26th Annual Meeting of the RSSB

Lorenz regressions

Alexandre Jacquemain
Sup. Cedric Heuchenne

UCL (ISBA)

alexandre.jacquemain@uclouvain.be

October 19, 2018

Overview

Context

The study of inequality

Social economists want to examine the inequality featured in some income distribution (Y)

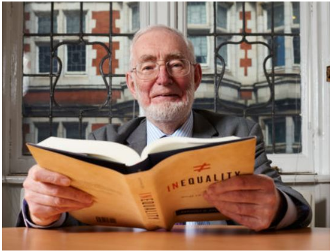


Figure: Prof. Anthony Atkinson

☑ **Measuring inequality**. We may use the Lorenz curve, or the Gini coefficient.

☒ **Explaining inequality**. We want to link inequality to a set of covariates

What do we have in mind ?

- To what extent can we attribute income inequality in Belgium to disparities in education?

The study of risk

In finance: Y is now the return of some financial asset

- We are interested in the risk related to Y
- To what extent can we attribute the risk to the type of asset (stock or bond) or macroeconomic conditions?

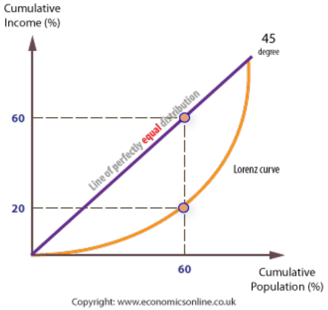
Figure: A finance worker

The Lorenz curve

Definition 1

The **Lorenz curve** (LC) of a continuous random variable Y with CDF F_Y is defined as

$$LC_Y(p) := \frac{E[Y\mathbb{1}\{F_Y(Y) \leq p\}]}{E[Y]}$$



- What share of income do the $p \times 100\%$-poorest individuals own?
- Scalar measure: the **Gini coefficient**

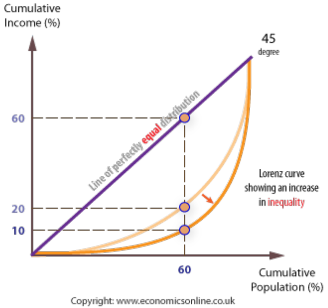
$$Gi_Y := 2\int_0^1 [p - LC_Y(p)]dp = \frac{2Cov[Y, F_Y(Y)]}{E[Y]}.$$

The Lorenz curve

Definition 1

The **Lorenz curve** (LC) of a continuous random variable Y with CDF F_Y is defined as

$$LC_Y(p) := \frac{E[Y\mathbb{1}\{F_Y(Y) \leq p\}]}{E[Y]}$$



- What share of income do the $p \times 100\%$-poorest individuals own?
- Scalar measure: the **Gini coefficient**

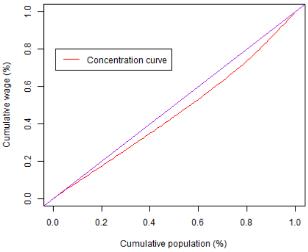
$$Gi_Y := 2\int_0^1 [p - LC_Y(p)]dp = \frac{2Cov[Y, F_Y(Y)]}{E[Y]}.$$

The concentration curve

Definition 2

The **concentration curve** (CC) of Y with respect to X, with CDF F_X is defined as

$$CC_{Y,X}(p) := \frac{E[Y\mathbb{1}\{F_X(X) \leq p\}]}{E[Y]}$$



- What share of wage do the $p \times 100\%$ least educated own?
- Scalar measure: the **concentration index**

$$Ci_{Y,X} := 2\int_0^1 [p - CC_{Y,X}(p)]dp = \frac{2Cov[Y, F_X(X)]}{E[Y]}.$$

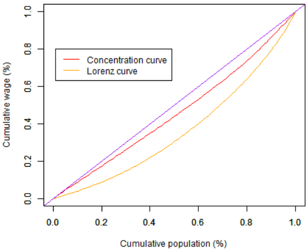
- Inequality that you can reproduce if you rank individuals in terms of education, not in terms of wage.

