rank granted to each group of observations

	Average rank		Random rank	
	x_1	x_2	x_1	x_2
θ_0	10.5	10.5	$U\{6,15\}$	$U\{6,15\}$
θ_1	13	8	$U\{11,15\}$	$U\{6,10\}$
θ_2	8	13	$U\{6,10\}$	$U\{11,15\}$

Estimating θ_0 requires to maximize the scalar product between the response and the vector of index ranks. Denote by $\overline{y}_1$ ($\overline{y}_2$) the mean response for observations with covariate vector x_1 (x_2). From the table, it is clear that θ_0 is never chosen by the average-rank solution, unless $\overline{y}_1 = \overline{y}_2$. Indeed, if $\overline{y}_1 > \overline{y}_2$, one would pick θ_1. Conversely, if $\overline{y}_1 < \overline{y}_2$, one would pick θ_2. This is not the case with the random rank solution since it depends on the generation of the uniform random variables. We performed 1000 Monte-Carlo simulations to confirm this intuition. More specifically, we generated 1000 samples from the DGP presented in Equation (2.5.4), with two binary covariates and parameter vector $\theta_0 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\intercal$. In each iteration, we recorded the maximizer(s) of the objective function. Table 2.3 counts the number of times θ_0, θ_1 and θ_2 maximize the objective function.

As we can observe, θ_0 is almost never chosen with the average-rank solution. The

Table 2.3: Number of times each parameter vector maximizes the objective function

	θ_0	θ_1	θ_2
Average rank	8	505	503
Random rank	177	435	420

8 instances where θ_0 maximizes the objective function correspond to situations where there is no observation either in x_1 or x_2. In this case, the three values of θ provide the same value for the objective function. This is not the case for the random solution. In total, θ_0 maximizes the objective function in 177 instances. Beyond the same 8 instances where θ_0, θ_1 and θ_2 all reach the optimum, there are 16 instances where θ_0 and one of either θ_1 or θ_2 maximize the objective function. This situation occurs when the generation of the uniform random variables ends up yielding the same rank vectors for θ_0 and for either θ_1 or θ_2. Hence, the 153 remaining instances correspond to situations where θ_0 is the sole maximizer of the objective function. Recall that it never occurs with the average rank method. The overall conclusion of this exercise runs as follows: the average-rank solution discriminates against certain types of solutions. These solutions are characterized by equal values of the index but made up with different values of the covariates. While the random solution remains neutral about this, the average-rank solution tries to disaggregate the index as much as possible.

Beyond statistical considerations, each procedure to solve the ties issue induces a different definition of the explained Gini coefficient and therefore a different value for this quantity. Facing one dataset, it would be interesting to have an idea of how small or large the coefficient can get. In this respect, our random assignment solution offers some help. If one repeats the generation of the vector of uniform variables a large number of times in the estimation of the explained Gini coefficient, the random solution sweeps through the different possibilities of attaching ranks to the ties. In this way, we internalize the variability related to the ties issue and are able to assess its influence. We formalize this procedure in the remaining of this section. Under mild assumptions, this variability is independent from the estimation of θ_0. As such, one does not need to re-estimate θ_0 each time.

In what follows, we work conditionally on having observed the data $(x_i^\intercal, y_i)_{i=1,\dots,n}^\intercal$ and suppose that ties in the index only happen if $x_i = x_j$. The corrected CDF was introduced in Equation (2.4.3). Here, we denote it by $\tilde{F}_i(\theta, V)$ to highlight the dependence on both θ and $V = (V_1, \dots, V_n)^\intercal$. We can decompose it as follows

$$\tilde{F}_i(\theta, V) = \dot{F}_i(\theta) + \ddot{F}_i(V), \tag{2.4.11}$$

where $\dot{F}_i(\theta)$ is the estimator of the CDF based on the mean-rank solution introduced in Equation (2.4.2), while $\ddot{F}_i(V)$ is given by

$$\ddot{F}_i(V) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{x_i = x_j\}\Big[\mathbb{1}\{V_i > V_j\} - \frac{1}{2}\Big].$$

The first term, $\dot{F}_i(\theta)$, is deterministic but depends on θ. The second term, $\ddot{F}_i(V)$, does not depend on θ thanks to our simplifying assumption, but it is a random variable with mean 0. This decomposition entails that we can separate out the impact of V from the impact of θ in the estimation of the CDF. This opens the door to the following procedure. Choose a number of generations M. For $m = 1, \ldots, M$, generate $V^m = (V_1^m, \ldots, V_n^m)^\intercal$, where $(V_i^m)_{i=1,\ldots,n}$ is a sample of i.i.d $U[0,1]$ random variables. For each iteration m, estimate the explained Gini coefficient with

$$\widehat{\mathrm{Gi}}_{Y,X}^m = \frac{2}{n}\sum_{i=1}^{n} \frac{y_i}{\overline{y}} \tilde{F}_i(\tilde{\theta}, V^m) - 1, \tag{2.4.12}$$

where $\tilde{\theta}$ is computed only once. We illustrate this procedure in Section 2.5.2.

2.5 Monte-Carlo simulations

This section is divided into three Monte-Carlo exercises. In the first, we implement the estimation procedure for θ_0 with a genetic algorithm and compare its performance with the WNLS estimator of Ichimura (1993) introduced in Section 1.5. In the second, we illustrate the procedure laid down in Subsection 2.4.3 to assess the impact of ties on the range of values taken by the explained Gini coefficient. Finally, we examine the performance of the bootstrap.

2.5.1 Performance of the estimation

Consider generic estimators $\hat{\theta}$ for θ_0 and $\widehat{\mathrm{Gi}}_{Y,X}$ for $\mathrm{Gi}_{Y,X}$. We evaluate the quality of the estimation via three criteria : the mean L2-distance between $\hat{\theta}$ and θ_0, the mean integrated squared error (MISE) of the regression curve and, most importantly, the mean squared error (MSE) of the explained Gini coefficient. These three quantities are respectively defined as

$$\text{Distance.}\theta \coloneqq \mathbb{E}[||\hat{\theta} - \theta_0||] \tag{2.5.1}$$

$$\text{MISE.Curve} \coloneqq \mathbb{E}\left[\int \left(H(x^\intercal\hat{\theta}) - H(x^\intercal\theta_0)\right)^2 dF_X(x)\right], \tag{2.5.2}$$

$$\text{MSE.Gini} \coloneqq \mathbb{E}\left[\left(\widehat{\mathrm{Gi}}_{Y,X} - \mathrm{Gi}_{Y,X}\right)^2\right], \tag{2.5.3}$$

where $F_X(\cdot)$ is the joint CDF of the X vector. These criteria are estimated with 1000 Monte-Carlo replications. We use the following DGP

$$Y_i = H\left(\theta_{0,1}X_i^1 + \ldots + \theta_{0,c}X_i^c + \theta_{0,c+1}Z_i^1 + \ldots + \theta_{0,c+d}Z_i^d\right)\epsilon_i, \tag{2.5.4}$$

where $i = 1, \ldots, n$ and $||\theta_0|| = 1$. For each $i = 1, \ldots, n$, the covariates $(X_i^k)_{k=1,\ldots,c}$ are i.i.d standard normal variables, while $(Z_i^k)_{k=1,\ldots,d}$ are i.i.d Bernoulli variables with probability of success equal to 0.5. In this subsection, we fix $c = 4$ and $d = 2$ and use the parameter vector

$$\theta_0 = (0.0923, -0.0889, 0.0506, 0.2838, -0.2074, 0.2270)^\intercal.$$

Finally, ϵ_i is a lognormal noise with mean 1 and a variance set to ensure a signal-to-noise ratio of 3. Throughout this section, we consider the following link function

$$H(t) = 1000 \exp\left[1 + \frac{1}{2}(t-1)^3\right].$$

Our competitor, the WNLS estimator, is available in the `R` package `np` and was introduced in Section 1.5. In the Lorenz regression, the monotonicity of the link function allows us to build an estimation procedure where only the ordering of the index matters. This corresponds to the passage from Equation (2.2.2) to Equation (2.2.3). The WNLS estimator does not rely on this monotonicity assumption so that Equation (2.2.3) no longer holds. Consequently, and contrary to our procedure, the ranks of $X^\intercal\hat{\theta}_{\text{WNLS}}$ and of $H(X^\intercal\hat{\theta}_{\text{WNLS}})$ are not identical. For the WNLS procedure, the estimator of the explained Gini coefficient is therefore based on the ranks of the fitted values, and not on those of the index.

Let us first focus on the explained Gini coefficient. The boxplot of the estimates obtained with each method is displayed in Figure 2.1. The horizontal line corresponds to the actual value of the parameter, i.e. around 23.88%. While the distribution of Lorenz estimates is centred on the true value, this is not the case for the WNLS estimator. Since the link function is not restricted to be increasing, the WNLS estimator tends to overfit the data and therefore to overestimate the explained Gini coefficient. Figure 2.2 reinforces our conclusion that the Lorenz regression provides a better estimator, even though the gap closes down with sample size.

We now compare the two estimators in terms of the other two criteria. Figures 2.3 and 2.4 show their evolution with sample size. They confirm the better performance of the Lorenz regression since the L2-distance is always lower in comparison with the WNLS estimator. A similar pattern emerges for the estimation of the regression curve.

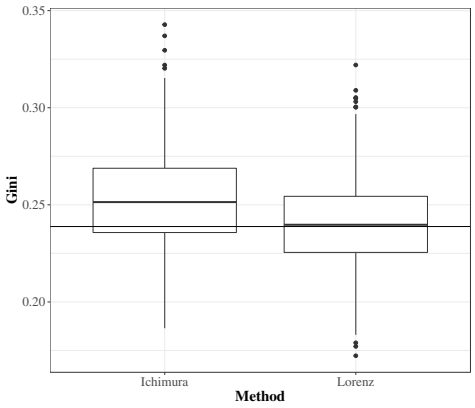

Figure 2.1: Distribution of $\widehat{\mathrm{Gi}}_{Y,X}$

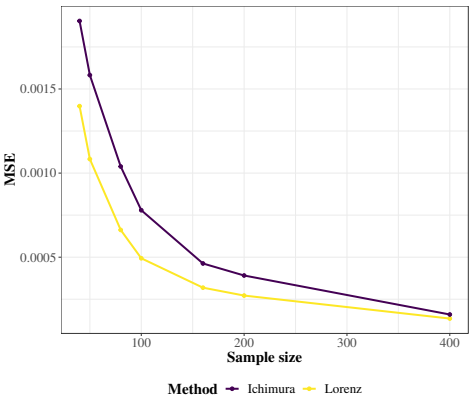

Figure 2.2: MSE of $\widehat{\mathrm{Gi}}_{Y,X}$

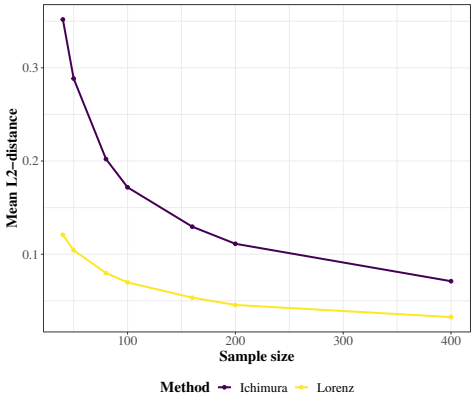

Figure 2.3: Mean L2-distance between $\hat{\theta}$ and θ_0

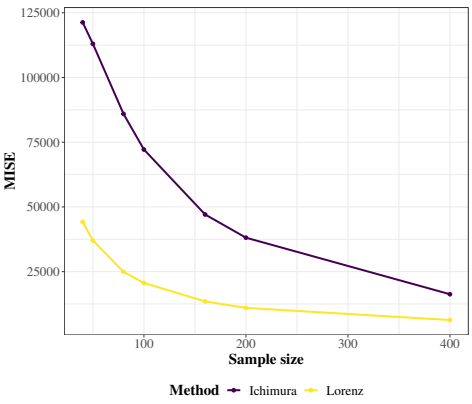

Figure 2.4: MISE of $\hat{H}(X^\intercal\hat{\theta})$

2.5.2 Range of the explained Gini coefficient in presence of ties

In this subsection, we strive first to highlight the definition problem of the explained Gini coefficient inherent to the discrete case. Second, we wish to illustrate the procedure explained at the end of Subsection 2.4.3. The point of the exercise is not to assess the estimation quality of the explained Gini coefficient. We work on one dataset, generated with the same DGP and link function as before and concentrate on a pure discrete scenario with $d = 4$. We generate $M = 1000$ uniform random vectors $V^1, \ldots, V^M$ and use these to estimate a vector of explained Gini coefficients, following Equation (2.4.12). This computation is undertaken using θ_0, i.e. the value that we used to generate the data, as well as using $\tilde{\theta}$. The two boxplots displayed in Figure 2.5 show the distribution of the explained Gini coefficient estimated with the random solution. In the first, we plug the estimated vector of parameters $\tilde{\theta}$. In the second, we rather use θ_0. The two boxplots are symmetric so that the median is close to the mean. The dashed and dotted lines represent the explained Gini coefficients obtained with the average-rank method for θ_0 and $\tilde{\theta}$ respectively. As expected, the two boxplots are centered on these lines. On average, the random method provides the same explained Gini coefficient as the average-rank solution. We also observe that the two boxplots display the same variability. This stems from the fact that the variable part of the explained Gini coefficient is independent from the value taken by the parameter vector. Finally, the total size of the boxplots indicate that the ties issue is not too severe. Indeed, their range is only slightly larger than 0.003.

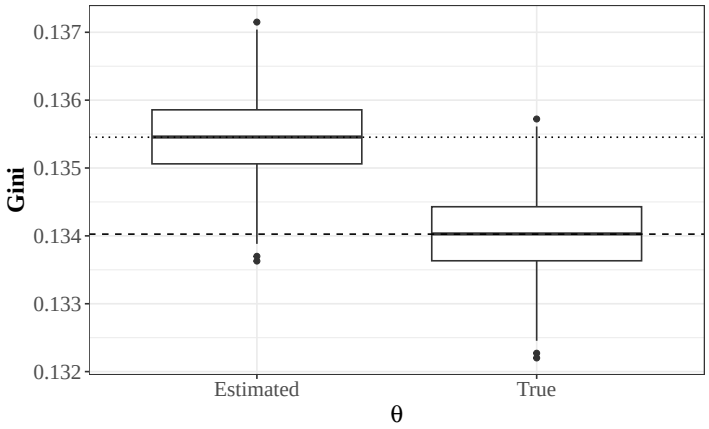

Figure 2.5: Distribution of the explained Gini coefficient for different generations of V

2.5.3 Performance of the bootstrap

We assess the performance of the bootstrap test of joint significance using the same DGP as in Section 2.5.1, again with 4 continuous and 2 discrete covariates. Even though the data are generated with a multiplicative error model, the bootstrap procedure described in Section 2.3 can still be used, by transforming the response Y into $\log(Y)$ before estimation. We test the joint significance of θ_4 and θ_5, with a significance level of $\alpha = 0.05$, and use 500 Monte-Carlo replications for the bootstrap. The vectors of parameters under the null and alternative hypotheses are given by

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
Under H_0	0.5504	0.1790	−0.1724	0	0	0.0982
Under H_1	0.44032	0.14320	−0.13792	0.1	0.1	0.07856

In what follows, we evaluate the power of our testing procedure. To do so, we generate 600 datasets using the aforementioned DGP with the parameter vector corresponding to the alternative hypothesis. We run the test of joint significance for each dataset and record the proportion of rejection. In parallel, we generate 600 datasets with the parameter vector corresponding to the null hypothesis. Based on these datasets, we approximate the distribution of the test statistic under the null hypothesis, allowing us to compute a *theoretical power curve*, isolating the power loss due to bootstrapping. Figure 2.6 displays the power curves thus obtained. As

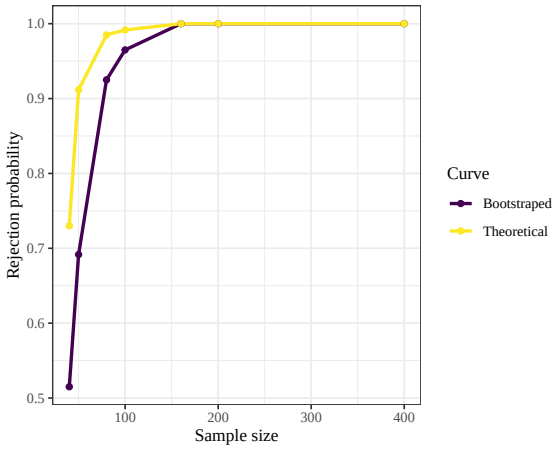

Figure 2.6: Theoretical and bootstrapped power curves

expected, both power curves converge to 1 as the sample size increases. The vertical distance between the two curves illustrate the power loss related to bootstrapping.

Though this difference is not negligible for small and medium sample sizes, it is never larger than 22% and it closes down relatively quickly. With a sample of 100 observations, our procedure already attains a power of 96.5%.

We evaluate the performance of the basic, percentile and parametric bootstrap confidence intervals with their lengths and coverages, using a confidence level of 95%. Similarly to the bootstrap test, we use 500 Monte-Carlo replications for the bootstrap and 600 replications for the estimation of the coverages and lengths. We use the same DGP as the one previously proposed for the alternative hypothesis.

The lengths of the intervals are displayed for different sample sizes in Table 2.4. The evolution of the coverages is displayed in Figure 2.7. The three methods yield intervals of similar lengths. In terms of coverage, parametric and percentile bootstrap offer the same results. They perform better than basic bootstrap in small samples while the picture is unclear for larger samples. The three methods seem to slightly undercover the true value of the parameter.

Table 2.4: Lengths of the 95% bootstrap confidence intervals for $\mathrm{Gi}_{Y,X}$

		Basic	Percentile	Parametric
Sample size	40	0.1561	0.1561	0.1566
	50	0.1419	0.1419	0.1420
	80	0.1144	0.1144	0.1146
	100	0.1040	0.1040	0.1042
	160	0.0830	0.0830	0.0832
	200	0.0748	0.0748	0.0750
	400	0.0532	0.0532	0.0533

Finally, we turn to confidence intervals on elements of the parameter vector θ_0. Without loss of generality, we focus on $\theta_{0,1}$. Results about the lengths are gathered in Table 2.5 while the evolution of the coverages with sample size is displayed in Figure 2.8. From these, we can conclude that the three methods yield intervals of similar lengths but basic bootstrap seems to undercover the true parameter.

2.6 Empirical illustration

In economics, attention has often been directed to identify the main determinants of wages. A classical starting point is the wage equation proposed by Mincer and Polachek (1974) of the form

$$\log(W) = \beta_0 + \beta_1 S + \beta_2 E + \beta_3 E^2 + \epsilon, \tag{2.6.1}$$

where W is the wage, S is schooling, E is professional experience and ϵ is an unobservable error term. Focusing on these two covariates is of course restrictive and

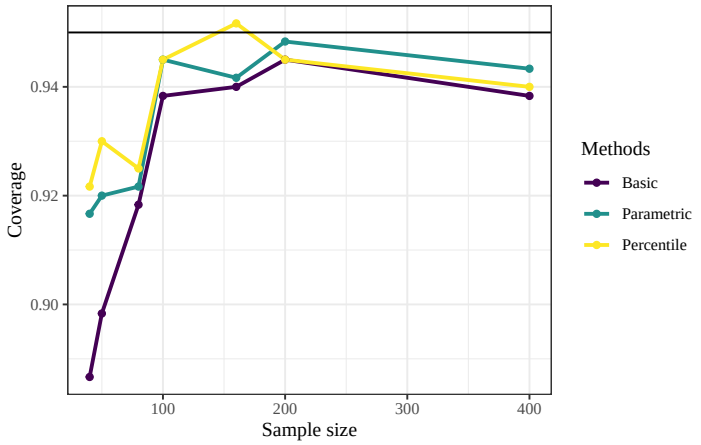

Figure 2.7: Coverages of the 95% bootstrap confidence intervals for $\text{Gi}_{Y,X}$

Table 2.5: Lengths of the 95% bootstrap confidence intervals for $\theta_{0,1}$

		Basic	Percentile	Parametric
Sample size	40	0.1912	0.1912	0.1893
	50	0.1630	0.1630	0.1618
	80	0.1216	0.1216	0.1211
	100	0.1080	0.1080	0.1075
	160	0.0797	0.0797	0.0793
	200	0.0698	0.0698	0.0696
	400	0.0472	0.0472	0.0472

many papers went beyond this basic setup. For example, Griliches (1976) drew attention on a bias which stems from ignoring ability while it is at the same time expected to be influenced by schooling and to have an effect on wages. In this section, we apply the Lorenz regression methodology on these determinants with the objective of visualizing the inequality that they help to reproduce.

Our discussion is based on data resulting from the Young Men's Cohort of the National Longitudinal Survey (NLS-Y), a survey started in 1966 on individuals of ages 14-24. The excerpt we use is available in the dataset `Griliches` contained in the R package **Ecdat** (Croissant and Graves, 2020). Besides wage, schooling and experience, we include the following variables in our analysis: age (A), ability (IQ), marital status (MS) and degree of urbanisation of the area where the subject lives (DU). Ability was computed as IQ scores collected in a school survey conducted in 1968. All the remaining variables were observed in 1980. The degree of urbanisation is a dummy determining whether the individual lives in a

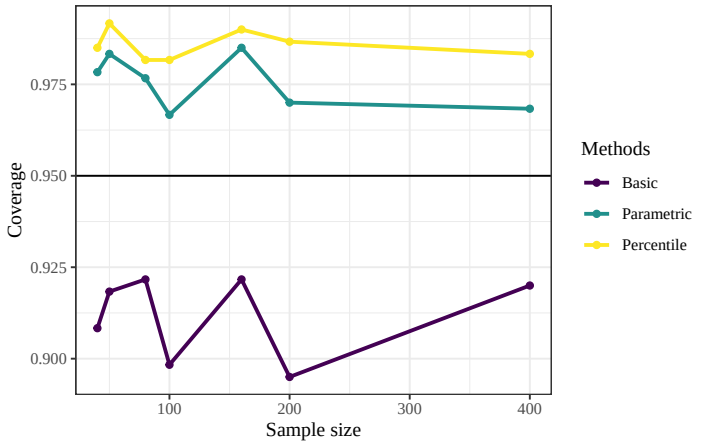

Figure 2.8: Coverages of the 95% bootstrap confidence intervals for θ_1

metropolitan area. Before delving into modelling, we note that the considered wage distribution exhibits a mean of 1000, a median of 948 and a Gini coefficient of 0.222.

We first restrict the attention to schooling and experience, allowing for a quadratic term on experience. The underlying models are presented in (2.6.2) and (2.6.3) and are more general than Mincer and Polachek (1974) equation since they do not impose a log link function.

$$\mathbb{E}[W|S,E] = H_{[1]}\left(\theta_S S + \theta_E E + \theta_{E^2} E^2\right), \tag{2.6.2}$$

$$\mathbb{E}[W|S,E] = H_{[2]}\left(\theta_S S + \theta_E E\right). \tag{2.6.3}$$

Table 2.6 displays the estimated parameters, bootstrap standard deviations and

Table 2.6: Lorenz regressions for (2.6.2) and (2.6.3)

	(2.6.2)			(2.6.3)		
	Estimate	Std dev	CI	Estimate	Std dev	CI
θ_S	0.552	0.133	[0.398;0.912]	0.786	0.055	[0.733;0.908]
θ_E	0.437	0.237	[-0.279;0.584]	0.214	0.055	[0.092;0.267]
θ_{E^2}	-0.011	0.011	[-0.018;0.020]			

95% percentile bootstrap confidence intervals corresponding to Equations (2.6.2) and (2.6.3). We observe that only schooling is significant when a quadratic experience term is included. However, experience becomes significant when the quadratic

term is dropped. As expected, the Lorenz-R^2 decreases very slightly from (2.6.2) to (2.6.3), i.e. from 0.391 to 0.389.

As a second step, we augment our model with several control variables. In Equation (2.6.4), we introduce marital status, degree of urbanisation, age and ability, as measured by IQ.

$$\mathbb{E}[W|S, MS, DU, A, E, IQ] = H_{[4]}(\theta_S S + \theta_{MS} MS + \theta_{DU} DU + \theta_A A + \theta_E E + \theta_{IQ} IQ). \tag{2.6.4}$$

Table 2.7 displays the output of the Lorenz regression corresponding to Equation

Table 2.7: Lorenz and log-linear regressions for (2.6.4)

	Lorenz regression			Log-linear regression		
	Estimate	Standard error	CI	Estimate	Standard error	CI
Intercept	/	/	/	4.689	0.182	[4.332;5.047]
θ_S	0.123	0.030	[0.086;0.206]	0.054	0.008	[0.038;0.069]
θ_{MS}	0.372	0.117	[0.034;0.490]	0.169	0.044	[0.082;0.255]
θ_{DU}	0.429	0.097	[0.344;0.686]	0.202	0.029	[0.145;0.260]
θ_A	0.029	0.021	[0.002;0.085]	0.014	0.005	[0.004;0.024]
θ_E	0.037	0.016	[0.008;0.067]	0.016	0.004	[0.008;0.024]
θ_{IQ}	0.01	0.005	[0.005;0.024]	0.004	0.001	[0.002;0.007]

(2.6.4), as well as the results of a linear regression on log-wages, using the same covariates. Both methods offer similar conclusions. However, the Lorenz regression provides us with further interpretations. The Lorenz-R^2 is of 0.48, indicating that we can reproduce 48% of the observed inequality with our covariates. Figure 2.9 displays the observed and explained Lorenz curves. By computing marginal rates of substitution, we can also compare the magnitude of two coefficients. For example, the MRS of DU with respect to MS is of 1.153 meaning that, all other things being equal, the degree of urbanisation has a marginal impact on wages 1.153 times more important than marital status. Also, the MRS of S with respect to E is of 3.324, meaning that, all other things being equal, spending one more year of schooling has three-times more impact than a further year of professional experience.

Interestingly, our dataset contains a variable recording the individual's residency. Namely, we know if he or she lives in a Southern or Northern state. Dividing our dataset on the basis of this information and using model (2.6.4), we fit two

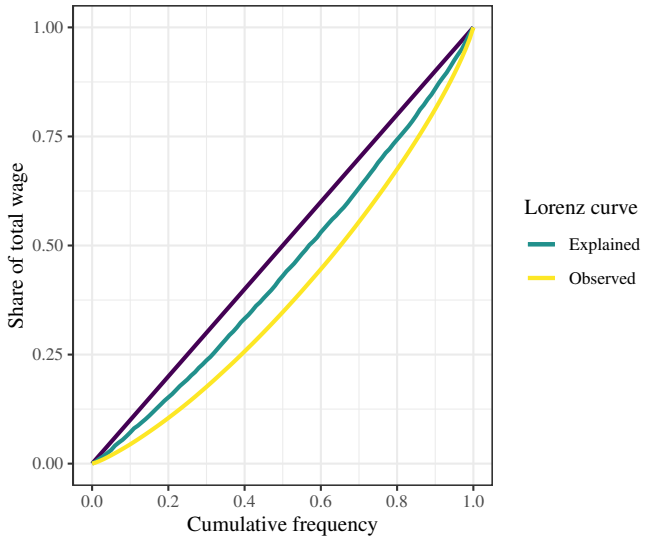

Figure 2.9: Observed and explained Lorenz curve

distinct Lorenz regressions. This allows us to compare the explained Gini coefficient in Northern and Southern states. Results are gathered in Table 2.8. The

Table 2.8: Lorenz regressions distinguishing between Northern and Southern states

	South	North		South	North
θ_S	0.212	0.097	Lorenz-R^2	0.535	0.4753
θ_{MS}	0.255	0.253	$\widehat{\text{Gi}}_Y$	0.2517	0.2096
θ_{DU}	0.431	0.570	$\widehat{\text{Gi}}_{Y,X}$	0.1347	0.0996
θ_A	0.003	0.056	$\widehat{\text{Gi}}_Y - \widehat{\text{Gi}}_{Y,X}$	0.117	0.11
θ_E	0.094	0.011			
θ_{IQ}	0.006	0.013			

Gini coefficient obtained when ranking individuals in terms of explained income is higher in Southern states (13.47% against 9.96% in Northern states. In some sense, this was expected since the Gini coefficient of wages is higher in Southern states

(25.17% against 20.96% in Northern states). Still, it is worth mentionning that the unexplained part of inequality, i.e. the difference between the actual and the explained Gini coefficients, is rather similar between the two situations (11.17% in Southern states against 11% in Northern states). This stems from the fact that the model explains a larger proportion of the observed inequality in the Northern states. Indeed, the Lorenz-R^2 there is of 53.5% against 47.53% in Southern states.

Having estimated the index in Equation (2.6.4), we may estimate the link function using the rearrangement method and compare it with an exponential fit. The latter is obtained from a classical linear regression of log-wages on the estimated index. More specifically, we assume $\log(W) = \alpha + \beta T + \varepsilon$, where T is the estimated index, $\mathbb{E}[\varepsilon|T] = 0$ and ε is independent from T. Denote by $\hat{\alpha}$ and $\hat{\beta}$ the OLS estimators of α and β. We have $\mathbb{E}[W|T] = \exp(\alpha + \beta T)\mathbb{E}[\exp(\varepsilon)]$. The first piece is estimated using $\exp(\hat{\alpha} + \hat{\beta}T)$ while the second is estimated by the empirical mean of the $W/\exp(\hat{\alpha} + \hat{\beta}T)$ vector. As we can observe in Figure 2.10, both curves exhibit the same behaviour.

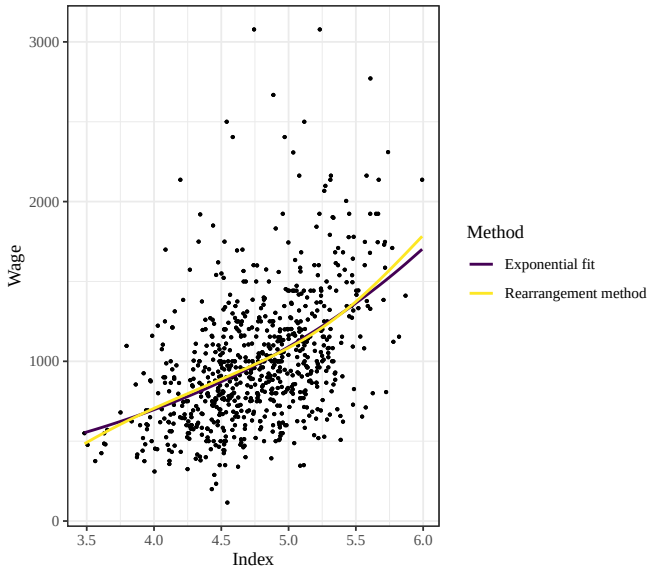

Figure 2.10: Nonparametric and parametric fit of the wage equation

The purpose of this empirical example was to provide a first illustration of the Lorenz regression and the interpretations that can be build upon. In this case, this methodology yields quite similar results to a linear regression on log-wages. This provides some evidence supporting the logarithm as link function in the wage

equation proposed by Mincer and Polachek (1974).

2.7 Discussion

The explained Gini coefficient introduced in this chapter provides information on the inequality of a response variable that a set of covariates helps reproducing. This measure benefits from several advantages. It is straightforward to interpret as it corresponds to the Gini coefficient of the conditional expectation of the response given the covariates. Also, it rests on a semiparametric single-index model, which allows for dimension reduction without sacrificing too much in flexibility. Finally, only the parametric part of the model needs to be estimated. Indeed, the explained Gini coefficient can be obtained as the solution of a maximization programme where the nonparametric link function does not intervene.

The inference procedure proposed in Section 2.3, called Lorenz regression, builds on this last point. The estimator obtained for the parameter vector of the single-index model is a special case of the monotone rank estimator of Cavanagh and Sherman (1998) and is therefore consistent and asymptotically normal. Since its covariance matrix can be difficult to estimate, we suggest to produce inference using bootstrap. We also propose a test of joint significance for several covariates. The test statistic involves the Lorenz-R^2, which measures the proportion of inequality explained by the covariates. In practice, the estimates can be numerically obtained using a genetic algorithm. In the simulation analysis performed in Section 2.5, the Lorenz regression performs at least as good as the proposed competitor, i.e. the WNLS estimator of Ichimura (1993).

The situation where all covariates are discrete is problematic, both in theory and in practice. In theory, because it breaks the identifiability of the single-index model and, in practice, because it induces ties in the index. Section 2.4 tackles these two issues. The empirical CDF is replaced with a corrected version, which breaks the ties on a random basis. This solution is compared to the mean-rank assignment. Theoretically, we show that the explained Gini coefficient is still identifiable and is consistently estimated with the corrected estimation procedure.

Two drawbacks arise from the form of the maximization programme underlying the Lorenz regression. First, it produces an estimator of the explained Gini coefficient that never decreases as one keeps introducing new covariates. Performing an estimation on many covariates may therefore lead to an overestimation of the explained inequality. Second, the objective function being discrete, we proposed a genetic algorithm to perform the optimization numerically. When the number of covariates is large, the algorithm may fail to converge. The procedure introduced in Chapter 3 tackles these problems by adapting the Lorenz regression in two ways. First, the discrete objective function is replaced with a smooth approximation. Second, a SCAD penalty is added to ensure sparsity.

2.8 Proofs

Proof of Proposition 2.4

We need to show that $P(\tilde{F}(Z) \leq q) = q$ for $q \in [0,1]$. Let $F^{-1}(q) \coloneqq \inf\{z \in \mathbb{R} : F(z) \geq q\}$. Obviously, we can restrict the proof to the region included between the first and last points of discontinuity of $F(\cdot)$. Three situations need to be covered.

(i) *q is located in the middle of the jump*, i.e. $F(F^{-1}(q)^-) < q < F(F^{-1}(q))$. This situation is treated first.

(ii) *q is located at the end of the jump*, i.e. $F(F^{-1}(q)) = q$. This situation is treated second.

(iii) *q is located at the start of the jump*, i.e. $F(F^{-1}(q)^-) = q$. This case can be treated in the exact same way as (ii) and is therefore not covered.

Let us start with situation (i). We have

$$P(\tilde{F}(Z) \leq q) = P\Big(\tilde{F}(Z) \leq F(F^{-1}(q)^-)\Big) + P\Big(F(F^{-1}(q)^-) < \tilde{F}(Z) \leq q\Big).$$

Regardless of the value of V, $\tilde{F}(Z) \leq F(F^{-1}(q)^-)$ occurs when $Z < F^{-1}(q)$. Hence, the first piece boils down to $F(F^{-1}(q)^-)$. After conditioning on $Z = F^{-1}(q)$ and using the definition of $\tilde{F}(\cdot)$, the second piece becomes

$$\begin{aligned} &P\Big(F(F^{-1}(q)^-) < F(F^{-1}(q)^-) + V[F(F^{-1}(q)) - F(F^{-1}(q)^-)] \leq q \Big| Z = F^{-1}(q)\Big) \\ &\qquad \times P(Z = F^{-1}(q)) \\ &= P\left(0 < V \leq \frac{q - F(F^{-1}(q)^-)}{F(F^{-1}(q)) - F(F^{-1}(q)^-)} \middle| Z = F^{-1}(q)\right) \\ &\qquad \times (F(F^{-1}(q)) - F(F^{-1}(q)^-)). \end{aligned}$$

Using the independence and uniformity of V, we obtain the desired result.

