# Income Inequality vs Index of Health and Social Problems

*coop711*

*2017-04-24*

## Data

Equality Trust에서 기부금을 받고 제공하는 두 종류의 자료 중 23개 국가의 각종 지표를 비교한 자료에 World Bank에서 발표하는 GDP자료 (https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita)를 추가하여 읽어들이면,

```
library(knitr)
# rm(list = ls())
# getwd()
load("Inequality_Index_HS.rda")
data.full <- read.csv("../data/international-inequality_GDP.csv", stringsAsFactors = FALSE)
str(data.full)
```

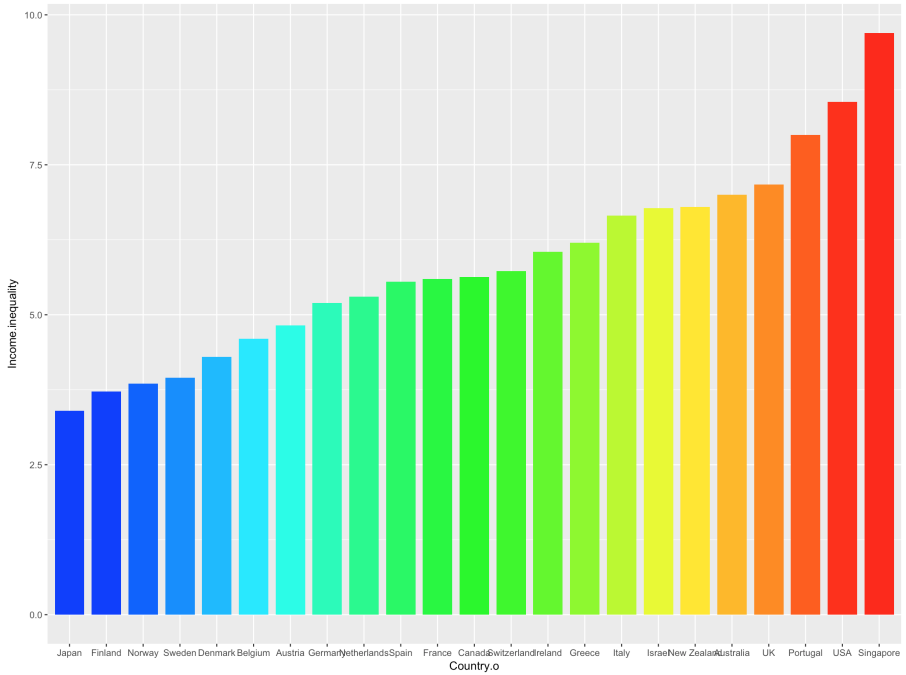
```
## 'data.frame':    23 obs. of  30 variables:
##  $ Country                          : chr  "Australia" "Austria" "Belgium" "Canad
a" ...
##  $ Income.inequality                : num  7 4.82 4.6 5.63 4.3 3.72 5.6 5.2 6.2 6.
05 ...
##  $ Trust                            : num  39.9 33.9 30.7 38.8 66.5 58 22.2 34.8 2
3.7 35.2 ...
##  $ Life.expectancy                  : num  79.2 78.5 78.8 79.3 76.6 78 79 78.3 78.
3 77 ...
##  $ Infant.mortality                 : num  4.9 4.8 5 5.3 5.3 3.7 4.4 4.4 5 5.9 ...
##  $ Obesity                          : num  18.4 14.5 13.5 12.8 15 ...
##  $ Mental.illness                   : num  23 NA 12 19.9 NA NA 18.4 9.1 NA NA ...
##  $ Maths.and.literacy.scores        : num  524 498 518 530 503 ...
##  $ Teenage.births                   : num  18.4 14 9.9 20.2 8.1 9.2 9.3 13.1 11.8
 18.7 ...
##  $ Homicides                        : num  16.9 11.6 13 17.3 12.7 28.2 21.5 13.7 1
3.9 8.6 ...
##  $ Imprisonment..log.               : num  4.61 4.52 4.28 4.77 4.17 4.11 4.5 4.51
 3.33 4.17 ...
##  $ Social.mobility                  : num  NA NA NA 0.14 0.14 0.15 NA 0.17 NA NA
 ...
##  $ Index.of.health...social_problems: num  0.07 0.01 -0.23 -0.07 -0.19 -0.43 0.05
 -0.06 0.38 0.25 ...
##  $ Child.overweight                 : num  NA 11.9 10.4 19.5 10.3 13.3 11.2 11.3 1
6 12.1 ...
##  $ Drugs.index                      : num  1.71 -0.02 -0.18 0.61 -0.09 -0.88 -0.35
 -0.3 -0.99 -0.03 ...
##  $ Calorie.intake                   : int  3142 3753 3632 3167 3405 3197 3576 3395
 3687 3656 ...
##  $ Public.health.expenditure        : num  67.9 69.3 71.7 70.8 82.4 75.6 76 74.9 5
6 76 ...
##  $ Child.wellbeing                  : num  -0.21 -0.07 0.05 0.04 0.21 0.34 -0.17 -
0.01 -0.04 -0.04 ...
##  $ Maths.education.science.score    : num  525 496 515 526 494 ...
##  $ Child.conflict                   : num  NA 0.31 0.33 0.24 -0.14 -1.25 0.59 -0.7
 0.4 -0.06 ...
##  $ Foreign.aid                      : num  0.25 0.52 0.53 0.34 0.81 0.47 0.47 0.35
 0.24 0.41 ...
##  $ Recycling                        : num  7.4 NA NA NA NA NA 6 3.4 NA NA ...
##  $ Peace.index                      : num  1.66 1.48 1.49 1.48 1.38 1.45 1.73 1.52
 1.79 1.4 ...
##  $ Maternity.leave                  : int  0 16 15 17 18 18 16 14 17 18 ...
##  $ Advertising                      : num  1.24 0.97 0.82 0.77 0.75 0.9 0.71 0.99
 1.04 1 ...
##  $ Police                           : int  304 305 357 186 192 160 NA 303 NA NA
 ...
##  $ Social.expenditure               : num  17.8 27.5 26.5 17.2 27.6 25.8 29 27.3 1
9.9 15.8 ...
##  $ Women.s_status                   : num  0.46 -0.81 0.61 0.56 0.83 1.08 -0.17 -
0.21 -0.85 -0.21 ...
##  $ Lone.parents                     : int  21 15 12 17 22 19 12 21 3 14 ...
##  $ GDP_WB                           : int  45926 47682 43435 45066 45537 40676 393
28 46401 26851 49393 ...
```