The concentration curve

Definition 2

The **concentration curve** (CC) of Y with respect to X, with CDF F_X is defined as

$$CC_{Y,X}(p) := \frac{E[Y\mathbb{1}\{F_X(X) \leq p\}]}{E[Y]}$$



- What share of wage do the $p \times 100\%$ least educated own?
- Scalar measure: the **concentration index**

$$Ci_{Y,X} := 2\int_0^1 [p - CC_{Y,X}(p)]dp = \frac{2Cov[Y, F_X(X)]}{E[Y]}.$$

- Inequality that you can reproduce if you rank individuals in terms of education, not in terms of wage.

Shortcomings of the existing tools

What tools to determine the contributions of X on inequality of Y?

Decomposition ideas using the Lorenz curve. Assuming $Y = \alpha_1 X_1 + \ldots + \alpha_p X_p$.

- [Lerman and Yitzhaki, 1985] decomposed the Gini coefficient of income Y in the contributions of its sources X_k
- Problem: not a regression idea.

Regression ideas. For example, $Y = \alpha_1 X_1 + \ldots + \alpha_p X_p + \epsilon$

- Problem 1: the classical linear regression is not flexible.
- Problem 2: no link with inequality measurement.

Goal

We want to develop a **regression procedure** . . .

1. which determines the contribution of covariates X on the inequality of Y;
2. and which allows more flexibility than the classical linear regression.

Methodology

Maximization programme

Basic idea: we maximize the concentration index of Y by $X^T\theta$.

Lorenz regression - maximization programme

$$\max_{\theta} Cov[Y, F_\theta(X^T\theta)] \qquad \text{s.t. } ||\theta|| = 1, \tag{1}$$

where F_θ is the CDF of $X^T\theta$.

1. Reproducing inequality: we find the vector of weights θ which reproduces as much as possible the inequality of Y (more later).
2. Regression procedure: more flexibility and robustness because ranks are taken for $X^T\theta$

A covariance inequality

▸ A reminder on the concentration curve

Question: could we reproduce more than the Gini coefficient?

- Could it be that $Ci_{Y,X} > Gi_Y$? No!

Lemma 3

Let $Z \in \mathbb{R}^+$ and $Y \in \mathbb{R}$ be two continuous random variables with respective CDFs F_Z and F_Y. Then, the following inequality holds

$$E[ZY] \leq \int_0^1 F_Z^{-1}(p)F_Y^{-1}(p)dp.$$

Theorem 4

Let $Y \in \mathbb{R}$ be a continuous random variable with CDF F_Y and $X \in \mathbb{R}$ be a continuous random variable with CDF F_X. Then, the following inequality holds

$$Cov[Y, F_X(X)] \leq Cov[Y, F_Y(Y)].$$

Definitions

Assume $X^T\theta$ is **continuous**. Recall that $F_\theta(.)$ is the CDF of $X^T\theta$.

Definition 5

The **explained Lorenz curve** of Y by $X^T\theta$ is defined as

$$LC_{Y,X^T\theta}(p) := CC_{Y,X^T\theta}(p) = \frac{E[Y\mathbb{1}\{F_\theta(X^T\theta) \leq p\}]}{E[Y]},$$

and similarly, the **explained Gini coefficient** is

$$Gi_{Y,X^T\theta} := Ci_{Y,X^T\theta} = \frac{2Cov[Y, F_\theta(X^T\theta)]}{E[Y]}.$$

Intuition: $Gi_{Y,X^T\theta}$ represents the inequality which we can reproduce if we rank individuals in terms of $X^T\theta$ instead of Y.

Note: $Gi_{Y,X^T\theta} \leq Gi_Y$ (theorem 4)

Maximization programme

Programme (1) chooses θ in order to maximize $Gi_{Y,X^T\theta}$.

- We summarize the information contained in X in an index $X^T\theta$, where $||\theta|| = 1$.
- We choose the weight vector θ so that $X^T\theta$ reproduces as much as possible the inequality of Y.

We can examine how much inequality we can reproduce by comparing $Gi_{Y,X^T\theta^*}$ to Gi_Y.