We now move to situation (ii). Let z_1 and z_2 respectively be the first and second points of discontinuity of F. Suppose $z_1 < F^{-1}(q) < z_2$. Then, it holds that

$$\begin{aligned} P(\tilde{F}(Z) \leq q) &= P(F(Z^-) + V(F(Z) - F(Z^-)) \leq q) \\ &= \int_{-\infty}^{z_1} dF(z^-) + P(F(z_1^-) \leq F(Z^-) + V(F(Z) - F(Z^-)) \leq F(z_1)) \\ &\qquad + \int_{z_1}^{F^{-1}(q)} dF(z). \end{aligned}$$

Besides,

$$\begin{aligned}
&P(F(z_1^-) \leq F(Z^-) + V(F(Z) - F(Z^-)) \leq F(z_1)) \\
&= P(F(z_1^-) \leq F(z_1^-) + V(F(z_1) - F(z_1^-)) \leq F(z_1) | Z = z_1) P(Z = z_1) \\
&= P(0 \leq V \leq 1 | Z = z_1)(F(z_1) - F(z_1^-)) = F(z_1) - F(z_1^-).
\end{aligned}$$

Hence,

$$P(\tilde{F}(Z) \leq q) = F(z_1^-) + F(z_1) - F(z_1^-) + F(F^{-1}(q)) - F(z_1) = F(F^{-1}(q)) = q.$$

By induction, this is true for all q such that $F(F^{-1}(q)) = q$. □

Proof of Proposition 2.5

We start by rewriting the objective function $\tilde{G}(\theta)$. Let $(X_1^\intercal, Y_1)^\intercal$ and $(X_2^\intercal, Y_2)^\intercal$ be two i.i.d copies of $(X^\intercal, Y)^\intercal$. As mentioned earlier, it holds that

$$\tilde{G}(\theta) = \mathbb{E}[Y_1 \mathbb{1}\{X_1^\intercal \theta > X_2^\intercal \theta\}] + \frac{1}{2}\mathbb{E}[Y_1 \mathbb{1}\{X_1^\intercal \theta = X_2^\intercal \theta\}].$$

Using the law of total expectation, we have

$$\begin{aligned}
2\tilde{G}(\theta) = \mathbb{E}[H(X_1^\intercal \theta_0)\mathbb{1}\{X_1^\intercal \theta > X_2^\intercal \theta\}] + \mathbb{E}[H(X_2^\intercal \theta_0)\mathbb{1}\{X_2^\intercal \theta > X_1^\intercal \theta\}] \\
+ E\left[\frac{H(X_1^\intercal \theta_0) + H(X_2^\intercal \theta_0)}{2}\mathbb{1}\{X_1^\intercal \theta = X_2^\intercal \theta\}\right].
\end{aligned}$$

If $p_j \coloneqq P(X = x_j)$ and $p_k \coloneqq P(X = x_k)$, we have

$$2\tilde{G}(\theta) = \sum_{j,k} \tilde{G}_{j,k}(\theta) p_j p_k, \tag{2.8.1}$$

where

$$\begin{aligned}
\tilde{G}_{j,k}(\theta) \coloneqq \Bigg[H(x_j^\intercal \theta_0)\mathbb{1}\{x_j^\intercal \theta > x_k^\intercal \theta\} + H(x_k^\intercal \theta_0)\mathbb{1}\{x_k^\intercal \theta > x_j^\intercal \theta\} \\
+ \frac{H(x_j^\intercal \theta_0) + H(x_k^\intercal \theta_0)}{2}\mathbb{1}\{x_j^\intercal \theta = x_k^\intercal \theta\}\Bigg]
\end{aligned}$$

We now strive to prove part (i). Consider θ in the same region as θ_0. Let

$$\begin{aligned}
\mathcal{I}_1 &\coloneqq \{j, k : x_j^\intercal \theta_0 > x_k^\intercal \theta_0 \text{ or } x_k^\intercal \theta > x_j^\intercal \theta\} \\
\mathcal{I}_2 &\coloneqq \{j, k : x_j^\intercal \theta_0 = x_k^\intercal \theta_0.
\end{aligned}$$

Start with $\{j, k\} \in \mathcal{I}_1$. Because θ is in the same region as θ_0, the two parameter vectors yield the same ordering and $G_{j,k}(\theta) - G_{j,k}(\theta_0) = 0$. Consider now

$\{j,k\} \in \mathcal{I}_1$. Since $x_j^\intercal \theta_0 = x_k^\intercal \theta_0$, whether one chooses $H(x_j^\intercal \theta_0)$, $H(x_k^\intercal \theta_0)$, or the average of the two yields the same value. Hence, we have $G_{j,k}(\theta) - G_{j,k}(\theta_0) = 0$ again, which concludes the proof of the first part.

We now turn to the proof of part (ii). From Equation (2.8.1) and by noticing that $H(x_j^\intercal \theta_0) > H(x_k^\intercal \theta_0)$ whenever $x_j^\intercal \theta_0 > x_k^\intercal \theta_0$, it is clear that $\tilde{G}(\theta)$ is maximized in θ_0^*. In order to show that the maximum is unique, suppose per contra that there exists $\theta_1^* \neq \theta_0^*$ which also maximizes $\tilde{G}(\theta)$. Then, there must exist at least one pair (x_j, x_k) such that

$$x_j^\intercal \theta_0^* > x_k^\intercal \theta_0^* \text{ and } x_j^\intercal \theta_1^* < x_k^\intercal \theta_1^* \qquad \text{if } \theta_1^* \in \{B_1, \ldots, B_l\}, \tag{2.8.2}$$

$$\text{or } x_j^\intercal \theta_1^* = x_k^\intercal \theta_1^* \qquad \text{if } \theta_1^* \in \{H_1, \ldots, H_m\}. \tag{2.8.3}$$

Since $x_j^\intercal \theta_0^* > x_k^\intercal \theta_0^*$, we have $H(x_j^\intercal \theta_0^*) > H(x_k^\intercal \theta_0^*)$. Let us now examine the contribution of (x_j, x_k) to Equation (2.8.1) for θ_0^* and for θ_1^*. It consists of $H(x_j^\intercal \theta_0^*)$ for θ_0^*. For θ_1^*, it amounts either to $H(x_k^\intercal \theta_1^*)$ in the situation described by (2.8.2), or to $\frac{H(x_j^\intercal \theta_1^*) + H(x_k^\intercal \theta_1^*)}{2}$ in (2.8.3). In any case, this contradicts the fact that $\tilde{G}(\theta)$ is maximized in θ_1. □

CHAPTER 3

The penalized Lorenz regression

The explained Gini coefficient is a measure of inequality of the conditional expectation of a response given a vector of covariates, assuming a single-index model. An estimator for this quantity was proposed in Chapter 2. However, it is prone to overestimation when many covariates are included. In this chapter, we propose a penalized bootstrap procedure which selects the relevant covariates and produces valid inference for the explained Gini coefficient. The obtained estimator achieves the oracle property. Numerically, it is computed by the SCAD-FABS algorithm, an adaptation of the FABS algorithm to the SCAD penalty. The performance of the procedure is ensured by theoretical guarantees and assessed via Monte-Carlo simulations. Finally, a real data example is presented. The content of this chapter is based on the manuscripts

> Alexandre Jacquemain, Cédric Heuchenne, and Eugen Pircalabelu. A lasso-type estimation for the lorenz regression. In *the Proceedings of the 22th European Young Statisticians Meeting*, pages 41–45, 2021.
>
> Alexandre Jacquemain, Cédric Heuchenne, and Eugen Pircalabelu. A penalised bootstrap estimation procedure for the explained Gini coefficient, 2022a.

3.1 Introduction

Facing large micro-data covering hundreds of covariates, the Lorenz regression methodology introduced in Chapter 2 may face several issues. First, the objective function in Equation (2.3.2) is non-differentiable, which complicates the numerical solution. A genetic algorithm was proposed in Section 2.3 but its performance may be altered in datasets of large dimension. Another issue is overfitting. In large datasets we might expect some covariates to be irrelevant but, at the same time,

to bear some empirical correlation with the response. Much like the R^2 in linear regression, the estimated explained Gini coefficient never decreases as one keeps introducing new covariates. Facing a sparse model, where some of the elements of θ_0 are equal to 0, an estimation of the full model will lead us to overestimate the explained Gini coefficient.

These concerns echo the justifications underlying the use of penalized regressions, introduced in Section 1.6. Indeed, these procedures provide proper inference of the model parameters and automatic selection of the covariates, even in a situation where the number of covariates is large. Procedures based on the SCAD penalty ensure a strong statistical guarantee: the oracle property of the estimator. We refer the reader to Fan and Li (2001) for the original procedure and to Lin and Peng (2013) for an adaptation to the single-index model.

In this chapter, we present a penalized Lorenz regression combined with a pairs bootstrap procedure. The proposed methodology presents several advantages. First, it makes it possible to include many covariates without inducing an overestimation of the explained Gini coefficient. Second, the penalization leads to a selection of the relevant covariates. Finally, the pairs bootstrap allows us to construct confidence intervals for the explained Gini coefficient without having to estimate the link function $H(\cdot)$ of the single-index model, see Equation (2.1.1) for the definition of the model. The proposed procedure comes with statistical and numerical guarantees. Through the use of a SCAD penalty, the estimator achieves the oracle property. On the numerical side, we adapt the FABS algorithm proposed by Shi et al. (2018) to a SCAD penalty. Hence, we benefit from a procedure enjoying good theoretical properties as well as a fast and efficient algorithm to obtain estimates. From the viewpoint of the estimation of a single-index model, the most obvious competitor is the penalized maximum smoothing rank correlation (PMSRC) estimator introduced in Equation (1.6.4). Compared to this procedure, the advantage of the penalized Lorenz regression consists in the fact that it uses more information at the level of the response. It is therefore expected to provide a better tradeoff between flexibility and efficiency. This point is confirmed by the favourable simulation results displayed in Subsection 3.4.1.

As we already motivated in Section 2.2, the Lorenz regression methodology has a natural application in the context of equality of opportunity. In the so-called ex-ante utilitarian approach, constructing a measure of IOP is characterized by a two-stage nature. First, the advantage variable is fitted in an econometric model where the economic advantage is the response variable and circumstances are covariates. Second, the estimated model is used in combination with a measure of inequality in order to evaluate the extent of IOP. Blending these two stages harmoniously remains an issue in the literature. A recent approach measures IOP with the Gini coefficient of fitted values obtained via machine learning methods, see for

example Brunori et al. (2018). Interestingly, this method produces an automatic selection of the relevant circumstances and does not rest on a restrictive parametric model. However, as discussed in Escanciano and Terschuur (2022), it is prone to lead to a biased estimation of IOP due to the absence of robustness of the Gini coefficient with respect to the derivation of the fitted values. Also, the method comes with no inferential procedure. Escanciano and Terschuur (2022) propose a debiased estimator and a valid inferential procedure based on orthogonal moments. In practice, the debiased estimator boils down to the empirical concentration index of the economic advantage with respect to the fitted values. This discussion connects with the chain of equations presented in Proposition 2.1. Instead of constructing an estimator of $\mathrm{Gi}_{Y,X}$ based on its original definition, i.e. Equation (2.2.1), Escanciano and Terschuur (2022) propose to exploit Equation (2.2.2) so that the fitted values are only used to rank observations. Using the assumption of the single-index model, our estimation goes one step further since the ranking is based on the index and the fitted values do not need to be obtained. This finding unveils a new advantage of our procedure. In the assumed single-index model, the estimated explained Gini coefficient corresponds to the debiased estimator proposed by Escanciano and Terschuur (2022).

This paper is organized as follows. In Section 3.2, we present the penalized bootstrap Lorenz regression and provide asymptotic results for the estimated parameter vector and for the explained Gini coefficient. We also discuss the selection of the regularization parameter. The SCAD-FABS algorithm is presented in Section 3.3, along with convergence properties. Similarly to the FABS algorithm, we show that each solution along the SCAD-FABS path is a δ-approximate solution to the penalized programme from Equation (3.2.1). Simulation results assessing the performance of the procedure are displayed in Section 3.4. We confront the penalized bootstrap Lorenz regression to real data in Section 3.5 and a discussion on the method is provided in Section 3.6. The proofs of the results exposed in this chapter are gathered in Section 3.7.

3.2 The penalized bootstrap Lorenz regression

This section develops as follows. First, we introduce the penalized Lorenz regression programme and provide asymptotic results for the estimated parameter vector and the explained Gini coefficient. Second, we turn to the bootstrap procedure and the choice of the regularization parameter.

Given n i.i.d samples $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$, the penalized Lorenz regression solves the

following optimization programme

$$\hat{\theta} \coloneqq \underset{\theta,||\theta||=1}{\arg\max} \left\{ G_n(\theta) - \sum_{k=1}^{p} p_\lambda(|\theta_k|) \right\}, \tag{3.2.1}$$

where $p_\lambda(\cdot)$ is a nonconcave penalty function and $\lambda > 0$ is a penalty parameter. The non-penalized objective function is a smooth approximation of the objective function displayed in Equation (2.3.2). It is given by

$$G_n(\theta) \coloneqq \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i K\left(\frac{X_i^\intercal \theta - X_j^\intercal \theta}{h} \right), \tag{3.2.2}$$

where $K(\cdot)$ is the integral of a kernel function and h is its corresponding bandwidth. In essence, Equation (3.2.1) is an adaptation of the programme characterizing the penalized maximum smoothing rank correlation (PMSRC) estimator proposed by Lin and Peng (2013) and introduced in Equation (1.6.4). The difference lies in the original objective function. While the PMSRC is an adaptation of the maximum rank correlation estimator introduced by Han (1987), our estimator is related to the monotone rank estimator proposed by Cavanagh and Sherman (1998). In both cases, the idea consists in combining smoothing techniques and penalization with the double advantage of facilitating the estimation procedure and avoiding overfitting.

While the main results of this section allow for a general penalty, we will develop an algorithm and implement our method with the SCAD penalty, defined in Equation (1.6.3). Without loss of generality, we order the covariates such that $\{1, \ldots, s\}$ are active and $\{s+1, \ldots, p\}$ are non-active. Let $\theta_0 = (\theta_0^{A\intercal}, \theta_0^{I\intercal})^\intercal$ be the unknown parameter vector. The vector θ_0^A is related to the s active covariates, while $\theta_0^I = 0$ corresponds to the non-active part. In Theorem 3.2, it will also be useful to write $\vartheta_0 = (\theta_{0,1}, \ldots, \theta_{0,s-1})^\intercal$, i.e. $\theta_0^A = (\vartheta_0^\intercal, \theta_{0,s})^\intercal$.

Call $f(\cdot)$ the joint density of X and $g(\cdot)$ the density of $Z = X^\intercal \theta_0$. Also, denote by $F_{Y,X}(\cdot)$ and $F_{Y,Z}(\cdot)$ the joint distribution functions of $(Y, X^\intercal)^\intercal$ and $(Y, Z)^\intercal$ respectively. We will use the following regularity conditions in order to prove the main results.

(RC1) X is bounded with compact support $\mathcal{X}$. Also, we denote the compact support of Z by $\mathcal{Z}$.

(RC2) $H(X^\intercal \theta_0)$ has finite first and second moment and $H(\cdot)$ is twice continuously differentiable.

(RC3) $g(\cdot)$ is twice continuously differentiable. Also, for each $l = 1, \ldots, p$ and $m < l$, the joint density of $(X^l, Z)^\intercal$ is four-times continuously differentiable and the joint density of $(X^l, X^m, Z)^\intercal$ is three-times continuously differentiable.

(RC4) $\kappa(\cdot)$ is a density function with compact support $[-1,1]$ such that $\kappa(-1) = \kappa(1) = 0$, and satisfying $\int v^l \kappa(v)dv = 0$ for $l = 1, 2$ and $\int v^q \kappa(v)dv \neq 0$ for some $q \geq 3$. Also, $\kappa(\cdot)$ is twice continuously differentiable. Denote by $K(\cdot)$ the CDF related to $\kappa(\cdot)$.

(RC5) $nh^6 \to 0$ and $nh^5(\log(h^{-1}))^{-1} \to \infty$.

(RC6) $p_\lambda(\cdot)$ has piecewise continuous second derivatives and bounded third derivatives.

(RC7) $\mathbb{E}[YF_\theta(X^\intercal\theta)]$ has a unique maximum with respect to θ. Also, the second derivative of the function $\theta \mapsto F_\theta(x^\intercal\theta)$ is continuous with respect to x and θ.

(RC1)-(RC5) are standard conditions to obtain strong consistency results for a kernel density estimator and its first and second derivatives, and for the numerator of a Nadaraya-Watson estimator, as in the works of Silverman (1978), Mack and Silverman (1982) and Mack and Müller (1989). (RC6) is necessary to perform Taylor expansions of the objective function displayed in (3.2.1) and is satisfied for the SCAD. Finally, (RC7) is needed to ensure the local consistency of $\hat{\theta}$. As already stated, the first part of the assumption is proven under mild conditions in Theorem 1 from Cavanagh and Sherman (1998).

Theorems 3.1 and 3.2 are adaptations of Theorems 1 and 2 from Lin and Peng (2013) to our context. Their proofs are deferred to Subsection 3.7.1.

Theorem 3.1. *Let* $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$ *be i.i.d random vectors satisfying Equation* (2.1.1) *with* $H(\cdot)$ *strictly increasing,* $||\theta_0|| = 1$, *and* $\mathbb{E}[Y_i^2] < \infty$. *If* $\max_k\{p''_\lambda(|\theta_{0,k}|)\} \to 0$ *and (RC1)-(RC7) hold, then there exists a local maximizer* $\hat{\theta}$ *of Equation* (3.2.1) *such that*

$$||\hat{\theta} - \theta_0|| = O_p(n^{-1/2} + a_n),$$

where $a_n := \max\{p'_\lambda(|\theta_{0,k}|) : \theta_{0,k} \neq 0\}$.

With the SCAD penalty, $a_n = 0$ with a proper choice of λ. This allows us to obtain a $\sqrt{n}$-consistent estimator for the coefficient vector θ_0. This is an advantage over the LASSO penalty, where the convergence rate is $O_p(n^{-1/2} + \lambda)$. Define

$$b := [p'_\lambda(|\theta_{0,1}|)\text{sign}(\theta_{0,1}), \dots, p'_\lambda(|\theta_{0,s-1}|)\text{sign}(\theta_{0,s-1})]^\intercal - \frac{p'_\lambda(\theta_{0,s})}{\theta_{0,s}}\vartheta_0 \tag{3.2.3}$$

Theorem 3.2. *Let* $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$ *be i.i.d random vectors satisfying Equation* (2.1.1) *with* $H(\cdot)$ *strictly increasing,* $||\theta_0|| = 1$, $\theta_{0,s} > 0$ *and* $\mathbb{E}[Y_i^2] < \infty$. *Assume that* $a_n = 0$ *and*

$$\liminf_{n\to\infty} \liminf_{x\to 0^+} p'_\lambda(x)/\lambda > 0.$$

If $\lambda \to 0$ *and* $\sqrt{n}\lambda \to \infty$ *as* $n \to \infty$*, and (RC1)-(RC7) hold, then the* $\sqrt{n}$*-consistent local maximizers* $\hat{\theta} = (\hat{\theta}^{A\intercal}, \hat{\theta}^{I\intercal})^\intercal$ *of Theorem 3.1, where* $\hat{\theta}^A = (\hat{\vartheta}^\intercal, \hat{\theta}_s)^\intercal$ *estimates* θ_0^A *and* $\hat{\theta}^I$ *estimates* θ_0^I*, satisfy*

1. *Sparsity :* $P(\hat{\theta}^I = 0) \to 1$*, as* $n \to \infty$*.*

2. *Asymptotic normality :* $\sqrt{n}[\hat{\vartheta} - \vartheta_0 + \Sigma^{-1}b] \xrightarrow{d} N(0, \Sigma^{-1}\Omega(\Sigma^{-1})^\intercal)$*, where* Σ *is an invertible* $(s-1)\times(s-1)$ *matrix defined in* (3.7.11)*,* b *is a vector of dimension* $s-1$ *defined in* (3.2.3) *and* Ω *is an* $(s-1)\times(s-1)$ *matrix defined in* (3.7.12)*.*

With the SCAD penalty, as $\lambda \to 0$ and in light of Equation (1.6.3), each component of the vector b is set to 0. In that case, the penalized Lorenz regression enjoys two major properties. The first is the property of sparsity, which indicates that the estimated coefficients attached to the non-active covariates are set to zero with a probability tending to one. The second is the asymptotic normality of the estimated parameter vector related to the active covariates. The convergence rate is the same as what one would obtain with the monotone rank estimator computed on the active set of covariates. The selection process does not lead to any loss of efficiency. Hence, our estimator enjoys the oracle property discussed in Section 1.6. Finally, notice that due to the constraint $||\hat{\theta}||= 1$, the last active covariate can be viewed as a function of the others, namely $\hat{\theta}_s = \sqrt{1 - \hat{\vartheta}^\intercal\hat{\vartheta}}$. As a consequence, the asymptotic normality is obtained on $\hat{\vartheta}$ rather than on $\hat{\theta}^A$.

An estimator $\widehat{\mathrm{Gi}}_{Y,X}$ for the explained Gini coefficient is obtained by plugging $\hat{\theta}$ into Equation (2.3.3), where $\hat{\theta}$ is the $\sqrt{n}$-consistent estimator of Theorem 3.2. Theorem 3.3 establishes the asymptotic normality of $\widehat{\mathrm{Gi}}_{Y,X}$. Its proof is deferred to Subsection 3.7.1.

Theorem 3.3. *Let* $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$ *be i.i.d random vectors with* $\mathbb{E}[Y_i] > 0$*. If the conditions of Theorem 3.2 hold, then* $\sqrt{n}[\widehat{\mathrm{Gi}}_{Y,X} - \mathrm{Gi}_{Y,X}] \xrightarrow{d} N(0, \sigma_\zeta^2)$*, where* $\sigma_\zeta^2 \coloneqq \mathbb{V}[\zeta_i]$ *and* ζ_i *is defined in Equation* (3.7.13)*.*

As in Theorem 3.2, we achieve the same rate of convergence as an unpenalized procedure undertaken on the set of active covariates. In the penalized procedure, the asymptotic normality and unbiasedness comes from the estimated index being itself asymptotically normal and unbiased. This, in turn, stems from the use of an unbiased penalty function, i.e. the SCAD. In this construction, an alternative would be to use the debiased LASSO (Zhang and Zhang, 2014). However, a more general comment is worth making. From Escanciano and Terschuur (2022), it appears that we could accommodate for some bias in the estimation of the index. This is due

to the construction of $\widehat{\text{Gi}}_{Y,X}$, that uses the index only through its ordering structure. Following this route, it would be possible to relax the requirements on the estimation of θ_0. Interestingly, one of the conditions on the estimated index would be that the sign of the difference in the estimated index converges in probability to the sign of the difference in the true index. Thanks to Theorem 3.1, it is easy to show that this is satisfied, even for the LASSO. In light of this, the LASSO is another candidate to provide an unbiased estimation of the explained Gini coefficient.

Even though the variance is difficult to estimate in practice, the asymptotic normality of the estimated explained Gini coefficient opens the door to a parametric bootstrap procedure where the quantiles of the normal are used and the variance is estimated via bootstrap. Section 3.4.2 evaluates the asymptotic normality of the estimated explained Gini coefficient for several sample sizes, while Section 3.4.3 compares different bootstrap procedures. To anticipate slightly the discussion, the conclusion will be that the parametric bootstrap produces the best performance.

The proposed bootstrap procedure is described in Algorithm 1. In a nutshell, the idea consists in constructing training bootstrap resamples $(X^{*\intercal}, Y^*)^\intercal$ drawn with replacement from the pairs $(X_i^\intercal, Y_i)^\intercal$, and validation bootstrap samples $(\tilde{X}^{*\intercal}, \tilde{Y}^*)^\intercal$ corresponding to the out-of-bag (OOB) samples, i.e. the data unused in the construction of the training samples. As we know from bootstrap aggregating techniques, the size of the validation samples will be approximately of $\frac{n}{e}$, where e is Euler's number. The training samples are then used to obtain samples of bootstrap estimators for θ_0 and for $\text{Gi}_{Y,X}$ on a grid of λ values. The validation samples can then be used to determine an optimal λ. At the end of the procedure, one has at disposal samples of bootstrap estimates $\widehat{\text{Gi}}^*(\lambda)$ that one can use to produce confidence intervals.

The choice of the regularization parameter λ is of great importance. An optimal value should strike the balance between under- and over-penalization and, hence, between underfit and overfit of the explained Gini coefficient. As we have just stated, the bootstrap procedure offers a first method to select it : λ^{OOB} is the value of λ which performs best in terms of explained Gini coefficient in the out-of-bag resamples. Another possibility lies in the use of a BIC-like criterion, as proposed in Lin and Peng (2013). Formally, λ^{BIC} is obtained as the value of λ which maximizes

$$\text{BIC}_\lambda \coloneqq \log(\widehat{\text{Gi}}_{Y,X;\lambda}) - k_\lambda \frac{\log(n)}{2n},$$

where k_λ is the number of covariates selected using λ. Also, we write $\widehat{\text{Gi}}_{Y,X;\lambda}$ in order to acknowledge the dependence of the estimated explained Gini coefficient on λ through $\hat{\theta}$. As the simulation results from Section 3.4 and the real data example

Algorithm 1: Bootstrap procedure

Data: $(\mathbf{X}^\intercal, \mathbf{Y})^\intercal \in \mathbb{R}^{n\times(p+1)}$
, where $\mathbf{X}$ denotes the matrix of covariates and $\mathbf{Y}$ is the response vector.
for $b = 1$ **to** B **do**
  Generate the training bootstrap sample $(\mathbf{X}_b^{*\intercal}, \mathbf{Y}_b^*)^\intercal \in \mathbb{R}^{n\times(p+1)}$
  Obtain $\hat{\theta}_b^*(\lambda)$ by solving Equation (3.2.1) on $(\mathbf{X}_b^{*\intercal}, \mathbf{Y}_b^*)^\intercal$ via the SCAD-FABS algorithm from Section 3.3, using a λ sequence;
  Plug $\hat{\theta}_b^*(\lambda)$ and $(\mathbf{X}_b^{*\intercal}, \mathbf{Y}_b^*)^\intercal$ into Equation (2.3.3) and call the obtained estimator $\widehat{\mathrm{Gi}}_b^*(\lambda)$;
  Generate the validation bootstrap sample $(\tilde{\mathbf{X}}_b^{*\intercal}, \tilde{\mathbf{Y}}_b^*)^\intercal \in \mathbb{R}^{\tilde{n}_b\times(p+1)}$;
  Plug $\hat{\theta}_b^*(\lambda)$ and $(\tilde{\mathbf{X}}_b^{*\intercal}, \tilde{\mathbf{Y}}_b^*)^\intercal$ into Equation (2.3.3) and call the obtained estimator $\widetilde{\mathrm{Gi}}_b^*(\lambda)$
end
Obtain λ^{OOB} as the solution to $\max_\lambda$ OOB-score$(\lambda) \coloneqq \frac{1}{B}\sum_{b=1}^B \widetilde{\mathrm{Gi}}_b^*(\lambda)$;
For each λ, retrieve $\widehat{\mathrm{Gi}}^*(\lambda) \coloneqq (\widehat{\mathrm{Gi}}_1^*(\lambda), \ldots, \widehat{\mathrm{Gi}}_B^*(\lambda))^\intercal \in \mathbb{R}^B$

from Section 3.5 indicate, λ^{OOB} tends to select fuller models, while λ^{BIC} favours sparser ones. In practice, the user might decide between these two options by balancing the OOB-scores attained at λ^{OOB} and λ^{BIC} with the simplicity of the models that they induce.

3.3 The SCAD-FABS algorithm

The FABS has been developed by Shi et al. (2018) to solve penalized problems with differentiable but typically non-convex loss functions and the adaptive LASSO penalty. Its fundamental logic resembles the coordinate-descent algorithm of Friedman et al. (2007) presented in Section 1.6. It starts with a very large value for λ, imposing full sparsity, and then allows at each step the penalty parameter to relax. Hence, it produces a whole solution path ranging from high to low sparsity. For each λ, the optimization problem is solved using a coordinate-descent algorithm. To facilitate the comparison, we review the construction of the FABS algorithm in a pure LASSO setting. Then, we introduce the SCAD-FABS algorithm and present its convergence properties.

3.3.1 The FABS algorithm

For a grid of penalty parameters, the FABS solves

$$\min_{\theta} Q(\theta) \coloneqq L(\theta) + \sum_{k=1}^{p} p_{\lambda}(|\theta_k|), \tag{3.3.1}$$

where $Q(\cdot)$ is the objective function and $L(\cdot)$ is a general loss function. We focus here on a pure LASSO setting, where $p'_{\lambda}(x) = \lambda$. An iteration in the FABS is characterized by two actions. First, it updates only one coefficient by a fixed amount. Either this update is a backward step, where the magnitude of the coefficient decreases. This operation is undertaken if it reduces the value of the objective function. Otherwise, the iteration consists in a forward step, where the magnitude of the coefficient increases. The second action concerns the update of the penalization parameter λ_t.

For a given index k, the updated coefficient vector is of the form

$$\theta^{t+1} = \theta^t - \text{sign}(\theta_k^t)\mathbf{1}_k\epsilon \qquad \text{(backward step)} \tag{3.3.2}$$

$$\theta^{t+1} = \theta^t - \text{sign}\left(\nabla_k L(\theta^t)\right)\mathbf{1}_k\epsilon \qquad \text{(forward step)} \tag{3.3.3}$$

where θ^t (θ^{t+1}) represents the vector of coefficients at iteration t ($t+1$), $\mathbf{1}_k$ denotes the vector of size p taking value 1 at the kth component and 0 everywhere else, ϵ is the step size of the algorithm and ∇ is the gradient vector. The fact that a forward update, as established in Equation (3.3.3), increases the magnitude of the coefficient is proven in Lemma 1 in Shi et al. (2018). The index k is chosen optimally, using a first-order Taylor expansion of $L(\theta^{t+1})$ around θ^t. Formally, the coordinate to be updated at $(t+1)$ is determined as follows

$$k = \underset{l \in \mathcal{A}^t}{\arg\min} \left\{-\nabla_l L(\theta^t)\text{sign}(\theta_l^t)\right\} \qquad \text{(backward step)}$$

$$k = \underset{l=1,\ldots,p}{\arg\max} |\nabla_l L(\theta^t)| \qquad \text{(forward step)}$$

where $\mathcal{A}^t \coloneqq \{k \in \{1,\ldots,p\} : \theta_k^t \neq 0\}$. Finally, λ^t is updated as follows

$$\lambda^{t+1} = \lambda^t \qquad \text{(backward step)}$$

$$\lambda^{t+1} = \min\{\lambda^t, L_{\epsilon}^{t,t+1}\} \qquad \text{(forward step)} \tag{3.3.4}$$

where $L_{\epsilon}^{t,t+1} \coloneqq \frac{L(\theta^t) - L(\theta^{t+1})}{\epsilon}$. The update rule for λ^t generates a path of values for the regularization parameter that is a decreasing step function. For a given value, the algorithm minimizes the objective function by searching for the best direction. However, at some iteration, the objective function can no longer be decreased with this value of the regularization parameter. Hence, λ^t needs to make a jump, and the process is reiterated to search for the best direction that minimizes the objective function with this new value of the regularization parameter. Therefore, (3.3.4) ensures that an update occurs only when the objective function can no longer be improved using λ^t. Then, λ^{t+1} is chosen such that $Q(\theta^{t+1}, \lambda^{t+1}) = Q(\theta^t, \lambda^{t+1})$.

3.3.2 The SCAD-FABS algorithm

The SCAD-FABS solves the minimization problem displayed in (3.3.1) using the SCAD penalty function. The vector θ^{t+1} is obtained according to Equations (3.3.2) and (3.3.3). The form of the forward and backward steps is therefore the same as in the FABS. Also, the index k is chosen optimally, based on Taylor expansions. In a backward step, a Taylor expansion of the objective function around θ^t gives

$$Q(\theta^{t+1}) = Q(\theta^t) - \nabla_k Q(\theta^t)\text{sign}(\theta_k^t)\epsilon + O(\epsilon^2).$$

The index of the updated coefficient at iteration $(t+1)$ is then given by

$$k = \underset{l \in \mathcal{A}^t}{\arg\min} \left\{ -\nabla_l Q(\theta^t)\text{sign}(\theta_l^t) \right\}.$$

In a forward step, the Taylor expansion gives

$$Q(\theta^{t+1}) = Q(\theta^t) - |\nabla_k L(\theta^t)|\epsilon + p'_\lambda(|\theta_k^t|)\epsilon + O(\epsilon^2),$$

where $p'_\lambda(|\theta_k^t|)$ is defined according to Equation (1.6.3) for $|\theta_k^t| > 0$ and $p'_\lambda(|\theta_k^t|)$ is set to λ if $\theta_k^t = 0$. For $k \notin \mathcal{A}^t$, the result is obtained with a Taylor expansion of the loss while, for $k \in \mathcal{A}^t$, the result is obtained using a Taylor expansion of the loss and of the penalty, combined with Lemma 3.1. The index of the updated coefficient is obtained as

$$k = \underset{l=1,\dots,p}{\arg\max} \left\{ |\nabla_l L(\theta^t)| - p'_\lambda(|\theta_l^t|) \right\}.$$

Note that the choice of the direction differs from the FABS. The reason is the following. Since the derivative of the LASSO penalty is constant, coupled with the fact that only one coefficient is updated by a fixed amount, the penalty part does not play a role in the Taylor expansion of the FABS in Section 3.3.1. Using a SCAD penalty, this no longer applies and the whole objective function must be considered.