## 소득불평등

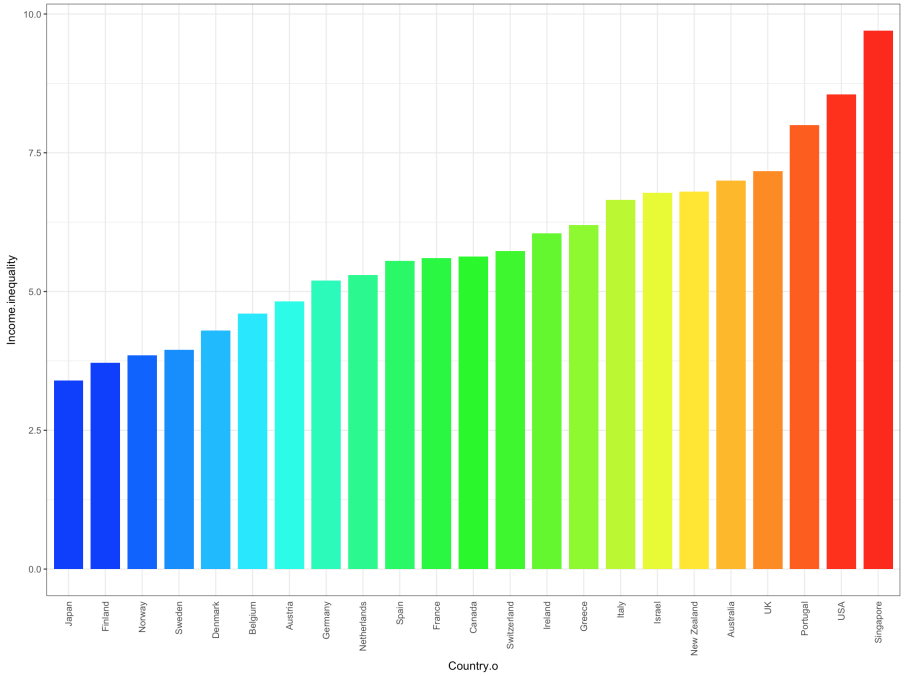
이 자료 중 소득불평등을 나타내는 지표는 5분위계수로서 두번째 컬럼에 `Income.inequality` 라는 이름으로 나와 있고, 건강과 사회문제 지표는 13번째 컬럼에 `Index.of.health...social_problems` 라는 이름으로 주어져 있다. 나라들은 `Country` 라는 변수명으로 첫번째 컬럼에 나와 있다.