Definition 6

We define the **proportion of explained inequality** (PEI) as

$$\text{PEI}_{Y,X^T\theta^*} := \frac{Gi_{Y,X^T\theta^*}}{Gi_Y} = \frac{Cov[Y, F_{\theta^*}(X^T\theta^*)]}{Cov[Y, F_Y(Y)]} \in [0,1].$$

The model underneath

What is the econometric model lying underneath our procedure?

- We need to find a model linking Y to $X^T\theta$ and for which maximization programme (1), once translated in the sample, would bring a good estimator of θ.
- Answer: the **single index model**.

The single-index model

Definition 7

Following [Horowitz, 2009], we define the single-index model as

$$E[Y|X = x] = H(x^T\theta_0)$$

where θ_0 is normalized (here $||\theta_0|| = 1$). Here, we furthermore assume that H is increasing.

It is a **semiparametric regression** procedure.

1. The functional form of H is left unspecified (hence, more flexible than parametric models).
2. The model displays a vector of parameters θ_0 (hence, avoid the curse of dimensionality of nonparametric regression).

Estimation of θ_0

We focus first on estimation of θ_0 (estimation of H will be discussed later on).

Several methods:

- Semiparametric least-squares (Ichimura 1993).
- Maximum likelihood (Klein and Spady 1993, Ai 1997).
- Average derivative (Powell et al. 1989, Hristache et al. 2001).

Common drawback: one or more subjective smoothing parameters to choose

The monotone rank estimator (MRE)

[Cavanagh and Sherman, 1998] introduced the **monotone rank estimator (MRE)**, obtained as

MRE - maximization programme

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} Y_i R_n(X_i^T \theta) \qquad \text{s.t. } ||\theta|| = 1, \qquad (2)$$

where $R_n(X_i^T \theta)$ denotes the rank of $X_i^T \theta$ in the vector $X^T \theta$.

Link with the reproduction of inequality.

- The MRE is a simple translation of maximization programme (1) in the sample!
- The MRE gives the vector of weights which reproduces as much as possible the observed inequality in Y.

Estimation of the regression curve (H)

Recall: we estimate θ_0 with the MRE, and we obtain the estimated index $T = X^T\hat{\theta}$. In order to estimate H, we should incorporate the assumption that it's an increasing function.

Idea: We rewrite H so that $H(.) = G(F_T(.))$. Hence,

$$\begin{aligned} P(H(T) \leq y) = P(G(F_T(T)) \leq y) &= P(F_T(T) \leq G^{-1}(y)) \\ &= G^{-1}(y) \end{aligned}$$

Three steps.

1. Provide an initial estimator $\hat{H}_1$ for H (Nadaraya-Watson, local polynomial, ...)
2. Estimate $P(\hat{H}_1(T_i) \leq y)$. This gives an estimator of $G^{-1}(y)$.
3. Invert this estimator in order to obtain $\hat{G}$. Finally, $\hat{H}(t) = \hat{G}(\hat{F}_T(t))$.

Two methods:

1. [Dette et al., 2006] use a Kernel estimator for $P(\hat{H}_1(T_i) \leq y)$.
2. [Chernozhukov et al., 2009] rather use the empirical CDF.

An inequality-based goodness of fit

In linear regression the R^2 measures the proportion of variability (as measured by the variance) that we can reproduce with the model.

Goal: we want to build a similar measure for Lorenz regressions. The PEI precisely does that in the population. We only need to translate it in the sample.

Definition 8

The Lorenz-R^2 (LR^2) is defined as

$$\mathrm{LR}^2 := \frac{\hat{G}i_{Y,X^T\theta}}{\hat{G}i_Y} = \frac{\frac{1}{n^2}\sum_{i=1}^n Y_i R_n^{\hat{\theta}}(X_i^T\hat{\theta}) - \frac{\overline{Y}}{2}}{\frac{1}{n^2}\sum_{i=1}^n Y_i R_n^Y(Y_i) - \frac{\overline{Y}}{2}} \in [0,1],$$

where $R_n^Y(.)$ corresponds to the rank in the Y vector while $R_n^{\hat{\theta}}(.)$ gives the rank in the $X^T\hat{\theta}$ vector.