The update rule for λ^t follows the same spirit as in the original FABS. An update is conducted when the objective function can no longer be improved using λ^t, i.e. $Q(\theta^{t+1}, \lambda^t) > Q(\theta^t, \lambda^t)$. Consider a forward update on coefficient k. Using a Taylor expansion of $p_{\lambda^t}(|\theta_k^{t+1}|)$ around $|\theta_k^t|$, it approximately holds

$$Q(\theta^{t+1}, \lambda^t) > Q(\theta^t, \lambda^t) \Leftrightarrow L(\theta^t) - L(\theta^{t+1}) < p'_{\lambda^t}(|\theta_k^t|)\epsilon.$$

This occurs when

$$\begin{cases} \lambda^t > L_\epsilon^{t,t+1} & \text{if } |\theta_k^t| \leq \lambda^t \\ \lambda^t > \frac{1}{a}\left[(a-1)L_\epsilon^{t,t+1} + |\theta_k^t|\right] & \text{if } \lambda^t < |\theta_k^t| \leq a\lambda^t \\ L_\epsilon^{t,t+1} < 0 & \text{if } |\theta_k^t| > a\lambda^t. \end{cases} \tag{3.3.5}$$

Notice that, in the last case, λ^t plays no role since $p'_{\lambda^t}(|\theta_k^t|) = 0$ and, hence, λ^t should not be updated. We now focus on the form of the update for λ^t. In the event of an update, λ^{t+1} is chosen to ensure that approximately $Q(\theta^{t+1}, \lambda^{t+1}) = Q(\theta^t, \lambda^{t+1})$, which happens whenever

$$\lambda^{t+1} = \lambda_A^{t+1} \coloneqq L_\epsilon^{t,t+1} \qquad \text{if } |\theta_k^t| \leq \lambda^{t+1} \tag{3.3.6}$$

$$= \lambda_B^{t+1} \coloneqq \frac{1}{a}\left[(a-1)L_\epsilon^{t,t+1} + |\theta_k^t|\right] \qquad \text{if } \lambda^{t+1} < |\theta_k^t| \leq a\lambda^{t+1}. \tag{3.3.7}$$

Using Equation (3.3.6), we choose $\lambda^{t+1} = \lambda_A^{t+1}$ if $|\theta_k^t| \leq \lambda_A^{t+1}$. Using Equations (3.3.6) and (3.3.7), we choose $\lambda^{t+1} = \lambda_B^{t+1}$ if $|\theta_k^t| > \lambda_A^{t+1}$ and $\lambda_B^{t+1} \leq |\theta_k^t| \leq a\lambda_B^{t+1}$. It is easy to prove that this boils down to choosing $\lambda^{t+1} = \max\{\lambda_A^{t+1}, \lambda_B^{t+1}\}$. The situation $|\theta_k^t| > a\lambda_B^{t+1}$ implies $L_\epsilon^{t,t+1} < 0$ and can be disregarded due to the loss improvement check introduced in the SCAD-FABS algorithm, presented in Algorithm 2. Taking our last results together with Equation (3.3.5), the update rule for λ^t becomes

$$\lambda^{t+1} = \begin{cases} \min\left(\max\{\lambda_A^{t+1}, \lambda_B^{t+1}\}, \lambda^t\right) & \text{if } |\theta_k^t| \leq a\lambda^t \\ \lambda^t & \text{otherwise.} \end{cases} \tag{3.3.8}$$

Algorithm 2: The SCAD-FABS algorithm

Data: $(\mathbf{X}^\intercal, \mathbf{Y})^\intercal \in \mathbb{R}^{n\times(p+1)}$

Initialization: start from the empty solution and compute

$k = \underset{l=1,\ldots,p}{\arg\max} |\nabla_l L(0)|; \mathcal{A}^0 = \{k\}$

$\theta^0 = -\text{sign}(\nabla_k L(0))\mathbf{1}_k\epsilon$

$\lambda^0 = \frac{1}{\epsilon}[L(0) - L(\theta^0)]$

Backward step: for each iteration $t \geq 1$, compute

$k = \underset{l\in\mathcal{A}^t}{\arg\min}\left\{-\nabla_l Q(\theta^t)\text{sign}(\theta_l^t)\right\}$

$\Delta_k = -\text{sign}(\theta_k^t)\mathbf{1}_k$

If $L(\theta^t + \Delta_k\epsilon) - L(\theta^t) - \epsilon p'_{\lambda^t}(|\theta_k^t|) < 0$, take a backward step. Then

$\theta^{t+1} = \theta^t + \Delta_k\epsilon$

$\lambda^{t+1} = \lambda^t$

Otherwise, take a forward step.

Forward step: set $\mathcal{B}^t = \{1, \ldots, p\}$ and compute

$k = \underset{l\in\mathcal{B}^t}{\arg\max}\left\{|\nabla_l L(\theta^t)| - p'_\lambda(|\theta_l^t|)\right\}$

If $L\left(\theta^t - \text{sign}\left(\nabla_k L(\theta^t)\right)\mathbf{1}_k\epsilon\right) > L(\theta^t)$, set $\mathcal{B}^t = \mathcal{B}^t\backslash\{k\}$ and return to the computation of the index k to be updated. Otherwise, set

$\theta^{t+1} = \theta^t - \text{sign}\left(\nabla_k L(\theta^t)\right)\mathbf{1}_k\epsilon$

$\lambda^{t+1} = \begin{cases} \min\left(\max\{\lambda_A^{t+1}, \lambda_B^{t+1}\}, \lambda^t\right) & \text{if } |\theta_k^t| \leq a\lambda^t \\ \lambda^t & \text{otherwise.} \end{cases}$

Stopping rule: Update $t \to t+1$, repeat the backward and forward steps and stop when $\lambda^{t+1} \leq 0$ or $\mathcal{B}^t = \emptyset$. The final iteration is then given by $T = t$.

Notice the presence of a loss improvement check in the forward step. This step is unnecessary in the classical FABS algorithm as proposed by Shi et al. (2018). Indeed, an increase in the loss would imply $L_\epsilon^{t,t+1} < 0$, which would cause $\lambda^{t+1} < 0$ and the algorithm would stop. In the SCAD-FABS, updates on coefficients whose amplitude exceed $a\lambda^t$ do not yield updates on λ^t. Hence, one needs to ensure that these coefficients are not updated if they yield an increase in the loss.

We use the SCAD-FABS algorithm in order to solve (3.2.1). Following Lin and Peng (2013), we incorporate the constraint $||\theta||= 1$ by considering the Lagrangian function related to our problem. The loss function is thus given by

$$L(\theta) = -G_n(\theta) + \gamma \sum_{k=1}^{p} \theta_k^2. \tag{3.3.9}$$

We then apply the SCAD-FABS with $a = 3.7$, as suggested by Fan and Li (2001), and $\gamma = 0.05$ throughout the paper. Extensive simulations in Section 3.4 have shown that these values worked well in practice. We close this section by reviewing the properties of the SCAD-FABS algorithm.

3.3.3 Properties of the SCAD-FABS algorithm

Throughout this section, we assume that $L(\cdot)$ has bounded second-order derivatives. The proofs of the following results are deferred to Subsection 3.7.2. Lemma 3.1 indicates that a forward step always increases the amplitude of the updated coefficient. As long as no backward step is taken, the solution path is therefore monotone.

Lemma 3.1. *Let θ_k^t with $k \in \mathcal{A}^t$ be updated via a forward step. Then, $\theta_k^{t+1} = \theta_k^t + \text{sign}(\theta_k^t)\epsilon$, i.e. $\text{sign}(\nabla_k L(\theta^t)) = -\text{sign}(\theta_k^t)$.*

The SCAD-FABS exploits the differentiability of the loss and penalty functions through Taylor expansions. As such, it produces approximation errors. Lemmas 3.6, 3.7 and 3.8, presented in Subsection 3.7.2, show that the approximation error is of order $O(\epsilon^2)$, where we remind the reader that ϵ is the step size of the algorithm. Proposition 3.1 implies that the objective function never increases from t to $t+1$, up to an approximation error of order ϵ^2. This stems from the fact that λ^t is updated whenever it becomes impossible to improve the score further with that value of the regularization parameter.

Proposition 3.1. *The SCAD-FABS algorithm ensures that, for all t, $Q(\theta^{t+1}, \lambda^{t+1}) \leq$*

$Q(\theta^t, \lambda^t)$. *More precisely, if k is the index of the updated coefficient, then it holds*

$$Q(\theta^{t+1}, \lambda^{t+1}) < Q(\theta^t, \lambda^t) - \frac{d_k^t \epsilon^2}{2(a-1)} \qquad \textit{(backward step)}$$
$$Q(\theta^{t+1}, \lambda^{t+1}) \leq Q(\theta^t, \lambda^t) - \frac{c_k^t \epsilon^2}{2(a-1)}, \qquad \textit{(forward step)}$$

where d_k^t and $c_k^t \in [0, 1]$.

In the remaining of this section, we show that the SCAD-FABS enjoys the same δ-optimality property as the FABS. A candidate is called a δ-approximate solution if it satisfies the KKT conditions associated to the constrained optimization problem, up to a tolerance δ. We adapt the definition introduced by Shi et al. (2018) to the SCAD penalty. The parameter vector $\theta = (\theta_1, \ldots, \theta_p)^\intercal$ is called a δ-approximate solution with regularization parameter λ if the following two conditions are met

$$\begin{aligned} |\nabla_l Q(\theta, \lambda)| &\leq \delta && \text{if } \theta_l \neq 0 \\ |\nabla_l L(\theta)| &\leq p'_\lambda(|\theta_l|) + \delta && \text{if } \theta_l = 0. \end{aligned}$$

In what follows, the path refers to the collection of iterations $\{t = 1, \ldots, T : \lambda^{t+1} < \lambda^t\}$, where T is the last iteration.

Theorem 3.4. *Every solution θ^t, with $t = 1, \ldots, T$, along the SCAD-FABS path is a δ-approximate solution with regularization parameter λ^t and $\delta = m\epsilon$, where m is the upper bound of the second-order derivatives of $L(\cdot)$.*

Theorem 3.4 shows that any point along the SCAD-FABS path is a δ-approximate solution, with δ proportional to the step size ϵ. The SCAD-FABS algorithm enjoys therefore the same optimality condition as the FABS. As $\epsilon \to 0$, the KKT conditions are recovered and each point along the path converges to a stationary point of (3.3.1).

3.4 Monte-Carlo simulations

In this section, we evaluate the performance of the penalized Lorenz regression combined with the proposed bootstrap procedure by means of Monte-Carlo simulations. As a first step, we focus on the quality of the estimation of the explained Gini coefficient and the performance of the model selection. Then, we evaluate the consistency and asymptotic normality of the estimated explained Gini coefficient. Finally, we turn to the coverage of the confidence intervals.

Throughout the simulations, we will use the following DGP

$$Y_i = \mathcal{Q}(F_{\theta_0}(X_i^\intercal \theta_0))\epsilon_i,$$

where $i = 1 \ldots, n$ and $||\theta_0|| = 1$. $\mathcal{Q}(\cdot)$ is the quantile function of the lognormal distribution, with parameters tailored to ensure an expected value of 2400 and an explained Gini coefficient of 0.11. This choice is justified by the popular use of the lognormal distribution in empirical work concerning income distributions, see Cowell and Flachaire (2015). Two scenarios are considered for the distribution of X. In a first case, X follows a multivariate normal distribution with mean 0, unit variance and a correlation matrix following an AR(1) process with correlation parameter $\rho = 0.3$. In a second case, X follows a multivariate Student distribution with 3 degrees of freedom, mean 0, variance of 3 and the same correlation matrix as before. The variable ϵ_i is a lognormal noise with mean 1 and a variance set to ensure that $\mathbb{V}[Y_i]/\mathbb{V}[H(X_i^\intercal \theta_0)] = 3/2$. Concerning the kernel, we use

$$K(u) = \begin{cases} 0 & \text{if } u < -1 \\ \frac{9}{8}u - \frac{5}{8}u^3 + \frac{1}{2} & \text{if } u \in [-1, 1] \\ 1 & \text{if } u > 1, \end{cases}$$

and set the bandwidth to $h = Cn^{-1/5.5}$. The constant C is chosen from the grid $(0.2, 0.5, 1, 2, 5)$ as the value that maximizes the OOB-score (bootstrap procedure) or that maximizes the BIC criterion (BIC procedure).

We consider two setups. The first is low-dimensional with $n = 100$ and $p = 20$ (Setup 1) while the second is high-dimensional with $n = 100$ and $p = 120$ (Setup 2). In both cases, 5 of the covariates are active and the parameter vector is $\theta_0 = (-\frac{\sqrt{3}}{5}, \frac{3}{5}, 0, 0, -\frac{\sqrt{7}}{5}, \frac{1}{5}, \frac{\sqrt{5}}{5}, 0^\intercal)^\intercal$, where $0^\intercal$ is a vector of zeroes of size 13 in Setup 1 and of size 113 in Setup 2. Unless specified otherwise, we sample $M = 400$ different datasets from the proposed DGPs and, for each simulation run, $B = 400$ bootstrap resamples are used.

3.4.1 Comparison with competitors

We start by examining the performance of the penalized bootstrap Lorenz regression in terms of estimation and model selection. The regularization parameter λ and the bandwidth h are either selected using the BIC-like criterion or the bootstrap procedure. We denote the results as SCAD-FABS (BIC) and SCAD-FABS (Bootstrap) accordingly. Our competitor is the PMSRC estimator, obtained via the procedure proposed by Lin and Peng (2013), denoted as PMSRC (LP) or computed using the FABS algorithm of Shi et al. (2018), denoted as PMSRC (FABS). In both cases, the optimal regularization parameter is obtained via the BIC-like criterion proposed by Lin and Peng (2013).

Table 3.1: Estimation quality

	Setup 1		Setup 2	
	Normal	Student	Normal	Student
	RMSE.Gini expressed in %			
PMSRC (LP)	1.39	1.62	1.56	2.02
PMSRC (FABS)	1.12	1.25	1.28	1.58
SCAD-FABS (Bootstrap)	1.07	1.09	1.18	1.34
SCAD-FABS (BIC)	1.09	1.15	1.04	1.13
	Distance.θ			
PMSRC (LP)	0.39	0.42	0.48	0.53
PMSRC (FABS)	0.28	0.31	0.40	0.45
SCAD-FABS (Bootstrap)	0.25	0.29	0.37	0.43
SCAD-FABS (BIC)	0.25	0.28	0.32	0.39

The estimation quality is evaluated by the square-root of MSE.Gini, where the latter criterion is defined in Equation (2.5.1). We also use the mean L2-distance between the estimated parameter vector and the true parameter vector, as defined by Distance.θ in Equation (2.5.2). These quantities were estimated with M simulation runs. In order to assess the performance of the model selection, we examine the number of correctly selected zero and nonzero coefficients.

As we can observe from Table 3.1, our procedure offers performances at least as good as the competitors. In Setup 1, the best results are offered by the bootstrap procedure. In Setup 2, the BIC-like criterion seems to perform slightly better. In both cases, the worst performer is the PMSRC estimator obtained with the procedure of Lin and Peng (2013). The two metrics provide similar conclusions.

Table 3.2 displays the number of correctly selected zero and nonzero coefficients. As to be expected, the bootstrap procedure is the best in terms of correctly selecting the non-zero coefficients. However, the procedures relying on BIC criteria perform better in terms of identification of the zero coefficients.

3.4.2 Consistency and asymptotic normality

In what follows, we provide an assessment of the consistency and asymptotic normality of the estimated explained Gini coefficient, using the SCAD-FABS on Setup 1 with multivariate normal covariates. These results were obtained using $M = 2000$

Table 3.2: Number of correctly selected zeroes and nonzeroes

	Normal		Student	
	Nonzeroes	Zeroes	Nonzeroes	Zeroes
Setup 1	5	15	5	15
PMSRC (LP)	3.83	14.68	3.70	14.59
PMSRC (FABS)	4.38	14.23	4.33	14.10
SCAD-FABS (Bootstrap)	4.86	10.92	4.76	10.59
SCAD-FABS (BIC)	4.35	14.83	4.36	14.66
Setup 2	5	115	5	115
PMSRC (LP)	3.30	114.30	3.11	114.21
PMSRC (FABS)	3.72	114.24	3.54	114.02
SCAD-FABS (Bootstrap)	4.46	105.26	4.23	105.63
SCAD-FABS (BIC)	4.20	114.44	4.12	113.84

datasets from the proposed DGP. Figure 3.1 displays the performance of the BIC and bootstrap procedures for sample sizes ranging from 50 to 1000. On the left side of the figure, we observe that, when the sample size increases, both procedures tend to identify correctly both the active and non-active sets. It also confirms that the BIC criterion yields the best performance for the identification of the zero coefficients. On the other hand, the bootstrap procedure is the best at correctly identifying the active set. Concerning the explained Gini coefficient, the right hand side of the figure indicates that both procedures yield a consistent estimator and follow very close trajectories in terms of mean squared error.
Figure 3.2 assesses the normality of the estimated explained Gini coefficient obtained on samples of size 1000. The qqplot on the left indicates that most quantiles of $\widehat{\text{Gi}}_{Y,X}$ follow the quantiles of a Normal distribution. We observe however some deviations for the largest quantiles, indicating that the distribution is slightly right-skewed. The histogram on the right is centred at 0.11, the true value of the explained Gini coefficient, and decays in a bell shaped way, as expected.

3.4.3 Coverages of the confidence intervals

We construct 95%-confidence intervals for the explained Gini coefficient using the bootstrap methods introduced in Section 2.3. As in the last section, we first focus on Setup 1 with multivariate normal covariates and perform our simulations on $M = 2000$ different datasets. We display results using the bootstrap and BIC procedures. The quality of the confidence intervals is assessed in terms of their lengths and coverages.

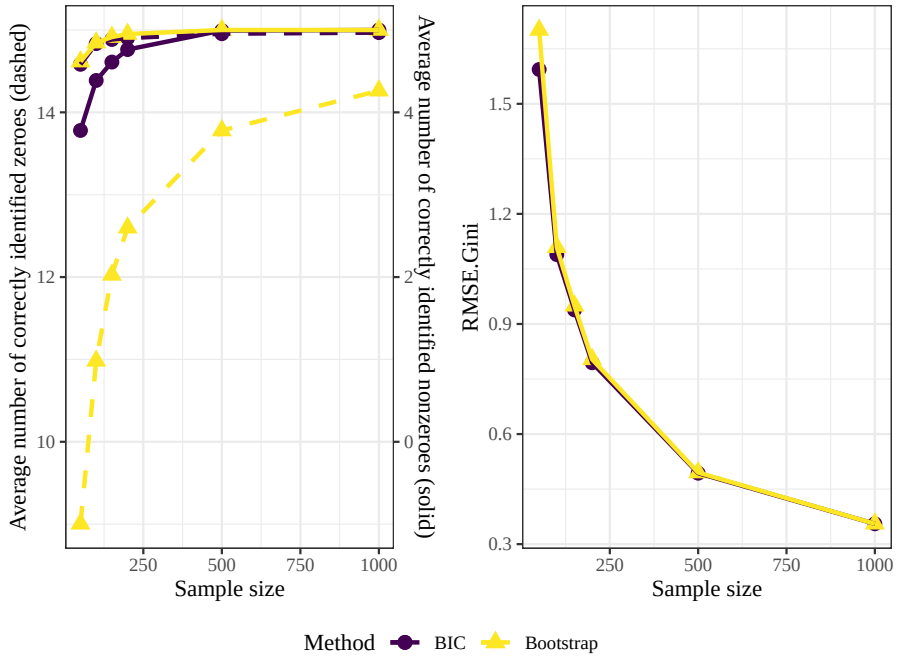

Figure 3.1: Consistency of the SCAD-FABS on Setup 1

Figure 3.3 displays the evolution of the coverages and lengths for different sample sizes. Solid lines refer to BIC while dashed lines correspond to the bootstrap procedure. Several observations are worth making. With low sample sizes, all confidence intervals undercover the true parameter. However, the coverages converge to the target level of 95% as the sample size increases. In terms of types of bootstrap, a clear ranking emerges. The basic bootstrap performs the worst. The percentile bootstrap comes second while the parametric bootstrap provides the best performance. Finally, BIC yields slightly wider confidence intervals with better coverages while the contrary goes for the bootstrap procedure.

We now consider results obtained on Setup 2, with multivariate normal covariates and a sample size of 100. Table 3.3 gathers the results of 95% parametric bootstrap confidence intervals for $\mathrm{Gi}_{Y,X}$. The first line displays the coverage of the confidence interval. As before, we observe a better performance of the BIC procedure. The second line shows the proportion of times the explained Gini coefficient is smaller than the whole confidence interval. In this respect, the BIC procedure performs the best. The third line shows the proportion of times it is

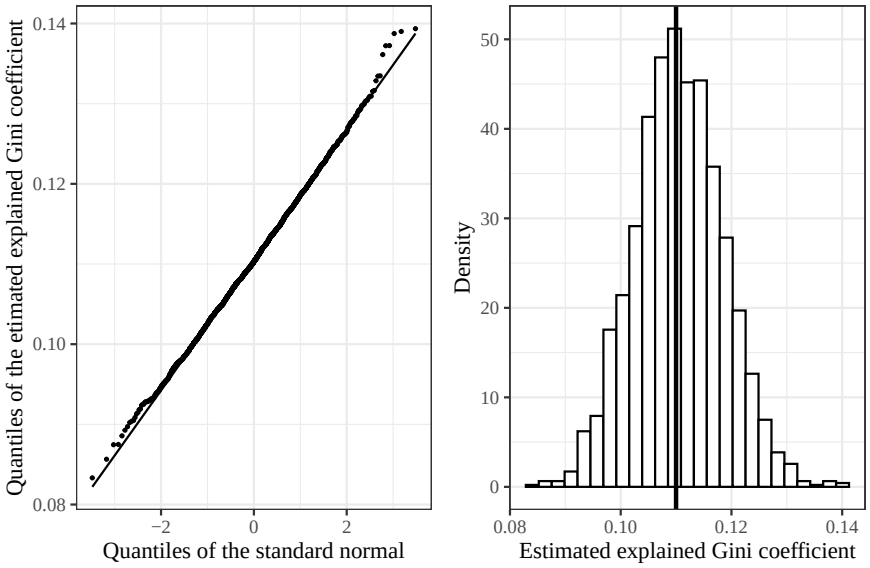

Figure 3.2: Asymptotic Normality of the SCAD-FABS on Setup 1, sample of size 1000

higher than the confidence interval. Here, the bootstrap procedure yields the best performance. This difference illustrates the idea that the BIC-like criterion favours sparser models and, hence, will tend to underestimate. On the other hand, the bootstrap procedure will tend to overestimate because it will select fuller models. The final line displays the length of the intervals. As before, BIC leads to wider confidence intervals.

3.5 Empirical illustration

Our discussion is based on the `Griliches` dataset introduced in Section 2.6. We recall that the response is wage and we include the following covariates: schooling (*Schooling*), experience (*Exp*), age (*Age*), ability (*IQ*), marital status (*Married*), degree of urbanisation (*Urban*) and a dummy indicating whether the individual lives in a Southern State (*South*). Some descriptive statistics were already given in Section 2.6. We recall that the Gini coefficient of wage is of 22%. The concentration indices of wage with respect to age, schooling, experience and ability are respectively of 4.34%, 7.21%, 0.71% and 6.14%.

We run a penalized bootstrap Lorenz regression including all the covariates as well

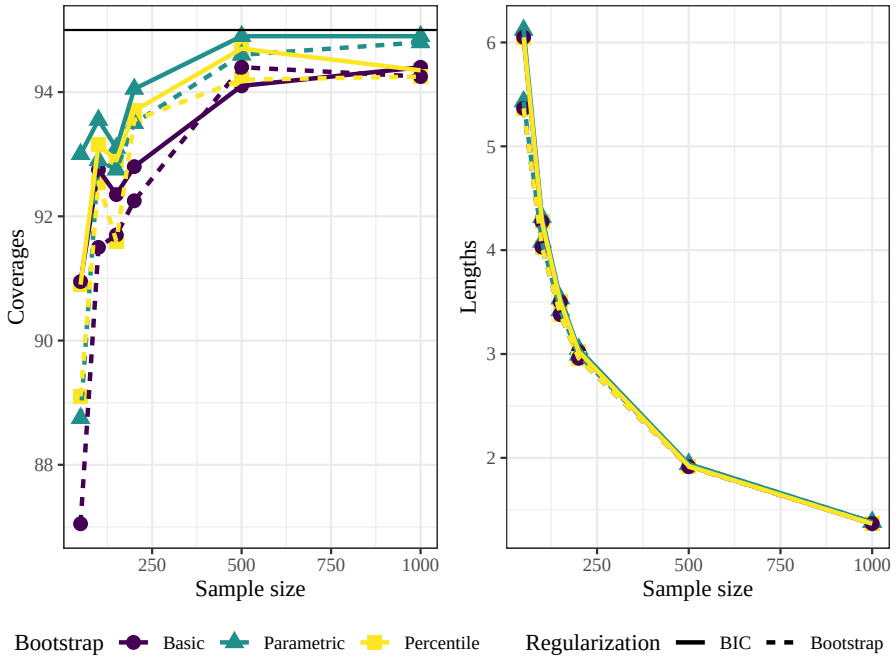

Figure 3.3: Coverages and Lengths of confidence intervals for $\text{Gi}_{Y,X}$ on Setup 1

as the interactions between the binary and the numeric ones. We let the procedure select the variables to include in the final model. In essence, it does so by balancing the complexity of the model with the extra amount of explained inequality that a deshrink in the coefficient would bring. Figure 3.4 plots the trace of the algorithm, with the value C selected by the bootstrap procedure, as in Section 3.4. The horizontal axis displays the evolution of the penalty parameter (in terms of $-\log(\lambda)$), while each line corresponds to the value taken by a given coefficient. At any time, the vector of coefficients is normalized in order to have a unit $L2$-norm. When the algorithm starts, penalization is the highest and schooling is the first covariate to enter the model. This makes sense since, as we have seen, this is the variable for which we observe the highest concentration index. As the algorithm proceeds, new variables enter the model but they may also shrink back to 0. We observe this phenomenon for variable *South* for example. Figure 3.5 plots the evolution of the estimated explained Gini coefficient. At the beginning of the algorithm, it is equal to 7.21% and coincides with the concentration index of income with respect to schooling. As sparsity reduces and new variables enter the model, the estimated explained Gini coefficient increases to reach 10.7% at the end of the algorithm.

Table 3.3: Results of 95% parametric bootstrap confidence intervals for $\text{Gi}_{Y,X}$, Setup 2 and sample size of 100

	Bootstrap	BIC
$\text{Gi}_{Y,X} \in \text{CI}$	91	95
$\text{Gi}_{Y,X} < \text{CI}$	7.25	0.50
$\text{Gi}_{Y,X} > \text{CI}$	1.75	4.50
Length	4.02	4.55

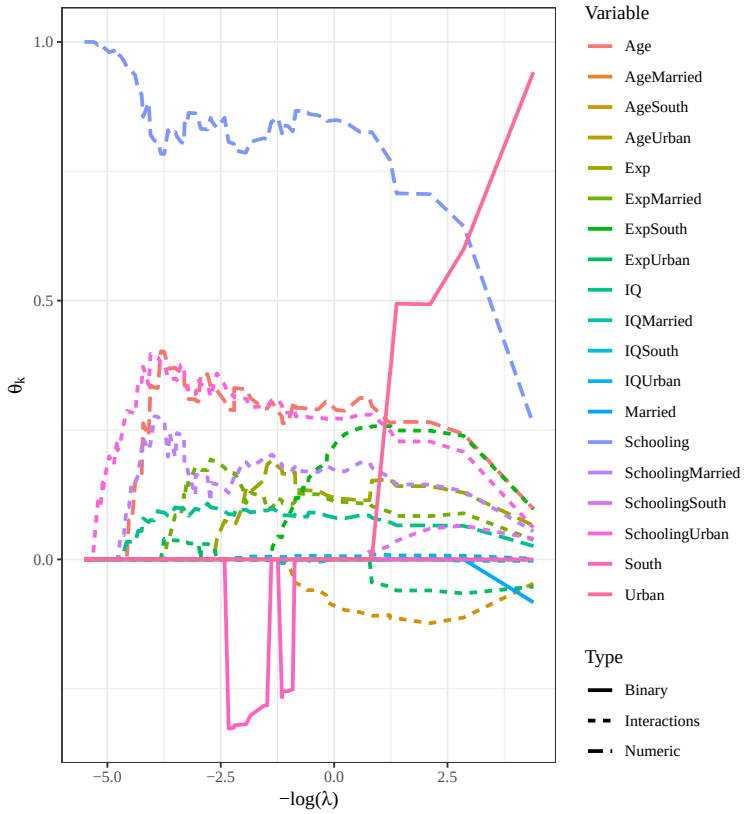

Figure 3.4: Trace plot of the SCAD-FABS on the Griliches data

We now compare the model fits obtained with the BIC and bootstrap procedures. Using the BIC-like criterion, 11 variables are included : *Age* (as well as interaction with *South*), *Schooling* (as well as interactions with *South*, *Urban* and *Married*), *Exp* (as well as interaction with *South* and *Married*) and *IQ* (as well as interaction

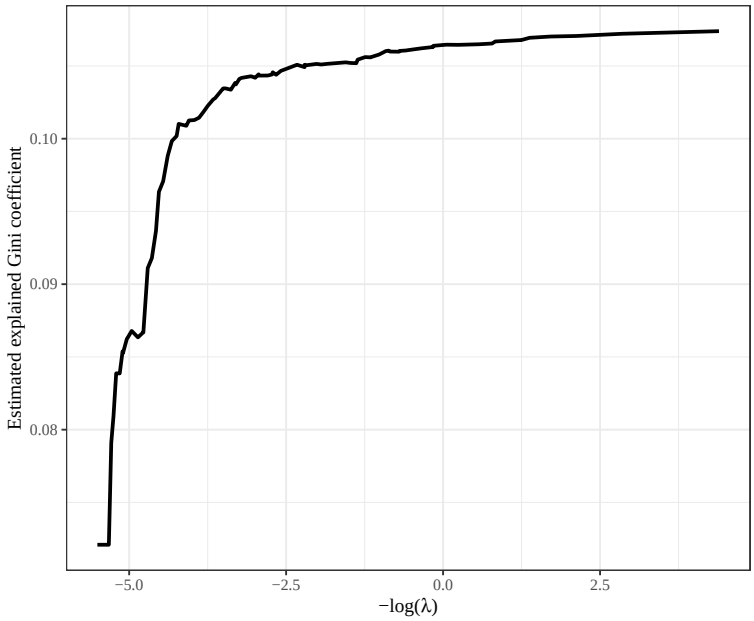

Figure 3.5: Estimated explained Gini coefficient along the SCAD-FABS on the Griliches data

with *South*). With the bootstrap procedure, 13 variables are selected. Compared to the BIC procedure, the interaction between IQ and *South* is not selected. Three extra covariates are selected: *Urban* (including interaction with IQ and Exp). Table 3.4 summarizes the model fits associated to each selection procedure. The results are extremely similar, with an estimated explained Gini coefficient of 10.7% and a 95% confidence interval around $[9.2\%, 12.2\%]$.

3.6 Discussion

The penalized Lorenz regression is proposed as a method to produce estimation and inference for the explained Gini coefficient in a high-dimensional setting. Several features make it a suitable choice. First, it shares the good properties of the Lorenz regression developed in Chapter 2. The statistical model underlying the procedure is a single-index and is therefore more flexible than fully parametric models. Also, the link function $H(\cdot)$, which corresponds to the nonparametric part of the model, does not need to be estimated.

Second, the penalized Lorenz regression provides some improvements to the original procedure. The addition of a SCAD penalty ensures that irrelevant co-

Table 3.4: Estimated explained Gini coefficient and OOB-score

	λ^{OOB}	λ^{BIC}
$\widehat{\text{Gi}}_{Y,X}$	10.702%	10.701%
OOB-score	9.90%	9.88%
Parametric Bootstrap 95% CI	[9.20%, 12.21%]	[9.15%, 12.25%]

variates are discarded with a probability tending to one and avoids to overestimate the explained Gini coefficient. The presence of a differentiable approximation in the objective function allows the use of the SCAD-FABS algorithm, for which convergence properties are established. The procedure also enjoys strong statistical guarantees. It yields an asymptotically normal estimator for the explained Gini coefficient, with a convergence rate unaffected by the selection process. The inference for the explained Gini coefficient and the selection of the regularization parameter can be dealt with in a single bootstrap procedure. In the simulations, the procedure proved to be at least as good as the proposed competitors.

What is missing is an assessment of the contribution of each covariate to the explained Gini coefficient. The estimated parameter vector $\hat{\theta}$ provides such a contribution but it has no direct interpretation in terms of inequality. Also, it is computed on the full model only and therefore measures a contribution of one covariate given the others. A procedure based on the so-called Shapley Value decomposition is proposed in Chapter 4 in order to fill this gap.

3.7 Proofs

3.7.1 Asymptotic properties of the penalized Lorenz regression

We provide the proofs of Theorems 3.1, 3.2 and 3.3. We also state and prove some technical results related to the asymptotic properties of the penalized Lorenz regression.

We rewrite Equation (3.2.2) as

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i \hat{F}_\theta(X_i^\intercal \theta),$$

where

$$\hat{F}_\theta(t) \coloneqq \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{t - X_i^\intercal \theta}{h}\right).$$

In what follows, for any function $\varphi_\theta(y)$, denote $\nabla_{\theta_0}\varphi_\theta(y) = [\partial\varphi_\theta(y)/\partial\theta_1, \ldots, \partial\varphi_\theta(y)/\partial\theta_p]^\intercal|_{\theta=\theta_0}$. Also, $\nabla^2_{\theta_0}\varphi_\theta(y)$ is a $[p\times p]$ matrix whose $[k,l]$ element is given by $[\partial^2\varphi_\theta(y)/\partial\theta_k\partial\theta_l]|_{\theta=\theta_0}$.

Lemma 3.2. *Assume (RC1)-(RC4). Then, for any $0 < \delta < 1$,*

$$\sup_{z\in\mathcal{Z}} |\hat{F}_{\theta_0}(z) - F_{\theta_0}(z)| = o_p(1) \tag{3.7.1}$$

$$\sup_{z\in\mathcal{Z}} |\hat{F}'_{\theta_0}(z) - F'_{\theta_0}(z)| = o_p(1) \tag{3.7.2}$$

$$\sup_{z,z'\in\mathcal{Z}} \frac{|\hat{F}'_{\theta_0}(z) - F'_{\theta_0}(z) - \hat{F}'_{\theta_0}(z') + F'_{\theta_0}(z')|}{|z-z'|^\delta} = O_p(1). \tag{3.7.3}$$

Proof. Relations (3.7.1) and (3.7.2) are direct consequences of Theorem 3 in Yamato (1973) and Theorem A in Silverman (1978) respectively. Hence, we focus on proving (3.7.3). Let $d_n(z) \coloneqq \hat{F}_{\theta_0}(z) - F_{\theta_0}(z)$ and $\beta_n(z,z') \coloneqq (d'_n(z) - d'_n(z'))/|z - z'|^\delta$. We have that $d'_n(z) - d'_n(z') = (z-z')d''_n(z^*)$ for some z^* between z and z'. Hence, it holds

$$\begin{aligned}|\beta_n(z,z')| &\leq |z-z'|^{1-\delta}\sup_{z\in\mathcal{Z}}|d''_n(z)| \\ &\leq C\sup_{z\in\mathcal{Z}}|d''_n(z)|,\end{aligned}$$

for some constant C. The supremum above is $O_p(1)$ by Theorem C in Silverman (1978). This finishes the proof. □

Proposition 3.2. *Assume (RC1)-(RC4). Then,*

$$\frac{1}{n}\sum_{i=1}^n \left\{Y_i\hat{F}_{\theta_0}(Z_i) - Y_iF_{\theta_0}(Z_i) - \mathbb{E}[Y\hat{F}_{\theta_0}(Z)|\mathcal{X}_n] + \mathbb{E}[YF_{\theta_0}(Z)]\right\} = o_p(n^{-1/2}),$$

where $\mathcal{X}_n = \{(X_i^\intercal, Y_i)^\intercal, i = 1, \ldots, n\}$ and $Z_i = X_i^\intercal\theta_0$.