## Barplot(geom_bar)

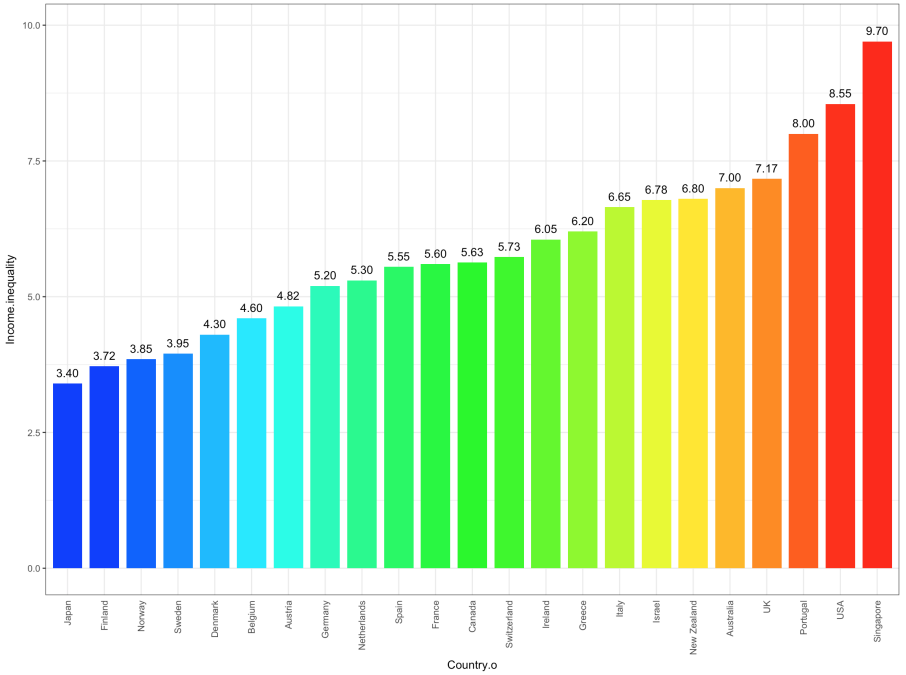
```
# par(mai = c(2.0, 0.8, 0.8, 0.4) + 0.02)
library(ggplot2)
o.ineq <- order(data.full$Income.inequality)
fifth_ratio <- data.full$Income.inequality
Country <- data.full$Country
data.full$Country.o <- factor(data.full$Country, levels = Country[o.ineq])
g1 <- ggplot(data = data.full[c("Income.inequality", "Country.o")])
g2 <- g1 + geom_bar(aes(x = Country.o, y = Income.inequality), stat = "identity", wid
th = 0.8, fill = rev(rainbow(23, start = 0, end = 2/3)))
g2
```



