A LASSO-TYPE ESTIMATION FOR THE LORENZ REGRESSION

Alexandre Jacquemain, Cédric Heuchenne, Eugen Pircalabelu

LIDAM Discussion Paper ISBA
2021 / 41

UCLouvain

LiDAM
ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

A lasso-type estimation for the Lorenz regression

Alexandre Jacquemain*[1], Cédric Heuchenne[2], and Eugen Pircalabelu[1]

[1] *Université catholique de Louvain, ISBA, B-1348 Louvain-la-Neuve, Belgium.*
[2] *HEC Liege, University of Liège, Rue Louvrex 14, B-4000 Liège, Belgium.*

September 23, 2021

Abstract: The Lorenz regression procedure introduced by [4] aims to estimate the explained Gini coefficient, a quantity with a natural application in the field of inequality of opportunity. In this paper, we introduce a lasso-type estimator for the explained Gini coefficient and discuss the selection of the regularization parameter. The performance of the procedure is compared to an oracle estimator on simulated data. Finally, an illustration on real-data is provided.

Keywords: Lorenz curve, Inequality of opportunity, Single-index models, LASSO, FABS algorithm

AMS subject classification: 62P20

1 Introduction

The concept of inequality of opportunity (IOP) has been defined by [6] in order to quantify unjust socioeconomic inequalities. More precisely, IOP measures the extent of inequality in an economic advantage which can be attributed to circumstances, i.e. variables over which individuals have no control. In this literature, one difficulty arises from balancing the normative measurement of inequality with the statistical modelling of the advantage. Indeed, the first aspect tends to call for simple linear or log-linear regression models. Before

*Corresponding author: alexandre.jacquemain@uclouvain.be

introducing the Lorenz regression procedure, we give the definition of the Gini coefficient. Denote the expected value by $E[\cdot]$ and let Y be a continuous random variable such that $0 < E[Y] < \infty$. The Gini coefficient is an index ranging from 0 (perfect equality) to 1 (perfect inequality), formally defined as

$$\mathrm{Gi}_Y := \frac{2C[Y, F_Y(Y)]}{E[Y]},$$

where F_Y is the CDF of Y and $C[\cdot,\cdot]$ is the covariance operator.

Let $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$ where Y is a response such that $0 < E[Y] < \infty$ and X is a vector of covariates. In our example, Y is the economic advantage and X gathers the circumstances. We start from the following single-index model

$$E[Y|X = x] = H(x^\intercal \theta_0), \tag{1}$$

where H is an increasing link function and θ_0 is a vector of unknown parameters, with parameter space Θ_0 constrained in order to ensure identifiability. Introduced by [4], the Lorenz regression estimates the *explained* Gini coefficient, defined as

$$\mathrm{Gi}_{Y,X} := \max_\theta \frac{2C[Y, F_\theta(X^\intercal\theta)]}{E[Y]} = \frac{2C[Y, F_{\theta_0}(X^\intercal\theta_0)]}{E[Y]} = \mathrm{Gi}_{H(X^\intercal\theta_0)},$$

where the last two equalities arise by using (1) and Theorem 3.1 from [3]. $E[Y|X = x]$ predicts the advantage of an individual with circumstances x. The extent of inequality in the distribution of $E[Y|X]$ thus provides information on IOP. The more inequality, the more the circumstances bear an impact on the expected income and hence, the greater IOP is. Since $\mathrm{Gi}_{Y,X}$ corresponds to the Gini coefficient of $E[Y|X]$ assuming the single-index model, we judge it a natural measure of IOP. Furthermore, the relationship between Y and X is semiparametric and, hence, more flexible than usual parametric methods. Given an iid sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ of size n with the same distribution as (Y, X), the vector θ_0 and $\mathrm{Gi}_{Y,X}$ are consistently estimated with

$$\overline{\theta} := \arg\max_\theta \frac{1}{n(n-1)} \sum_{i \neq j} Y_i \mathbf{1}\{X_i^\intercal\theta > X_j^\intercal\theta\}, \tag{2}$$

$$\overline{\mathrm{Gi}}_{Y,X} := \frac{2}{n^2} \sum_{i \neq j} \frac{Y_i}{\overline{Y}} \mathbf{1}\{X_i^\intercal\overline{\theta} > X_j^\intercal\overline{\theta}\} - \frac{n-1}{n}, \tag{3}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function and $\overline{Y}$ is the sample average. The estimator $\overline{\theta}$ is a special case of the monotone rank estimator proposed by [2], which derived its asymptotic properties.