Proof. We are going to use Lemma 19.24 from van der Vaart (1998), with

$$\begin{aligned}\hat{f}_n &: (y,z) \mapsto y\hat{F}_{\theta_0}(z) = yd_n(z) + yF_{\theta_0}(z) \\ f_0 &: (y,z) \mapsto yF_{\theta_0}(z),\end{aligned}$$

where $d_n(z) \coloneqq \hat{F}_{\theta_0}(z) - F_{\theta_0}(z)$. Define the class

$$\mathcal{F} = \{(z,y) \to y[d(z) + F_{\theta_0}(z)], d \in C_1^{1+\delta}(\mathcal{Z})\},$$

where $C_1^{1+\delta}(\mathcal{Z})$ is the class of differentiable functions d defined on $\mathcal{Z}$ satisfying $||d||_{1+\delta} \leq 1$, where $||d||_{1+\delta} = \max\{\sup_z|d(z)|, \sup_z|d'(z)|\} + \sup_{z,z'}|d'(z) - d'(z')|/|z - z'|^\delta$. From Lemma 3.2, we have that $P(d_n \in C_1^{1+\delta}(\mathcal{Z})) \to 1$ as $n \to \infty$. We start

by showing that $\mathcal{F}$ is Donsker. From Theorem 19.5 in van der Vaart (1998), this reduces to showing that

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{P_{Y,Z}})}d\epsilon < \infty, \tag{3.7.4}$$

where $N_{[]}$ is the bracketing number, $P_{Y,Z}$ is the probability measure related to $F_{Y,Z}$ and $||\cdot||_{P_{Y,Z}}$ denotes the $L2(P_{Y,Z})$-norm. By Corollary 2.7.2 in van der Vaart and Wellner (1996), $O(\exp(\epsilon^{-2/(1+\delta)}))$ brackets are needed for $d(z)$. Hence, the bracketing entropies $\log N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{P_{Y,Z}})$ are growing with order slower than ϵ^{-2}. Hence, (3.7.4) holds and $\mathcal{F}$ is Donsker. To be able to use Lemma 19.24 from van der Vaart (1998), it remains to show that

$$\int\int y^2[\hat{F}_{\theta_0}(z) - F_{\theta_0}(z)]^2 dF_{Y,Z}(y,z) = o_p(1).$$

This is a direct consequence of Lemma 3.2 and $\mathbb{E}[Y^2] < \infty$. □

Proposition 3.3. *Assume (RC1)-(RC4). Then,*

$$\frac{1}{n}\sum_{i=1}^n Y_i \hat{F}_{\theta_0}(Z_i) = \frac{1}{n}\sum_{i=1}^n \left\{ Y_i F_{\theta_0}(Z_i) + \int\int y\eta(Z_i, z) dF_{Y,Z}(y,z) \right\} + o_p(n^{-1/2}),$$

with $\eta(Z_i, z) = K((z - Z_i)/h) - F_{\theta_0}(z)$. *Also,* $\mathbb{E}[\eta(Z_i, z)] = O(h^3)$.

Proof. Using Proposition 3.2, we have

$$\frac{1}{n}\sum_{i=1}^n Y_i \hat{F}_{\theta_0}(Z_i) = \frac{1}{n}\sum_{i=1}^n Y_i F_{\theta_0}(Z_i) + \mathbb{E}[Y[\hat{F}_{\theta_0}(Z) - F_{\theta_0}(Z)]|\mathcal{X}_n] + o_p(n^{-1/2}).$$

Also, we can write

$$\mathbb{E}[Y[\hat{F}_{\theta_0}(Z) - F_{\theta_0}(Z)]|\mathcal{X}_n] = \frac{1}{n}\sum_{i=1}^n \int\int y\eta(Z_i, z) dF_{Y,Z}(y,z),$$

To finish the proof, we show that $\mathbb{E}[\eta(Z_i, z)] = O(h^3)$. Let $\mathcal{Z} = [z_1, z_2]$, we have

$$\begin{aligned}
\mathbb{E}\left[K\left(\frac{z - Z_i}{h}\right)\right] &= \int_{z_1}^{z_2} K\left(\frac{z-t}{h}\right) dF_{\theta_0}(t) \\
&= \left[K\left(\frac{z-t}{h}\right) F_{\theta_0}(t)\right]_{z_1}^{z_2} + \int_{z_1}^{z_2} \frac{1}{h}\kappa\left(\frac{z-t}{h}\right) F_{\theta_0}(t) dt \\
&= K\left(\frac{z-z_2}{h}\right) + \int_{z-h}^{z+h} \frac{1}{h}\kappa\left(\frac{t-z}{h}\right) F_{\theta_0}(t) dt - \left[1 - K\left(\frac{z_2 - z}{h}\right)\right] \\
&= \int_{-1}^{1} \kappa(u) F_{\theta_0}(z + hu) du \\
&= F_{\theta_0}(z) + O(h^3),
\end{aligned}$$

where the last equality comes from a Taylor expansion of $F_{\theta_0}(z+hu)$ around z. This closes the proof. □

Define next

$$\hat{g}(z) \coloneqq \frac{1}{nh}\sum_{i=1}^{n}\kappa\left(\frac{z-Z_i}{h}\right)$$
$$\hat{r}(z) \coloneqq \frac{1}{nh}\sum_{i=1}^{n}\kappa\left(\frac{z-Z_i}{h}\right)X_i^\intercal.$$

Hence, $\hat{g}(z)$ is a kernel density estimator of $g(z)$ and $\hat{r}(z)$ is an estimator of $r(z) \coloneqq g(z)\mathbb{E}[X|Z=z]^\intercal$. Let $\hat{r}_k(z)$ denote the kth component of the vector $\hat{r}(z)$, $\hat{r}_k(z)$ then corresponds to the numerator of a Nadaraya-Watson estimator of X^k on Z. These objects will play a role in the asymptotic representation of $\nabla G_n(\theta_0)$. Indeed, we can write

$$\hat{D}(x, x^\intercal\theta_0) \coloneqq \nabla_{\theta_0}\hat{F}_\theta(x^\intercal\theta) = x\hat{g}(x^\intercal\theta_0) - \hat{r}(x^\intercal\theta_0)$$
$$D(x, x^\intercal\theta_0) \coloneqq \nabla_{\theta_0}F_\theta(x^\intercal\theta) = xg(x^\intercal\theta_0) - r(x^\intercal\theta_0).$$

Notice that $\sup_{z\in\mathcal{Z}}|\hat{g}(z)-g(z)| = o_p(1)$ by (3.7.2).

Lemma 3.3. *Assume (RC1)-(RC4). Then, for any* $0<\delta<1$,

$$\sup_{z\in\mathcal{Z}}|\hat{g}'(z)-g'(z)| = o_p(1) \tag{3.7.5}$$
$$\sup_{z,z'\in\mathcal{Z}}\frac{|\hat{g}'(z)-g'(z)-\hat{g}'(z')+g'(z')|}{|z-z'|^\delta} = O_p(1) \tag{3.7.6}$$

Proof. Relation (3.7.5) is proven in Theorem C from Silverman (1978). The proof of (3.7.6) follows that of (3.7.3) and uses again Theorem C from Silverman (1978), applied to the second derivative of $\hat{g}(z)$. □

Lemma 3.4. *Assume (RC1)-(RC4). Then, for any* $0<\delta<1$,

$$\sup_{z\in\mathcal{Z}}|\hat{r}(z)-r(z)| = o_p(1) \tag{3.7.7}$$
$$\sup_{z\in\mathcal{Z}}|\hat{r}'(z)-r'(z)| = o_p(1) \tag{3.7.8}$$
$$\sup_{z,z'\in\mathcal{Z}}\frac{|\hat{r}'(z)-r'(z)-\hat{r}'(z')+r'(z')|}{|z-z'|^\delta} = O_p(1) \tag{3.7.9}$$

Proof. We start with the proof of (3.7.7). Write $|\hat{r}(z)-r(z)| \leq |\hat{r}(z)-\mathbb{E}[\hat{r}(z)]|+|\mathbb{E}[\hat{r}(z)]-r(z)|$. Using Proposition 4 from Mack and Silverman (1982), we have that $\sup_{z\in\mathcal{Z}}|\hat{r}(z)-$

$\mathbb{E}[\hat{r}(z)]|= o_p(1)$. Also,

$$\begin{aligned}\mathbb{E}[\hat{r}(z)] &= \frac{1}{h}\mathbb{E}\left[\kappa\left(\frac{z-Z}{h}\right)X\right] \\ &= \frac{1}{h}\int_{z_1}^{z_2}\kappa\left(\frac{z-t}{h}\right)\mathbb{E}[X|Z=t]g(t)dt \\ &= \frac{1}{h}\int_{z-h}^{z+h}\kappa\left(\frac{t-z}{h}\right)r(t)dt \\ &= \int_{-1}^{1}\kappa(u)r(z+hu)du \\ &= r(z)+O(h^3),\end{aligned}$$

where the last step uses a Taylor expansion of $r(z+hu)$ around z and the fact that $r(\cdot)$ is three times continuously differentiable. This closes the proof of (3.7.7). The same structure can be applied to the proof of (3.7.8). First, $\sup_{z\in\mathcal{Z}}|\hat{r}'(z) - \mathbb{E}[\hat{r}'(z)]|= o_p(1)$ follows as a special case of Theorem 2 from Mack and Müller (1989). Then,

$$\begin{aligned}\mathbb{E}[\hat{r}'(z)] &= \frac{1}{h}\int_{-1}^{1}\kappa'(u)r(z+hu)du \\ &= \int_{-1}^{1}\kappa(u)r'(z+hu)du \\ &= r'(z)+O(h^3)\end{aligned}$$

where the second equality uses integration by parts and the fact that $\kappa(-1) = \kappa(1) = 0$. The last equality is obtained using a Taylor expansion of $r'(z+hu)$ around z and the fact that $r(\cdot)$ is four times continuously differentiable. This finishes the proof of (3.7.8). The proof of (3.7.9) follows that of (3.7.3) and uses again Theorem 2 from Mack and Müller (1989), applied to the second derivative of $\hat{r}(z)$. □

Proposition 3.4. *Assume (RC1)-(RC4). Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i\hat{D}(X_i,Z_i) - Y_iD(X_i,Z_i) - \mathbb{E}[Y\hat{D}(X,Z)|\mathcal{X}_n] + \mathbb{E}[YD(X,Z)]\right\} = o_p(n^{-1/2}).$$

Proof. As in the proof of Proposition 3.2, we use Lemma 19.24 from van der Vaart (1998), now with

$$\begin{aligned}\hat{f}_n &: (x,z,y) \mapsto y\hat{D}(x,z) = y[xd_n^1(z) + d_n^2(z)] + yD(x,z) \\ f_0 &: (x,z,y) \mapsto yD(x,z),\end{aligned}$$

where $d_n^1(z) \coloneqq \hat{g}(z) - g(z)$ and $d_n^2(z) \coloneqq \hat{r}(z) - r(z)$. Define the class

$$\mathcal{F} \coloneqq \{(x,z,y) \to y[xd^1(z) + d^2(z) + D(x,z)], (d^1,d^2) \in C_1^{1+\delta}(\mathcal{Z})\}.$$

From Lemmas 3.2, 3.3 and 3.4, we have that $P(d_n^1 \in C_1^{1+\delta}(\mathcal{Z})) \to 1$ and $P(d_n^2 \in C_1^{1+\delta}(\mathcal{Z})) \to 1$ as $n \to \infty$. With this in mind, the proof that $\mathcal{F}$ is Donsker is the same as in Proposition 3.2. It remains to show that

$$\int\int y^2[xd_n^1(x^\intercal\theta_0) + d_n^2(x^\intercal\theta_0)]^2 dF_{Y,X}(y,x) = o_p(1).$$

This is a direct consequence of Lemmas 3.3, 3.4, and the fact that $\mathbb{E}[Y^2] < \infty$. □

Proposition 3.5. *Assume (RC1)-(RC4). Then,*

$$\frac{1}{n}\sum_{i=1}^n Y_i\hat{D}(X_i,Z_i) = \frac{1}{n}\sum_{i=1}^n \left\{ Y_i D(X_i,Z_i) + \int\int y\eta_1(X_i,x)dF_{Y,X}(y,x) \right\} + o_p(n^{-1/2}),$$

with

$$\eta_1(X_i,x) \coloneqq \frac{x}{h}\kappa\left(\frac{x^\intercal\theta_0 - X_i^\intercal\theta_0}{h}\right) - \frac{X_i}{h}\kappa\left(\frac{x^\intercal\theta_0 - X_i^\intercal\theta_0}{h}\right) - D(x,x^\intercal\theta_0).$$

Also, $\mathbb{E}[\eta_1(X_i,x)] = O(h^3)$.

Proof. Using Proposition 3.4, we have

$$\frac{1}{n}\sum_{i=1}^n Y_i\hat{D}(X_i,Z_i) = \frac{1}{n}\sum_{i=1}^n Y_i D(X_i,Z_i) + \mathbb{E}[Y[\hat{D}(X,Z) - D(X,Z)]|\mathcal{X}_n] + o_p(n^{-1/2}).$$

The expectation above can be written as

$$\frac{1}{n}\sum_{i=1}^n \int\int y\eta_1(X_i,x)dF_{Y,X}(y,x).$$

Hence, $\mathbb{E}[\eta_1(X_i,x)] = x^\intercal\mathbb{E}[\hat{g}(x^\intercal\theta_0)] - \mathbb{E}[\hat{r}(x^\intercal\theta_0)] - D(x,x^\intercal\theta_0)$. By Taylor expansions similar to that operated in the proof of Lemma 3.4, we have $\mathbb{E}[\hat{g}(z)] = g(z) + O(h^3)$ and $\mathbb{E}[\hat{r}(z)] = r(z) + O(h^3)$. This implies that $\mathbb{E}[\eta_1(X_i,x)] = O(h^3)$ and closes the proof. □

Denote by $\hat{D}^2(x,x^\intercal\theta_0) \coloneqq \nabla^2_{\theta_0}\hat{F}_\theta(x^\intercal\theta)$ and by $D^2(x,x^\intercal\theta_0) \coloneqq \nabla^2_{\theta_0}F_\theta(x^\intercal\theta)$.

Proposition 3.6. *Assume (RC1)-(RC4). If* $\mathbb{E}[Y^2] < \infty$*, then it holds that*

$$\frac{1}{n}\sum_{i=1}^n Y_i\hat{D}^2(X_i,Z_i) - \mathbb{E}[YD^2(X,Z)] = o_p(1). \tag{3.7.10}$$

Proof. The left-hand side of (3.7.10) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i D^2(X_i,Z_i) - \mathbb{E}[Y D^2(X,Z)]\} + \frac{1}{n}\sum_{i=1}^{n} Y_i[\hat{D}^2(X_i,Z_i) - D^2(X_i,Z_i)].$$

The first term is $o_p(1)$ by the weak law of large numbers since $\mathbb{E}[|Y D^2(X,Z)|] < \infty$. We now focus on the second term. By Markov's inequality, it is sufficient to show that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i[\hat{D}^2(X_i,Z_i) - D^2(X_i,Z_i)]\right|\right] \to 0,$$

as $n \to \infty$. Since $\mathbb{E}[Y^2] < \infty$, it is enough to show that

$$\sup_{x\in\mathcal{X}} |\hat{D}^2(x, x^\intercal\theta_0) - D^2(x, x^\intercal\theta_0)| = o_p(1).$$

Notice that

$$\hat{D}^2(x, x^\intercal\theta_0) = \frac{1}{nh^2}\sum_{i=1}^{n} \kappa'\left(\frac{x^\intercal\theta_0 - X_i^\intercal\theta_0}{h}\right)[x - X_i][x - X_i]^\intercal.$$

The $[k,l]$ element of this matrix can be decomposed into $x^k x^l \hat{g}'(x^\intercal\theta_0) - x^k \hat{r}_l'(x^\intercal\theta_0) - x^l \hat{r}_k'(x^\intercal\theta_0) + \hat{m}_{kl}'(x^\intercal\theta_0)$, where

$$\hat{m}_{kl}(z) \coloneqq \frac{1}{nh}\sum_{i=1}^{n} \kappa\left(\frac{z - Z_i}{h}\right) X_i^k X_i^l.$$

Similarly, we can decompose the $[k,l]$ element of $D^2(x, x^\intercal\theta_0)$ into $x^k x^l g'(x^\intercal\theta_0) - x^k r_l'(x^\intercal\theta_0) - x^l r_k'(x^\intercal\theta_0) + m_{kl}'(x^\intercal\theta_0)$, where

$$m_{kl}(z) \coloneqq g(z)\mathbb{E}[X^k X^l | Z = z].$$

Using Lemmas 3.3 and 3.4, we are left with showing that $\sup_{z\in\mathcal{Z}} |\hat{m}_{kl}'(z) - m_{kl}'(z)| = o_p(1)$ for any k, l, which is proven in the same way as (3.7.8). This closes the proof. □

Proof of Theorem 3.1

Let $\alpha_n \coloneqq 1/\sqrt{n} + a_n$ and $J_n(\theta) \coloneqq G_n(\theta) - \sum_{k=1}^{p} p_\lambda(|\theta_k|)$. Following Lin and Peng (2013), we show that for any given $\epsilon > 0$, there exists a constant C such that

$$P\left(\sup_{||u||=1, u^\intercal\theta_0=0} J_n\Big((1 - C^2\alpha_n^2)^{1/2}\theta_0 + C\alpha_n u\Big) < J_n(\theta_0)\right) \geq 1 - \epsilon.$$

Let $\theta^* := (1 - C^2\alpha_n^2)^{1/2}\theta_0 + C\alpha_n u$. Since $p_\lambda(|\theta_{0,k}|) = 0$ for $k \in \{s+1, \ldots, d\}$, it holds

$$J_n(\theta^*) - J_n(\theta_0) \leq G_n(\theta^*) - G_n(\theta_0) - \sum_{k=1}^{s} \Big[p_\lambda(|\theta_k^*|) - p_\lambda(|\theta_{0,k}|) \Big].$$

After Taylor expansions of $G_n(\theta^*)$ around θ_0 and of $p_\lambda(|\theta_k^*|)$ around $|\theta_{0,k}|$, we have

$$J_n(\theta^*) - J_n(\theta_0) \leq I_1 + I_2 + I_3,$$

with

$$\begin{aligned}
I_1 &= [\nabla G_n(\theta_0)]^\intercal [\theta^* - \theta_0] \\
I_2 &= \frac{1}{2}[\theta^* - \theta_0]^\intercal \nabla^2 G_n(\theta_0)[\theta^* - \theta_0][1 + o_p(1)] \\
I_3 &= -\sum_{k=1}^{s} \Big[p'_\lambda(|\theta_{0,k}|)\text{sign}(\theta_{0,k})[\theta_k^* - \theta_{0,k}] + \frac{1}{2} p''_\lambda(|\theta_{0,k}|)[\theta_k^* - \theta_{0,k}]^2[1 + o(1)] \Big].
\end{aligned}$$

We first focus on I_1. Using Proposition 3.5, the weak law of large numbers and $\mathbb{E}[|YD^2(X,Z)|] < \infty$, we have that $\nabla G_n(\theta_0) = O_p(n^{-1/2})$. Using the definition of θ^* and a Taylor expansion of $(1 - C^2\alpha_n^2)^{\frac{1}{2}}$ around 1, we conclude that $I_1 = O_p(C\alpha_n/\sqrt{n})$. Let us now consider I_2. Using Proposition 3.6, $\nabla^2 G_n(\theta_0) = O_p(1)$. Hence, $I_2 = O_p(C^2\alpha_n^2)$. Finally, I_3 is bounded above by

$$s\alpha_n C a_n + s\alpha_n^2 C^2 \max\{|p''_\lambda(|\theta_{0,k}|)|\colon \theta_{0,k} \neq 0\}.$$

Hence, choosing a sufficiently large C, I_2 is the dominating term. The fact that it is asymptotically negative stems from (RC7). □

Proof of Theorem 3.2

Take any θ^A such that $||\theta^A - \theta_0^A|| = O_p(1/\sqrt{n})$ and $||\theta^A|| = 1$, and take any constant C. We show that

$$J_n((\theta^{A\intercal}, 0^\intercal)^\intercal) = \max_{||\theta^I|| \leq C/\sqrt{n}} J_n((\theta^{A\intercal}, \theta^{I\intercal})^\intercal).$$

It is sufficient to show that for some small $\epsilon_n = C/\sqrt{n}$ and for $k = s+1, \ldots, p$, we have

$$\begin{aligned}
\nabla_k J_n(\theta) &< 0 \qquad \text{for } 0 < \theta_k < \epsilon_n \\
&> 0 \qquad \text{for } -\epsilon_n < \theta_k < 0.
\end{aligned}$$

We have

$$\begin{aligned}
\nabla_k J_n(\theta) &= \nabla_k G_n(\theta) - p'_\lambda(|\theta_k|)\text{sign}(\theta_k) \\
&= O_p(n^{-1/2}) - p'_\lambda(|\theta_k|)\text{sign}(\theta_k),
\end{aligned}$$

using a Taylor expansion of $\nabla_k G_n(\theta)$ around θ_0 as well as Propositions 3.5 and 3.6. Hence,

$$\nabla_k J_n(\theta) = \lambda\Big[- \frac{p'_\lambda(|\theta_k|)}{\lambda}\text{sign}(\theta_k) + O_p\Big(\frac{1}{\sqrt{n}\lambda}\Big)\Big].$$

Recall that $\liminf_{n\to\infty} \liminf_{x\to 0^+} p'_\lambda(x)/\lambda > 0$. Hence, asymptotically, the sign of $\nabla_k J_n(\theta)$ is fully determined by the sign of θ_k. This closes the first part of this proof. Now we move on with the proof of the asymptotic normality. Since $\theta_{0,s} > 0$, it is easy to show that there exists $(\hat\theta_1, \ldots, \hat\theta_{s-1})^\intercal$ that is a $\sqrt{n}$-consistent local solution of the following maximization programme

$$\max_{(\theta_1,\ldots,\theta_{s-1})} J_n\left(\left(\theta_1, \ldots, \theta_{s-1}, \sqrt{1 - \sum_{k=1}^{s-1}\theta_k^2}, 0^\intercal\right)^\intercal\right).$$

Let $\hat\theta_s = \sqrt{1 - \sum_{k=1}^{s-1}\hat\theta_k^2}$, $\hat\theta^A = (\hat\theta_1, \ldots, \hat\theta_s)$ and $\hat\theta = (\hat\theta^{A\intercal}, 0^\intercal)^\intercal$. For all $k = 1, \ldots, s-1$, $\hat\theta_k$ must satisfy

$$\frac{\partial}{\partial\hat\theta_k} J_n\left(\left(\hat\theta_1, \ldots, \hat\theta_{s-1}, \sqrt{1 - \sum_{k=1}^{s-1}\hat\theta_k^2}, 0^\intercal\right)^\intercal\right) = 0.$$

We develop this derivative, starting with the part related to the non-penalized objective function. We have

$$\begin{aligned}\frac{\partial}{\partial\hat\theta_k} G_n\left(\left(\hat\theta_1, \ldots, \hat\theta_{s-1}, \sqrt{1 - \sum_{k=1}^{s-1}\hat\theta_k^2}, 0^\intercal\right)^\intercal\right) &= \frac{1}{n}\sum_{i=1}^n Y_i\left[\frac{1}{n}\sum_{j=1}^n \kappa\Big(\frac{X_i^\intercal\hat\theta - X_j^\intercal\hat\theta}{h}\Big)\frac{[X_i^k - X_j^k]}{h}\right] \\ &\quad - \frac{\hat\theta_k}{\hat\theta_s}\frac{1}{n}\sum_{i=1}^n Y_i\left[\frac{1}{n}\sum_{j=1}^n \kappa\Big(\frac{X_i^\intercal\hat\theta - X_j^\intercal\hat\theta}{h}\Big)\frac{[X_i^s - X_j^s]}{h}\right].\end{aligned}$$

Using a Taylor expansion on the function $(\hat\theta_1, \ldots, \hat\theta_{s-1})^\intercal \mapsto \kappa(\frac{X_i^\intercal\hat\theta - X_j^\intercal\hat\theta}{h})$ around $(\theta_{0,1}, \ldots, \theta_{0,s-1})^\intercal$ and the previous notations, the partial derivative above can be written as

$$\begin{aligned}\nabla_k G_n(\theta_0) - \frac{\hat\theta_k}{\hat\theta_s}\nabla_s G_n(\theta_0) + \sum_{l=1}^{s-1}\nabla^2_{kl} G_n(\theta_0)[\hat\theta_l - \theta_{0,l}] - \sum_{l=1}^{s-1}\frac{\theta_{0,l}}{\theta_{0,s}}\nabla^2_{ks} G_n(\theta_0)[\hat\theta_l - \theta_{0,l}] \\ - \frac{\hat\theta_k}{\hat\theta_s}\sum_{l=1}^{s-1}\nabla^2_{ls} G_n(\theta_0)[\hat\theta_l - \theta_{0,l}] + \frac{\hat\theta_k}{\hat\theta_s}\sum_{l=1}^{s-1}\frac{\theta_{0,l}}{\theta_{0,s}}\nabla^2_{ss} G_n(\theta_0)[\hat\theta_l - \theta_{0,l}] + o_p(1).\end{aligned}$$

We focus now on the part related to the penalty. Using the consistency of $\hat{\theta}_k$, we have

$$\frac{\partial}{\partial \hat{\theta}_k}\left\{\sum_{k=1}^{s-1} p_\lambda(|\hat{\theta}_k|) + p_\lambda(\sqrt{1-\sum_{k=1}^{s-1}\hat{\theta}_k^2})\right\} = p'_\lambda(|\hat{\theta}_k|)\text{sign}(\hat{\theta}_k)\mathbb{1}\{\hat{\theta}_k \neq 0\} \\ - p'_\lambda(\hat{\theta}_s)\frac{\hat{\theta}_k}{\hat{\theta}_s} + o_p(1).$$

Using again the consistency of $\hat{\theta}_k$, Taylor expansions of $p'_\lambda(|\hat{\theta}_k|)$ around $\theta_{0,k}$ and of $p'_\lambda(\hat{\theta}_s)$ around $\theta_{0,s}$, the partial derivative above writes as

$$\left[p'_\lambda(|\theta_{0,k}|)\text{sign}(\theta_{0,k}) - p'_\lambda(\theta_{0,s})\frac{\hat{\theta}_k}{\hat{\theta}_s} + p''_\lambda(|\theta_{0,k}|)[\hat{\theta}_k - \theta_{0,k}]\right]\mathbb{1}\{\hat{\theta}_k \neq 0\} \\ + \frac{\hat{\theta}_k}{\hat{\theta}_s}\sum_{l=1}^{s-1}\left[p''_\lambda(\theta_{0,s})\frac{\theta_{0,l}}{\theta_{0,s}}\right](\hat{\theta}_l - \theta_{0,l}) + o_p(1).$$

Finally, we can get rid of the indicator since $\mathbb{1}\{\hat{\theta}_k \neq 0\}$ converges in probability to one. Notice that in the case of the SCAD, the penalty is only piecewise differentiable. However, this is not an issue since $\lambda \to 0$ as $n \to \infty$. Hence, for a sufficiently large n, the value of $\theta_{0,k}$ will not lie at a discontinuity point. In what follows, we adopt the subsequent notations. For a vector v, we denote by $\tilde{v}$ the vector obtained by taking the first $s-1$ elements of v. For a matrix M, we denote by $\tilde{M}$ the matrix formed by taking the first $s-1$ rows and columns of M. Also, we denote by $\tilde{M}_{.,s}$ the vector formed by taking the first $s-1$ rows and fixing the s-th column of M. Define

$\Sigma_1 \coloneqq \mathbb{E}[Y\tilde{D}^2(X,Z)]$.

$\Sigma_2 \coloneqq \mathbb{E}[Y\tilde{D}^2_{.,s}(X,Z)]\frac{\tilde{\theta}_0^\intercal}{\theta_{0,s}}$.

$\Sigma_3 \coloneqq \frac{\tilde{\theta}_0\tilde{\theta}_0^\intercal}{\theta_{0,s}^2}\mathbb{E}[YD^2_{ss}(X,Z)]$.

Σ_4 is an $(s-1)\times(s-1)$ diagonal matrix where the diagonal is given by $[p''_\lambda(|\theta_{0,1}|)\text{sign}(\theta_{0,1}), \ldots, p''_\lambda(|\theta_{0,s-1}|)\text{sign}(\theta_{0,s-1})]^\intercal$.

$\Sigma_5 \coloneqq \frac{\tilde{\theta}_0\tilde{\theta}_0^\intercal}{\theta_{0,s}^2}p''_\lambda(\theta_{0,s})$.

$\Omega_1 \coloneqq \mathbb{E}[\tilde{\xi}_i\tilde{\xi}_i^\intercal]$.

$\Omega_2 \coloneqq \frac{\tilde{\theta}_0\tilde{\theta}_0^\intercal}{\theta_{0,s}^2}\mathbb{E}[(\xi_i^s)^2]$.

$\Omega_3 \coloneqq \mathbb{E}[\tilde{\xi}_i\xi_i^s]\frac{\tilde{\theta}_0^\intercal}{\theta_{0,s}}$

where

$$\xi_i \coloneqq Y_i D(X_i, Z_i) + \int\int y\eta_1(X_i, x) dF_{Y,X}(y, x).$$

Also, let

$$\Sigma \coloneqq -\Sigma_1 + \Sigma_2 + \Sigma_2^\intercal - \Sigma_3 + \Sigma_4 - \Sigma_5 \tag{3.7.11}$$
$$\Omega \coloneqq \Omega_1 + \Omega_2 - \Omega_3 - \Omega_3^\intercal. \tag{3.7.12}$$

In order to avoid any confusion with the notations of Section 2.4, we define $\vartheta_0 = \tilde{\theta}_0$ and $\hat{\vartheta} = \hat{\tilde{\theta}}$. Recall that b is defined in Equation (3.2.3). Then, we have

$$\begin{aligned}\sqrt{n}\Sigma[\hat{\vartheta} - \vartheta_0 + \Sigma^{-1}b] &= \sqrt{n}\tilde{\nabla}G_n(\theta_0) - \frac{\vartheta_0}{\theta_{0,s}}\sqrt{n}\nabla_s G_n(\theta_0) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\tilde{\xi}_i - \frac{\vartheta_0}{\theta_{0,s}}\xi_{is}\right) + o_p(1),\end{aligned}$$

where the last equality uses Proposition 3.5. Notice that $\mathbb{E}[Y_i D(X_i, Z_i)] = 0$ because the function $\theta \mapsto \mathbb{E}[Y F_\theta(X^\intercal\theta)]$ is maximized in θ_0. Hence, using Proposition 3.5 and the central limit theorem, we have

$$\sqrt{n}\Sigma[\hat{\vartheta} - \vartheta_0 + \Sigma^{-1}b] \xrightarrow{d} N(0, \Omega),$$

which concludes the proof. □

Proposition 3.7. *Assume (RC1)-(RC7). Then,*

$$\frac{1}{n}\sum_{i=1}^n \Big[\mathbb{1}\{X_i^\intercal\theta_n \le x^\intercal\theta_n\} - \mathbb{1}\{X_i^\intercal\theta_0 \le x^\intercal\theta_0\} - P(X^\intercal\theta_n \le x^\intercal\theta_n|\mathcal{X}_n) + P(X^\intercal\theta_0 \le x^\intercal\theta_0)\Big] = o_p(n^{-1/2})$$

uniformly in $x \in \mathcal{X}$ and in $\theta_n \xrightarrow{p} \theta_0$, with $\theta_{n,s} > 0$.

