Stat 5361 Final Project

A Simple and Effective Inequality Measure

Chamundeswari Koppisetti* Shynggys Magzanov†

12 17 2020

Abstract

Economists use various income inequality metrics to get a better picture of the economic inequality in the country. Numerous research results suggest that there is a strong relationship between income inequalities amid individuals and levels of poverty, mental illnesses, crime and social unrest. At the same time most of the public policies like welfare benefits, taxation, health care, etc. have direct influence on how the income is distributed. So, to properly plan how to address these issues one has to have a good measure of current income inequality. In this project we discussed on a non-parametric method, ratios of symmetric quantiles that can be very useful to get more accurate result than it can be provided by the Gini coefficient or the Lorenz curves. Ratios of quantiles are often computed for income distributions as rough measures of inequality, and symmetric quantiles is of special interest here because the graph of such ratios, plotted as a function of p (p/2 quantile) over the unit interval, yields an informative inequality curve. The area above the curve and less than the horizontal line at one is an easily interpretable measure of inequality.

1. Introduction

1.1. Measures of Inequality

There exist various types of metrics used by economists to measure the income inequality. The most widely exploited ones are the Lorenz curves and the Gini coefficient, each of which have their own advantages and limitations. Those obvious limitations include the requirement for the population mean and variance to exist, as well as down-weighting smaller incomes and that way stressing way more attention to the middle incomes. This paper analyzes another income inequality measure, namely using ratios of symmetric quantiles, that proves to be helpful to overcome previously mentioned disadvantages. In addition to that, it is proven that the given metric satisfies the median preserving principle and applicable to widely used income distributions. But the major benefit is that there is no need of parametric model assumption to work with the given inequality measure.

*chamundeswari.koppisetti@uconn.edu
†shynggys.magzanov@uconn.edu

2. Concepts and Definitions

2.1. Lorenz curve and Gini coefficient

A Lorenz curve is a graph used in economics to show inequality in income spread or wealth. The x-axis on a Lorenz curve typically shows the portion or percentage of the total population, and the y-axis shows the portion of total income/ wealth, or whatever is being analyzed.

Since perfect equality would mean that a $1/k$ portion of the population controlled $1/k$ of the wealth, perfect equality on the graph would be shown by a straight line with a slope of 1. This line is often drawn on the graph as a point of reference, alongside the curved line which represents the actual wealth/income/size distribution.Any point on the curve can be read to tell us what percentage or portion of the population command what percent of the wealth, income, or whatever variable is being studied.

Gini ratio (or Gini coefficient) is a measure of inequality, based on the Lorenz curve, that goes from 0 (absolute equality) up to 1 (absolute inequality). It's calculated as a ratio of the areas on the Lorenz curve diagram. If the area between the line of perfect equality and Lorenz curve is A, and the area underneath the Lorenz curve is B, then the Gini coefficient is $A/(A+B)$.

Greater inequality shows up as a larger area between the Lorenz curve and the diagonal line of absolute equality.

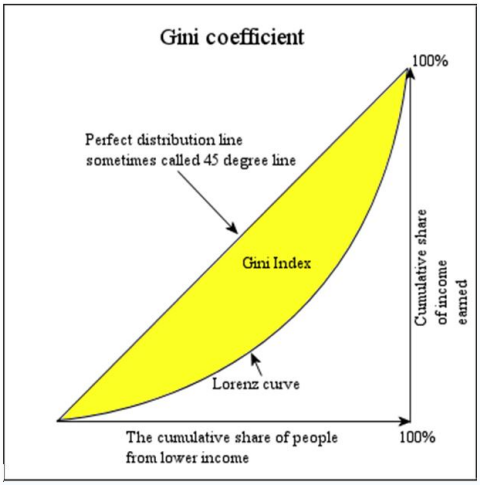

Figure 1: Lorenz curve and Gini Coefficient

Estimating G

Since $A + B = 0.5$, the Gini coefficient, G = 2A = 1-2B. If the Lorenz curve is represented by the function Y = L(X), the value of B can be found with integration and:

