Gini Index and Inequality

ID, Last Name, First Name

2024/01/24

## Short Abstract

In this note, we study GINI index using WDI and compare with other index. In an OECD report, ‘OECD Regions and Cities at a Glance 2022’ [Link](https://read.oecd-ilibrary.org/urban-rural-and-regional-development/oecd-regions-and-cities-at-a-glance-2022_14108660-en#page1), S80/S20 ratios are used. We consider a question if the ratio is related to GINI index.

**Definition S80/S20 ratio**: The total income received by the 20% of people with the highest income in a region divided by the total income received by the 20% of people with the lowest income in the same region.

## Information of data

### Poverty and Inequality

**Distribution of income or consumption**

Gini Index: SI.POV.GINI [[Link](https://data.worldbank.org/indicator/SI.POV.GINI)]

Income share held by lowest 20%: SI.DST.FRST.20 [[Link](https://data.worldbank.org/indicator/SI.DST.FRST.20)]

Income share held by second 20%: SI.DST.02ND.20 [[Link](https://data.worldbank.org/indicator/SI.DST.02ND.20)]

Income share held by third 20%: SI.DST.03RD.20 [[Link](https://data.worldbank.org/indicator/SI.DST.03RD.20)]

Income share held by fourth 20%: SI.DST.04TH.20 [[Link](https://data.worldbank.org/indicator/SI.DST.04TH.20)]

Income share held by highest 20%: SI.DST.05TH.20 [[Link](https://data.worldbank.org/indicator/SI.DST.05TH.20)]

## Setup

Install a package DescTools first.

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(broom)  
library(WDI)  
library(DescTools)

## Importing Data

df\_gini <- WDI(indicator = c(gini = "SI.POV.GINI",  
 `0-20` = "SI.DST.FRST.20",  
 `20-40` = "SI.DST.02ND.20",  
 `40-60` = "SI.DST.03RD.20",  
 `60-80` = "SI.DST.04TH.20",  
 `80-100` = "SI.DST.05TH.20"))

write\_csv(df\_gini, "data/gini.csv")

df\_gini <- read\_csv("data/gini.csv")

## Rows: 16758 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): country, iso2c, iso3c  
## dbl (7): year, gini, 0-20, 20-40, 40-60, 60-80, 80-100  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

REGION <- c("1A", "1W", "4E", "6F", "6N", "6X", "7E", "8S", "A4", "A5",   
"A9", "B1", "B2", "B3", "B4", "B6", "B7", "B8", "C4", "C5", "C6",   
"C7", "C8", "C9", "D2", "D3", "D4", "D5", "D6", "D7", "EU", "F1",   
"F6", "M1", "M2", "N6", "OE", "R6", "S1", "S2", "S3", "S4", "T2",   
"T3", "T4", "T5", "T6", "T7", "V1", "V2", "V3", "V4", "XC", "XD",   
"XE", "XF", "XG", "XH", "XI", "XJ", "XL", "XM", "XN", "XO", "XP",   
"XQ", "XT", "XU", "XY", "Z4", "Z7", "ZB", "ZF", "ZG", "ZH", "ZI",   
"ZJ", "ZQ", "ZT")

## Viewing Data

df\_gini

## # A tibble: 16,758 × 10  
## country iso2c iso3c year gini `0-20` `20-40` `40-60` `60-80` `80-100`  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Afghanistan AF AFG 1960 NA NA NA NA NA NA  
## 2 Afghanistan AF AFG 1961 NA NA NA NA NA NA  
## 3 Afghanistan AF AFG 1962 NA NA NA NA NA NA  
## 4 Afghanistan AF AFG 1963 NA NA NA NA NA NA  
## 5 Afghanistan AF AFG 1964 NA NA NA NA NA NA  
## 6 Afghanistan AF AFG 1965 NA NA NA NA NA NA  
## 7 Afghanistan AF AFG 1966 NA NA NA NA NA NA  
## 8 Afghanistan AF AFG 1967 NA NA NA NA NA NA  
## 9 Afghanistan AF AFG 1968 NA NA NA NA NA NA  
## 10 Afghanistan AF AFG 1969 NA NA NA NA NA NA  
## # ℹ 16,748 more rows