Proof. In this proof, we will make use of Corollary 2.3.12 in van der Vaart and Wellner (1996). We first show that $f_n(X_i) \coloneqq \mathbb{1}\{X_i^\intercal\theta_n \le x^\intercal\theta_n\} - \mathbb{1}\{X_i^\intercal\theta_0 \le x^\intercal\theta_0\} - P(X_i^\intercal\theta_n \le x^\intercal\theta_n|\mathcal{X}_n) + P(X_i^\intercal\theta_0 \le x^\intercal\theta_0)$ belongs to a Donsker class. Define $\theta_0^C = (\frac{\theta_{0,1}}{\theta_{0,s}}, \dots, \frac{\theta_{0,s-1}}{\theta_{0,s}}, \frac{\theta_{0,s+1}}{\theta_{0,s}}, \dots, \frac{\theta_{0,p}}{\theta_{0,s}})^\intercal$ and $\theta_n^C = (\frac{\theta_{n,1}}{\theta_{n,s}}, \dots, \frac{\theta_{n,s-1}}{\theta_{n,s}}, \frac{\theta_{n,s+1}}{\theta_{n,s}}, \dots, \frac{\theta_{n,p}}{\theta_{n,s}})^\intercal$, where $\theta_n \xrightarrow{p} \theta_0$. Also, let $x^c = (x^1, \dots, x^{s-1}, x^{s+1}, \dots, x^p)^\intercal$. Notice that the first term of $f(X_i)$ can be rewritten as $\mathbb{1}\{X_i^s \le x^\intercal\theta_n^C - X_i^{C\intercal}\theta_n^C\}$. Define the class

$$\mathcal{F} \coloneqq \Big\{x \mapsto [\mathbb{1}\{x^s \le t - x^{C\intercal}\theta^C\} - \mathbb{1}\{x^s \le u - x^{C\intercal}\theta_0^C\} - P(X^s \le t - X^{C\intercal}\theta^C) + P(X^s \le u - x^{C\intercal}\theta_0^C)], t \in \mathbb{R}, u \in \mathbb{R}, \theta^C \in \Theta^C\Big\},$$

where θ^C is a vector of size $(p-1)$ defined on a compact set Θ^C. We prove that $\mathcal{F}$ is Donsker using the same reasoning as in Proposition 3.2. We focus on the class $\mathcal{F}_1 \coloneqq \{x \mapsto \mathbb{1}\{x^s \leq t - (x^c)^\intercal\theta^c\}, \theta^c \in \Theta^c, t \in \mathbb{R}\}$. The other terms can be dealt with in a similar way. Embed θ^C into a hypercube $[\theta_1^l, \theta_1^u] \times \ldots \times [\theta_{p-1}^l, \theta_{p-1}^u]$ of dimension $(p-1)$. For all j, partition $[\theta_j^l, \theta_j^u]$ into $O(\epsilon^{-2})$ intervals of length $O(\epsilon^2)$. Hence, we partitioned θ^C into $O(\epsilon^{-2(p-1)})$ hypercubes. Denote by R_i one such hypercube. For each non-empty R_i, let

$$\Gamma_i^l(X^C) \coloneqq \min_{\theta^C \in (R_i \cap \Theta^C)} (X^c)^\intercal\theta^C$$
$$\Gamma_i^u(X^C) \coloneqq \max_{\theta^C \in (R_i \cap \Theta^C)} (X^c)^\intercal\theta^C.$$

Now, notice that, for all t, there exists an i such that

$$\mathbb{1}\{X^s \leq t - \Gamma_i^u(X^C)\} \leq \mathbb{1}\{X^s \leq t - (X^c)^\intercal\theta^C\} \leq \mathbb{1}\{X^s \leq t - \Gamma_i^l(X^C)\}.$$

Define $P_i^u(t) \coloneqq P(X^s \leq t - \Gamma_i^u(X^C))$ and partition the line into segments t_{ik}^u, with $k = 1, \ldots, O(\epsilon^{-2})$, and with P_i^u-probability less than a fraction of ϵ^2. Similarly, define $P_i^l(t) \coloneqq P(X^s \leq t - \Gamma_i^l(X^C))$ and partition the line into segments t_{ik}^l, with $k = 1, \ldots, O(\epsilon^{-2})$, and with P_i^l-probability less than a fraction of ϵ^2. Denote by $t_{ik_1}^l$ the largest $t_{ik}^l \leq t$ and denote by $t_{ik_2}^l$ the smallest $t_{ik}^u \geq t$. Brackets for $\mathbb{1}\{X^s \leq t - (X^c)^\intercal\theta^C\}$ are then given by $[\mathbb{1}\{X^s \leq t_{ik_1}^l - \Gamma_i^u(X^C)\}, \mathbb{1}\{X^s \leq t_{ik_2}^u - \Gamma_i^l(X^C)\}]$. Let us now compute their size. Denoting $||\cdot||_{P_X}$ the $L2(P_X)$-norm, we have

$$\begin{aligned}
&||\mathbb{1}\{X^s \leq t_{ik_2}^u - \Gamma_i^l(X^C)\} - \mathbb{1}\{X^s \leq t_{ik_1}^l - \Gamma_i^u(X^C)\}||_{P_X}^2 \\
&\quad = P(X^s \leq t_{ik_2}^u - \Gamma_i^l(X^C)) - P(X^s \leq t_{ik_1}^l - \Gamma_i^u(X^C)) \\
&\quad = P_i^l(t) - P_i^u(t) + O(\epsilon^2) \\
&\quad = \int \left[F_{X^s|X^C}(t - \Gamma_i^l(x^C)|x^C) - F_{X^s|X^C}(t - \Gamma_i^u(x^C)|x^C)] dF_{X^C}(x^C) \right] + O(\epsilon^2) \\
&\quad = \int \left[\sup_{x^s, x^C} f_{X^1|X^C}(x^s|x^C)\epsilon^2 \right] dF_{X^C}(x^C) + O(\epsilon^2) = O(\epsilon^2),
\end{aligned}$$

where $f_{X^s|X^C}$ and $F_{X^s|X^C}$ denote the conditional density and the conditional CDF of X^s with respect to X^C, and F_{X^C} is the CDF of X^C. Hence, the brackets are of size $O(\epsilon)$. In conclusion, the bracketing number associated to $\mathcal{F}_1$ is $O(\epsilon^{-2p})$. Using Theorem 19.5 in van der Vaart (1998), it follows that $\mathcal{F}_1$ is Donsker. By straightforward computations, $\mathcal{F}$ is also Donsker. We have

$$\begin{aligned}
\mathbb{V}[f_n(X)|\mathcal{X}_n] &= \mathbb{V}[\mathbb{1}\{X^\intercal\theta_n \leq x^\intercal\theta_n\} - \mathbb{1}\{X^\intercal\theta_0 \leq x^\intercal\theta_0\}|\mathcal{X}_n] \\
&\leq \mathbb{E}[(\mathbb{1}\{X^\intercal\theta_n \leq x^\intercal\theta_n\} - \mathbb{1}\{X^\intercal\theta_0 \leq x^\intercal\theta_0\})^2|\mathcal{X}_n] \\
&\leq \mathbb{E}[(\mathbb{1}\{X^\intercal\theta_n \leq x^\intercal\theta_n\} - \mathbb{1}\{X^\intercal\theta_0 \leq x^\intercal\theta_0\})|\mathcal{X}_n]^{2/3} \\
&= o_p(1),
\end{aligned}$$

where the last line uses the continuous mapping theorem and the fact that $\theta_n \xrightarrow{p} \theta_0$. Since $\mathcal{F}$ is Donsker, it follows from Corollary 2.3.12 in van der Vaart and Wellner (1996) that

$$\lim_{\alpha \downarrow 0} \limsup_{n \to \infty} P\left(\sup_{f \in \mathcal{F}, \mathbb{V}[f] < \alpha} n^{-1/2} \left| \sum_{i=1}^{n} f(X_i) \right| > \bar{\epsilon} \right) = 0,$$

for each $\bar{\epsilon} > 0$. We obtain the desired result by restricting the above probability to elements in $\mathcal{F}$ satisfying $\theta^C = \theta_n^C$. □

Proof of Theorem 3.3

We can rewrite

$$\widehat{\mathrm{Gi}}_{Y,X} = \frac{1}{\overline{Y}} \frac{2}{n} \sum_{i=1}^{n} Y_i \left[\frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j^\intercal \hat{\theta} \leq X_i^\intercal \hat{\theta}\} \right] - \frac{n+1}{n}$$

where $\hat{\theta}$ is the estimator of Theorem 3.2. Using Proposition 3.7 and the same notations as in Theorem 3.2, we can write

$$\frac{2}{\sqrt{n}} \sum_{i=1}^{n} Y_i \left[\frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j^\intercal \hat{\theta} \leq X_i^\intercal \hat{\theta}\} \right] = A_{n1} + A_{n2} + o_p(1),$$

with

$$\begin{aligned} A_{n1} &= \frac{2}{\sqrt{n}} \sum_{i=1}^{n} Y_i \left[\frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j^\intercal \theta_0 \leq X_i^\intercal \theta_0\} \right] \\ A_{n2} &= \frac{2}{n} \sum_{i=1}^{n} Y_i \int \sqrt{n} \left[\mathbb{1}\left\{ x^s \leq [\tilde{X}_i - \tilde{x}]^\intercal \frac{\hat{\vartheta}}{\hat{\theta}_s} + [X_i^I - x^I]^\intercal \frac{\hat{\theta}^I}{\hat{\theta}_s} + X_i^s \right\} \right. \\ &\qquad \left. - \mathbb{1}\left\{ x^s \leq [\tilde{X}_i - \tilde{x}]^\intercal \frac{\vartheta_0}{\theta_{0,s}} + X_i^s \right\} \right] dF_X(\tilde{x}, x^s), \end{aligned}$$

where we recall that $\tilde{X}_i$ is the vector obtained by taking the first $s-1$ elements of X_i. Let us first focus on A_{n2}. We can rewrite the integral displayed inside the sum as

$$\int \sqrt{n} \left[F_{X^s}\left([\tilde{X}_i - \tilde{x}]^\intercal \frac{\hat{\vartheta}}{\hat{\theta}_s} + [X_i^I - x^I]^\intercal \frac{\hat{\theta}^I}{\hat{\theta}_s} + X_i^s \right) - F_{X^s}\left([\tilde{X}_i - \tilde{x}]^\intercal \frac{\vartheta_0}{\theta_{0,s}} + X_i^s \right) \right] dF_{\tilde{X}}(\tilde{x}),$$

where F_{X^s} is the CDF of X^s and $F_{\tilde{X}}$ is the CDF of $\tilde{X}$. Using Theorem 3.1, a Taylor expansion on the piece inside the integral yields

$$f_{X^s}\left([\tilde{X}_i - \tilde{x}]^\intercal \frac{\vartheta_0}{\theta_{0,s}} + X_i^s \right) \left([\tilde{X}_i - \tilde{x}]^\intercal \sqrt{n} \left[\frac{\hat{\vartheta}}{\hat{\theta}_s} - \frac{\vartheta_0}{\theta_{0,s}} \right] + [X_i^I - x^I]^\intercal \sqrt{n} \frac{\hat{\theta}^I}{\hat{\theta}_s} \right) + o_p(1).$$

By Theorem 3.1 and the property of sparsity proven in Theorem 3.2, the part related to $\hat{\theta}^I$ is negligible. Using the same development as in Theorem 3.2, we have

$$\begin{aligned}\sqrt{n}\left[\frac{\hat{\vartheta}}{\hat{\theta}_s}-\frac{\vartheta_0}{\theta_{0,s}}\right] &= \frac{\sqrt{n}[\hat{\vartheta}-\vartheta_0]}{\theta_{0,s}}-\frac{\hat{\vartheta}}{\hat{\theta}_s}\frac{\sqrt{n}[\hat{\theta}_s-\theta_{0,s}]}{\theta_{0,s}} \\ &= \frac{1}{\theta_{0,s}^3}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{[\theta_{0,s}^2+\vartheta_0\vartheta_0^\intercal]\tilde{\xi}_i-\frac{\vartheta_0}{\theta_{0,s}}\xi_i^s\right\}+o_p(1).\end{aligned}$$

Hence,

$$A_{n2}=\frac{1}{\theta_{0,s}^3}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\mathbb{E}[\rho_i^\intercal]\left([\theta_{0,s}^2+\vartheta_0\vartheta_0^\intercal]\tilde{\xi}_i-\frac{\vartheta_0}{\theta_{0,s}}\xi_i^s\right)\right\}+o_p(1),$$

where

$$\rho_i=2Y_i\int f_{X^s}\left([\tilde{X}_i-\tilde{x}]^\intercal\frac{\vartheta_0}{\theta_{0,s}}+X_i^s\right)[\tilde{X}_i-\tilde{x}]dF_{\tilde{X}}(\tilde{x}).$$

We now focus on A_{n1}. We can write

$$A_{n1}=\sqrt{n}U_{n1}+o_p(1),$$

where

$$U_{n1}=\frac{2}{n(n-1)}\sum_{i<j}Y_i\mathbb{1}\{X_j^\intercal\theta_0\leq X_i^\intercal\theta_0\}+Y_j\mathbb{1}\{X_i^\intercal\theta_0\leq X_j^\intercal\theta_0\}$$

is a U-statistic. Let $m_H(t)\coloneqq\mathbb{E}[Y_i\mathbb{1}\{X_i^\intercal\theta_0\geq t\}]$. Using Theorem 12.3 in van der Vaart (1998) and the fact that $\mathbb{E}[Y_i^2]<\infty$, we have

$$\sqrt{n}[U_{n1}-2G(\theta_0)]=\frac{2}{n}\sum_{i=1}^{n}\left[Y_iF_{\theta_0}(X_i^\intercal\theta_0)+m_H(X_i^\intercal\theta_0)-2G(\theta_0)\right]+o_p(1),$$

and has mean zero. Using the last developments and the fact that $\mathbb{E}[Y_i]>0$, we can write

$$\sqrt{n}[\widehat{\mathrm{Gi}}_{Y,X}-\mathrm{Gi}_{Y,X}]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_i+o_p(1),$$

where

$$\zeta_i\coloneqq\frac{1}{\mathbb{E}[Y_i]}\left(2Y_iF_{\theta_0}(X_i^\intercal\theta_0)+2m_H(X_i^\intercal\theta_0)-4G(\theta_0)+\frac{\mathbb{E}[\rho_i^\intercal]}{\theta_{0,s}^3}\left([\theta_{0,s}^2+\vartheta_0\vartheta_0^\intercal]\tilde{\xi}_i-\frac{\vartheta_0}{\theta_{0,s}}\xi_i^s\right)\right). \tag{3.7.13}$$

Hence, $\sqrt{n}[\widehat{\mathrm{Gi}}_{Y,X}-\mathrm{Gi}_{Y,X}]\xrightarrow{d}N(0,\sigma_\zeta^2)$, with $\sigma_\zeta^2\coloneqq\mathbb{V}[\zeta_i]$. □

3.7.2 Properties of the SCAD-FABS algorithm

Proof of Lemma 3.1. Let $k \in \mathcal{A}^t$ be updated via a forward step. Then, it both holds that

$$L(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon) - L(\theta^t) - \epsilon p'_{\lambda^t}(|\theta_l^t|) \geq 0 \qquad \forall l \in \mathcal{A}^t \tag{3.7.14}$$

$$L(\theta^{t+1}) < L(\theta^t). \tag{3.7.15}$$

(3.7.14) holds because $t \to t+1$ is not a backward step and (3.7.15) comes from the fact that a forward step always decreases the loss. Now, assume per contra that $\theta_k^{t+1} = \theta_k^t - \text{sign}(\theta_k^t)\epsilon$. Equation (3.7.15) implies $L(\theta^t - \text{sign}(\theta_k^t)\mathbf{1}_k\epsilon) - L(\theta^t) < 0$, which contradicts (3.7.14). □

Lemma 3.5. *Consider that coefficient k is updated via a forward step.*

(a) If $\lambda^{t+1} = \lambda^t$ and $|\theta_k^t| \leq \lambda^t$, then $\lambda^t \leq \lambda_A^{t+1}$.

(b) If $\lambda^{t+1} = \lambda^t$ and $\lambda^t < |\theta_k^t| \leq a\lambda^t$, then $\lambda^t \leq \lambda_B^{t+1}$.

Proof. Recall that λ^t is updated according to (3.3.4). Also, λ_A^{t+1} and λ_B^{t+1} are determined by (3.3.6) and (3.3.7).

We start with the proof of part (a). Let $\lambda^{t+1} = \lambda^t$ and $|\theta_k^t| \leq \lambda^t$. We may face two scenarios

$$\lambda_A^{t+1} \geq \lambda_B^{t+1} \tag{3.7.16}$$

$$\lambda_B^{t+1} > \lambda_A^{t+1}. \tag{3.7.17}$$

In situation (3.7.16), $\lambda^t \leq \lambda_A^{t+1}$ holds trivially because $\lambda^t = \lambda^{t+1}$. Assume now that (3.7.17) holds. Using the definition of λ_A^{t+1} and λ_B^{t+1}, we have $\lambda_B^{t+1} < |\theta_k^t|$. Hence, we must have $\lambda_B^{t+1} < |\theta_k^t| \leq \lambda^t$, which contradicts $\lambda^t \leq \lambda_B^{t+1}$. Hence, (3.7.17) cannot happen, which closes the proof of part (a).

We move to the proof of part (b). Let $\lambda^{t+1} = \lambda^t$ and $\lambda^t < |\theta_k^t| \leq a\lambda^t$. Once again, we may face either (3.7.16) or (3.7.17). If (3.7.17) holds, then $\lambda^t \leq \lambda_B^{t+1}$ arises trivially from $\lambda^{t+1} = \lambda^t$. Assume now (3.7.16). Using the definition of λ_A^{t+1} and λ_B^{t+1}, we now have $\lambda_B^{t+1} \geq |\theta_k^t|$. Hence, we must have $\lambda^t < |\theta_k^t| \leq \lambda_B^{t+1}$, which closes the proof of part (b). □

Lemma 3.6. *Let $k \in \{1, \ldots, p\}$ be updated via a forward step and $\lambda > 0$. Then, there exists some $c_k^t \in [0, 1]$ such that*

$$Q(\theta^{t+1}, \lambda) - Q(\theta^t, \lambda) = L(\theta^{t+1}) - L(\theta^t) + p'_\lambda(|\theta_k^t|)\epsilon - \frac{c_k^t \epsilon^2}{2(a-1)}.$$

Proof. Let $\lambda > 0$ and consider that $k \in \{1, \dots, p\}$ is updated via a forward step. By Lemma 3.1, we have $\theta^{t+1} = \theta^t + \text{sign}(\theta_k^t)\mathbf{1}_k\epsilon$. We face the following situations:

$$|\theta_k^{t+1}| \leq \lambda \tag{3.7.18}$$
$$|\theta_k^t| \leq \lambda < |\theta_k^{t+1}| \leq a\lambda \tag{3.7.19}$$
$$\lambda < |\theta_k^t| < |\theta_k^{t+1}| \leq a\lambda \tag{3.7.20}$$
$$\lambda < |\theta_k^t| \leq a\lambda < |\theta_k^{t+1}| \tag{3.7.21}$$
$$|\theta_k^t| > a\lambda. \tag{3.7.22}$$

In situation (3.7.18), there is no approximation error since

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = \lambda\epsilon = p'_\lambda(|\theta_k^t|)\epsilon.$$

Let us move to situation (3.7.19). Then, we have

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = p'_\lambda(|\theta_k^t|)\epsilon - c_{k,1}^2 \frac{\epsilon^2}{2(a-1)},$$

where $c_{k,1} \coloneqq \frac{|\theta_k^{t+1}| - \lambda}{\epsilon} \in [0, 1]$. Consider now situation (3.7.20). Then, we have

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = p'_\lambda(|\theta_k^t|)\epsilon - \frac{\epsilon^2}{2(a-1)}.$$

In cases covered by (3.7.21), we have

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = p'_\lambda(|\theta_k^t|)\epsilon - (1 - c_{k,2}^2)\frac{\epsilon^2}{2(a-1)},$$

where $c_{k,2} \coloneqq \frac{|\theta_k^{t+1}| - a\lambda}{\epsilon} \in [0, 1]$. Obviously, there is no approximation error in (3.7.22) since

$$p_\lambda(|\theta_k^{t+1}|) - p_\lambda(|\theta_k^t|) = 0 = p'_\lambda(|\theta_k^t|)\epsilon.$$

□

Lemma 3.7. *Let $k \in \mathcal{A}^t$ be updated via a backward step and $\lambda > 0$. Then, there exists some $d_k^t \in [0, 1]$ such that*

$$Q(\theta^{t+1}, \lambda) - Q(\theta^t, \lambda) = L(\theta^{t+1}) - L(\theta^t) - p'_\lambda(|\theta_k^t|)\epsilon - \frac{d_k^t \epsilon^2}{2(a-1)}.$$

This result can be proven by the same reasoning as the one used in the proof of Lemma 3.6.

Lemma 3.8. *Let $k \in \{1, \dots, p\}$ be updated via a forward step. Then, $\forall l \neq k$ such that $L(\theta^t - sign(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) - L(\theta^t) < 0$, it holds*

$$L(\theta^{t+1}) - L(\theta^t - sign(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) + \left[p'_{\lambda^t}(|\theta_k^t|) - p'_{\lambda^t}(|\theta_l^t|)\right]\epsilon \leq \frac{e_k^t - e_l^t}{2}\epsilon^2,$$

where $e_k^t = \nabla^2_{kk} L(\dot{\theta})$, with $\dot{\theta}$ between θ^t and $\theta^t - \text{sign}(\nabla_m L(\theta^t))\mathbf{1}_k\epsilon$ and ∇^2 denotes the Hessian matrix.

Proof. Consider that $k \in \{1, \dots, p\}$ is updated via a forward step. Using a Taylor expansion, we have

$$L(\theta^{t+1}) - L(\theta^t) = -\nabla_k L(\theta^t)\text{sign}(\nabla_k L(\theta^t))\epsilon + \frac{e_k^t}{2}\epsilon^2.$$

This implies that

$$[|\nabla_k L(\theta^t)| - p'_{\lambda^t}(|\theta_k^t|)]\epsilon = L(\theta^t) - L(\theta^{t+1}) - p'_{\lambda^t}(|\theta_k^t|)\epsilon + \frac{e_k^t}{2}\epsilon^2.$$

Remark that, in a forward step, the left hand side of the previous equation is maximized on the set $\{l \in \{1, \dots, p\} : L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) - L(\theta^t) < 0\}$. Hence, for such coefficients, it holds that

$$L(\theta^t) - L(\theta^t - \text{sign}(\nabla_l L(\theta^t))) - p'_{\lambda^t}(|\theta_k^t|)\epsilon + \frac{e_l^t}{2}\epsilon^2 \leq L(\theta^t) - L(\theta^{t+1}) - p'_{\lambda^t}(|\theta_k^t|)\epsilon + \frac{e_k^t}{2}\epsilon^2,$$

which leads us to the desired result. □

Proof of Proposition 3.1

Let k be the index of the updated coefficient. We first show that, in a backward step, it holds

$$Q(\theta^{t+1}, \lambda^{t+1}) < Q(\theta^t, \lambda^t) - \frac{d_k^t \epsilon^2}{2(a-1)}.$$

Since we are considering a backward step, it both holds $\lambda^{t+1} = \lambda^t$ and

$$L(\theta^{t+1}) - L(\theta^t) - \epsilon p'_{\lambda^t}(|\theta_k^t|) < 0.$$

Using Lemma 3.7, this implies

$$Q(\theta^{t+1}, \lambda^{t+1}) - Q(\theta^t, \lambda^t) < -d_k^t \frac{\epsilon^2}{2(a-1)}.$$

We now prove that, in a forward step, it holds

$$Q(\theta^{t+1}, \lambda^{t+1}) \leq Q(\theta^t, \lambda^t) - \frac{c_k^t \epsilon^2}{2(a-1)}. \tag{3.7.23}$$

Assume first that $\lambda^{t+1} = \lambda_A^{t+1}$. Recall that $\lambda_A^{t+1} = L_\epsilon^{t,t+1}$ and $|\theta_k^t| \leq \lambda_A^{t+1}$. Using Lemma 3.6, it holds

$$Q(\theta^{t+1}, \lambda_A^{t+1}) - Q(\theta^t, \lambda_A^{t+1}) = -\frac{c_k^t \epsilon^2}{2(a-1)}.$$

Since $\lambda_A^{t+1} \leq \lambda^t$, we also have $Q(\theta^t, \lambda_A^{t+1}) \leq Q(\theta^t, \lambda^t)$. Taken together, the last two results yield (3.7.23). Second, consider that $\lambda^{t+1} = \lambda_B^{t+1}$. Recall that $\lambda_B^{t+1} = \frac{1}{a}[(a-1)L_\epsilon^{t,t+1} + |\theta_k^t|]$ and $|\theta_k^t| > \lambda_B^{t+1}$. Using the same reasoning as in the last point yields (3.7.23). Third, we assume that $\lambda^{t+1} = \lambda^t$ and $|\theta_k^t| \leq \lambda^t$. Using Lemma 3.6, it holds

$$\begin{aligned} Q(\theta^{t+1}, \lambda^t) - Q(\theta^t, \lambda^t) &= L(\theta^{t+1}) - L(\theta^t) + \epsilon\lambda^t - \frac{c_k^t \epsilon^2}{2(a-1)} \\ &\leq L(\theta^{t+1}) - L(\theta^t) + \epsilon\lambda_A^{t+1} - \frac{c_k^t \epsilon^2}{2(a-1)} \\ &= -\frac{c_k^t \epsilon^2}{2(a-1)}, \end{aligned}$$

where the inequality in the second line is due to Lemma 3.5. Fourth, let $\lambda^{t+1} = \lambda^t$ and $\lambda^t < |\theta_k^t| \leq a\lambda^t$. Relation (3.7.23) is proven similarly to the last situation. Finally, assume $|\theta_k^t| > a\lambda^t$. The result emerges using Lemma 3.6 and the fact that a forward step never increases the loss. □

Proposition 3.8. *If $\lambda^{t+1} < \lambda^t$ and k is the index of the updated coefficient, then $\forall l \in \mathcal{A}^t$, it holds*

$$\begin{aligned} Q(\theta^t, \lambda^t) &\leq Q(\theta^t + \operatorname{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{\epsilon^2}{2}\left[\frac{c_l^t}{a-1} + (e_k^t - e_l^t)\right] \\ Q(\theta^t, \lambda^t) &\leq Q(\theta^t - \operatorname{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{d_l^t \epsilon^2}{2(a-1)}. \end{aligned}$$

Proof. Let $\lambda^{t+1} < \lambda^t$ and k be the index of the updated coefficient. Consider any $l \in \mathcal{A}^t$. We first show that

$$Q(\theta^t, \lambda^t) \leq Q(\theta^t + \operatorname{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{\epsilon^2}{2}\left[\frac{c_l^t}{a-1} + (e_k^t - e_l^t)\right]. \tag{3.7.24}$$

We can safely restrict to the set $\{l \in \mathcal{A}^t : L(\theta^t + \operatorname{sign}(\theta_l^t)\mathbf{1}_l\epsilon) - L(\theta^t) < 0\}$. Indeed, on the complementary set, (3.7.24) is obviously true since θ^t leads to lower penalty and loss values. Write $\tilde{\theta} \coloneqq \theta^t + \operatorname{sign}(\theta_l^t)\mathbf{1}_l\epsilon$. We prove the result by showing that it both holds

$$Q(\theta^t, \lambda^t) \leq Q(\theta^{t+1}, \lambda^t) + \frac{c_k^t \epsilon^2}{2(a-1)} \tag{3.7.25}$$

$$Q(\theta^{t+1}, \lambda^t) - Q(\tilde{\theta}, \lambda^t) \leq \frac{\epsilon^2}{2}\left(\frac{(c_l^t - c_k^t)}{a-1} + (e_k^t - e_l^t)\right). \tag{3.7.26}$$

Let us first focus on (3.7.25). Suppose first that $\lambda^{t+1} = \lambda_A^{t+1}$. In that case, $\lambda^t \geq \lambda_A^{t+1} \geq |\theta_k^t|$. Hence

$$\begin{aligned} Q(\theta^t, \lambda^t) &= Q(\theta^{t+1}, \lambda^t) + L(\theta^t) - L(\theta^{t+1}) - \epsilon\lambda^t + \frac{c_k^t \epsilon^2}{2(a-1)} \\ &\leq Q(\theta^{t+1}, \lambda^t) + L(\theta^t) - L(\theta^{t+1}) - \epsilon\lambda_A^{t+1} + \frac{c_k^t \epsilon^2}{2(a-1)} \\ &= Q(\theta^{t+1}, \lambda^t) + \frac{c_k^t \epsilon^2}{2(a-1)}. \end{aligned}$$

Consider now that $\lambda^{t+1} = \lambda_B^{t+1}$. We have $\lambda^t \geq \lambda_B^{t+1} \geq \lambda_A^{t+1}$ and $\lambda_B^{t+1} < |\theta_k^t| \leq a\lambda_B^{t+1}$. Hence, $|\theta_k^t| \leq a\lambda^t$. But it might either be that $|\theta_k^t| \leq \lambda^t$ or $\lambda^t < |\theta_k^t| \leq a\lambda^t$. In the first situation, relation (3.7.25) follows from the exact same reasoning as above. If $\lambda^t < |\theta_k^t| \leq a\lambda^t$, we have

$$\begin{aligned} Q(\theta^t, \lambda^t) &= Q(\theta^{t+1}, \lambda^t) + L(\theta^t) - L(\theta^{t+1}) - \epsilon\frac{a\lambda^t - |\theta_k^t|}{a-1} + \frac{c_k^t \epsilon^2}{2(a-1)} \\ &\leq Q(\theta^{t+1}, \lambda^t) + L(\theta^t) - L(\theta^{t+1}) - \epsilon\frac{a\lambda_B^{t+1} - |\theta_k^t|}{a-1} + \frac{c_k^t \epsilon^2}{2(a-1)} \\ &= Q(\theta^{t+1}, \lambda^t) + \frac{c_k^t \epsilon^2}{2(a-1)}. \end{aligned}$$

We now turn to proving (3.7.26). We have

$$\begin{aligned} Q(\tilde{\theta}, \lambda^t) - Q(\theta^t, \lambda^t) &= L(\tilde{\theta}) - L(\theta^t) + p'_{\lambda^t}(|\theta_l^t|)\epsilon - \frac{c_l^t \epsilon^2}{2(a-1)} \\ Q(\theta^{t+1}, \lambda^t) - Q(\theta^t, \lambda^t) &= L(\theta^{t+1}) - L(\theta^t) + p'_{\lambda^t}(|\theta_k^t|)\epsilon - \frac{c_k^t \epsilon^2}{2(a-1)}. \end{aligned}$$

Together, these two equations imply that

$$\begin{aligned} Q(\theta^{t+1}, \lambda^t) - Q(\tilde{\theta}, \lambda^t) &= L(\theta^{t+1}) + p'_{\lambda^t}(|\theta_k^t|)\epsilon - [L(\tilde{\theta}) + p'_{\lambda^t}(|\theta_l^t|)\epsilon] + \frac{\epsilon^2}{2(a-1)}(c_l^t - c_k^t) \\ &\leq \frac{\epsilon^2}{2}\left(\frac{(c_l^t - c_k^t)}{a-1} + (e_k^t - e_l^t)\right). \end{aligned}$$

The inequality in the last equation is justified as follows. First, following the proof of Lemma 3.1, it is easy to show that $\text{sign}(\theta_l^t) = -\text{sign}(\nabla_l L(\theta^t))$. We can then use Lemma 3.8 to obtain the desired inequality. Now, we show that

$$Q(\theta^t, \lambda^t) \leq Q(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{d_l^t \epsilon^2}{2(a-1)}.$$

Since $t \mapsto t+1$ is a forward step, it means that no backward step is taken. Hence, for all $l \in \mathcal{A}^t$, we have

$$L(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon) - L(\theta^t) - \epsilon p'_{\lambda^t}(|\theta_l^t|) \geq 0.$$

Using Lemma 3.7 leads to the desired result. □

Proof of Theorem 3.4

Let t be such that $\lambda^t < \lambda^{t+1}$ and consider that k is the index of the updated coefficient. Recall that m is the upper bound of the second order derivatives of $L(\cdot)$.

Suppose first that $l \in \mathcal{A}^t$, we want to show that $|\nabla_l Q(\theta^t, \lambda^t)| \leq m\epsilon$. Remark that $\nabla_l Q(\theta^t, \lambda^t) = \nabla_l L(\theta^t) + p'_{\lambda^t}(|\theta_l^t|)\text{sign}(\theta_l^t)$. Using Taylor expansions of $L(\theta^t + \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon)$ and of $L(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon)$, both around θ^t, we have

$$L(\theta^t + \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon) - L(\theta^t) = \nabla_l L(\theta^t)\text{sign}(\theta_l^t)\epsilon + \frac{e_l^t}{2}\epsilon^2 \tag{3.7.27}$$

$$L(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon) - L(\theta^t) = -\nabla_l L(\theta^t)\text{sign}(\theta_l^t)\epsilon + \frac{e_l^t}{2}\epsilon^2. \tag{3.7.28}$$

Using Lemma 3.6, we can rewrite (3.7.27) as

$$Q(\theta^t + \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) = Q(\theta^t, \lambda^t) + p'_{\lambda^t}(|\theta_l^t|)\epsilon + \nabla_l L(\theta^t)\text{sign}(\theta_l^t)\epsilon + \frac{e_l^t}{2}\epsilon^2 - \frac{c_l^t\epsilon^2}{2(a-1)}.$$

Using Proposition 3.8, we have

$$Q(\theta^t, \lambda^t) \leq Q(\theta^t + \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{\epsilon^2 c_l^t}{2(a-1)} + \frac{\epsilon^2}{2}[e_k^t - e_l^t].$$

Putting the last two results together, we obtain that

$$\epsilon p'_{\lambda^t}(|\theta_l^t|) + \nabla_l L(\theta^t)\text{sign}(\theta_l^t)\epsilon + \frac{e_k^t}{2}\epsilon^2 \geq 0.$$

Since $e_k^t \leq m$, we get

$$-\left[\nabla_l L(\theta^t)\text{sign}(\theta_l^t) + p'_{\lambda^t}(|\theta_l^t|)\right] \leq \frac{m\epsilon}{2}. \tag{3.7.29}$$

We now focus on (3.7.28). Using Lemma 3.7, we rewrite it as

$$Q(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) = Q(\theta^t, \lambda^t) + p'_{\lambda^t}(|\theta_l^t|)\epsilon - \nabla_l L(\theta^t)\text{sign}(\theta_l^t)\epsilon + \frac{e_l^t}{2}\epsilon^2 - \frac{d_l^t\epsilon^2}{2(a-1)}.$$

Again, using Proposition 3.8, we have that

$$Q(\theta^t, \lambda^t) \leq Q(\theta^t - \text{sign}(\theta_l^t)\mathbf{1}_l\epsilon, \lambda^t) + \frac{\epsilon^2 d_l^t}{2(a-1)}.$$

These last two results imply that

$$\nabla_l L(\theta^t)\text{sign}(\theta_l^t) + p'_{\lambda^t}(|\theta_l^t|) \leq \frac{m\epsilon}{2}. \tag{3.7.30}$$

Together, (3.7.29) and (3.7.30) lead to the desired result.

Consider now that $l \notin \mathcal{A}^t$. We want to show that $|\nabla_l L(\theta^t)| \leq p'_{\lambda^t}(|\theta_l^t|) + m\epsilon$. Using a Taylor expansion of $L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon)$ around θ^t, we have

$$L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) - L(\theta^t) = -\nabla_l L(\theta^t)\text{sign}(\nabla_l L(\theta^t))\epsilon + \frac{e_l^t}{2}\epsilon^2,$$

which we can rewrite as

$$|\nabla_l L(\theta^t)| = \frac{L(\theta^t) - L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon)}{\epsilon} + \frac{e_l^t}{2}\epsilon. \tag{3.7.31}$$

In what follows, we restrict to the set $\{l \notin \mathcal{A}^t : L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon) - L(\theta^t) < 0\}$. Notice that in the complementary set, we directly have $|\nabla_l L(\theta^t)| \leq m\epsilon$. Using Lemma 3.8, we have

$$\frac{L(\theta^{t+1}) - L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon)}{\epsilon} \leq p'_{\lambda^t}(|\theta_l^t|) - p'_{\lambda^t}(|\theta_k^t|) + \frac{e_k^t - e_l^t}{2}\epsilon. \tag{3.7.32}$$

Returning to Equation (3.7.31), we can write

$$\begin{aligned}|\nabla_l L(\theta^t)| &= \frac{L(\theta^t) - L(\theta^{t+1})}{\epsilon} + \frac{L(\theta^{t+1}) - L(\theta^t - \text{sign}(\nabla_l L(\theta^t))\mathbf{1}_l\epsilon)}{\epsilon} + \frac{e_l^t\epsilon}{2} \\ &\leq L_\epsilon^{t,t+1} + p'_{\lambda^t}(|\theta_l^t|) - p'_{\lambda^t}(|\theta_k^t|) + \frac{e_k^t\epsilon}{2},\end{aligned}$$

where the inequality is implied by (3.7.32). Suppose that $|\theta_k^t| \leq \lambda^t$. Then, it holds

$$L_\epsilon^{t,t+1} - p'_{\lambda^t}(|\theta_k^t|) = \lambda_A^{t+1} - \lambda^t \leq 0,$$

where the inequality is due to $\lambda_A^{t+1} = \lambda^{t+1} < \lambda^t$. Consider instead that $\lambda^t < |\theta_k^t| \leq a\lambda^t$. In that case,

$$L_\epsilon^{t,t+1} - p'_{\lambda^t}(|\theta_k^t|) = \frac{a\lambda_B^{t+1} - |\theta_k^t|}{a-1} - \frac{a\lambda^t - |\theta_k^t|}{a-1} \leq 0,$$

where the inequality is due to $\lambda_B^{t+1} = \lambda^{t+1} < \lambda^t$. Finally, the situation covered by $|\theta_k^t| > a\lambda^t$ can be disregarded since k is the index of the updated coefficient and $\lambda^{t+1} < \lambda^t$. In conclusion, we have

$$\begin{aligned}|\nabla_l L(\theta^t)| - p'_{\lambda^t}(|\theta_l^t|) &\leq \frac{e_k^t\epsilon}{2} \\ &\leq \frac{m\epsilon}{2},\end{aligned}$$

which closes the second part of the proof. □

CHAPTER 4

An application to equality of opportunity

We use the methodology developed in this thesis to estimate inequality of opportunity in three European countries in 2011 and 2019. The explained Gini coefficient is proposed as an ex-ante utilitarian measure of inequality of opportunity and the penalized Lorenz regression is used to estimate it. A procedure based on the Shapley Value decomposition is proposed to break down the explained Gini coefficient into contributions of each selected circumstance. Inequality of opportunity is estimated using data from the European Union Statistics on Income and Living Conditions. Results are compared with estimates obtained from a penalized OLS regression. In many instances, our procedure offers at least similar estimates of inequality of opportunity but with a sparser model, including less circumstances. The content of this chapter is based on the manuscript

> Alexandre Jacquemain, Cédric Heuchenne, and Philippe van Kerm. A Penalized Lorenz regression analysis of inequality of opportunity, 2022b.

4.1 Introduction

The notion of inequality of opportunity (IOP) has been introduced in Section 1.2 and the issue of its measurement has been tackled in Section 1.3. We recall that, in the IOP context, the response variable Y is a measure of economic advantage and the covariates $X = (X_1, \ldots, X_p)^\intercal$ are circumstances beyond the individual's control. An *ex-ante utilitarian measure* of IOP is obtained by estimating the inequality of $\mathbb{E}[Y|X]$. The explained Gini coefficient is the ex-ante utilitarian measure of IOP when a single-index model with strictly increasing link function is assumed and the Gini coefficient is used as measure of inequality. Throughout the last sections, several arguments established the relevance of the Lorenz regression in the measurement of IOP. We gather them in the following paragraphs.