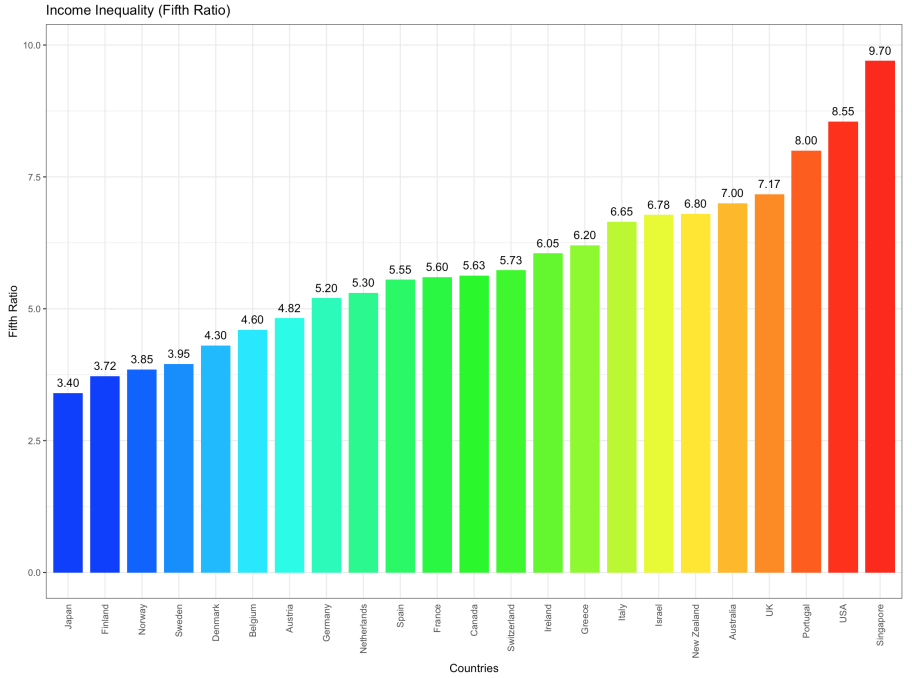

```
g3 <- g2 +
  theme_bw()
g4 <- g3 +
#  theme(axis.text.x = element_blank()) +
  theme(axis.ticks.x = element_blank()) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
g4
```



```
g5 <- g4 +
  geom_text(aes(x = Country.o, y = Income.inequality + 0.2, label = format(Income.ine
quality, digits = 3)))
g5
```




```
g6 <- g5 +
  labs(title = "Income Inequality (Fifth Ratio)", x = "Countries", y = "Fifth Ratio")
g6
```




```
ggsave("../pics/Fifth_Ratio_ggplot.png", g6, width = 12, height = 6)
```

# 소득불평등과 건강 및 사회문제

## data management

그리고, 건강과 사회문제 지표에 결측치들이 있기 때문에 먼저 이 나라들을 제외하고 분석작업을 수행하여야 한다. `which()`를 이용하여 해당 인덱스를 찾고, 나라명을 추출한다.

```
(country.na <- which(is.na(data.full$Index.of.health...social_problems)))
```

```
## [1] 11 18
```

```
data.full$Country[country.na]
```

```
## [1] "Israel"    "Singapore"
```

결측치가 있는 나라를 빼고, 필요한 변수만 챙겨서 새로운 data frame 을 구성하기 위하여 건강과 사회문제 지표의 위치를 찾아보자.

```
names(data.full)
```

```
##  [1] "Country"
##  [2] "Income.inequality"
##  [3] "Trust"
##  [4] "Life.expectancy"
##  [5] "Infant.mortality"
##  [6] "Obesity"
##  [7] "Mental.illness"
##  [8] "Maths.and.literacy.scores"
##  [9] "Teenage.births"
## [10] "Homicides"
## [11] "Imprisonment..log."
## [12] "Social.mobility"
## [13] "Index.of.health...social_problems"
## [14] "Child.overweight"
## [15] "Drugs.index"
## [16] "Calorie.intake"
## [17] "Public.health.expenditure"
## [18] "Child.wellbeing"
## [19] "Maths.education.science.score"
## [20] "Child.conflict"
## [21] "Foreign.aid"
## [22] "Recycling"
## [23] "Peace.index"
## [24] "Maternity.leave"
## [25] "Advertising"
## [26] "Police"
## [27] "Social.expenditure"
## [28] "Women.s_status"
## [29] "Lone.parents"
## [30] "GDP_WB"
## [31] "Country.o"
```

```
which(names(data.full) == "Index.of.health...social_problems")
```

```
## [1] 13
```