Inference on θ_0

Asymptotic distribution. [Cavanagh and Sherman, 1998] showed that $\sqrt{n}[\hat{\theta} - \theta_0] \xrightarrow{d} N(0, \Sigma)$. However, estimation of Σ appears to be a tedious task. Hence, we turn to bootstrapping procedures.

Bootstrap. [Subbotin, 2007] established the convergence of the asymptotic distribution of θ^* to that of $\hat{\theta}$. It also proves the consistency of the bootstrap estimator of the variance, Σ^*. Two options

- Hybrid bootstrap: we retain the asymptotic normality and only bootstrap $\hat{\Sigma}$.
- Basic bootstrap: we bootstrap the whole distribution of $\hat{\theta}$.

We can use both methods to build confidence intervals or tests.

Simulations

Performance of the estimation

We compare the estimation error of our procedure with the SLS estimator of [Ichimura, 1993]. Formally, we look at

1. MISE of the index

$$MISE[X^T\theta] = E\left[\int \left(X^T\hat{\theta} - X^T\theta\right)^2 dx\right]$$

2. MISE of the regression curve

$$MISE[H(X^T\theta)] = E\left[\int \left(\hat{H}(X^T\hat{\theta}) - H(X^T\theta)\right)^2 dx\right]$$

Data generating process:

$$Y_i = H\left(\theta_1 X_i^1 + \ldots + \theta_c X_i^c + \theta_{c+1} Z_i^1 + \ldots + \theta_{c+d} Z_i^d\right) + \epsilon_i,$$

where $i = 1, \ldots, n$. $H(t) = 3 + t + t^3$, the X_i's are c continuous $N(0,1)$ and the Z_i's are d discrete $Be(0.5)$.

Sample size

Fix $c = 3$ and $d = 1$ and examine how the MISE evolves with n.

		n=25	n=50	n=100	n=200	n=500
Index	Ichimura	0.1940	0.1155	0.0563	0.0239	0.0061
	Lorenz	0.0614	0.0393	0.0211	0.0110	0.0051
Curve	Ichimura	1.0135	0.6320	0.3468	0.1849	0.0804
	Lorenz	0.9236	0.6200	0.3469	0.1839	0.0910

- **Index**: Lorenz outperforms Ichimura, but slows down with sample size.
- **Curve**: sensibly the same performances.

Continuous covariates

Fix $n = 100$, and consider only continuous variables $(c = 2, c = 10$ and $c = 20)$.

		c=2	c=10	c=20
Index	Ichimura	0.009	0.041	0.047
	Lorenz	0.009	0.008	0.009
Curve	Ichimura	0.459	0.060	0.044
	Lorenz	0.466	0.031	0.021

- **Two covariates**: sensibly the same performances.
- **More covariates**: Lorenz outperforms Ichimura, especially for the index.

Questions?

References I

Christopher Cavanagh and Robert P. Sherman (1998)
Rank Estimators for Monotonic Index Models
Journal of Econometrics, 84(2):351-381

Victor Chernozhukov, Ivan Fernandez-Val and Alfred Galichon (2009)
Improving point and interval estimators of monotone functions by rearrangement
Biometrika, 96(3):559-575

Holger Dette, Natalie Neumeyer and Kay F. Pilz (2006)
A Simple Nonparametric Estimator of a Strictly Monotone Regression Function
Bernoulli, 12(3):469-490

Joel L. Horowitz (2009)
Semiparametric and Nonparametric Methods in Econometrics
Springer Series in Statistics.

References II

Hidehiko Ichimura (1993)
Semiparametric least squares (SLS- and weighted SLS estimation of single-index models
Journal of Econometrics, 58(1):71-120.

Robert L. Lerman and Shlomo Yitzhaki (1985)
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

Viktor Subbotin (2007)
Asymptotic and Bootstrap Properties of Rank Regressions
Ph.D. dissertation, Department of Economics, Northwestern University, Evanston, IL