- The Lorenz regression offers a good compromise between the efficiency obtained with parametric methods and the flexibility guaranteed by nonparametric ones. Indeed, it rests on a single-index model where the parameter vector allows for dimension reduction and the nonparametric link function guarantees flexibility. Furthermore, the link function does not need to be estimated. This first argument was presented in depth in Section 2.1
- The estimated explained Gini coefficient is a robust measure of IOP because the estimated index is only used to rank observations. This is not the case for the empirical Gini coefficient of fitted values, as shown in Escanciano and Terschuur (2022) and discussed in Section 3.1
- As shown in Chapter 3, the penalized Lorenz regression produces valid inference even when many circumstances are included. It avoids the overestimation of the explained Gini coefficient and provides an automatic selection of the relevant circumstances.
- In this chapter, we propose a decomposition procedure for the explained Gini coefficient based on the Shapley Value. With this descriptive procedure, the researcher is able to assess the individual contribution of each circumstance to the IOP estimated by the Lorenz regression.

This chapter is organized as follows. Section 4.2 introduces the Shapley decomposition procedure. The methodology is then applied on data obtained from the European Union Statistics on Income and Living Conditions. The construction of the datasets is explained in Section 4.3, where we also discuss specific challenges to be faced. For example, we adapt the procedure to take into account sample weights. The results are presented and interpreted in Section 4.4. Section 4.5 summarizes the main findings of this chapter. Finally, Section 4.6 is an appendix gathering auxiliary results.

4.2 A decomposition procedure for the explained Gini coefficient

The penalized Lorenz regression helps us disentangle between relevant and irrelevant circumstances based on whether estimated weights are set to zero or not. As a next step, we aim to decompose $\widehat{\text{Gi}}_{Y,X}$ into contributions of each relevant circumstance. The Shapley Value decomposition is often used in linear regression in order to determine the relative importance of covariates. This technique decomposes the R^2 of the regression into contributions of each covariate. It does so by fitting all combinations of submodels, leaving out one or more covariates, and recording losses of R^2. Shorrocks (2013) generalized this procedure in the context of distributional analysis, while Hoyos and Narayan (2011) applied it in the specific

case of IOP.

In this paper, we decompose the explained Gini coefficient estimated on the set of selected circumstances. Several advantages stem from this procedure. The decomposition is additive in the sense that individual contributions add up to the estimated explained Gini coefficient. Therefore, contributions can be interpreted as shares of IOP attributable to each selected circumstance. Contributions can be obtained for a categorical covariate as a whole. Leaving out a categorical predictor means to exclude all the corresponding dummies. While the Shapley procedure is computationally intensive as it requires fitting many submodels, this issue is reduced by focusing on the set of selected circumstances. Even when the original penalized Lorenz regression starts from a large set of circumstances, the number of submodels to consider may still be moderate if sparsity is large, i.e. if few circumstances are selected.

We now explain the procedure formally. In what follows, PLR refers to the penalized Lorenz regression. First, we fit the original PLR on the full set of circumstances. Let K be the number of selected covariates, where a categorical covariate counts as one irrespective of the number of categories. Each model is indexed by a couple (k, m), where $k \in \{0, \ldots, K\}$ denotes the number of covariates left out by this model and $m \in \{1, \ldots, M_k = \binom{K}{k}\}$ indexes the model. Let $\widehat{\mathrm{Gi}}_m^k$ be the estimated explained Gini coefficient related to (k, m), i.e. $\widehat{\mathrm{Gi}}_1^0$ corresponds to the original PLR. Fit the PLR on each m and $k = 1, \ldots, K-1$, and retrieve $\widehat{\mathrm{Gi}}_m^k$. For $k = K$, set $\widehat{\mathrm{Gi}}_m^k = 0$. The contribution of the kth covariate to IOP is measured with

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

where N_k^l is the set of all models excluding l covariates and where the kth covariate is included. Also, $(m_{-k}, l+1)$ is the model with the same covariates as (m, l), except that the kth covariate is now excluded. By construction, $\widehat{\mathrm{Gi}}_{Y,X} = c_1 + \ldots + c_K$.

4.3 Data

We apply the penalized Lorenz regression on data obtained from the European Union Statistics on Income and Living Conditions (EU-SILC) survey. The bandwidth and the regularization parameter, i.e. the couple (h, λ), are determined by the bootstrap procedure developed in Section 3.2. This section discusses the methodological challenges peculiar to the data, as well as the considered variables.

The next section provides an interpretation of the results and compares the estimates with more traditional approaches.

EU-SILC is an official micro-data source for calculation of social indicators in European countries. It is available annually since 2003. EU-SILC 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 November 2020 release of EU-SILC containing cross-sectional data up to 2019 (https://doi.org/10.2907/EUSILC2004-2019V.1). We work with two cross-sections of the data, collected in 2011 and 2019. In those two years, EU-SILC collected retrospective information about childhood circumstances of respondents aged from 24 to 59. Retrospective information mostly refers to when the respondent was aged around 14 (Andreoli and Fusco, 2019).

Our analysis focuses on three countries (Belgium, Spain and Norway) and two years (2011 and 2019). In each case, we further distinguish a sample of young individuals, i.e. observations in the $[25, 45]$ year interval, and a sample of old individuals, i.e. observations in the $(45, 60]$ year interval. In tables and figures, we will code each case as follows : two letters representing the country (BE, ES and NO), two digits representing the year (11 and 19) and one letter for the age cohort (Y and O).

The advantage variable is a measure of single-adult equivalent household annual disposable income (as in, e.g., Brzezinski, 2020; Ramos and Van de gaer, 2021) and is denoted by Y. An alternative would consist in using personal wages. The use of the former presents two advantages. First, it better tracks the actual standard of living of individuals since it also encompasses property income and transfers. Second, it alleviates the question of how to model individuals for which the wage is zero.

Table 4.1 lists the variables included in the analysis. The first row describes the advantage variable. The remaining ones present the included circumstances as well as their categories, except for the number of siblings, which is treated as a quantitative variable. Variables observed at the level of the parents are available both for the mother and for the father. As indicated in the second column of the table, these variables are coded with two letters, the first one indicating whether it provides information about the mother (M) or the father (F), while the second identifies the covariate. For the remaining variables, only one letter is used. Categories are identified with one digit, indicated in the third column of the table. For example, MA2 indicates that the mother is self-employed. We restrict to observations without any missing value and for which both parents are known.

The performance of the Lorenz regression can be harmed by categories with ex-

Table 4.1: Variables considered in the analysis

Variable	Name	Categories
Income	Y	/
Sex	G	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)
Number of siblings	S	/

tremely low frequencies. One simple illustration appears in the bootstrap procedure. If only a very few individuals share 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. To avoid this, we exclude categories with less than 10 individuals as well as the corresponding observations. Categorical covariates are introduced as dummies and each dummy is penalized separately. Several papers have acknowledged 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. As an illustration of the issue, consider the covariate ME and assume that unequal opportunities are mainly stemming from a disadvantage of individuals with a mother with low education ($ME1$). Suppose first that $ME1$ is used as reference category and, therefore, not included in the regression. In this situation, it takes two updates for the SCAD-FABS algorithm to give observations with $ME1$ a lower index rank with respect to observations in the other two categories (one update on the coefficient of $ME2$ and one on the coefficient of $ME3$). If another category is used as reference, this ranking is obtained with only one update (on $ME1$ directly). Obtaining the same ordering therefore requires a coefficient vector with a higher L2-norm. This problem is alleviated by introducing all the dummies in the regression. As discussed at the end of Subsection 2.4.2, the presence of multicolinearity is not a threat to the identifiability and consistent estimation of the explained Gini coefficient. To the main effects, we add all two-way interactions with, again, no less than 10 individuals. Descriptive statistics of the final datasets are relegated to Section 4.6.

Since our method can accommodate for a nonpositive response variable, we keep observations with a negative income. However, we delete outliers based on the empirical influence function of the Gini coefficient. For an observation i, it is defined as $(n-1)(\widehat{\mathrm{Gi}}_Y - \widehat{\mathrm{Gi}}_{Y,-i})$, where $\widehat{\mathrm{Gi}}_{Y,-i}$ is the estimate of the Gini coefficient of Y obtained after deleting observation i. We keep observations with an influence value smaller than 2.5. This threshold is determined empirically; it ensures the stability

of the analyses without removing too much information. Across the samples, this operation deletes at most 0.8% of the observations. Table 4.2 provides information about the sample sizes. The first column gives the number of observations before removing the outliers. The second column displays the number of outliers. The number of observations with a nonpositive income in the final datasets is given in the last column.

Table 4.2: Information about sample sizes

	Number of observations		
	before removing outliers	outliers	nonpositive incomes
BE11O	1734	1	0
BE11Y	2619	3	4
BE19O	2406	5	4
BE19Y	3365	1	3
ES11O	6139	15	46
ES11Y	8683	16	61
ES19O	7357	8	38
ES19Y	8409	6	26
NO11O	901	3	1
NO11Y	1253	1	2
NO19O	499	4	0
NO19Y	1130	1	2

We take into account sample weights by adapting the definitions of the objective function of the penalized Lorenz regression and of the estimated explained Gini coefficient. Let π be the vector of sample weights, satisfying $\pi_i \geq 0 \ \forall i = 1, \ldots, n$ and $\sum_{i=1}^{n} \pi_i = 1$. The estimators of θ_0 and the explained Gini coefficient are given by

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

$$\widehat{\mathrm{Gi}}_{Y,X} = \frac{2}{\overline{Y}} \sum_{i=1}^{n} \pi_i Y_i \hat{F}_i(\hat{\theta}) - 1,$$

where $\hat{F}_i(\theta)$ is now given by

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

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

To preserve the dependence between individuals of the same household, the resampling is achieved at the level of the household.

4.4 Results

In this section, we apply the PLR to the EU-SILC data. All computations are performed using the statistical software `R`. The PLR is compared with two benchmarks. In the first case (OLS), fitted values are obtained from a linear regression of log-incomes on the circumstances. In the second (Penalized OLS), they are obtained from a penalized log-regression, where the penalty is the LASSO and the regularization parameter is chosen via cross-validation. To perform this analysis, we use the function `cv.glmnet` from the library `glmnet`. In the PLR, inequality of opportunity is estimated with the explained Gini coefficient ($\widehat{\text{Gi}}_{Y,X}$). In the other procedures, it is either estimated with the Gini coefficient of the fitted values ($\widehat{\text{Gi}}_{\hat{Y}}$) or, as proposed by Escanciano and Terschuur (2022), by the concentration index of the advantage with respect to the fitted values ($\widehat{\text{CI}}_{Y,\hat{Y}}$) . As we previously mentioned, the latter is a more robust estimator of IOP. Because the logarithm is a strictly increasing transformation, it corresponds to an explained Gini coefficient, where the index is estimated via (penalized) OLS. It is therefore more comparable to $\widehat{\text{Gi}}_{Y,X}$. While negative values for the advantage variable are not an issue in the PLR, they are not allowed in the other two procedures because of the log-transformation. We therefore delete observations with negative incomes for the OLS and Penalized OLS.

Our analysis develops as follows. Section 4.4.1 provides inference for IOP. Namely, we compare the estimates of the different methods and provide confidence intervals obtained by the PLR. We also compare the penalized procedures in terms of selected circumstances. Shapley decompositions of the estimated explained Gini coefficients are performed in Section 4.4.2.

4.4.1 Inequality of opportunity estimates

For each case, the first row of Table 4.3 provides the estimates of IOP obtained by the different methods. The second row provides Lorenz-R^2 measures, which represent shares of total inequality that are linked to unequal opportunities.

We first focus on the OLS procedures. We observe a noticeable difference between $\widehat{\text{Gi}}_{\hat{Y}}$ and $\widehat{\text{CI}}_{Y,\hat{Y}}$, usually of around one percentage point. In almost all cases, $\widehat{\text{Gi}}_{\hat{Y}} \geq \widehat{\text{CI}}_{Y,\hat{Y}}$. In absence of penalization, we expect overfitting since the regression coefficients are unconstrained. When regression coefficients are penalized, both measures of IOP diminish. However, the decrease is more pronounced for $\widehat{\text{Gi}}_{\hat{Y}}$ than for $\widehat{\text{CI}}_{Y,\hat{Y}}$. Moving from the unpenalized to the penalized OLS induces an

Table 4.3: Estimates of inequality of opportunity and (Lorenz-R^2)

		PLR	OLS		Penalized OLS	
	$\widehat{\text{Gi}}_Y$	$\widehat{\text{Gi}}_{Y,X}$	$\widehat{\text{Gi}}_{\hat{Y}}$	$\widehat{\text{CI}}_{Y,\hat{Y}}$	$\widehat{\text{Gi}}_{\hat{Y}}$	$\widehat{\text{CI}}_{Y,\hat{Y}}$
BE11O	24.5	9	10.5	9.4	9.5	8.7
		(36.7)	(42.9)	(38.4)	(38.8)	(35.5)
BE11Y	23.4	10.2	11.9	10.7	11	10.1
		(43.6)	(50.9)	(45.7)	(47)	(43.2)
BE19O	24.9	8.3	9.8	9.1	8.9	8.3
		(33.3)	(39.4)	(36.5)	(35.7)	(33.3)
BE19Y	21.9	9.3	10.9	10	10.2	9.4
		(42.5)	(49.8)	(45.7)	(46.6)	(42.9)
ES11O	33.1	11.6	11.9	10.8	11.5	10.7
		(35)	(36)	(32.6)	(34.7)	(32.3)
ES11Y	32.3	13.1	14.8	13.4	14.2	13.4
		(40.6)	(45.8)	(41.5)	(44)	(41.5)
ES19O	33.9	12.1	13.3	12.6	12.4	12.1
		(35.7)	(39.2)	(37.2)	(36.6)	(35.7)
ES19Y	30.9	13.6	16.7	14.4	16.3	14.3
		(44)	(54)	(46.6)	(52.8)	(46.3)
NO11O	21.3	4.4	6.3	6.3	3.4	4
		(20.7)	(29.6)	(29.6)	(16)	(18.8)
NO11Y	20.3	5.6	8	6	6.1	5.1
		(27.6)	(39.4)	(29.6)	(30)	(25.1)
NO19O	23.2	7.4	11.7	9.4	7.5	6.2
		(31.9)	(50.4)	(40.5)	(32.3)	(26.7)
NO19Y	21.7	8.4	11.1	8.3	7.1	5.5
		(38.7)	(51.2)	(38.2)	(32.7)	(25.3)

average change of -1.6 percentage points in the former and of -1 percentage point in the latter.

In general, the estimates obtained by the PLR follow quite closely those obtained by $\widehat{\text{CI}}_{Y,\hat{Y}}$ in the penalized OLS, but are slightly higher. Only in four instances, i.e. the old-age cohort of Spain in 2011, the young-age cohort of Spain in 2019, and both age cohorts of Norway 2019, is the difference of more than half a percentage point. In most cases, the estimate obtained with PLR is higher: $+0.9$ percentage point for ES11O, $+1.2$ percentage point for NO19O and $+2.9$ percentage points for NO19Y. Only for ES19Y is the difference negative: -0.7 percentage point.

We now focus on the PLR and start with a comparison between the countries.

In each combination of year and age cohort, the same ranking emerges: Spain is the most unequal country, Belgium comes second and Norway is the least, as measured by the Gini coefficient. This ranking is preserved when we turn to the explained Gini coefficient. The Lorenz-R^2 obtained in Belgium and Spain are quite similar. As such, there is more inequality of opportunity in Spain than in Belgium but it amounts to the same share of total inequality. In Norway, the Lorenz-R^2 is lower. It is therefore in this country that unequal opportunities are both the lowest, and represent the smallest share in total inequality.

In every situation, the Gini coefficient is lower in the young age cohort compared to the old one, but this situation is reversed when we turn to the explained Gini coefficient. At the start of the work life, income inequality is lower but is more related to circumstances. Advantages among older people reveal less bound to their circumstances. Even though total inequality grows, IOP reduces. The results provided so far would be incomplete without a proper inference analysis. In this respect, Table 4.4 provides 95% confidence intervals for the explained Gini coefficient and for the Lorenz-R^2 obtained by the PLR.

Table 4.4: Confidence intervals for the explained Gini coefficient and for the Lorenz-R^2

	Explained Gini coefficient		Lorenz-R^2	
	Lower bound	Upper bound	Lower bound	Upper bound
BE11O	7.7	10.4	31.8	41.9
BE11Y	9.1	11.2	39.7	47.1
BE19O	7	9.6	28.9	37.9
BE19Y	8.4	10.2	38.9	45.9
ES11O	10.4	12.8	31.8	38.5
ES11Y	11.9	14.3	37.5	43.6
ES19O	10.3	13.9	31.2	40.3
ES19Y	12.4	14.8	41	47.3
NO11O	3	5.9	14.5	27.2
NO11Y	4.5	6.8	22.5	33
NO19O	4.9	9.9	22.4	41.1
NO19Y	7	9.9	33.2	44.6

Table 4.5 gives an idea about the size of the relevant opportunity set. The first column indicates the number of variables included at the start of the regression. This number depends on each dataset since variables are automatically excluded when they have too low frequencies. The next two columns indicate the number of circumstances selected by each procedure, i.e. the number of non-zero estimated coefficients. In general, the PLR leads to a sparser model, meaning that it retains

fewer circumstances, while providing estimates of IOP comparable to the penalized OLS. In some situations, it retains more circumstances, but this is always justified by a larger estimate of IOP. Interestingly, there are also cases where it provides a larger estimate of IOP while retaining less circumstances. Let us take BE11O as an example. The PLR estimates that inequality of opportunity is of 9% and includes 15 circumstances. In contrast, the estimate of IOP with the penalized OLS is of 8.7%, with 21 circumstances included. This happens because the logic surrounding the Lorenz regression is to find the circumstances which maximize the concentration index of Y with respect to fitted values from a single-index model. Alternatively, the size of the opportunity set can be evaluated with the number of profiles, displayed in the last two columns. In a given profile, all the individuals have the same index values. As such, it is comparable to terminal nodes in a tree-based method. The analysis of this measure reinforces our conclusion that the PLR generally offers a more parsimonious opportunity set without substantial loss in estimates of IOP. We illustrate this with BE19Y, for which the estimates of IOP are of 9.4% for penalized OLS and of 9.3% for the PLR, while the number of profiles are respectively of 1301 and of 103.

Table 4.5: Number of circumstances and profiles selected by each penalized procedure

	Number of circumstances			Number of profiles	
	included	selected			
		OLS	PLR	OLS	PLR
BE11O	215	21	15	247	175
BE11Y	275	23	11	806	330
BE19O	235	28	10	705	214
BE19Y	316	31	10	1301	103
ES11O	217	28	34	611	676
ES11Y	325	37	8	1059	108
ES19O	241	25	11	722	194
ES19Y	334	40	8	1232	45
NO11O	121	7	7	19	26
NO11Y	164	13	8	177	111
NO19O	133	4	4	24	13
NO19Y	250	11	19	109	367

Tables 4.10, 4.11 and 4.12, displayed in Section 4.6, present the selected circumstances. Each row corresponds to a circumstance, while the columns relate to the datasets. A + (−) sign is indicated if the estimated weight is positive (negative). The cell is left blank if the circumstance is not selected. The most selected cir-

cumstances are the interaction between ME1 and FE1 (6 times), the main effect of FI3 and its interaction with FN2 (5 times each) and finally, with 4 times each, the following interactions: FE3×H1, FI3×MN2 and ME1×S.

4.4.2 Contribution of each circumstance

Tables 4.6 and 4.7 display the contribution of each circumstance to the explained Gini coefficient. For each case, the first line displays the absolute contribution, summing to the explained Gini coefficient, while the second line sums to 100% and displays a relative contribution.

In the selected models, most of the estimated IOP is attributable to a few circumstances. In all cases, the nationalities (MN and FN), education levels (ME and FE) and types of occupation (MI and FI) of the parents contribute to more than half of the explained Gini coefficient. In ten cases, the figure rises to more than 70%. Education is the most important driver, accounting for more than 30% in the majority of the cases, and up to 49% (BE19O). In seven cases, nationality contributes to more than 20% of the explained Gini coefficient. In NO19O, it rises to 87% because all the selected covariates are either the main effect of FN, or interactions with other circumstances. Finally, the type of occupation (most importantly of the father) contributes to more than 20% of the explained Gini coefficient in four cases. In both age cohorts of Norway 2011, the relative contribution rises to more than 40%.

Behind these key circumstances, the number of siblings contributes moderately to IOP. In four cases, it accounts for more than 10% of the explained Gini coefficient. The remaining circumstances contribute only slightly, except in very specific cases. For example, the household composition accounts for more than 10% in both age cohorts of Belgium 2011. Gender never accounts for more than 6% of the estimated IOP. This is not necessarily surprising as our choice of using household incomes overshadows potential gender inequalities.

We propose to visualize the impact of a given circumstance with the average rank of the index values, computed across all individuals sharing that circumstance. The rank is then normalized to live in the unit interval. The closer the average rank to 1 (0), the more this circumstance contributes positively (negatively) to the advantage variable. In Figure 4.1, we illustrate this procedure for the interaction between MN and ME. We focus the interpretation on Belgium. Overall, the average rank increases as the education of the mother increases and as we move from a non-native to a native mother. However, we observe a difference between the young and old-age cohorts. For the old-age cohorts, three clusters emerge. Individuals with a non-native mother of low education ($MN1 \times ME1$) fare the worst with an average rank around 12.5%. The second cluster is formed of individuals

Table 4.6: (Relative) contribution of each selected circumstance to IOP - First part

	G	FE	ME	FA	MA
BE11O	0.18	0.89	1.91	0.14	0.22
	(1.95)	(9.88)	(21.11)	(1.52)	(2.44)
BE11Y	0	2.45	2.17	0	0.2
	(0)	(24.08)	(21.35)	(0)	(1.96)
BE19O	0.25	2.51	1.59	0.44	0
	(2.98)	(30.26)	(19.1)	(5.34)	(0)
BE19Y	0.34	1.6	1.17	0.47	1
	(3.64)	(17.26)	(12.58)	(5.08)	(10.78)
ES11O	0.48	1.72	0.96	0.49	0.27
	(4.1)	(14.81)	(8.28)	(4.25)	(2.34)
ES11Y	0	3.2	1.53	0	0
	(0)	(24.37)	(11.69)	(0)	(0)
ES19O	0.37	1.59	3.38	0.09	0.84
	(3.06)	(13.12)	(27.86)	(0.74)	(6.9)
ES19Y	0	4.29	0.37	0	0
	(0)	(31.45)	(2.68)	(0)	(0)
NO11O	0	0.74	1.4	0	0.21
	(0)	(16.68)	(31.44)	(0)	(4.69)
NO11Y	0.33	0.49	0	0	0.1
	(5.91)	(8.76)	(0)	(0)	(1.86)
NO19O	0	0.71	0.11	0	0
	(0)	(9.58)	(1.46)	(0)	(0)
NO19Y	0.42	0.64	0.67	0.62	0.08
	5.01	7.61	7.96	7.39	0.95

with a native mother of low education ($MN2 \times ME1$), with an average rank around 50%. Finally, the remaining profiles have an average rank between 75% and 90%. In conclusion, the effect of mother education on unequal opportunities is mainly through a distinction between low education versus the rest, and it is aggravated if the mother is non-native. Concerning the young-age cohorts, the effect of education is more gradual since the average rank steadily increases as we move to higher education levels. Also, the impact of the mother nationality is felt at each education level. Approximately, an individual with a non-native mother faces similar prospects as an individual with a native mother but with one level less of education.

Table 4.7: (Relative) contribution of each selected circumstance to IOP - Second part

	FI	MI	FN	MN	H	S
BE11O	2.4	0.38	0.41	0.64	1.07	0.81
	(26.51)	(4.19)	(4.54)	(7.1)	(11.79)	(8.97)
BE11Y	0.39	0.22	0.47	0.71	1.75	1.81
	(3.82)	(2.19)	(4.65)	(6.97)	(17.17)	(17.8)
BE19O	1.44	0.08	1.28	0	0.12	0.59
	(17.37)	(0.91)	(15.47)	(0)	(1.5)	(7.08)
BE19Y	0.42	0.47	1.93	1.47	0.42	0
	(4.53)	(5.08)	(20.78)	(15.76)	(4.5)	(0)
ES11O	3.8	0.31	1.38	0.9	0.3	1.01
	(32.67)	(2.66)	(11.89)	(7.75)	(2.59)	(8.66)
ES11Y	2.01	0	1.91	2.54	0.88	1.04
	(15.36)	(0)	(14.55)	(19.37)	(6.7)	(7.96)
ES19O	1.62	0.08	1.84	0.7	0	1.61
	(13.4)	(0.66)	(15.15)	(5.81)	(0)	(13.3)
ES19Y	0.94	0.8	4.43	2.22	0.58	0
	(6.93)	(5.89)	(32.53)	(16.27)	(4.24)	(0)
NO11O	1.29	0.64	0.17	0	0	0
	(29.07)	(14.38)	(3.74)	(0)	(0)	(0)
NO11Y	2.01	0.56	1.05	0.32	0	0.76
	(35.65)	(9.97)	(18.68)	(5.7)	(0)	(13.48)
NO19O	0	0	6.41	0	0	0.15
	(0)	(0)	(86.92)	(0)	(0)	(2.03)
NO19Y	0.56	0.79	1.76	0.53	0.73	1.61
	6.66	9.42	20.92	6.31	8.69	19.07

4.5 Discussion

In this chapter, we applied the Lorenz regression toolbox to analyze inequality of opportunity. This procedure produces inference for the explained Gini coefficient, an ex-ante utilitarian measure of IOP. The methodology offers statistical guarantees: it is flexible, efficient and more robust as it uses ranks. It also offers an automatic selection of the relevant circumstances. Finally, it is possible to decompose the measure of IOP thus obtained into contributions of each circumstance.

In the three countries analyzed, our procedure performs favorably compared to a penalized OLS on log-incomes. Typically, it proposes a sparser model while providing comparable estimates of IOP. In absolute terms, inequality of opportunity

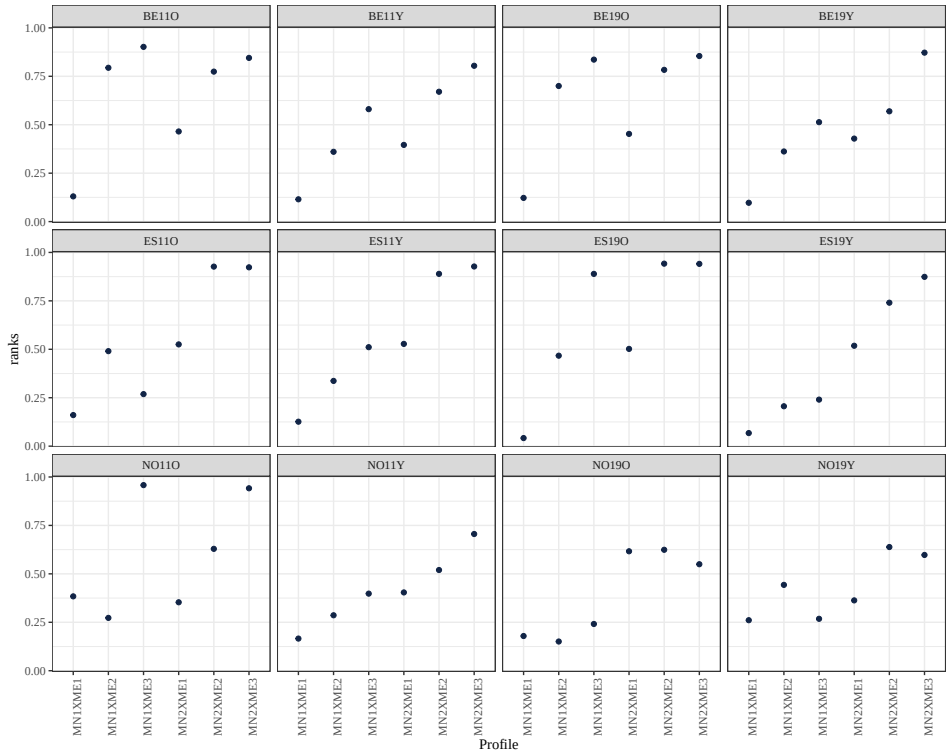

Figure 4.1: Average rank for each category of $MN \times ME$

ranges from 4.4% to 13.6%. It can also be expressed as a share of total inequality. Thus measured, unequal opportunities represent from 20.7% to 44% of the observed inequality. Norway scores best, both in absolute and relative terms. In Belgium and Spain, unequal opportunities amount to similar shares of total inequality. However, Spain is the most unequal in absolute terms. Opportunities are consistently more unequal for the young-age cohorts, even though total inequality is higher in the old-age cohorts. The most important circumstances are the parents levels of education, types of occupation and nationalities.

We end this discussion with a comment on the possible misspecification of the model. As discussed in Ramos and Van de gaer (2016), the omission of relevant circumstances yields a downward bias in the estimation of IOP. Borrowing from their terminology, the analysis performed in this chapter restricts circumstances to *social background luck*, i.e. variables related to the family or social origin of the individual. We are therefore leaving from consideration aspects, for example talent, which could nevertheless be considered as circumstances beyond the individual

control. In this respect, Björklund et al. (2012) found out that IQ measured at the age of 18 was the strongest contributor to IOP in Sweden. Even in the set of social background information, our data might miss important features. For example, it does not contain information on parental income or social networks. The results obtained with our method may therefore be lower bound estimates of the actual IOP.

4.6 Appendix

Tables 4.8 and 4.9 provide descriptive statistics for each variable included in the study, see Table 4.1 for a description of these variables. The first column displays the median income, while the remaining columns provide information about the circumstances. For categorical variables, the proportion in each category is displayed in percentages. The last column of Table 4.9 displays the mean number of siblings.

Table 4.8: Descriptive statistics - First part

	Y	G1	G2	ME1	ME2	ME3	FE1	FE2	FE3	MA1	MA2	MA3	FA1	FA2	FA3
BE11O	23,962	50.3	49.7	76.1	15.9	8.0	66.0	17.5	16.5	76.9		23.1	3.7	13.0	83.3
BE11Y	23,855	51.7	48.3	50.3	28.7	21.0	46.4	27.5	26.1	51.2		48.8	5.0	14.5	80.5
BE19O	25,517	50.7	49.3	68.5	21.6	9.9	61.6	21.3	17.0	58.1	11.5	30.4	3.9	17.0	79.1
BE19Y	25,479	50.9	49.1	44.2	29.4	26.4	42.9	29.0	28.1	35.0	9.7	55.4	4.5	17.4	78.0
ES11O	16,355	51.4	48.6	95.6	2.4	2.0	89.8	3.9	6.3	82.1	8.0	9.8	2.9	27.6	69.5
ES11Y	15,202	51.2	48.8	87.0	7.1	5.9	80.9	8.6	10.5	70.6	7.0	22.4	3.1	22.8	74.2
ES19O	15,628	51.0	49.0	91.0	4.8	4.1	84.0	7.3	8.7	74.8	9.2	16.0	2.6	24.5	72.9
ES19Y	15,319	50.7	49.3	76.9	12.7	10.3	71.5	13.9	14.6	57.0	8.6	34.4	2.4	22.1	75.6
NO11O	46,524	44.8	55.2	48.6	41.3	10.1	41.3	36.4	22.3	39.0	12.2	48.8	1.4	29.2	69.4
NO11Y	40,364	47.0	53.0	22.9	47.7	29.4	20.6	41.1	38.3	17.3	9.3	73.3	1.1	23.2	75.6
NO19O	42,397	45.3	54.7	29.7	46.7	23.6	26.3	42.6	31.1	19.4		80.6	2.4		97.6
NO19Y	37,215	48.6	51.4	20.4	39.4	40.2	15.8	41.4	42.9	14.9		85.1	4.3		95.7

Table 4.9: Descriptive statistics - Second part

	MI1	MI2	MI3	FI1	FI2	FI3	MN1	MN2	FN1	FN2	H1	H2	S
BE11O	8.7	4.6	9.9	22.0	33.5	40.8	17.0	83.0	16.7	83.3	95.2	4.8	2.0
BE11Y	10.0	8.5	30.4	16.9	30.8	47.2	23.2	76.8	23.5	76.5	90.6	9.4	1.5
BE19O	7.2	19.3	15.4	20.1	37.4	38.6	25.4	74.6	24.8	75.2	95.6	4.4	1.8
BE19Y	8.7	22.0	34.3	17.5	31.8	46.1	30.3	69.7	30.8	69.2	91.1	8.9	1.5
ES11O	5.8	9.4	2.7	30.0	46.4	20.6	6.5	93.5	6.7	93.3	96.9	3.1	1.5
ES11Y	8.7	11.6	9.1	27.6	42.4	26.9	12.1	87.9	12.1	87.9	96.0	4.0	1.3
ES19O	8.0	12.8	4.4	28.6	50.3	18.4	8.9	91.1	9.0	91.0	98.8	1.2	1.7
ES19Y	11.7	16.9	14.3	24.7	48.5	24.4	16.9	83.1	16.6	83.4	96.6	3.4	1.3
NO11O	11.7	27.4	21.9	17.7	42.0	38.9	8.0	92.0	6.6	93.4			1.3
NO11Y	10.0	26.6	46.1	12.6	35.1	51.1	13.0	87.0	12.5	87.5	98.9	1.1	0.9
NO19O	9.5	24.4	46.7	14.7	30.9	51.9	22.0	78.0	21.6	78.4	93.3	6.7	1.7
NO19Y	5.3	24.9	54.9	13.2	22.1	60.4	19.8	80.2	19.0	81.0	83.9	16.1	1.5

Table 4.10: Circumstances selected by the PLR - First part

	BE				ES				NO			
	11O	11Y	19O	19Y	11O	11Y	19O	19Y	11O	11Y	19O	19Y
G1×FE2					+							
G1×FI2					+							-
G1×ME1			-									
G1×S										-		
G2×FA1							-					
G2×FE3	+											
G2×FI2												+
G2×FI3					+							
G2×FN2				+								
G2×MA1					-							
G2×ME2												+
G2×ME3					+							
G2×S							-					
ME1		-					-					
ME1×FA1												-
ME1×FE1	-	-			-	-	-		-			
ME1×FN1			-	-	-							
ME1×H2												-
ME1×MN1	-						-	-				
ME1×S	-		-		-	-						
ME2×FI3												+
ME2×FN1							-		-		-	
ME2×MA1												+
ME2×S					+							
ME3									+			
ME3×FE1	+											
ME3×FE3									+			
ME3×FN2				+								
ME3×MN1					-							
ME3×MN2				+								

Table 4.11: Circumstances selected by the PLR - Second part

	BE				ES				NO			
	11O	11Y	19O	19Y	11O	11Y	19O	19Y	11O	11Y	19O	19Y
FE1		-	-					-				
FE1×FI2												+
FE1×FN1					-							
FE1×FN2					-							
FE1×MA1				-								
FE1×MI1										-		
FE1×MN1				-								
FE1×MN2					+							
FE1×S							-					
FE2×FN1						-						
FE2×FN2											+	
FE2×H1						+						
FE2×H2					-							
FE2×MI3			+									
FE2×MN2								-				
FE2×S												-
FE3			+			+						
FE3×FA1												-
FE3×FI1												+
FE3×FN2								+		+		
FE3×H1		+	+	+	+							
FE3×MA1	+											
FE3×MA3									+			
FE3×MN1					-							-
MA1×FI1							-					
MA1×FN1							-					
MA2×FN1										-		
MA3×FN1					-							
MA3×H1		+										
MA3×H2	-											
MA3×MN1					-							
MA3×MN2				+								
FA1					-							
FA1×FN2					-							
FA1×MN2	-											
FA2×H2					+							
FA3×FN2			+									-
FA3×H1												+
FA3×MI3				+								
FA3×S					-							

Table 4.12: Circumstances selected by the PLR - Third part

	BE				ES				NO			
	11O	11Y	19O	19Y	11O	11Y	19O	19Y	11O	11Y	19O	19Y
MI1									-			
MI1×MN1							-					
MI2×MN1												+
MI2×MN2					+							
MI2×S					-							
MI3×FN1	+											
MI3×FN2					+					+		+
MI3×H1		+										
MI3×MN1	+											
MI3×MN2								+				
MI3×S					+							
FI2										-		
FI3	+		+		+				+	+		
FI3×FN2	-			+	+		+	+				
FI3×H1						+						
FI3×H2					+							
FI3×MN1					+							
FI3×MN2	+	+			+	+						
MN1×FN1					-	-						
MN1×S					+					-		
MN2×H1		+						+				
FN1×S		-									-	-
FN2								+			-	
FN2×H1	+	+										+
H2	-											
H2×S			-		-							
S		-										-

CHAPTER 5

The LorenzRegression R package

In this chapter, we describe the **LorenzRegression** package, which implements the non-penalized and penalized Lorenz regressions in R. The non-penalized procedure is implemented via a genetic algorithm, making use of the **GA** package. For the penalized case, the user can choose between working with a LASSO or SCAD penalty. In the former case, the estimation procedure is performed with the FABS algorithm, while the latter is implemented with the SCAD-FABS algorithm. Computationally intensive parts of the code are written in C++. The main functions and S3 methods are carefully described and their use is illustrated on income data.