## Transforming Data

We add a new column with the value s80/s20 = 80-100/0-20.

df\_gini <- df\_gini |> mutate(`s80/s20` = `80-100`/`0-20`)  
df\_gini

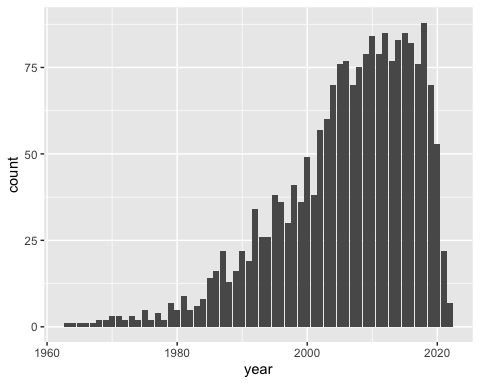
## # A tibble: 16,758 × 11  
## country iso2c iso3c year gini `0-20` `20-40` `40-60` `60-80` `80-100`  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Afghanistan AF AFG 1960 NA NA NA NA NA NA  
## 2 Afghanistan AF AFG 1961 NA NA NA NA NA NA  
## 3 Afghanistan AF AFG 1962 NA NA NA NA NA NA  
## 4 Afghanistan AF AFG 1963 NA NA NA NA NA NA  
## 5 Afghanistan AF AFG 1964 NA NA NA NA NA NA  
## 6 Afghanistan AF AFG 1965 NA NA NA NA NA NA  
## 7 Afghanistan AF AFG 1966 NA NA NA NA NA NA  
## 8 Afghanistan AF AFG 1967 NA NA NA NA NA NA  
## 9 Afghanistan AF AFG 1968 NA NA NA NA NA NA  
## 10 Afghanistan AF AFG 1969 NA NA NA NA NA NA  
## # ℹ 16,748 more rows  
## # ℹ 1 more variable: `s80/s20` <dbl>

## Visualization and Analysis

### Number of Data in Each Year

Check the number of data available in year year.

df\_gini |> drop\_na(gini, `0-20`, `80-100`) |>   
 ggplot(aes(year)) + geom\_bar()



### Correlation of Three Indicators

We calculate the correlations among three indicators, GINI, top 20% and s80/s20 ratio.

1. Correlation using all available values.
2. Correlation using all available values of countries.
3. Correlation using all available values of countries in 2018.

df\_gini |> drop\_na(gini, `0-20`, `80-100`) |> select(gini, `80-100`, `s80/s20`) |>  
 cor() |> as.data.frame()

## gini 80-100 s80/s20  
## gini 1.0000000 0.9943488 0.8663291  
## 80-100 0.9943488 1.0000000 0.8592673  
## s80/s20 0.8663291 0.8592673 1.0000000

df\_gini |> drop\_na(gini, `0-20`, `80-100`) |>   
 filter(!(iso2c %in% REGION)) |> select(gini, `80-100`, `s80/s20`) |>  
 cor() |> as.data.frame()

## gini 80-100 s80/s20  
## gini 1.0000000 0.9943488 0.8663291  
## 80-100 0.9943488 1.0000000 0.8592673  
## s80/s20 0.8663291 0.8592673 1.0000000

df\_gini |> drop\_na(gini, `0-20`, `80-100`) |> filter(year == 2018) |>   
 filter(!(iso2c %in% REGION)) |> select(gini, `80-100`, `s80/s20`) |>  
 cor() |> as.data.frame()