$$G = 1 - 2\int_0^1 L(X)dX$$

In some cases, this equation can be applied to calculate the Gini coefficient without direct reference to the Lorenz curve. For example, for a population with values $y_i, i = 1...n$, that are indexed in non-decreasing order ($y_i \leq y_{i+1}$), with $S_i = \sum_{j=1}^{i} f(y_j)y_j, S_0 = 0$:

$$G = 1 - \frac{\sum_{i=1}^{n} f(y_i)(S_{i-1} + S_i)}{S_n}$$

For a cumulative distribution function $F(y)$ that is piecewise differentiable, has a mean μ, and is zero for all negative values of y:

$$G = 1 - \frac{1}{\mu} \int_0^{\infty} (1 - F(y))^2 dy$$

Sometimes the entire Lorenz curve is not known, and only values at certain intervals are given. In that case, the Gini coefficient can be approximated by using various techniques for interpolating the missing values of the Lorenz curve. If (X_k, Y_k) are the known points on the Lorenz curve, with the X_k indexed in increasing order $(X_{k-1} < X_k)$, so that: X_k is the cumulated proportion of the population variable, for $k = 0, ..., n$ with $X_0 = 0$, $X_n = 1$ and Y_k is the cumulated proportion of the income variable, for $k = 0, ..., n$, with $Y_0 = 0$, $Y_n = 1$.

If the Lorenz curve is approximated on each interval as a line between consecutive points, then the area B can be approximated with trapezoids (Brown Formula) and:

$$G_1 = |1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1})|$$

is the resulting approximation for G.

More accurate results can be obtained using other methods to approximate the area B, such as approximating the Lorenz curve with a quadratic function across pairs of intervals, or building an appropriately smooth approximation to the underlying distribution function that matches the known data. If the population mean and boundary values for each interval are also known, these can also often be used to improve the accuracy of the approximation.

Advantages: The Gini coefficient's main advantage is that it is a measure of inequality, not a measure of average income or some other variable which is unrepresentative of most of the population, such as gross domestic product.It can be used to compare income distributions across different population sectors as well as countries, for example the Gini coefficient for urban areas differs from that of rural areas in many countries.

Disadvantages: The Lorenz curve may understate the actual amount of inequality if richer households are able to use income more efficiently than lower income households. Lorenz curves may intersect, reflecting differing patterns of income distribution, but nevertheless resulting in very similar Gini coefficient values.This troubling property of the Lorenz framework complicates comparisons of Gini coefficient values and may confound tests of the income inequality hypothesis.It is also claimed that the Gini coefficient is more sensitive to the income of the middle classes than to that of the extremes.

2.2. Basic properties of the ratio of symmetric quantiles

Say F satisfy $F(0-) = 0$ and the pth quantile $x_p = Q(p) = F^{-1}(p) = inf(x : F(x) \geq p), 0 < p < 1$.The symmetric ratio of quantiles for $0 < p < 1$ is given by $R(p) = x_{p/2}/ = x_{1-p/2}$

So, for each p, $R(p)$ gives the ratio of the typical (median) income of the lowest proportion p of incomes to the typical (median) income of the largest proportion p. Extend R to [0,1] by defining $R(0) = 0$ and $R(1) = 1$. The graph $(p; R(p))$ of R has the following properties:

1. $0 \leq R(p) \leq 1$
2. $R(p)$ is monotone increasing from $R(0) = 0$ to $R(1) = 1$
3. $R(p) = 1$ for all $0 < p < 1$ if and only if all incomes are equal.
4. $R(p)$ is scale invariant.
5. After any median preserving transformation of funds from the upper half of incomes to the lower half of incomes, $R(p)$ can only increase.