새로운 data frame 을 `data.21` 으로 저장하자. 시각적 가독성을 높이기 위하여 자릿수를 조정한다.

```
options(digits = 2)
v.names <- c("Country", "Income.inequality", "Index.of.health...social_problems", "GD
P_WB")
data.21 <- data.full[-c(11, 18), v.names]
names(data.21)[3] <- "Index.HS"
```

```
kable(data.21)
```

| | Country | Income.inequality | Index.HS | GDP_WB |
|---|---|---|---|---|
| 1 | Australia | 7.0 | 0.07 | 45926 |
| 2 | Austria | 4.8 | 0.01 | 47682 |
| 3 | Belgium | 4.6 | -0.23 | 43435 |
| 4 | Canada | 5.6 | -0.07 | 45066 |
| 5 | Denmark | 4.3 | -0.19 | 45537 |
| 6 | Finland | 3.7 | -0.43 | 40676 |
| 7 | France | 5.6 | 0.05 | 39328 |
| 8 | Germany | 5.2 | -0.06 | 46401 |
| 9 | Greece | 6.2 | 0.38 | 26851 |
| 10 | Ireland | 6.0 | 0.25 | 49393 |
| 12 | Italy | 6.7 | -0.12 | 35463 |
| 13 | Japan | 3.4 | -1.26 | 36319 |
| 14 | Netherlands | 5.3 | -0.51 | 48253 |
| 15 | New Zealand | 6.8 | 0.29 | 37679 |
| 16 | Norway | 3.9 | -0.63 | 65615 |
| 17 | Portugal | 8.0 | 1.18 | 28760 |
| 19 | Spain | 5.5 | -0.30 | 33629 |
| 20 | Sweden | 4.0 | -0.83 | 45297 |
| 21 | Switzerland | 5.7 | -0.46 | 59540 |
| 22 | UK | 7.2 | 0.79 | 40233 |
| 23 | USA | 8.6 | 2.02 | 54630 |