Alexandre Jacquemain and Cédric Heuchenne. LorenzRegression: An implementation of the Lorenz and penalized Lorenz regressions in R, 2022.

5.1 Introduction

Inequality measurement does not benefit from a widespread implementation in R (R Core Team, 2021), even though a few packages are publicly available from the Comprehensive R Archive Network (CRAN). The **laeken** package (Alfons and Templ, 2012) estimates indicators of social exclusion from complex surveys and can be used to compute Gini coefficients. The **ineq** package (Zeileis, 2014) estimates several inequality measures, including the Gini coefficient and the Lorenz curve. The **wINEQ** package (Machowska et al., 2022) estimates inequality measures for weighted data, including the Gini coefficient. While we are not aware of any package available on CRAN, an open source implementation of the concentration index is proposed in the **rineq** package (Devleesschauwer et al., 2017). It also comes with a parametric decomposition method, following the procedure proposed by Speybroeck et al. (2010).

The **LorenzRegression** package can be used to measure inequality, through the Gini coefficient and concentration index, and to graphically represent it, via the Lorenz and concentration curves. We refer the reader to Section 1.3 for a definition of these objects. The computation of all these quantities may include sample weights. More importantly, it is the first and only implementation so far of the Lorenz regression methodology in R. In the non-penalized case, the estimation is based on maximizing a discrete objective function. Due to lack of differentiability, the numerical solution is obtained through a genetic algorithm, implemented via the R package **GA** (Scrucca, 2013). In the penalized Lorenz regression, the issue is circumvented by replacing the discrete objective function with a smooth approximation. The numerical solution is obtained via the FABS algorithm described in Subsection 3.3.1 in the case of a LASSO penalty and via the SCAD-FABS algorithm introduced in Subsection 3.3.2 if the SCAD penalty is chosen instead. In the **LorenzRegression** package, the main functions are implemented in C++ and integrated in R, via the **Rcpp** (Eddelbuettel and Balamuta, 2018) and **RcppArmadillo** (Eddelbuettel and Sanderson, 2014) packages. To further reduce computation time, parallel computing is implemented with the **foreach** package (Microsoft and Weston, 2022). Finally, all plots are produced with **ggplot2** (Wickham, 2016).

The paper is organized as follows. Section 5.2 describes how the characteristics of the Lorenz and penalized Lorenz regressions are implemented inside the package. In Section 5.3, we present the end-user functions and the associated S3 methods. Section 5.4 illustrates the use of the package on a real-data example. Finally, Section 5.5 concludes.

5.2 Implementation of the Lorenz regression

The numerical solution of the Lorenz and penalized Lorenz regressions is governed by the following functions

`Lorenz.GA()`	(non-penalized Lorenz regression)
`Lorenz.SCADFABS()`	(penalized Lorenz regression - SCAD penalty)
`Lorenz.FABS()`	(penalized Lorenz regression - LASSO penalty)

Before we explain their functioning in detail, we turn the attention to two arguments used throughout the package. First, the argument `ties.method` allows the user to specify his or her preferred method to handle ties in the computation of empirical CDFs. It takes two possible values: `"mean"` corresponds to the average-rank solution described in Equation (2.4.2) and `"random"` relates to the random assignment set out in Equation (2.4.3). Second, the argument `weight` is used to specify sample weights. By default, all observations are given the same weight.

5.2.1 The Lorenz regression

The genetic algorithm described in Section 2.3 is implemented in the function `Lorenz.GA()`. Its usage is summarized as follows.

```
Lorenz.GA(YX_mat, ties.method=c("random","mean"),
          ties.Gini=c("random","mean"), seed.random=NULL, weights=NULL)
```

The only required argument `YX_mat` is a matrix of dimensions $n \times (p+1)$ where the first column is the response vector Y and the remaining ones are the covariates X. The arguments `ties.method` and `ties.Gini` indicate how ties are handled. By default, both arguments are set to `"random"`. The existence of two separate arguments stems from the fact that ties influence first the estimation of θ_0 through Equation (2.3.2) and, second, the estimation of $\mathrm{Gi}_{Y,X}$ through Equation (2.4.10). To ensure the comparability with other methods, one could choose to estimate the explained Gini coefficient with the mean rank solution, even though the random allocation is used to estimate θ_0. In that case, one would set `ties.method="random"` but `ties.Gini="mean"`. The argument `seed.random` imposes a fixed seed before generation of the sample of uniform data when the random allocation is used. The default is `NULL`, in which case no seed is imposed. Finally, a vector of sample weights can be provided using the argument `weights`. By default, it is set to `NULL`, in which case all observations are given a weight of $1/n$. We refer the reader to the help file of `Lorenz.GA()` for details on additional arguments. To speed up computations, the fitness function displayed in Equation (2.3.9) is written in C++ and integrated in R using **RcppArmadillo**.

5.2.2 The penalized Lorenz regression

In the **LorenzRegression** package, the penalized Lorenz regression may be implemented using either the LASSO or the SCAD penalty. The former is solved with the FABS algorithm, described in Subsection 3.3.1, and implemented in the function `Lorenz.FABS()`. The latter is solved using the SCAD-FABS algorithm proposed in Subsection 3.3.2 and implemented in the function `Lorenz.SCADFABS()`. In what follows, we focus on the `Lorenz.SCADFABS()` function. Its main arguments are described in the following code excerpt.

```
Lorenz.SCADFABS(YX_mat, weights=NULL, h, eps, a = 3.7,
                lambda="Shi", gamma = 0.05)
```

As in the function `Lorenz.GA()`, `YX_mat` and `weights` provide the data and optional sample weights. The remaining arguments correspond to tuning parameters of either the penalized problem or of the SCAD-FABS algorithm. The smoothness of the objective function is determined by the kernel and the bandwidth. Concerning the former, we impose $K(u) = [\frac{9}{8}u - \frac{5}{8}u^3 + \frac{1}{2}]\mathbb{1}\{-1 \leq u \leq 1\}$ to ensure good theoretical properties, where $K(\cdot)$ is the integral of the kernel, see Equation (3.2.2). The bandwidth h is determined via the argument `h`. The step size ϵ of the

SCAD-FABS algorithm is set by the argument `eps`. The SCAD penalty depends on two parameters, a and λ. The former is ruled by the argument `a`, with default value of 3.7, as advised in Fan and Li (2001). The argument `lambda` takes three possible values. By default, it is set to `"Shi"`, where the path of λ values is determined within the algorithm, as described in Subsestion 3.3.2. If set to `"grid"`, the path is defined by a grid, equidistant on the log scale. Otherwise, the user may provide a vector of values. Finally, the argument `gamma` corresponds to the Lagrange multiplier of Equation (3.3.9), with a default value of 0.05.

The smoothness of the kernel and the extent of penalization are crucial parameters of the estimation procedure. A careful choice for the bandwidth h and regularization parameter λ is therefore needed. The SCAD-FABS algorithm returns vectors of estimated coefficients for a path of λ values and a fixed h. As proposed in Jacquemain et al. (2022a), we advise to repeat this process over a grid of values for the bandwidth satisfying $h = Cn^{-1/5.5}$, where C is a positive constant. In the **LorenzRegression** package, three methods are available to select the pair (λ,h): BIC, bootstrap and cross-validation. The first two methods were discussed in Section 2.3 so we focus here on the last one. In K-fold cross-validation, the data are split into K folds of roughly equivalent sizes. For each iteration, $(K-1)$ folds are used as training samples and the remaining one is used as validation sample. The procedure is then similar to the bootstrap selection. A cross-validation score is defined for each iteration, call it $\widetilde{\text{Gi}}_{-k}(\lambda, h)$. The cross-validation selection procedure chooses the pair (λ, h) that maximizes the cross-validation score

$$\text{CV-score}_{(\lambda,h)} = \frac{1}{K}\sum_{k=1}^{K} \widetilde{\text{Gi}}_{-k}(\lambda, h).$$

5.3 End-user functions and methods

The **Lorenz.Regression** package offers several functions to analyze the inequality pattern characterizing the input data. We first describe the function `Gini.coef()`, which computes a Gini coefficient or a concentration index. Its arguments are

- `y`: vector gathering the variable of interest for each observation.
- `x`: vector gathering the variable to use for the ranking of each observation. By default, it is set to `y` and the function computes the Gini coefficient of `y`. If the user supplies a vector `x`, the output becomes the concentration index of `y` with respect to `x`.
- `na.rm`: should missing values be deleted. Default value is `TRUE`. If `FALSE` is selected, missing values generate an error message.

- `ties.method`: determines how ties are handled. By default, the argument is set to `"mean"` and the average-rank is used. If set to `random`, the random allocation is used instead.
- `seed`: fixes what seed is used if the random allocation is chosen. Default is `NULL`, in which case no seed is imposed.
- `weights`: vector of sample weights. Default is NULL, in which case each observation is given a weight of $1/n$, where n is the number of observations.

With similar arguments, the function `Lorenz.curve()` can either produce the Lorenz curve of `y` or the concentration curve of `y` with respect to `x`. The output is a function, whose unique argument is a proportion between 0 and 1. The function `Lorenz.curve()` admits an extra logical argument `graph`. If set to TRUE, a plot of the curve is displayed. The default value is FALSE, in which case no plot is produced.

Finally, the function `Lorenz.graphs()` allows the user to plot simultaneously the Lorenz curve of a variable, as well as several concentration curves. It takes the following two main arguments:

- `formula`: standard formula object of the form `Y ~ X1 + X2 + ...`
- `data`: data frame containing the variables displayed in the formula.

The output of the function is a graph of the Lorenz curve of `Y` and of each of the concentration curves of `Y` with respect to the covariates `X1`, `X2`, ... indicated in the formula.

In what follows, we discuss the use of the function `Lorenz.Reg()`. It is the most important function of the package as it allows the user to fit a Lorenz or penalized Lorenz regression. Its usage is similar to other regression functions and several methods such as `print()`, `coef()`, `summary()`, `plot()` and `confint()` have been defined to provide detailed information on the fitted model. It uses the following arguments.

- `formula`: standard formula object of the form `Y ~ X1 + ...` determining the response variable and the covariates.
- `data`: data frame containing the variables displayed in the formula.
- `standardize`: whether covariates should be standardized prior to estimation. Default value is TRUE, in which case covariates are standardized. We advise not to set it to FALSE, especially if a penalized Lorenz regression is fitted. Without standardization, covariates are not treated on equal footing by the penalized procedure.
- `weights`: vector of sample weights. Default is NULL, in which case each observation is given a weight of $1/n$, where n is the number of observations.

- `parallel`: determines whether parallel computing should be used to distribute the computations between different CPUs. Default value is FALSE, in which case no parallel computing is performed. If set to TRUE, the number of core is set to `detectCores()`-1. The user can also supply a numerical value determining the number of cores to use.

- `penalty`: determines whether a non-penalized (`"none"`, default) or a penalized Lorenz regression should be fitted. In the latter case, the user has the choice between the SCAD (`"SCAD"`) and the LASSO (`"LASSO"`) penalty.

The following arguments are particular to the penalized Lorenz regression.

- `h.grid`: grid of values for the bandwidth. The default is determined as

  ```
  h.grid=c(0.1,0.2,1,2,5)*nrow(data)^(-1/5.5)
  ```

- `eps`: step size of the algorithm. Default value is 0.005. Typically, lower values of `eps` lead to longer and finer paths but require more computation time.

- `sel.choice`: determines the selection method for the bandwidth and regularization parameter. As explained in Section 5.2, three methods are available: BIC, cross-validation and bootstrap. Possible values are any subset of the vector `c("BIC","CV","Boot")`. The default is `"BIC"`.

The next three parameters are only used if `sel.choice` contains `"CV"`.

- `nfolds`: number of folds. Default value is 10.

- `seed.CV`: value used as seed before the folds are formed. Default value is NULL, in which case no seed is imposed.

- `foldID`: determines how the folds are formed. The default value is NULL, in which case the folds are determined randomly. Otherwise, the user can supply a vector of size n determining the index of the fold for each observation.

Several parameters are particular to the bootstrap. In the case of a non-penalized Lorenz regression, bootstrap is only used to perform inference. If a penalized Lorenz regression is fitted instead, bootstrap is used both as selection method and to perform inference.

- `Boot.inference`: determines whether bootstrap inference should be produced. Since bootstrap may require a lot of computation time, the default value is FALSE. In the case of a penalized Lorenz regression, this parameter is automatically turned to TRUE if `sel.choice` contains `"Boot"`. Conversely, if `Boot.inference` is set to TRUE, `"Boot"` is added to `sel.choice`.

- `B`: number of bootstrap repetitions. Default value is 500.

- `alpha`: significance level for the bootstrap confidence intervals. Default is 0.05.
- `bootID`: determines how the bootstrap resamples are formed. The default value is NULL, in which case the samples are drawn with replacement from the original data. Otherwise, the user can supply a matrix where each row provides the ID of the observations selected in each bootstrap resample.
- `seed.boot`: value used as seed before the bootstrap resamples are formed. Default value is NULL, in which case no seed is imposed.

The last two arguments are mainly useful when the estimation is performed in several steps. For example, the original estimation is fitted on a single computer and the bootstrap iterations are distributed between different machines.

- `LR`: output of a call to `Lorenz.GA()` or to `PLR.wrap()`. The latter is a wrapper of functions `Lorenz.FABS()` and `Lorenz.SCADFABS()`. We refer the reader to the help of this function for extra information.
- `LR.boot`: output of a call to `Lorenz.boot()`, which allows the user to perform bootstrap for the Lorenz and penalized Lorenz regressions. Again we refer the reader to the help for further details.

Finally, additional parameters corresponding to arguments of functions `Lorenz.GA()`, `Lorenz.FABS()` and `Lorenz.SCADFABS()` can be passed in the `...` argument.

If the `penalty` argument is set to `none`, the function outputs an object of class `LR`. Otherwise, it outputs an object of class `PLR`. The associated S3 methods are illustrated in the next section.

5.4 Illustration

We start by loading the packages required in this example.

```
R> library(LorenzRegression)
R> library(ineq)
```

Throughout this illustration, we use the dataset `Ilocos` introduced in Section 1.1. The data are loaded with

```
R> data("Ilocos", package="ineq")
```

5.4.1 Descriptive analysis

The Gini coefficient of `income` and the concentration index of `income` with respect to `family.size` are computed using `Gini.coef()`. Since `family.size` is a discrete

covariate, ties appear in the ordering generated by this variable. In what follows, we repeat the analysis performed in Subsection 2.5.2, i.e. we compare the values of the concentration index obtained with the mean-rank and the random assignments. In the function `Gini.coef()`, the mean-rank is used by default.

```
R> Gi <- Gini.coef(Ilocos$income)
R> Gi
[1] 0.43
R> CI.mean <- Gini.coef(Ilocos$income,Ilocos$family.size)
R> CI.mean
[1] 0.089
```

In what follows, we use the random allocation instead, repeating the operation 1000 times with a different seed.

```
R> CI.random <- sapply(1:1000, function(i)Gini.coef(Ilocos$income,
+                                                   Ilocos$family.size,
+                                                   ties.method = "random",
+                                                   seed = i))
R> boxplot(CI.random)
R> abline(h=CI.mean)
```

Figure 5.1 displays a boxplot of the concentration indices thus obtained. As we explained at the end of Subsection 2.5.2, the distribution is centred on the solution obtained with the mean-rank, indicated by the horizontal line. The size of the boxplot indicates how the concentration index varies as we change the rule determining how ties are broken. We now compute the Lorenz curve of `income` and evaluate it on a small grid of values.

```
R> LC <- Lorenz.curve(Ilocos$income)
R> p <- seq(0,1,length.out=5)
R> LC(p)
[1] 0.000 0.079 0.214 0.446 1.000
```

Finally, we display the Lorenz curve of `income` and the concentration curve of `income` with respect to `family.size` in Figure 5.2. They are obtained using the function `Lorenz.graphs()`.

```
R> Lorenz.graphs(income ~ family.size, data = Ilocos)
```

5.4.2 The Lorenz regression

We fit the Lorenz regression to the `Ilocos` data, with `income` as response variable. The covariates include `sex`, `family.size`, `urbanity` and `province`. We also include interactions between `sex` and `family.size`, and between `sex` and `urbanity`. We use the mean rank in the estimation of the explained Gini coefficient, by setting `ties.Gini = "mean"`.

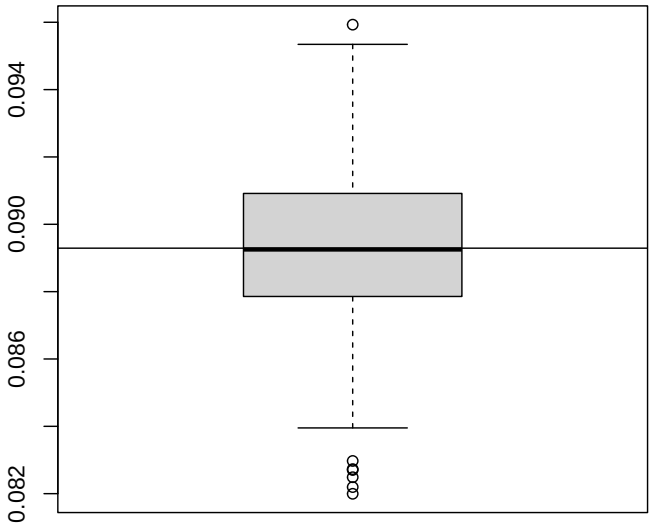

Figure 5.1: Concentration index of `income` with respect to `family.size` - Random allocation

```
R> LR <- Lorenz.Reg(income ~ sex*family.size + sex*urbanity + province,
+                   data = Ilocos, ties.Gini = "mean", seed.random = 456,
+                   Boot.inference = TRUE)
```

The vector of estimated coefficients is retrieved with the method `coef()`.

```
R> coef(LR)
              sexmale           family.size         urbanityurban
                0.318                 0.126                 0.845
   provinceIlocos Sur      provinceLa Union    provincePangasinan
                0.071                -0.219                -0.226
  sexmale:family.size sexmale:urbanityurban
               -0.028                -0.254
```

A summary of the model is provided with the method `summary()`. The estimated explained Gini coefficient is of 15.4%, which corresponds to 36.2% of the observed inequality. The summary also displays a table with the estimated coefficients. Let us focus on covariate `family.size`. In Figure 5.2, we observed a positive association between `income` and `family.size`. This is confirmed by the regression model since the estimated coefficient is positive. In the descriptive analysis, we also computed a concentration index of 8.9%. This would be the value of the explained Gini coefficient if we included only `family.size` in the regression. Including the remaining covariates helps us rise explained inequality from 8.9% to 15.5%.

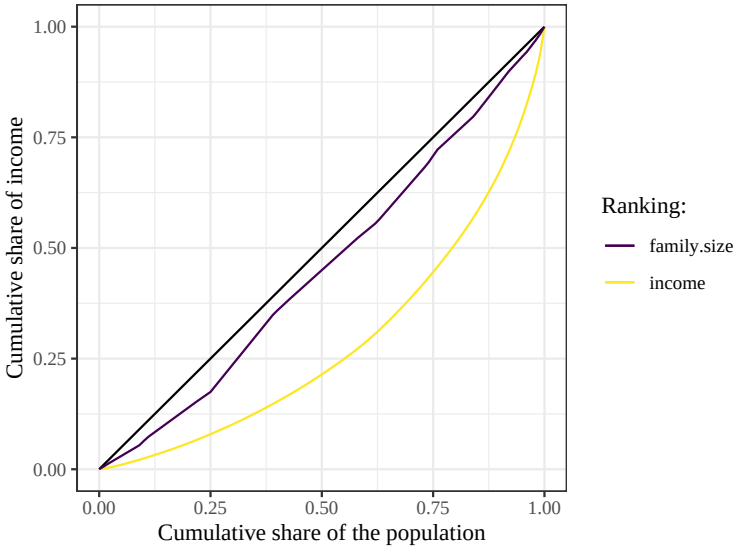

Figure 5.2: Lorenz curve of `income` and concentration curve of `income` with respect to `family.size`

```
R> summary(LR)
The explained Gini coefficient is of 0.15452

The Lorenz-R2 is of 0.36191

Estimated coefficients and associated p-values
|                       | estimate| p-value|
|:----------------------|--------:|-------:|
|sexmale                |     0.32|    0.21|
|family.size            |     0.13|    0.01|
|urbanityurban          |     0.84|    0.00|
|provinceIlocos Sur     |     0.07|    0.75|
|provinceLa Union       |    -0.22|    0.20|
|provincePangasinan     |    -0.23|    0.12|
|sexmale:family.size    |    -0.03|    0.43|
|sexmale:urbanityurban  |    -0.25|    0.33|
```

Because we set the argument `Boot.inference` to `TRUE`, p-values were computed for each coefficient. At a 5%-level of significance, we see that only `family.size` and `urbanityurban` are significant. A confidence interval for the explained Gini coefficient is obtained with the method `confint()`.

```
R> confint(LR)
Lower bound Upper bound
       0.12        0.19
```

By default, 95% confidence intervals are constructed, but the level can be changed using the argument `level`. The three bootstrap methods introduced in Section 2.3 are available and are chosen via the argument `boot.method`. The default is `"Param"`, which corresponds to parametric bootstrap. The other options are `"Perc"` for percentile bootstrap and `"Basic"` for basic bootstrap.

Confidence intervals for each individual coefficient can be obtained by changing the `parm` argument to `"theta"`. We also change the significance level and the bootstrap method to illustrate the use of arguments `level` and `boot.method`.

```
R> confint(LR, parm = "theta", level = 0.9, boot.method = "Perc")
                      Lower bound Upper bound
sexmale                    -0.373       0.480
family.size                 0.022       0.175
urbanityurban               0.601       0.918
provinceIlocos Sur         -0.245       0.470
provinceLa Union           -0.375       0.179
provincePangasinan         -0.399       0.106
sexmale:family.size        -0.056       0.059
sexmale:urbanityurban      -0.520       0.343
```

We can also display explained inequality via concentration and Lorenz curves. In the following piece of code, we make use of the function `Lorenz.graphs()` to produce Figure 5.3. The green curve is the Lorenz curve of `income`. The purple curve is the concentration curve of `income` with respect to `family.size`. Finally, the yellow curve is the concentration curve of `income` with respect to the index estimated by the Lorenz regression. The closer the yellow curve to the green curve, the greater the Lorenz-R^2, i.e. the more observed inequality is explained by the regression model.

```
R> data.plot <- cbind(LR$Fit,Ilocos$family.size)
R> colnames(data.plot) <- c("income","index_LR","family_size")
R> Lorenz.graphs(income ~ ., data.plot)
```

5.4.3 The penalized Lorenz regression

We fit the penalized Lorenz regression to the `Ilocos` data, with the same response variable and covariates as before. We use the SCAD penalty and choose the couple

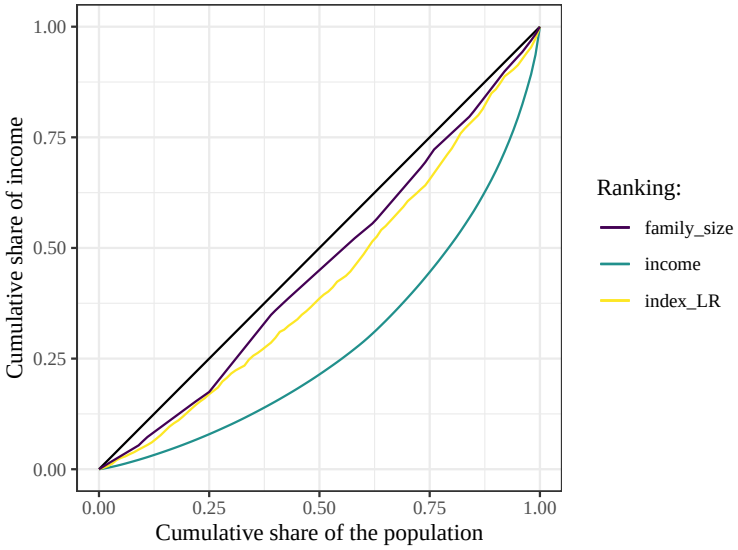

Figure 5.3: Lorenz curve of `income`, concentration curve of `income` with respect to `family.size` and concentration curve of `income` with respect to the estimated index

(h, λ) by BIC and bootstrap. The step size of the SCAD-FABS algorithm is set to $\epsilon = 0.01$. Finally, parallel computing is used to distribute the bootstrap iterations across the available CPUs.

```
R> PLR <- Lorenz.Reg(income ~ sex*family.size + sex*urbanity + province,
+                     data = Ilocos,
+                     penalty = "SCAD",
+                     sel.choice = c("BIC","Boot"),
+                     seed.boot = 123,
+                     eps = 0.01,
+                     parallel = TRUE)
```

The coefficients estimated by the SCAD-FABS algorithm are obtained with the method `coef()`. The output is a list where each element corresponds to a method of selection for the couple (λ, h).

```
R> coef(PLR, renormalize = FALSE)
$BIC
               sexmale             family.size          urbanityurban
                 0.000                   0.119                   0.767
   provinceIlocos Norte      provinceIlocos Sur        provinceLa Union
                 0.118                   0.247                   0.000
     provincePangasinan   sexfemale:family.size     sexmale:family.size
```

```
                -0.178                   0.023                   0.000
sexfemale:urbanityrural   sexmale:urbanityrural sexfemale:urbanityurban
                -0.539                   0.000                   0.000
  sexmale:urbanityurban
                 0.000

$Boot
                sexmale             family.size           urbanityurban
                  0.000                   0.099                   0.995
   provinceIlocos Norte     provinceIlocos Sur        provinceLa Union
                  0.000                   0.000                   0.000
     provincePangasinan   sexfemale:family.size     sexmale:family.size
                  0.000                   0.000                   0.000
sexfemale:urbanityrural   sexmale:urbanityrural sexfemale:urbanityurban
                  0.000                   0.000                   0.000
  sexmale:urbanityurban
                  0.000
```

In presence of categorical covariates, the fit of the penalized Lorenz regression would depend on the chosen reference category, see Section 4.3 for a discussion. We avoid this issue by introducing all dummies in the regression, except for variables with only two categories. This is the reason why the four provinces are included in the output of `coef(PLR, renormalize = FALSE)`. Once the estimation is performed, the vector of coefficients can be transformed to match a specific choice of reference categories. By setting `renormalize=TRUE` (the default), the reference categories are chosen as in the non-penalized case, i.e. as the first categories in the alphabetical order. The method `summary()` displays these transformed coefficients as well as information about the model fit.

```
R> summary(PLR)
Summary of the model fit
|      | Explained Gini| Lorenz-R2| lambda|     h| Number of variables| BIC score| Boot score|
|:-----|--------------:|---------:|------:|-----:|-------------------:|---------:|----------:|
|BIC   |         0.1541|    0.3609|  114.1| 1.548|                   7|    -1.906|     0.1306|
|Boot  |         0.1435|    0.3362| 3442.9| 1.548|                   2|    -1.951|     0.1346|

Estimated coefficients
|                       |     BIC|   Boot|
|:----------------------|-------:|------:|
|sexmale                |  0.3462| 0.0000|
|family.size            |  0.0910| 0.0989|
|urbanityurban          |  0.8386| 0.9951|
|provinceIlocos Sur     |  0.0832| 0.0000|
|provinceLa Union       | -0.0756| 0.0000|
|provincePangasinan     | -0.1897| 0.0000|
|sexmale:family.size    | -0.0147| 0.0000|
|sexmale:urbanityurban  | -0.3462| 0.0000|
```

The summary of the model fit includes the estimated explained Gini coefficient and Lorenz-R^2, the selected couple (λ, h), the number of included variables, as well as the BIC and bootstrap scores. Both selection methods lead to the same choice for

the bandwidth, i.e. $h = 1.548$. The bootstrap selection method favours a larger penalty: $\lambda = 3273.96$, against $\lambda = 114.13$ for the BIC criterion, and therefore yields a sparser model. The amount of sparsity is represented by the number of non-zero coefficients before choosing a reference category, which are displayed in column `Number of variables`. With the bootstrap selection, only 2 variables have a non-zero coefficient, against 7 variables with the BIC procedure. The latter yields an output very similar to the non-penalized regression. The estimated Gini coefficient is of 15.41%, against 15.45% in the non-penalized case. The estimated coefficients are also of similar magnitudes. The bootstrap procedure yields an estimated Gini coefficient of 14.38%.

```
R> plot(PLR)
```

The method `plot()` displays several graphs. The first plot is the Lorenz curve of the response, along with the concentration curves of the response with respect to the index obtained by each selection method. We do not display this graph here as we want to compare with the fit obtained with the non-penalized Lorenz regression. Therefore, we display instead Figure 5.4, obtained by running the following piece of code.

```
R> data.plot$index_PLR.BIC <- PLR$Fit$Index.BIC
R> data.plot$index_PLR.Boot <- PLR$Fit$Index.Boot
R> Lorenz.graphs(income ~ ., data.plot)
```

As expected, the concentration curve obtained with BIC is very close to that obtained with the non-penalized procedure. The concentration curve obtained with bootstrap reproduces less inequality, since it lies closer to the 45° line. The second plot, given in Figure 5.5, displays the evolution of the estimated coefficients (before transformation related to the choice of reference categories) as the penalty relaxes. At the beginning, the penalty is the largest and only one coefficient is included. This coefficient corresponds to `urbanityurban` and it takes a value of 1 because of the norm constraint. As the algorithm proceeds, the penalty decreases and new variables enter the model. For example, `family.size` is the second variable to enter. The third plot, provided in Figure 5.6, shows the evolution along the path of the scores corresponding to each selection method. To ease the comparison, the scores are normalized so that the optimum is attained at 1. The bootstrap score yields a well-defined maximum in the beginning of the path, producing a highly sparse model. The BIC score, however, is relatively flat along the path. Therefore, we would favour using the former. Figures 5.5 and 5.6 indicate that the bandwidth was selected by BIC. Since the whole path depends on the choice of h, these two plots are reproduced for each selection method. In this illustration, the same bandwidth is selected with both methods. The plots related to the bootstrap are exactly the same as those obtained for BIC and are therefore not reproduced.

Finally, we obtain confidence intervals for the explained Gini coefficient using the method `confint()`.

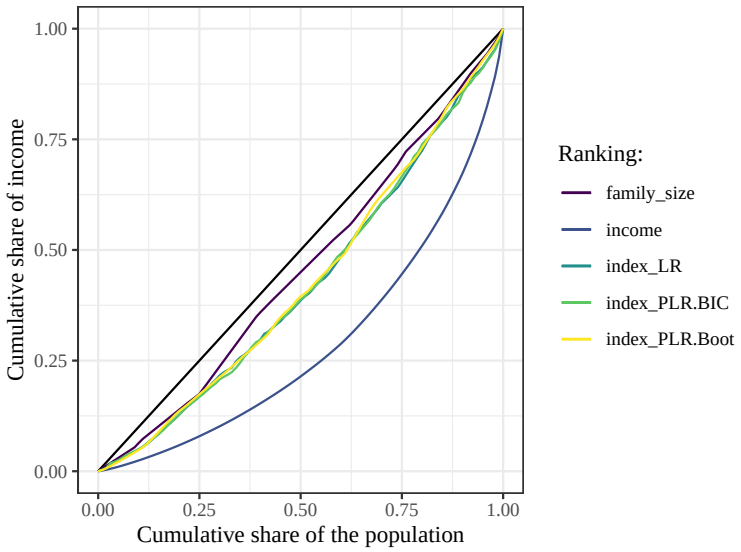

Figure 5.4: Lorenz curve of `income`, concentration curve of `income` with respect to `family.size` and concentration curves of `income` with respect to each estimated index

```
R> confint(PLR)
     Lower bound Upper bound
BIC         0.12        0.19
Boot        0.10        0.18
```

5.5 Discussion

In this chapter, we demonstrate the use of the **LorenzRegression** package, which implements the non-penalized and penalized Lorenz regressions. The fitting of the model is similar to standard regression techniques. It is performed with the function `Lorenz.Reg()` and it benefits from several S3 methods. Information on the model fit is obtained numerically with `summary()`, or graphically with `plot()`. Confidence intervals for the explained Gini coefficient are obtained with `confint()`. As a side advantage, the package allows the computation and graphical representation of Lorenz and concentration curves, as well as the computation of a Gini coefficient or a concentration index. All functions allow the user to specify sample weights.