We define ratio of inequality by $I = I(F) = 1 - \int_0^1 R(p)dp$ and interpreted as "if one selects an income at random from those below the median and divides it by its symmetric quantile, on average one obtains $1 - I(F)$. Therefore, $I(F)$ has the simple interpretation as the average relative distance $(Y - X)/Y$ of X from its symmetric quantile Y. These properties of I lead to explore the inequality measure as an alternative to the Gini Index which is defined as:

$$G = 1 - \frac{1}{E(X)} \int_0^\infty (1 - F(x))^2 dx$$

3. Methodology

Approximation of I

First, let us introduce the simple yet powerful approximation of I. Provided some integer J, a grid $\{p_j\}$ is defined on a unit interval by $p_j = (j - 1/2)/J$, for $j = 1, 2, ..., J$. Then $R(p_j)$ is evaluated for each p_j and $I^{(J)} = J^{-1} \sum_j \{1 - R(p_j)\}$ is calculated. It is clear that if J is chosen large enough, we can make $I^{(J)}$ as close to I as we want. In our project we set $J = 1000$ since it provided sufficiently good approximation.

Estimating I and Confidence Interval

For every $0 < p < 1$ let $\hat{R}(p) = \frac{\hat{x}_{p/2}}{\hat{x}_{1-p/2}}$. According to Prendergast & Staudte (2015b), $\sqrt{n}\{\hat{R}(p) - R(p)\}$ in distribution converges to $N(0, \sigma_p^2)$, where the variance is:

$$\sigma_p^2 = a_0 + a_1 R(p) + a_2 R^2(p)$$

with $a_0 = (p/2)(1 - p/2)q^2(p/2)/x^2_{1-p/2}$,$a_1 = -2(p/2)^2 q(p/2)q(1 - p/2)/x^2_{1-p/2}$, $a_2 = (p/2)(1 - p/2)q^2(1 - p/2)/x^2_{1-p/2}$. This result makes it possible to find the $100(1 - \alpha)\%$ confidence interval for $R(p)$, i.e. $\hat{R}(p) \pm z_{1-\alpha}\hat{\sigma}_p/\sqrt{n}$. As we have seen calculation of the standard deviation in this case requires one to find the quantile density q at $p/2$ and $1 - p/2$. Details can be found in Prendergast & Staudte (2016).
Now let us proceed to estimation of I. From the definitions provided above it is clear that the estimate of I will be $\hat{I} = 1 - \int_0^1 \hat{R}(p)dp$. But since it is not always possible to get a closed form of I, we use numerical approximation. Using the previously mentioned result, we can define the numerical approximation of I using I^J.

$$\hat{I}^{(J)} = J^{-1} \sum_j \{1 - \hat{R}(p_j)\}$$

It's confidence interval is $\hat{I}^{(J)} \pm z_{1-\alpha/2}\sqrt{Var[\hat{I}^{(J)}]}$, where $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. Details about $Var[\hat{I}^{(J)}]$ can be found in Prendergast & Staudte (2016).

4. Application

The given concept was applied to three sets of data below. Namely, Disposable Personal Income, Gross Domestic Income, Earnings Data of Women in the USA.

4.1. Disposable Personal Income.

The data of Disposable Personal Income between 1959 and 2020 can be obtained at https://fred.stlouisfed.org/series/DSPI. Firstly, I and G were calculated for the Disposable Personal Income Data. It appears that $\hat{I} = 0.7561$, which is an estimate of I, with a standard error of 0.0106. Additionally, 95% confidence interval of $\hat{I}$ was calculated to be $(0.7352, 0.7769)$. While $\hat{G} = 0.4872$

is significantly lower, with a standard error of 0.0095 and respective 95% confidence interval of $(0.4686, 0.5059)$. Below is the Inequality curve of the data plotted over it's Lorenz curve.

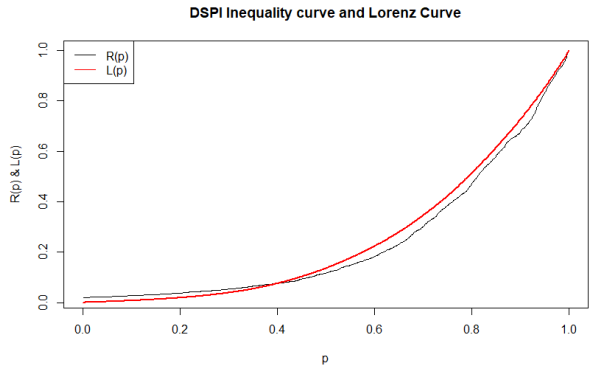

Figure 2: Income inequality curve and Lorenz curve of Disposable Personal Income in the US between 1959 and 2020

One can observe that the Lorenz curve underestimates the income inequality relative to income inequality curve after 40% of population mark.