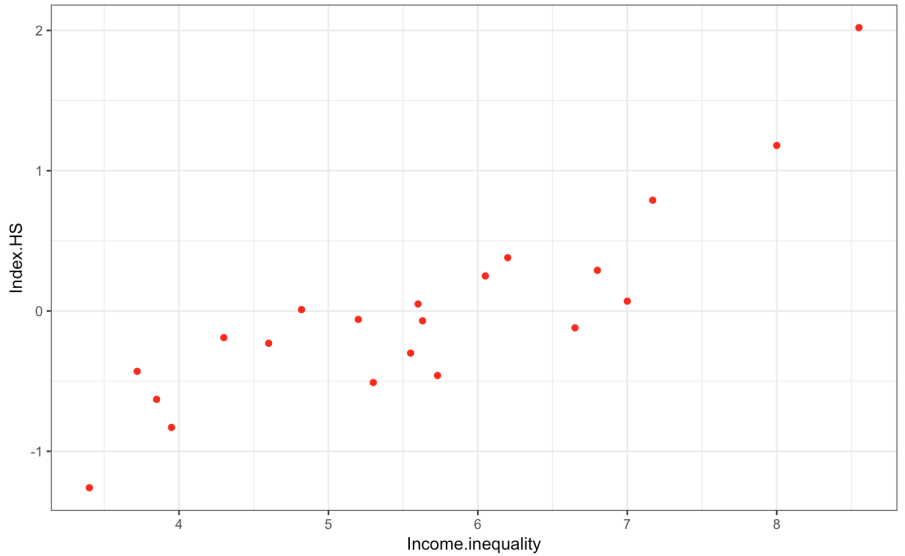
# Plots

우선 소득불평등과 건강 및 사회문제 지표의 관계를 대략적으로 살펴보면,

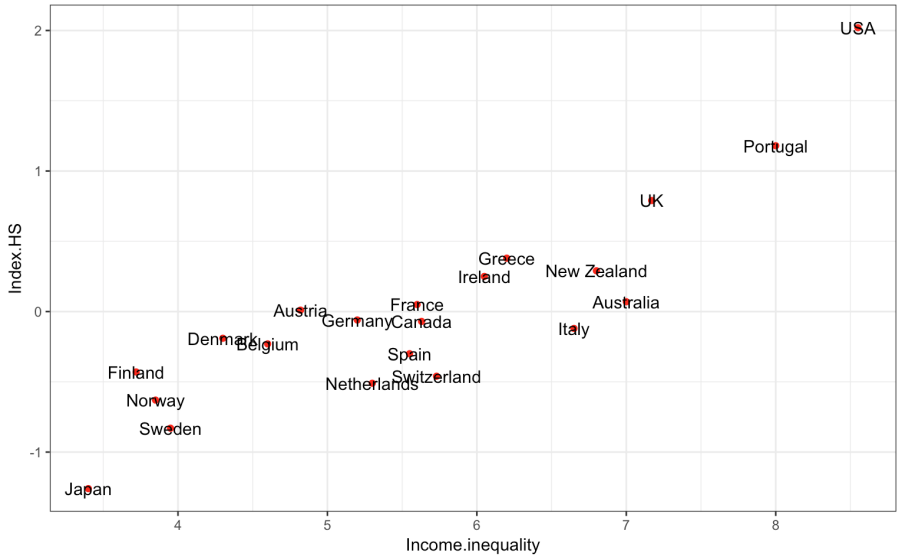
```
Index_inequality.df <- data.21[c("Country", "Income.inequality", "Index.HS")]
# plot(Index_inequality.df)
# plot(data.21[c("Income.inequality", "Index.HS")])
cor.1 <- cor(data.21["Income.inequality"], data.21["Index.HS"])
cor.1
```

```
##                   Index.HS
## Income.inequality     0.87
```

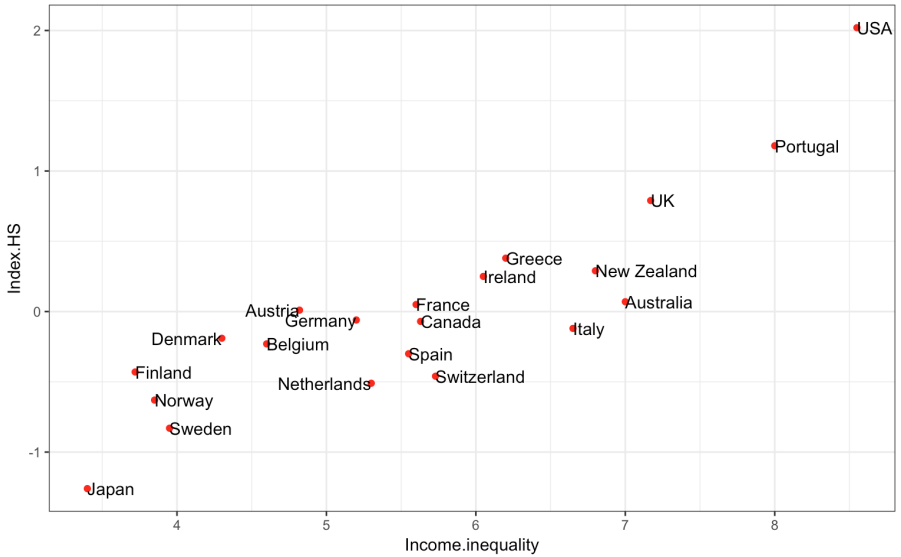
```
p1 <- ggplot(data = Index_inequality.df, aes(x = Income.inequality, y = Index.HS, lab
el = Country)) + theme_bw()
p2 <- p1 + geom_point(col = "red")
p2
```



```
p3 <- p2 + geom_text()
p3
```




```
p4 <- p2 + geom_text(hjust = hjust.text)
p4
```




텍스트 위치를 옮겨 보자. 점의 왼편으로 옮겨야 할 나라들( `ggplot` 에서는 `hjust = "right"` )을 먼저 찾아보자.

```
Country <- data.21$Country
which(Country %in% c("Austria", "Denmark", "Germany", "Netherlands"))
```

```
## [1]  2  5  8 13
```

```
text.left <- which(Country %in% c("Austria", "Denmark", "Germany", "Netherlands"))
text.left
```

```
## [1]  2  5  8 13
```

```
text.right <- setdiff(1:nrow(data.21), text.left)
text.right
```

```
##  [1]  1  3  4  6  7  9 10 11 12 14 15 16 17 18 19 20 21
```

```
hjust.text <- ifelse(1:nrow(data.21) %in% text.left, "right", "left")
```