Similarly to the R^2 in linear regression, it is easy to see that $\overline{\mathrm{Gi}}_{Y,X}$ will never decrease as we introduce new covariates. In a sparse setup, where some of the covariates have no influence on the response, an estimation on the full model will lead to an overfit of the explained Gini coefficient. In this paper, we tackle this issue by using an L_1 penalized procedure. We argue that it suits perfectly the setup explained above. First, it makes it possible for an economist to include many covariates without inducing an overestimation of the explained Gini coefficient. Second, it also provides a selection method for the relevant features.

The rest of this paper is structured as follows. In Section 2, we introduce a lasso-type estimator for the Lorenz regression and discuss the choice of the regularization parameter. Section 3 compares the performance of the method with an oracle estimator. Finally, a real-data example is presented in Section 4.

2 The penalized Lorenz regression

In this section, we introduce the penalized Lorenz regression. We borrow the idea developped by [5] for the maximum rank correlation estimator. It consists in replacing the original discrete objective function by a differentiable approximation and adding a penalty part to it. This procedure has the double advantage of introducing penalization and facilitating the numerical computation.

A lasso-type estimator for θ_0 is given by

$$\hat{\theta}(\lambda) := \arg\max_{\theta} \left\{ G(\theta) - \lambda \sum_{k=1}^{p} |\theta_k| \right\},$$

where $\lambda > 0$ is a regularization parameter and $G(\theta)$ is a smooth approximation of the indicator function displayed in (2). Formally,

$$G(\theta) := \frac{1}{n(n-1)} \sum_{i \neq j} Y_i S_\sigma(X_i^\intercal \theta - X_j^\intercal \theta),$$

where $S_\sigma(t) = 1/(1+\exp(-t/\sigma))$, and σ is a tuning parameter determining the accuracy of the approximation. Following [7], we allow σ to depend on n and set it to $\sigma = 1/\sqrt{n}$. An estimator for the explained Gini coefficient is then obtained by plugging $\hat{\theta}(\lambda)$ in (3). The numerical implementation can be carried out using the FABS algorithm developped by [7] for nonconvex loss functions and the adaptive Lasso penalty. The method starts with a very large value for λ, imposing full sparsity, and then allows the penalty parameter to decrease. 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.

In practice, an optimal regularization parameter can be determined via cross-validation. Another possibility lies in the use of an information criterion, as proposed by [5]. In this respect, we set λ_n as the one which maximizes

$$\mathrm{IC}_\lambda := \log(G(\hat{\theta}(\lambda))) - p_{\mathrm{IC}}(n)k_\lambda,$$

where k_λ is the number of covariates selected using λ and $p_{IC}(n)$ is the penalty associated to the information criterion. For a BIC-like criterion, $p_{IC}(n) = \log(n)/(2n)$. For an AIC-like criterion, $p_{IC}(n) = 1/n$. We evaluate the performance of these alternative methods in Section 3.