4.2. Gross Domestic Income.

The data on Gross Domestic Income can be found here: https://fred.stlouisfed.org/series/GDI. As results reveal, ratio of symmetric quantiles of the GDI is $\widehat{I} = 0.8204$ with a standard error 0.0162 and confidence interval $(0.7886, 0.8523)$. At the same time $\widehat{G} = 0.5499$ with a standard error of 0.0161 and CI $(0.5182, 0.5815)$.

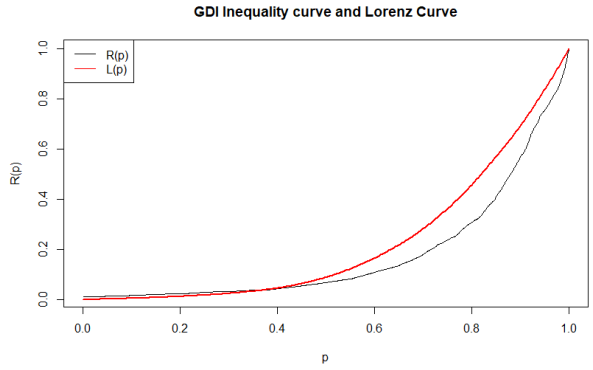

Figure 3: Income inequality curve and Lorenz curve of Gross Domestic Income in the US between 1947 and 2020

Similarly to the previous data set we can see the relative underestimation of income inequality by Lorenz curve in GDI as well. However this time the discrepancy seems to be larger which is reflected in a slightly bigger difference between $\widehat{I}$ and $\widehat{G}$ in this data set.

4.3. Earnings Data of Women in the USA.

The data on Earnings of women in the US between 1979 and 2020 can be otbained from: https://fred.stlouisfed.org/series/LES1252882700Q. Results show that $\hat{I} = 0.4666$ with a standard error 0.0183 and CI $(0.4307, 0.5025)$. While Gini coefficient is $\hat{G} = 0.2199 \pm 0.0094$ with CI $(0.2015, 0.2384)$.
Figure 4 shows that this data set is different from the previous two. According to our results, Lorenz curve dramatically overestimates the income inequality in this case. It can also be noticed from the fact that $\hat{G}$ significantly exceeds $\hat{I}$.

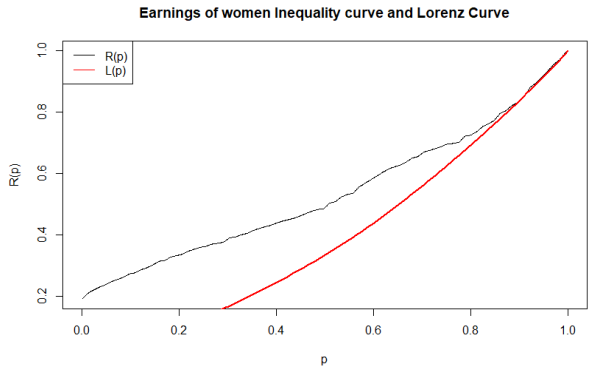

Figure 4: Income inequality curve and Lorenz curve of Earnings Data of Women in the US between 1979 and 2020

Let us take a look at boxplots of three data sets to see how this data set can be different from previous ones.

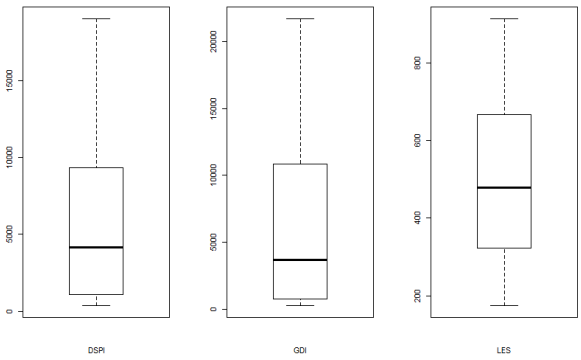

Figure 5: Boxplots of Data sets

One can notice that all three data sets have outliers at the upper part of values. The apparent difference might be resulted from the last data set having outliers not only in the upper part of the values but at the lower part as well. And in the work of Luke A. Prendergast and Robert G. Staudte (2016) relative resistance to outliers of Inequality curves relative to Lorenz curves was proven by means of various simulations.