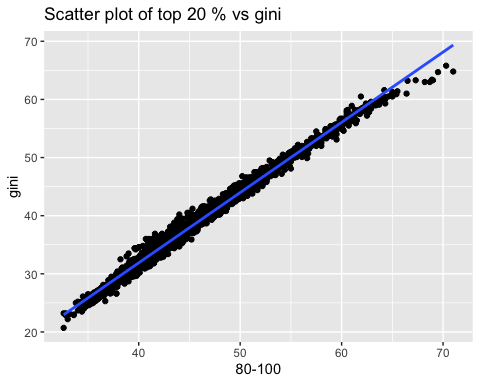
## gini 80-100 s80/s20  
## gini 1.0000000 0.9894834 0.9343159  
## 80-100 0.9894834 1.0000000 0.9074783  
## s80/s20 0.9343159 0.9074783 1.0000000

**Observations:**

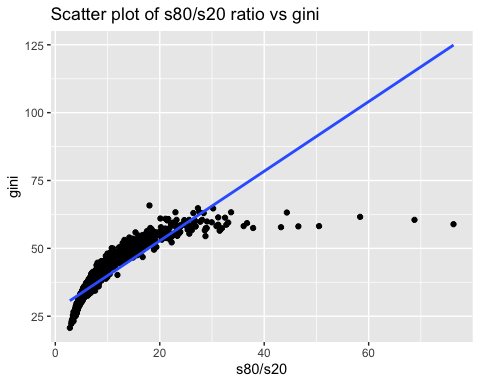
* The correlation between GINI index and the top 20% share of income is very close to 1.
* We chose 2018 as it is the year we have the most available values.
* There are no regional values of these three indices. So the values of the first two coincide.

### Scatter Plots

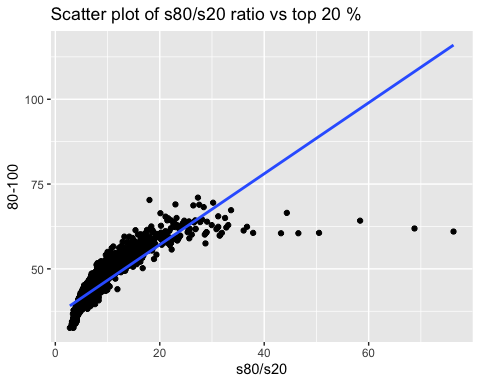
df\_gini |> drop\_na(gini, `0-20`, `80-100`) |>   
 ggplot(aes(`80-100`, gini)) + geom\_point() +   
 geom\_smooth(formula = 'y~x', method = "lm", se = FALSE) +   
 labs(title = "Scatter plot of top 20 % vs gini")



df\_gini |> drop\_na(gini, `0-20`, `80-100`) |>   
 ggplot(aes(`s80/s20`, gini)) + geom\_point() +   
 geom\_smooth(formula = 'y~x', method = "lm", se = FALSE) +  
 labs(title = "Scatter plot of s80/s20 ratio vs gini")



df\_gini |> drop\_na(gini, `0-20`, `80-100`) |>   
 ggplot(aes(`s80/s20`, `80-100`)) + geom\_point() +   
 geom\_smooth(formula = 'y~x', method = "lm", se = FALSE) +  
 labs(title = "Scatter plot of s80/s20 ratio vs top 20 %")



## Models

We set three models.

model\_gini\_top20 <- df\_gini |> lm(gini ~ `80-100`, data = \_)  
model\_gini\_8020 <- df\_gini |> lm(gini ~ `s80/s20`, data = \_)  
model\_8020\_top20 <- df\_gini |> lm(`s80/s20` ~ `80-100`, data = \_)

### Summary of the model gini ~ top 20%

model\_gini\_top20 |> summary()

##   
## Call:  
## lm(formula = gini ~ `80-100`, data = df\_gini)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.5592 -0.6513 -0.0618 0.5784 3.4748   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -16.456298 0.131171 -125.5 <2e-16 \*\*\*  
## `80-100` 1.208670 0.002879 419.8 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9416 on 2005 degrees of freedom  
## (14751 observations deleted due to missingness)  
## Multiple R-squared: 0.9887, Adjusted R-squared: 0.9887   
## F-statistic: 1.762e+05 on 1 and 2005 DF, p-value: < 2.2e-16