3 Monte-Carlo Simulations

In this section, we evaluate the performance of the procedure presented in Section 2 by means of Monte-Carlo simulations. We use the following data generating process (DGP)

$$Y_i = H(X_i^\intercal \theta)\epsilon_i,$$

where $i = 1\ldots, n = 100$ and $\sum_k |\theta_k| = 1$. X is a multivariate normal with mean 0, unit variance and a correlation matrix following an AR(1) process with correlation parameter $\rho = 0.3$. The variable ϵ_i is a lognormal noise with mean 1 and a variance set to ensure a signal-to-noise ratio of 3. Finally, we use the following link function

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

We consider a low-dimensional scenario with $p = 20$ and a high-dimensional one with $p = 120$. In both cases, 5 covariates are active. We simulate from the DGP 2000 different datasets and focus on selecting λ by BIC and AIC-like scores, CV and benchmark the performance of the penalized procedure against an oracle estimator that has knowledge about which covariates are active.

The performance of the procedure is assessed first by the square-root of the MSE of the explained Gini coefficient (RMSE.Gini). Moreover, we judge the quality of the model selection by the true positive rate (TPR), giving the percentage of selected covariates which are truly active, and by the false positive rate (FPR), which represents the percentage of selected variables which are not active. The results are displayed in Table 1. In terms of model selection, we observe a clear tradeoff between TPR and FPR. Cross-validation yields the best performance in terms of TPR but the worst in terms of FPR. The converse story holds for the BIC. Finally, the AIC selection seems to offer the best balance between these two aspects. Concerning the estimation of the explained Gini, the AIC yields once again the best performance, and even slightly outperforms the oracle.

Table 1: Comparison with an oracle estimator

	$p = 20$			$p = 120$		
	RMSE.Gini	TPR	FPR	RMSE.Gini	TPR	FPR
BIC	2.25	82.4	2.0	2.27	75.3	0.6
AIC	2.23	90.6	9.0	2.22	82.5	2.2
CV	2.28	92.5	23.8	2.37	83.6	5.0
Oracle	2.25	/	/	2.26	/	/

4 Real data example

This illustration is based on a random sample of wages data from the 1985 Current Population Survey constructed in [1]. Besides hourly wages, the data provide information on 14 covariates for 534 individuals. A description of the covariates is given in Table 2. The Gini coefficient of wages is of 29.5%. Table 2 displays the estimated vector of coefficients for the penalized and unpenalized Lorenz regression. Taking into account the results of the last section, the regularization parameter is chosen by AIC. As we can observe, only one covariate is not included in the model selection. The estimated explained Gini coefficient is of 18.02% in the unpenalized case and of 17.99% in the penalized regression.

Table 2: Estimation of θ_0 in the penalized and unpenalized cases

	Unpenalized	Penalized
Years of education	.027	.028
Sex (reference is male)	-.100	-.101
Married (reference is not married)	.051	.046
Ethnicity (reference is Caucasian)		
Hispanic	-.066	-.065
Black	-.057	-.055
Region (reference is South)	-.044	-.043
Union (reference is nonunionized)	.146	.146
Occupation (reference is tradesperson or assembly line worker)		
Technical or professional	.139	.142
Services	-.028	-.029
Office and clerical	.041	.041
Sales	-.001	
Management and administration	.195	.203
Sector (reference is other)		
manufacturing and mining	.053	.053
construction	.051	.048

References

[1] E. R. Berndt. *The Practice of Econometrics: Classical and Contemporary*. Addison-Wesley, Boston, 1991.

[2] C. Cavanagh and R. P. Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351-381, 1998.

[3] S. Das Gupta. Gini association and pseudo Lorenz curve. *Commun. Stat. Theory Methods*, 28(9):2181-2199, 1999.

[4] C. Heuchenne and A. Jacquemain. Inference for monotone single-index conditional means: a Lorenz regression approach. (Submitted).

[5] H. Lin and H. Peng. Smoothed rank correlation of the linear transformation regression model. *Comput Stat Data Anal*, 57:615-630, 2013.

[6] J. E. Roemer. *Equality of Opportunity*. Harvard University Press, Cambridge, 1998.

[7] X. Shi, Y. Huang, J. Huang and S. Ma. A Forward an Backward Stagewise Algorithm for Nonconvex Loss Functions with Adaptive Lasso. *Comput Stat Data Anal*, 124:235-251, 2018.