5. Conclusion

In this project the article of Luke A. Prendergast and Robert G. Staudte (2016) was reproduced using three data sets: Disposable Personal Income, Gross Domestic Income and Earnings Data of Women in the US. The main idea is the proposition of a new Inequality measure I which is the difference between the horizontal line at 1 and the symmetric ratio of quantiles curve $\{p, R(p)\}$. Despite it's calculational simplicity it has numerous advantages over Gini coefficient and Lorenz curves like resistance to outliers (Prendergast,Staudte, 2016) and non-reliance on income distribution. Our results agree with the research results of Luke A. Prendergast and Robert G. Staudte (2016). R-script for calculating I and G, plotting Inequality and Lorenz curves is provided.

Appendix

Codes used to generate the results:

```
DSPI <- read.csv("data/DSPI.csv")
```

```
plot(DSPI, type = "l")
```

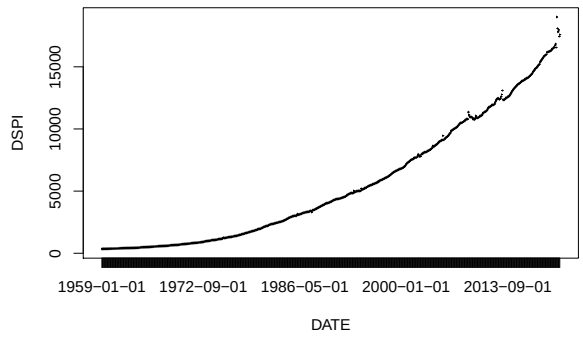

Figure 6: Disposable Personal Income

```
QOR.ln <- function(u){
  # QOR function for the log-normal
  q.n.0 <- 1/dnorm(qnorm(u))
  q.n.1 <- qnorm(u)*q.n.0^2
  q.n.2 <- (1 + 2*qnorm(u)^2)*q.n.0^3
  1/(q.n.0^2 + 3*q.n.1 + q.n.2/q.n.0)
}

QOR.gld <- function(u, lambda = NULL){
  # QOR function for the GLD FKML paramaterization
  if(is.null(lambda)) lambda <- fit.fkml(x)$lambda
  l3 <- lambda[3]
  l4 <- lambda[4]
  (u^(l3 - 1) + (1 - u)^(l4 - 1))/(u^(l3 - 3)*(l3 - 2)*(l3 - 1) +
                                     (1 - u)^(l4 - 3)*(l4 - 2)*(l4 - 1))
}

hatI <- function(x, J = 1000, conf.level = 0.95,
                 bw.correct = TRUE, QOR.FUN = QOR.ln, ...){

  n <- length(x)
```

```
us <- ((1:J) - 0.5)/J
Rs <- (xu2 <- quantile(x, us/2))/(x1u2 <- quantile(x, 1 - us/2))
I <- sum(1 - Rs)/J

if(!is.null(conf.level)){
  v <- c(us/2, 1 - us/2)
  qor <- QOR.FUN(v, ...)
  bw <- 15^(1/5)*abs(qor)^(2/5)/n^(1/5)
  if (bw.correct) bw[v <= bw] <- v[v <= bw]

  kernepach <- function(u) 3/4*(1 - u^2)*(abs(u) <= 1)
  m1 <- matrix(v, nrow = 2*J, ncol = n, byrow = FALSE)
  m2 <- matrix(1:n, nrow = 2*J, ncol = n, byrow = TRUE)

  consts <- kernepach((m1 - (m2 - 1)/n)*(1/bw))*(1/bw) -
    kernepach((m1 - m2/n)*(1/bw))*(1/bw)

  x.sorted <- sort(x)
  q.hat <- c(consts%*%x.sorted)