### Summary of the model gini ~ s80/s20

model\_gini\_8020 |> summary()

##   
## Call:  
## lm(formula = gini ~ `s80/s20`, data = df\_gini)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -66.009 -2.487 0.140 2.871 15.560   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 27.12397 0.17022 159.35 <2e-16 \*\*\*  
## `s80/s20` 1.28242 0.01652 77.65 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.429 on 2004 degrees of freedom  
## (14752 observations deleted due to missingness)  
## Multiple R-squared: 0.7505, Adjusted R-squared: 0.7504   
## F-statistic: 6029 on 1 and 2004 DF, p-value: < 2.2e-16

### Summary of the model s80/s20 ~ top 20%

model\_8020\_top20 |> summary()

##   
## Call:  
## lm(formula = `s80/s20` ~ `80-100`, data = df\_gini)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.242 -1.450 -0.068 0.975 56.544   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -23.327946 0.427285 -54.6 <2e-16 \*\*\*  
## `80-100` 0.705481 0.009382 75.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.064 on 2004 degrees of freedom  
## (14752 observations deleted due to missingness)  
## Multiple R-squared: 0.7383, Adjusted R-squared: 0.7382   
## F-statistic: 5655 on 1 and 2004 DF, p-value: < 2.2e-16

### broom::tidy and broom::glance

tidy(model\_gini\_top20) |> rbind(tidy(model\_gini\_8020)) |> rbind(tidy(model\_8020\_top20))

## # A tibble: 6 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -16.5 0.131 -125. 0  
## 2 `80-100` 1.21 0.00288 420. 0  
## 3 (Intercept) 27.1 0.170 159. 0  
## 4 `s80/s20` 1.28 0.0165 77.6 0  
## 5 (Intercept) -23.3 0.427 -54.6 0  
## 6 `80-100` 0.705 0.00938 75.2 0

glance(model\_gini\_top20) |> rbind(glance(model\_gini\_8020)) |> rbind(glance(model\_8020\_top20))

## # A tibble: 3 × 12  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.989 0.989 0.942 176193. 0 1 -2726. 5458. 5475.  
## 2 0.751 0.750 4.43 6029. 0 1 -5831. 11667. 11684.  
## 3 0.738 0.738 3.06 5655. 0 1 -5092. 10189. 10206.  
## # ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

## Conclusion

The GINI index and the income share held by highest 20% is strongly correlated. The relation is even stronger than the correlation between the GINI index and the s80/s20 ratio.

## Calculation Model of Gini Index

df\_gini\_calc <- df\_gini |>   
 mutate(`0` = 0, `20` = `0-20`,  
 `40` = `0-20` + `20-40`,   
 `60` = `0-20` + `20-40` + `40-60`,   
 `80` = `0-20` + `20-40` + `40-60` + `60-80`,   
 `100` = 100) |>  
 select(-c(`0-20`:`60-80`))   
df\_gini\_calc %>% drop\_na()

## # A tibble: 2,003 × 13  
## country iso2c iso3c year gini `80-100` `s80/s20` `0` `20` `40` `60`  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Albania AL ALB 1996 27 36.1 3.92 0 9.2 22.9 40.6  
## 2 Albania AL ALB 2002 31.7 40.4 4.81 0 8.4 21 37.5  
## 3 Albania AL ALB 2005 30.6 39.2 4.67 0 8.4 21.3 38.3  
## 4 Albania AL ALB 2008 30 39 4.38 0 8.9 22 38.8  
## 5 Albania AL ALB 2012 29 37.8 4.25 0 8.9 22.1 39.4  
## 6 Albania AL ALB 2014 34.6 41.7 5.96 0 7 18.5 34.9  
## 7 Albania AL ALB 2015 32.8 40.6 5.27 0 7.7 19.8 36.3  
## 8 Albania AL ALB 2016 33.7 41.2 5.64 0 7.3 19.1 35.5  
## 9 Albania AL ALB 2017 33.1 40.7 5.36 0 7.6 19.6 36.1  
## 10 Albania AL ALB 2018 30.1 38.2 4.84 0 7.9 20.9 38.4  
## # ℹ 1,993 more rows  
## # ℹ 2 more variables: `80` <dbl>, `100` <dbl>

df\_gini\_calc\_long <- df\_gini\_calc |> pivot\_longer(`0`:`100`, names\_to = "classes", values\_to = "cumulative\_share") |> mutate(classes = as.numeric(classes))  
df\_gini\_calc\_long %>% drop\_na()