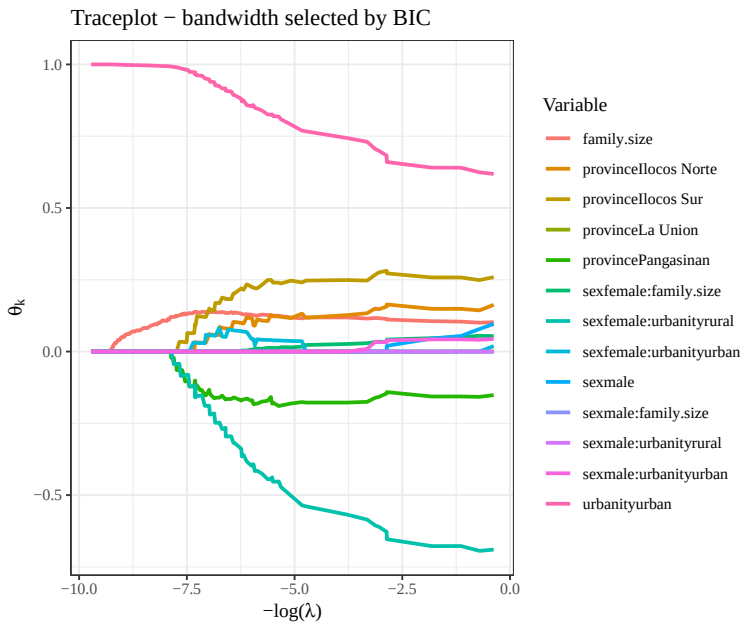

Figure 5.5: Evolution of the estimated coefficients as the penalty relaxes

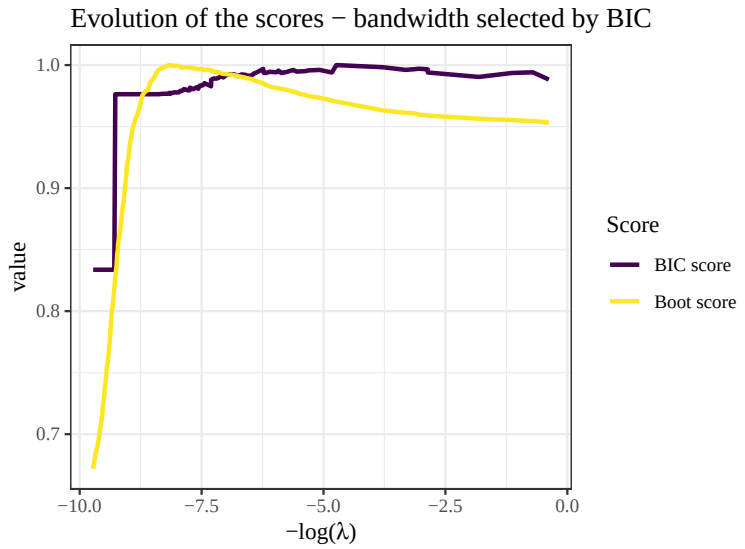

Figure 5.6: Evolution of the scores for each selection method as the penalty relaxes

CHAPTER 6

Discussion and extensions

In this thesis, we developed the Lorenz regression, an inference procedure for the explained Gini coefficient, a measure of the inequality of a response that is explained by a set of covariates. Besides providing inference, the procedure can be used to select the relevant covariates and to determine their contributions to the inequality thus measured. In Chapter 4, we argued that the Lorenz regression is a promising technique to measure inequality of opportunity. More generally, one may resort to this procedure whenever one is interested in analyzing the inequality of "fitted values", where the latter is formalized by the conditional expectation of Y given X. This chapter proposes three directions in which the methodology may be expanded.

(i) *Beyond inequality measurement.* Throughout this work, we suppose that the response variable Y is a continuous random variable with a positive expectation. This assumption is justified by the context of application, as Y is a measure of economic advantage, e.g. wealth, income or wage. Section 6.1 considers the very opposite situation, where Y is a binary variable. Assuming a single-index model, the index $X^\intercal\theta$ is a binary classifier for Y. It turns out that the Lorenz regression may still be applied. It looks for the binary classifier which maximizes the area under the receiver operating characteristic curve.

(ii) *Beyond the Gini coefficient.* As a scalar measure of inequality, the Gini coefficient is characterized by a certain degree of inequality aversion. With the extended Gini coefficient, the level of inequality aversion is specified by the researcher. It is therefore natural to determine whether the Lorenz regression methodology can be extended in this way. This is the subject of Section 6.2.

(iii) *Beyond the single-index model.* The Lorenz regression is built on the premise of the single-index model. This assumption is elegant in the sense that it allows for more flexibility than parametric methods without suffering from the

curse of dimensionality. Furthermore, the nonparametric part of the model does not need to be estimated. Still, the structure of the model may seem restrictive, as it assumes that the ordering of the conditional expectation is a linear combination of the covariates. Alternatively, one may think of using machine learning methods to obtain fitted values and plug these in an estimator of the Gini coefficient. In Section 6.3 we start from the idea, already expressed in Escanciano and Terschuur (2022) that such a procedure is likely to yield biased estimates. Instead, the concentration index of the response with respect to the fitted values should be used. We propose a regression tree procedure where each split is chosen so as to maximize the concentration index of Y with respect to the fitted values.

6.1 An application of the Lorenz regression in binary classification

This section develops as follows. In Subsection 6.1.1, we give a brief introduction to binary classification in order to introduce the relevant material. The most important notions are the receiver operating characteristic curve and its summary measure, the area under the curve. In Subsection 6.1.2, we show that, assuming a single-index model, the Lorenz regression entails to find the binary classifier which maximizes the area under the curve.

6.1.1 A small introduction to binary classification

Consider a binary response variable $Y \in \{0, 1\}$ and a vector of covariates $X = (X_1, \ldots, X_p)^\intercal \in \mathbb{R}^p$, also called *predictors*. In the context of economics, Y could measure the employment status of an individual. The value $Y = 0$ is attributed to unemployment, while $Y = 1$ is used to denote employment. The goal is then to determine a *decision rule* $\delta(X)$, where $\delta : \mathbb{R}^p \to \{0, 1\}$, that determines whether the individual is employed, i.e. $\delta(X) = 1$, or unemployed, i.e. $\delta(X) = 0$. Finally, we denote by $\{\delta_c(X)\}_c$ a series of decision rules depending on a *threshold* c.

We assume the single-index model used throughout this thesis. Namely,

$$P(Y = 1|X = x) = H(x^\intercal \theta_0), \tag{6.1.1}$$

where $H(\cdot)$ is a strictly increasing function and θ_0 is a vector of parameters satisfying $||\theta_0|| = 1$. The same model is assumed in Pepe et al. (2006). As we will formalize later on, it turns out that the series of decision rules of the type

$$\delta_{c^*}(X) = 1 \quad \text{if } X^\intercal \theta_0 \geq c^* \qquad \text{and} \qquad \delta_{c^*}(X) = 0 \quad \text{if } X^\intercal \theta_0 < c^*, \tag{6.1.2}$$

where c^* is a constant, is optimal. The performance of the decision can be assessed by turning the attention to the *true-positive rate* (TPR) and *false-positive rate*

(FPR), formally defined as

$$\begin{aligned}\mathrm{TPR}(c) &\coloneqq P(\delta(X) = 1|Y = 1) = P(X^\intercal\theta_0 \geq c|Y = 1)\\ \mathrm{FPR}(c) &\coloneqq P(\delta(X) = 1|Y = 0) = P(X^\intercal\theta_0 \geq c|Y = 0),\end{aligned}$$

where the last equality in each line only holds if the decision rule displayed in Equation (6.1.2) is chosen.

For a given series of decision rules $\{\delta_c(X)\}_c$, the receiver operating characteristic (ROC) curve is defined as

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

where $p \in [0, 1]$. Evaluated at p, the ROC curve provides the true-positive rate of the decision rule ensuring a false-positive rate equal to p. Typical ROC curves are represented in red and green in Figure 6.1. In what follows, we summarize the

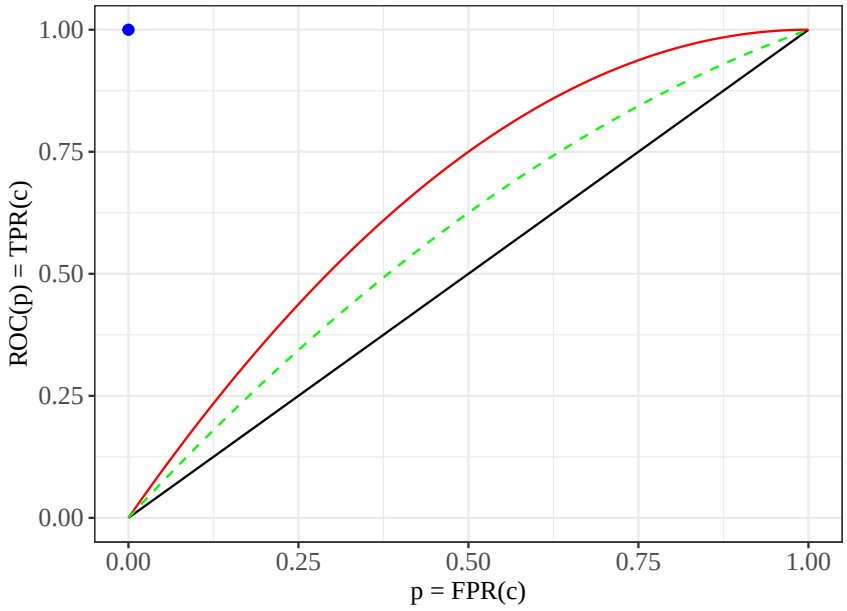

Figure 6.1: Typical ROC curves

main features governing the interpretation of a ROC curve. For a more complete review, we refer the reader to Chapter 4 of Pepe (2004).

Two particular scenarios are of interest. First, the *perfect* test would ensure an FPR of 0 and a TPR of 1. In figure 6.1, this situation is represented by the blue point with coordinates $(0, 1)$. Second, an *uninformative* test with FPR and TPR

both equal to p is obtained by randomly assigning $\delta(X) = 1$ with probability p and $\delta(X) = 0$ with probability $1 - p$. The ROC curve traced by this series of decisions coincides with the 45° line. In our context, this is the ROC curve obtained for the rule based on the binary classifier $X^\intercal\theta_0$ if Y is independent from X.

Typically, the ROC curve is an increasing and concave curve nested between the two scenarios mentioned above. The closer the curve is to the upper left, the better is the classification. In Figure 6.1, the series of decision rules corresponding to the solid red ROC curve are better than the ones yielding the dashed green ROC curve. We say that the red ROC curve *dominates* the green one. In what follows, we consider that a series of decision rules $\delta_c(X)_c$ is optimal when its ROC curve dominates that of any other rule based on X. A summary measure of the ROC curve is the *area under the curve* (AUC), defined as

$$\text{AUC} \coloneqq \int_0^1 \text{ROC}(p)dp.$$

By the Neyman-Pearson lemma (Neyman and Pearson, 1933), the series of decision rules defined by

$$\delta_c(X) = 1 \quad \text{if } \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} \geq c$$
$$\delta_c(X) = 0 \quad \text{if } \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} < c$$

is optimal. Now, due to Equation (6.1.1) and the fact that $H(\cdot)$ is strictly increasing, these decisions boil down to those displayed in Equation (6.1.2). Hence, the decision based on the index $X^\intercal\theta_0$ is optimal. In practice θ_0 is unknown. Pepe et al. (2006) propose an estimation procedure which entails to maximize the empirical area under the curve. In the context of this thesis, we saw that the estimator $\overline{\theta}$ obtained from the Lorenz regression, see Equation (1.5.4), is also a suitable estimator of θ_0. In the following subsection, we explore the possible links between the Lorenz regression methodology and procedures based on the ROC curve and the AUC.

6.1.2 Link with the Lorenz regression

In what follows, we provide a small comparison between the estimator proposed by Pepe et al. (2006), based on the maximization of the AUC, and the estimator arising from the Lorenz regression.

Throughout this section, we consider that $(X^\intercal, Y)^\intercal \in \mathbb{R}^p \times \{0, 1\}$ satisfy Equation (6.1.1) with $H(\cdot)$ strictly increasing. We also assume that the conditions (SI1) to (SI4) introduced in Section 1.5 hold. Finally, we restrict our attention to the series of decision rules described in Equation (6.1.2).

Let $(X_1^\intercal, Y_1)^\intercal$ and $(X_2^\intercal, Y_2)^\intercal$ be two i.i.d copies of $(X^\intercal, Y)^\intercal$. As established in Result 4.6 from Pepe (2004), the AUC can be rewritten as

$$\text{AUC}(\theta) = P(X_1^\intercal\theta > X_2^\intercal\theta | Y_1 = 1, Y_2 = 0),$$

where the notation $\text{AUC}(\theta)$ is used to acknowledge explicitly the dependence on θ. As we pointed out in Section 2.2, the Lorenz regression aims to find the parameter vector which maximizes the concentration index of Y with respect to $X^\intercal\theta$. The latter is

$$\text{CI}_{Y,X^\intercal\theta} = \frac{2\mathbb{E}[Y_1\mathbb{1}\{X_1^\intercal\theta > X_2^\intercal\theta\}]}{\mathbb{E}[Y_1]} - 1.$$

But

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

Hence,

$$\text{CI}_{Y,X^\intercal\theta} = P(Y = 0)(2\text{AUC}(\theta) - 1).$$

Therefore, the maximization programme underlying the Lorenz regression in the population boils down to a maximization of the AUC. We now show that the equivalence extends to the sample.

Let $(X_i^\intercal, Y_i)_{i=1,\dots,n}^\intercal$ be a sample of i.i.d copies of $(X^\intercal, Y)^\intercal$. Let $\mathcal{I}_1 \coloneqq \{i = 1, \dots, n : Y_i = 1\}$ and $\mathcal{I}_0 \coloneqq \{i = 1, \dots, n : Y_i = 0\}$. The estimator of Pepe et al. (2006) is defined as

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

In the Lorenz regression, the estimator $\overline{\theta}$ is the maximizer of $G_n(\theta)$, where

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

Notice that the second sum is a constant since $\sum_{j \in \mathcal{I}_1} \mathbb{1}\{X_i^\intercal \theta > X_j^\intercal \theta\}$ evaluated at each $i \in \mathcal{I}_1$ provides the rank vector $\{0, \ldots, |\mathcal{I}_1|-1\}$, assuming that ties cannot occur. Hence, $\hat{\theta}_{\text{AUC}} = \overline{\theta}$.

The potential of the Lorenz regression in the context of binary classification therefore arises from these findings. The whole machinery developed in this thesis can be used to estimate the binary classifier $X^\intercal \theta_0$ which predicts Y in the best way. The Lorenz regression is suitable because it looks for the parameter vector which maximizes the area under the ROC curve. The penalized Lorenz regression may be applied if the researcher is facing many predictors and looks for an automatic selection procedure. The Shapley value decomposition may be reframed to decompose the AUC of the final model into individual contributions of each predictor.

6.2 A generalization of the Lorenz regression to the extended Gini family

The extended Gini coefficient is a family of inequality measures indexed by an inequality aversion parameter. In this Section, we provide the backbone for an extended Lorenz regression. In Subsection 6.2.1, we define the extended Gini coefficient and provide the intuition underlying its use. The extended Lorenz regression is sketched in Subsection 6.2.2.

6.2.1 The extended Gini family of inequality measures

The following discussion is a brief introduction to the extended Gini family of inequality measures. For a more complete review, we refer the reader to Yitzhaki and Schechtman (2005). As a starting point, notice that the Gini coefficient is characterized by a certain degree of inequality aversion. To see this, remark that the Gini coefficient of Y can be written as

$$\text{Gi}_Y = 1 - \frac{\mathbb{E}[\min(Y_1, Y_2)]}{\mathbb{E}[Y]},$$

where Y_1 and Y_2 are i.i.d copies of Y. The Gini coefficient compares the expected income received by the poorest member of a pair of individuals randomly drawn from the population with the overall expected income. A different degree of inequality aversion can be set by extending the number of elements in the numerator. The extended Gini coefficient is

$$\text{Gi}_{Y;\nu} \coloneqq 1 - \frac{\mathbb{E}[\min(Y_1, Y_2, \ldots, Y_\nu)]}{\mathbb{E}[Y]},$$

where ν is a parameter determined by the researcher. The Gini coefficient is recovered when $\nu = 2$. A larger value of the parameter ν corresponds to more inequality

aversion since the numerator retrieves the expected income received by the poorest member from a set of ν individuals randomly drawn from the population. Back to our usual formulation, the extended Gini coefficient is defined as

$$\mathrm{Gi}_{Y;\nu} \coloneqq \frac{-\nu\mathbb{C}[Y,(1-F_Y(Y))^{\nu-1}]}{\mathbb{E}[Y]},$$

where $\nu \geq 1$. This formulation is more general than the previous one as ν is not restricted to be integer-valued. The extended concentration index of Y with respect to W is given by

$$\mathrm{CI}_{Y,W;\nu} \coloneqq \frac{-\nu\mathbb{C}[Y,(1-F_W(W))^{\nu-1}]}{\mathbb{E}[Y]}$$

6.2.2 The extended Lorenz regression

In this section, we provide the skeleton of an extended Lorenz regression. We call *extended explained Gini coefficient* the extended Gini coefficient of the conditional expectation of Y given X in the assumed single-index model. This quantity is denoted by $\mathrm{Gi}_{Y,X;\nu}$. Proposition 6.1 reproduces the chain of equations presented in Proposition 2.1 for the Lorenz regression.

Proposition 6.1. *Let $(X^\intercal, Y)^\intercal \in \mathbb{R}^p \times \mathbb{R}$, with $0 < \mathbb{E}[Y] < \infty$, be continuous random variables satisfying Equation (2.1.1) with $H(\cdot)$ strictly increasing. Let $\nu \geq 1$ be an inequality aversion parameter. Then, it holds*

$$\begin{aligned}
\mathrm{Gi}_{Y,X;\nu} &\coloneqq \mathrm{Gi}_{H(X^\intercal\theta_0);\nu} \\
&= \frac{-\nu\mathbb{C}\Big[Y,(1-F_H(H(X^\intercal\theta_0)))^{\nu-1}\Big]}{\mathbb{E}[Y]} && (6.2.1)\\
&= \frac{-\nu\mathbb{C}\Big[Y,(1-F_{\theta_0}(X^\intercal\theta_0))^{\nu-1}\Big]}{\mathbb{E}[Y]} && (6.2.2)\\
&= \max_{\theta,||\theta||=1} \frac{-\nu\mathbb{C}\Big[Y,(1-F_\theta(X^\intercal\theta))^{\nu-1}\Big]}{\mathbb{E}[Y]}. && (6.2.3)
\end{aligned}$$

The proof of (6.2.1) and (6.2.2) is the same as in Proposition 2.1. At this stage, Equation (6.2.3) remains a conjecture. Notice that the final optimization boils down to maximizing $\mathbb{E}[Y(1-F_\theta(X^\intercal\theta))^{\nu-1}]$. Similarly to the Lorenz regression, the *extended proportion of explained inequality* is given by

$$\mathrm{PEI}_{Y,X;\nu} \coloneqq \frac{\mathrm{Gi}_{Y,X;\nu}}{\mathrm{Gi}_{Y;\nu}}.$$

An estimator of θ_0 is

$$\overline{\theta}_\nu \coloneqq \underset{\theta,||\theta||=1}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} Y_i(1-\hat{F}_i(\theta))^{\nu-1},$$

where we recall that

$$\hat{F}_i(\theta) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_i^\intercal\theta \geq X_j^\intercal\theta\}.$$

It is then possible to generalize the whole toolbox of the Lorenz regression to the extended Gini coefficient. The generalization allows the researcher to impose his or her own parameter of inequality aversion in the estimation process.

6.3 A nonparametric Lorenz regression

In this section, we take a step back from the assumption of the single-index model. Let Y be a response variable and $X = (X_1, \ldots, X_p)^\intercal$ a vector of covariates. We consider that

$$\mathbb{E}[Y|X = x] = m(x),$$

where $m : \mathbb{R}^p \to \mathbb{R}$ is an unknown function. Let $F_m(\cdot)$ be the CDF of $m(X)$. We define the explained Gini coefficient obtained on this more general model as $\mathrm{Gi}_{Y,X} \coloneqq \mathrm{Gi}_{m(X)}$. In typical applications, as we saw in Chapter 4, p tends to be quite large. Consequently, usual nonparametric methods will face the curse of dimensionality. An estimator of $\mathrm{Gi}_{Y,X}$ could be obtained from a two-stages procedure. In the first stage, the conditional expectation $m(X)$ is estimated using a machine-learning method, producing fitted values $(\hat{m}(X_1), \ldots, \hat{m}(X_n))^\intercal$. In the second stage, the fitted values are plugged in an estimator of the Gini coefficient. However, this procedure is expected to perform poorly. As mentioned in Escanciano and Terschuur (2022), machine-learning methods tend to produce biased predictions and the estimator of the Gini coefficient is not robust to these deviations. The biases therefore propagate into the final estimator. Alternatively, we can rewrite the explained Gini coefficient as the concentration index of Y with respect to $m(X)$, i.e.

$$\mathrm{Gi}_{Y,X} = \frac{\mathbb{C}[Y, F_m(m(X))]}{2\mathbb{E}[Y]} = \mathrm{CI}_{Y,m(X)}. \tag{6.3.1}$$

The estimator built upon this equation is therefore

$$\widehat{\mathrm{Gi}}_{Y,X} = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{Y_i}{\overline{Y}} \mathbb{1}\{\hat{m}(X_i) \geq \hat{m}(X_j)\} - \frac{n+1}{n} \tag{6.3.2}$$

where the fitted value $\hat{m}(X_i)$ is obtained in the same way as before. Notice that the usual semiparametric Lorenz regression goes one step further than Equation (6.3.1). Exploiting the structure of the single-index model, the ranking is produced by the index and the fitted values do not need to be obtained. In the pure nonparametric context considered in this chapter, we cannot avoid to estimate $m(X)$. What Equation (6.3.1) indicates is that the fitted values only need to reproduce the ordering of $m(X)$; they do not need to be good predictions per se. The estimator displayed in Equation (6.3.2) is exactly the debiased estimator proposed by Escanciano and Terschuur (2022) in the context of inequality of opportunity. Interestingly, this discussion echoes the finding developed in the context of insurance pricing by Denuit et al. (2021) that machine-learning methods may produce biased predictions for $m(X)$ but reproduce quite well its ordering structure.

We propose to estimate $m(X)$ with a regression tree procedure, where the cost function is given by

$$C_n(\hat{m}) = \sum_{i=1}^{n}\sum_{j=1}^{n} Y_i\Big(\mathbb{1}\{\hat{m}(X_i) > \hat{m}(X_j)\} + \frac{1}{2}\mathbb{1}\{\hat{m}(X_i) = \hat{m}(X_j)\}\Big)$$

where we use the average-rank to deal with ties in the fitted values. The random assignment developed in Section 2.4 could also be used. Using this cost function, one looks for fitted values whose ranks correlate the most with the response. The procedure is then based on a series of binary splits, recursively dividing the sample into groups. For each possible split, the cost function is computed, with fitted values given by the mean responses of the observations inside the groups. The best split is obtained as the one which maximizes the cost function. A stopping criterion and a pruning method should be developed in order to avoid overfitting. Bagging and random forests are straightforward extensions.

References

Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, editors, *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, pages 199–213. Springer, New York, NY, 1998. ISBN 978-1-4612-1694-0. doi: 10.1007/978-1-4612-1694-0_15.

Andreas Alfons and Matthias Templ. Estimation of Social Exclusion Indicators from Complex Surveys: The R Package Laeken, 2012.

Francesco Andreoli and Alessio Fusco. 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, 2019. doi: 10.1016/j.dib.2019.104301.

Anders Björklund, Markus Jäntti, and John E. Roemer. Equality of opportunity and the distribution of long-run income in Sweden. *Social Choice and Welfare*, 39(2/3):675–696, 2012. ISSN 0176-1714.

François Bourguignon, Francisco H. G. Ferreira, and Marta Menéndez. Inequality of Opportunity in Brazil. *Review of Income and Wealth*, 53(4):585–618, 2007. ISSN 1475-4991.

Leo Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1032181158.

Paolo Brunori, Paul Hufe, and Daniel Gerszon Mahler. The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees. SSRN Scholarly Paper ID 3127234, Social Science Research Network, Rochester, NY, 2018.

Michal Brzezinski. The evolution of inequality of opportunity in Europe. *Applied Economics Letters*, 27(4):262–266, 2020. doi: 10.1080/13504851.2019.1613493.

Christopher Cavanagh and Robert P. Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351–381, 1998. ISSN 0304-4076.

V. Chernozhukov, I. Fernández-Val, and A. Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009. ISSN 0006-3444.

Julien Chiquet, Yves Grandvalet, and Guillem Rigaill. On coding effects in regularized categorical regression. *Statistical Modelling*, 16(3):228–237, 2016. ISSN 1471-082X. doi: 10.1177/1471082X16644998.

Franck A. Cowell and Emmanuel Flachaire. Statistical Methods for Distributional Analysis. Technical Report 1507, Aix-Marseille School of Economics, France, 2015.

Yves Croissant and Spencer Graves. *Ecdat: Data Sets for Econometrics*, 2020. URL https://CRAN.R-project.org/package=Ecdat. R package version 0.3-9.

Somesh Das Gupta. Gini association and pseudo lorenz curve. *Communications in Statistics - Theory and Methods*, 28(9):2181–2199, 1999. doi: 10.1080/03610929908832414.

Michel Denuit, Rachel J. Huang, and Christine Wang. Almost marginal conditional stochastic dominance. *Journal of Banking & Finance*, 41:57, 2014. ISSN 0378-4266.

Michel Denuit, Arthur Charpentier, and Julien Trufin. Autocalibration and Tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, 101:485–497, November 2021. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2021.09.001.

Brecht Devleesschauwer, Saveria Willimes, Carine Van Malderen, Peter Konings, and Niko Speybroeck. *rineq: Statistical Analysis of Health Inequalities*, 2017. URL https://github.com/brechtdv/rineq/. R package version 0.0.1.

Dirk Eddelbuettel and James Joseph Balamuta. Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician*, 72(1):28–36, 2018. ISSN 0003-1305. doi: 10.1080/00031305.2017.1375990.

Dirk Eddelbuettel and Conrad Sanderson. RcppArmadillo. *Computational Statistics & Data Analysis*, 71(C):1054–1063, 2014. ISSN 0167-9473. doi: 10.1016/j.csda.2013.02.005.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32(2):407–499, 2004. ISSN 0090-5364. doi: 10.1214/009053604000000067.

Juan Carlos Escanciano and Joël Robert Terschuur. Debiased Semiparametric U-Statistics: Machine Learning Inference on Inequality of Opportunity, 2022.

Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, December 2001. ISSN 0162-1459. doi: 10.1198/016214501753382273.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, December 2007. ISSN 1932-6157, 1941-7330. doi: 10.1214/07-AOAS131.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

Thesia I. Garner. Consumer Expenditures and Inequality: An Analysis Based on Decomposition of the Gini Coefficient. *The Review of Economics and Statistics*, 75(1):134–138, 1993. ISSN 0034-6535.

Zvi Griliches. Wages of Very Young Men. *Journal of Political Economy*, 84(4): S69–S85, 1976. ISSN 0022-3808.

Aaron K. Han. Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, 35(2):303–316, 1987. ISSN 0304-4076. doi: 10.1016/0304-4076(87)90030-3.

Tristen Hayfield and Jeffrey S. Racine. Nonparametric Econometrics: The np Package. *Journal of Statistical Software*, 27:1–32, 2008. ISSN 1548-7660. doi: 10.18637/jss.v027.i05.

Cédric Heuchenne and Alexandre Jacquemain. Inference for monotone single-index conditional means: A Lorenz regression approach. *Computational Statistics & Data Analysis*, 167(C), 2022. ISSN 0167-9473. doi: 10.1016/j.csda.2021.107347.

Joel Horowitz. *Semiparametric and Nonparametric Methods in Econometrics*, volume 692 of *Springer Series in Statistics*. Springer-Verlag, New York, 2009. ISBN 978-0-387-92869-2. doi: 10.1007/978-0-387-92870-8.

Alejandro Hoyos and Ambar Narayan. Inequality of Opportunities Among Children: How Much Does Gender Matter? Working Paper, World Bank, Washington, DC, 2011.

Hidehiko Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1):71–120, 1993.

Alexandre Jacquemain and Cédric Heuchenne. LorenzRegression: An implementation of the Lorenz and penalized Lorenz regressions in R, 2022.

Alexandre Jacquemain, Cédric Heuchenne, and Eugen Pircalabelu. A lasso-type estimation for the lorenz regression. In *the Proceedings of the 22th European Young Statisticians Meeting*, pages 41–45, 2021.

Alexandre Jacquemain, Cédric Heuchenne, and Eugen Pircalabelu. A penalised bootstrap estimation procedure for the explained Gini coefficient, 2022a.

Alexandre Jacquemain, Cédric Heuchenne, and Philippe van Kerm. A Penalized Lorenz regression analysis of inequality of opportunity, 2022b.

Andrew M. Jones and Angel López Nicolás. Measurement and explanation of socioeconomic inequality in health with longitudinal data. *Health Economics*, 13 (10):1015–1030, 2004. ISSN 1099-1050.

Nanak Kakwani, Adam Wagstaff, and Eddy van Doorslaer. Socioeconomic inequalities in health: Measurement, computation, and statistical inference. *Journal of Econometrics*, 77(1):87–103, 1997. ISSN 0304-4076. doi: 10.1016/S0304-4076(96)01807-6.

Roger Koenker and Gilbert Bassett. Regression Quantiles. *Econometrica*, 46(1): 33–50, 1978. ISSN 0012-9682. doi: 10.2307/1913643.

Robert I. Lerman and Shlomo Yitzhaki. Income Inequality Effects by Income Source: A New Approach and Applications to the United States. *The Review of Economics and Statistics*, 67(1):151–156, 1985. ISSN 0034-6535.

Robert I. Lerman and Shlomo Yitzhaki. Improving the accuracy of estimates of Gini coefficients. *Journal of Econometrics*, 42(1):43–47, September 1989. ISSN 0304-4076.

Hua Liang, Xiang Liu, Runze Li, and Chih-Ling Tsai. Estimation and Testing for Partially Linear Single-Index Models. *The Annals of Statistics*, 38(6):3811–3836, 2010. ISSN 0090-5364.

Huazhen Lin and Heng Peng. Smoothed rank correlation of the linear transformation regression model. *Computational Statistics & Data Analysis*, 57:615–630, 2013. doi: 10.1016/j.csda.2012.07.012.

Katarzyna Machowska, Jaroslaw Napora, and Sebastian Wójcik. *wINEQ: Inequality Measures for Weighted Data*, 2022. URL https://CRAN.R-project.org/package=wINEQ. R package version 1.0.1.

Y. P. Mack and Hans-Georg Müller. Derivative Estimation in Nonparametric Regression with Random Predictor Variable. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 51(1):59–72, 1989. ISSN 0581-572X.

Y. P. Mack and B. W. Silverman. Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 61(3):405–415, 1982. ISSN 1432-2064. doi: 10.1007/BF00539840.

Microsoft and Steve Weston. *foreach: Provides Foreach Looping Construct*, 2022. URL https://CRAN.R-project.org/package=foreach. R package version 1.5.2.

Branko Milanovic. *Global Inequality: A New Approach for the Age of Globalization*. Harvard University Press, April 2016. ISBN 978-0-674-73713-6.

Jacob Mincer and Solomon Polachek. Family Investment in Human Capital: Earnings of Women. *Journal of Political Economy*, 82(2):S76–S108, 1974.

É. A. Nadaraya. On Non-Parametric Estimates of Density Functions and Regression Curves. *Theory of Probability & Its Applications*, 10(1):186–190, 1965. ISSN 0040-585X. doi: 10.1137/1110024.

J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231: 289–337, 1933. ISSN 0264-3952.

Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series. Oxford University Press, Oxford, New York, 2004. ISBN 978-0-19-856582-6.

Margaret Sullivan Pepe, Tianxi Cai, and Gary Longton. Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve. *Biometrics*, 62(1):221–229, 2006. ISSN 0006-341X.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Xavier Ramos and Dirk Van de gaer. Approaches to inequality of opportunity: Principles, measures and evidence. *Journal of Economic Surveys*, 30(5):855–883, 2016. doi: 10.1111/joes.12121.

Xavier Ramos and Dirk Van de gaer. Is inequality of opportunity robust to the measurement approach? *Review of Income and Wealth*, 67(1):18–36, 2021. doi: https://doi.org/10.1111/roiw.12448.

John E. Roemer. *Equality of Opportunity*. Harvard University Press, 1998. ISBN 978-0-674-25991-1.

Edna Schechtman and Ricardas Zitikis. Gini indices as areas and covariances: What is the difference between the two representations? *Metron - International Journal of Statistics*, LXIV(3):385–397, 2006.

Luca Scrucca. GA: A Package for Genetic Algorithms in R. *Journal of Statistical Software*, 53:1–37, April 2013. ISSN 1548-7660. doi: 10.18637/jss.v053.i04.

Haim Shalit and Shlomo Yitzhaki. An Asset Allocation Puzzle: Comment. *American Economic Review*, 93(3):1002–1008, 2003.

Xingjie Shi, Yuan Huang, Jian Huang, and Shuangge Ma. A Forward and Backward Stagewise Algorithm for Nonconvex Loss Functions with Adaptive Lasso. *Computational Statistics & Data Analysis*, 124:235–251, 2018. ISSN 0167-9473. doi: 10.1016/j.csda.2018.03.006.

Anthony F. Shorrocks. Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *The Journal of Economic Inequality*, 11(1):99–126, 2013. ISSN 1573-8701. doi: 10.1007/s10888-011-9214-z.

Bernard W. Silverman. Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives. *The Annals of Statistics*, 6(1):177–184, 1978. ISSN 0090-5364.

Niko Speybroeck, Peter Konings, John Lynch, Sam Harper, Dirk Berkvens, Vincent Lorant, Andrea Geckova, and Ahmad Reza Hosseinpoor. Decomposing socioeconomic health inequalities. *International Journal of Public Health*, 55(4):347–351, August 2010. ISSN 1661-8564. doi: 10.1007/s00038-009-0105-z.

Charles J. Stone. Optimal Rates of Convergence for Nonparametric Estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345206.

Viktor Subbotin. Asymptotic and Bootstrap Properties of Rank Regressions. SSRN Scholarly Paper ID 1028548, Social Science Research Network, Rochester, NY, November 2007.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246.

Dirk Van de Gaer. *Equality of Opportunity and Investment in Human Capital*. Ph.D. Dissertation. KULeuven, Leuven, 1993.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. ISBN 978-0-521-78450-4.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, New York, NY, 1996. ISBN 978-1-4757-2545-2. doi: 10.1007/978-1-4757-2545-2_15.

Eddy van Doorslaer and Xander Koolman. Explaining the differences in income-related health inequalities across European countries. *Health Economics*, 13(7): 609–628, 2004. ISSN 1099-1050.

Adam Wagstaff, Pierella Paci, and Eddy van Doorslaer. On the measurement of inequalities in health. *Social Science & Medicine*, 33(5):545–557, 1991. ISSN 0277-9536. doi: 10.1016/0277-9536(91)90212-U.

Adam Wagstaff, Eddy van Doorslaer, and Naoko Watanabe. On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam. *Journal of Econometrics*, 112(1):207–223, 2003. ISSN 0304-4076.

Geoffrey S. Watson. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics; Series A*, 26, 1964. ISSN 0581-572x.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL http://ggplot2.org.

Hajime Yamato. Uniform convergence of an estimator of a distribution function. *Bulletin of Mathematical Statistics*, 15(3/4):69–78, 1973. ISSN 0007-4993. doi: info:doi/10.5109/13073.

Shlomo Yitzhaki. Economic distance and overlapping of distributions. *Journal of Econometrics*, 61(1):147–159, March 1994a. ISSN 0304-4076. doi: 10.1016/0304-4076(94)90081-7.

Shlomo Yitzhaki. On The Progressivity of Commodity Taxation. In Wolfgang Eichhorn, editor, *Models and Measurement of Welfare and Inequality*, pages 448–466. Springer Berlin Heidelberg, 1994b. ISBN 978-3-642-79037-9.

Shlomo Yitzhaki and Edna Schechtman. The Properties of the Extended Gini Measures of Variability and Inequality. *Metron - International Journal of Statistics*, LXIII:401–433, 2005. doi: 10.2139/ssrn.815564.

Shlomo Yitzhaki and Edna Schechtman. *The Gini Methodology*, volume 272 of *Springer Series in Statistics*. Springer New York, New York, NY, 2013. ISBN 978-1-4614-4719-1 978-1-4614-4720-7.

Achim Zeileis. *ineq: Measuring Inequality, Concentration, and Poverty*, 2014. URL https://CRAN.R-project.org/package=ineq. R package version 0.2-13.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):217–242, 2014. ISSN 1369-7412.