  q.hat.1 <- q.hat[1:(length(q.hat)/2)]
  q.hat.2 <- q.hat[-(1:(length(q.hat)/2))]

  rc <- matrix(Rs, ncol = J, nrow = J, byrow = FALSE)

  covm <- ((1/x1u2)%*%t(1/x1u2))*(((us/2)%*%t(1 - us/2))*(q.hat.1%*%t(q.hat.1)
                                        + Rs%*%t(Rs)*(q.hat.2%*%t(q.hat.2))) -
                                ((us/2)%*%t(us/2))*((q.hat.1%*%t(q.hat.2))*t(rc)
                                                    + (q.hat.2%*%t(q.hat.1))*rc))/n

  sigma.p2 <- (us/2)*(1 - us/2)*q.hat.1^2
  sigma.q2 <- (1 - us/2)*(us/2)*q.hat.2^2
  sigma.pq <- (us/2)^2*q.hat.1*q.hat.2
  a0 <- sigma.p2/x1u2^2
  a1 <- -2*sigma.pq/x1u2^2
  a2 <- sigma.q2/x1u2^2
  Vs <- (a0 + a1*Rs + a2*Rs^2)/n

  V <- (sum(Vs) + 2*sum(covm[row(covm) < col(covm)]))/J^2
  SE <- sqrt(V)
  conf.int <- I + c(-1, 1)*qnorm(1 - (1 - conf.level)/2)*sqrt(V)
} else{
  V <- NULL
  SE <- NULL
  conf.int <- NULL
}
```

```
  list(I = I, SE = SE, conf.int = conf.int)
}

hatG <- function(x, conf.level = 0.95){
  indices <- 1:(n <- length(x))
  ordered.x <- sort(x)
  sx <- sum(ordered.x*(indices - 1/2))
  mu.hat <- mean(x)
  Gv <- 2/mu.hat/n^2*sx - 1

  Z.hat <- -(Gv + 1)*ordered.x + (2*indices - 1)/n*ordered.x -
                                          2/n*cumsum(ordered.x)
  Z.bar <- mean(Z.hat)

  V <- 1/n^2/mu.hat^2*sum((Z.hat - Z.bar)^2)
  conf.int <- Gv + c(-1, 1)*qnorm(1 - (1 - conf.level)/2)*sqrt(V)

  list(G = Gv, SE = sqrt(V), conf.int = conf.int)
}
```

```
DSPI <- DSPI$DSPI
I <- hatI(DSPI)
I
```

Disposable Personal Income

```
## $I
## [1] 0.7560647
##
## $SE
## [1] 0.01063773
##
## $conf.int
## [1] 0.7352151 0.7769143
```

```
G <- hatG(DSPI)
G
```

```
## $G
## [1] 0.487252
##
## $SE
## [1] 0.009500346
##
## $conf.int
## [1] 0.4686317 0.5058723
```

```r
R_p <- function(x, J = 1000){
  us <- ((1:J) - 0.5)/J
  Rs <- (quantile(x, us/2))/(quantile(x, 1 - us/2))
  return(cbind(us, Rs))
}

ineq_DSPI <- R_p(DSPI)
#plot(ineq_DSPI, type = "l", main = "DSPI Inequality curve and Lorenz Curve",
 #    xlab = "p", ylab = "R(p) & L(p)")
#lines(ineq::Lc(DSPI), col = "red")
#legend( "topleft", legend=c("R(p)", "L(p)"),
 #      col=c("black", "red"), lty=1:1, cex=1)
```

```r
GDI <- read.csv("data/GDI.csv")
```

```r
plot(GDI, type = "l")
```