## # A tibble: 12,018 × 9  
## country iso2c iso3c year gini `80-100` `s80/s20` classes cumulative\_share  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Albania AL ALB 1996 27 36.1 3.92 0 0   
## 2 Albania AL ALB 1996 27 36.1 3.92 20 9.2  
## 3 Albania AL ALB 1996 27 36.1 3.92 40 22.9  
## 4 Albania AL ALB 1996 27 36.1 3.92 60 40.6  
## 5 Albania AL ALB 1996 27 36.1 3.92 80 63.9  
## 6 Albania AL ALB 1996 27 36.1 3.92 100 100   
## 7 Albania AL ALB 2002 31.7 40.4 4.81 0 0   
## 8 Albania AL ALB 2002 31.7 40.4 4.81 20 8.4  
## 9 Albania AL ALB 2002 31.7 40.4 4.81 40 21   
## 10 Albania AL ALB 2002 31.7 40.4 4.81 60 37.5  
## # ℹ 12,008 more rows

df\_gini\_f <- df\_gini\_calc\_long |> group\_by(country, year) |>   
 drop\_na(gini) |>  
 reframe(gini, gini\_spline = round(100-AUC(classes, cumulative\_share, method = "spline")/50, digits = 1), gini\_trapezoid = round(100-AUC(classes, cumulative\_share)/50, digits = 1), `80-100`, `s80/s20`) |>   
 distinct(country, year, gini, gini\_spline, gini\_trapezoid, `80-100`, `s80/s20`)  
df\_gini\_f

## # A tibble: 2,009 × 7  
## country year gini gini\_spline gini\_trapezoid `80-100` `s80/s20`  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Albania 1996 27 26.4 25.4 36.1 3.92  
## 2 Albania 2002 31.7 30.6 29.4 40.4 4.81  
## 3 Albania 2005 30.6 29.7 28.5 39.2 4.67  
## 4 Albania 2008 30 28.9 27.7 39 4.38  
## 5 Albania 2012 29 28.1 27 37.8 4.25  
## 6 Albania 2014 34.6 33.9 32.6 41.7 5.96  
## 7 Albania 2015 32.8 32 30.8 40.6 5.27  
## 8 Albania 2016 33.7 33.1 31.8 41.2 5.64  
## 9 Albania 2017 33.1 32.2 30.9 40.7 5.36  
## 10 Albania 2018 30.1 29.6 28.4 38.2 4.84  
## # ℹ 1,999 more rows

df\_gini\_f |> drop\_na(gini, gini\_spline, gini\_trapezoid, `80-100`, `s80/s20`) |> select(gini, gini\_spline, gini\_trapezoid, `80-100`, `s80/s20`) |> cor() |> as.data.frame()

## gini gini\_spline gini\_trapezoid 80-100 s80/s20  
## gini 1.0000000 0.9993752 0.9992505 0.9943488 0.8663291  
## gini\_spline 0.9993752 1.0000000 0.9999799 0.9913027 0.8666667  
## gini\_trapezoid 0.9992505 0.9999799 1.0000000 0.9908249 0.8665828  
## 80-100 0.9943488 0.9913027 0.9908249 1.0000000 0.8592673  
## s80/s20 0.8663291 0.8666667 0.8665828 0.8592673 1.0000000

**Observation:**

* Since gini\_spline and gini\_trapezoid are calculated using the definition of the gini index, they are strongly correlated, though they are not exactly equal.