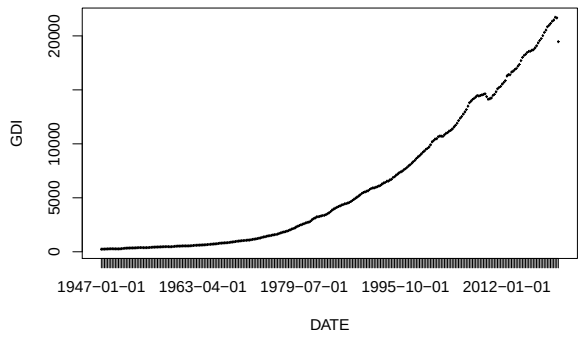

Figure 7: Gross Domestic Income

```r
GDI <- GDI$GDI
I_gdi <- hatI(GDI)
I_gdi
```

Gross Domestic Income

```
## $I
## [1] 0.8203284
##
## $SE
## [1] 0.01623472
##
## $conf.int
```

```
## [1] 0.7885089 0.8521478
```

```
G_gdi <- hatG(GDI)
G_gdi
```

```
## $G
## [1] 0.549877
##
## $SE
## [1] 0.01613956
##
## $conf.int
## [1] 0.5182440 0.5815099
```

```
ineq_GDI <- R_p(GDI)
#plot(ineq_GDI, type = "l", main = "GDI Inequality curve and Lorenz Curve",
 #    xlab = "p", ylab = "R(p)")
#lines(ineq::Lc(GDI), col = "red")
#legend( "topleft", legend=c("R(p)", "L(p)"),
 #      col=c("black", "red"), lty=1:1, cex=1)
```

```
LES <- read.csv("data/LES.csv")
```

```
plot(LES, type = "l")
```

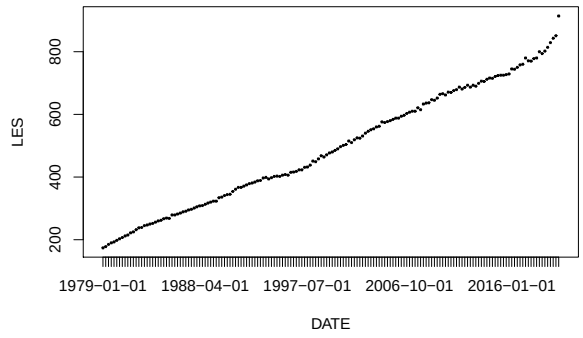

Figure 8: Earnings of Women

```
LES <- LES$LES
I_les <- hatI(LES)
I_les
```

Earnings data of women

```
## $I
```

```
## [1] 0.466592
##
## $SE
## [1] 0.0183179
##
## $conf.int
## [1] 0.4306896 0.5024945
```

```
G_les <- hatG(LES)
G_les
```

```
## $G
## [1] 0.2199384
##
## $SE
## [1] 0.009415175
##
## $conf.int
## [1] 0.2014850 0.2383918
```

```
ineq_LES <- R_p(LES)
#plot(ineq_LES, type = "l", main = "Earnings of women Inequality curve and Lorenz Curve",
 #    xlab = "p", ylab = "R(p)")
#lines(ineq::Lc(LES), col = "red")
#legend( "topleft", legend=c("R(p)", "L(p)"),
 #      col=c("black", "red"), lty=1:1, cex=1)

#par(mfrow=c(1,3))
#boxplot(DSPI, xlab = "DSPI")
#boxplot(GDI, xlab = "GDI")
#boxplot(LES, xlab = "LES")
```

Reference List

- Luke A. Prendergast & Robert G. Staudte (2018), "A Simple and Effective Inequality Measure," The American Statistician, 72:4, 328-343, DOI: 10.1080/00031305.2017.1366366
- Joseph L. Gastwirth (2017), "Is the Gini Index of Inequality Overly Sensitive to Changes in the Middle of the Income Distribution?," Statistics and Public Policy, 4:1, 1-11, DOI: 10.1080/2330443X.2017.1360813
- De Maio FG, "Income inequality measures," J Epidemiol Community Health. 2007;61(10):849-852. doi:10.1136/jech.2006.052969
- Sitthiyot, T., Holasut, K. "A simple method for measuring inequality," Palgrave Commun 6, 112 (2020). https://doi.org/10.1057/s41599-020-0484-6
- Trapeznikova, I. "Measuring income inequality," IZA World of Labor 2019: 462 doi: 10.15185/izawol.462