독일의 라벨을 위로 붙이면 보기가 나아질 것으로 생각되므로,

```
which(Country %in% "Germany")
```

```
## [1] 8
```

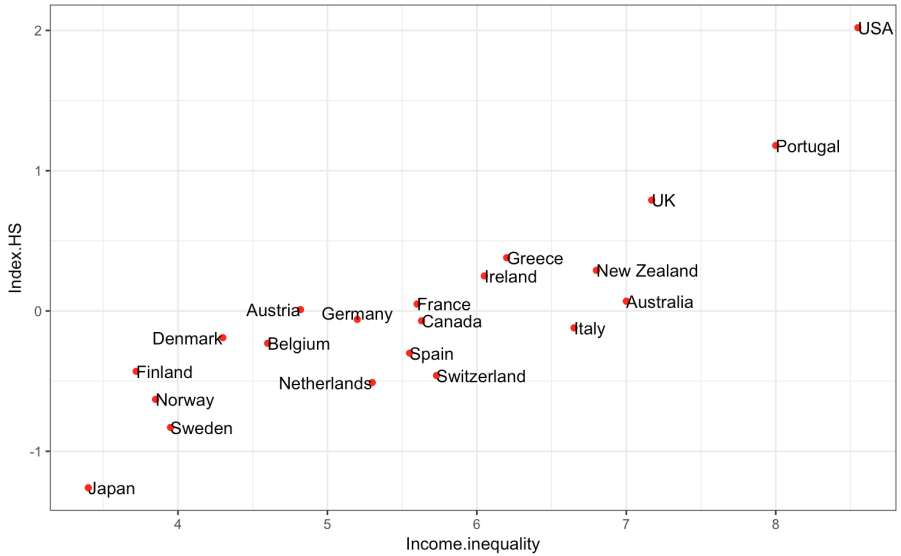
```
text.up <- which(Country %in% "Germany")
text.up
```

```
## [1] 8
```

```
text.left <- setdiff(1:nrow(data.21), c(text.right, text.up))
text.left
```

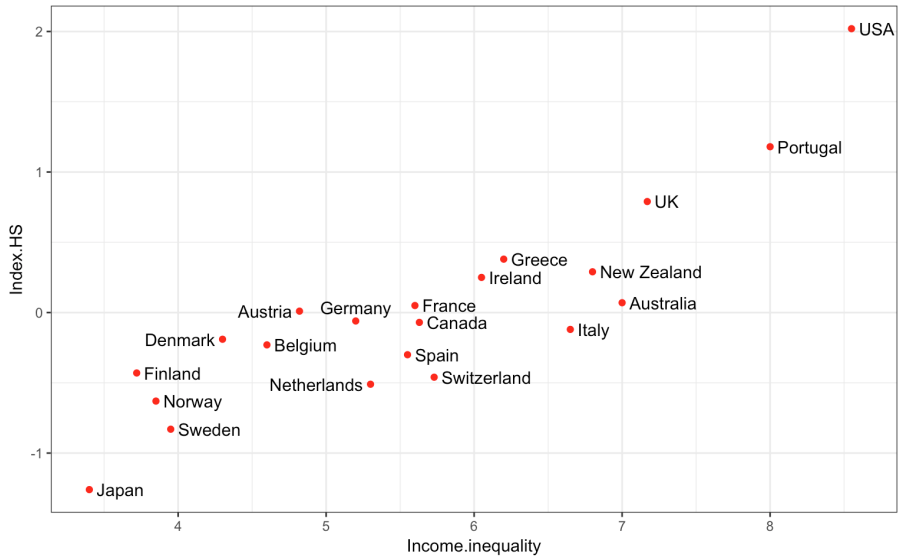
```
## [1]  2  5 13
```

```
vjust.text <- ifelse(1:nrow(data.21) %in% text.up, "bottom", "center")
hjust.text <- ifelse(1:nrow(data.21) %in% text.up, "middle", hjust.text)
p5 <- p2 + geom_text(hjust = hjust.text, vjust = vjust.text)
p5
```



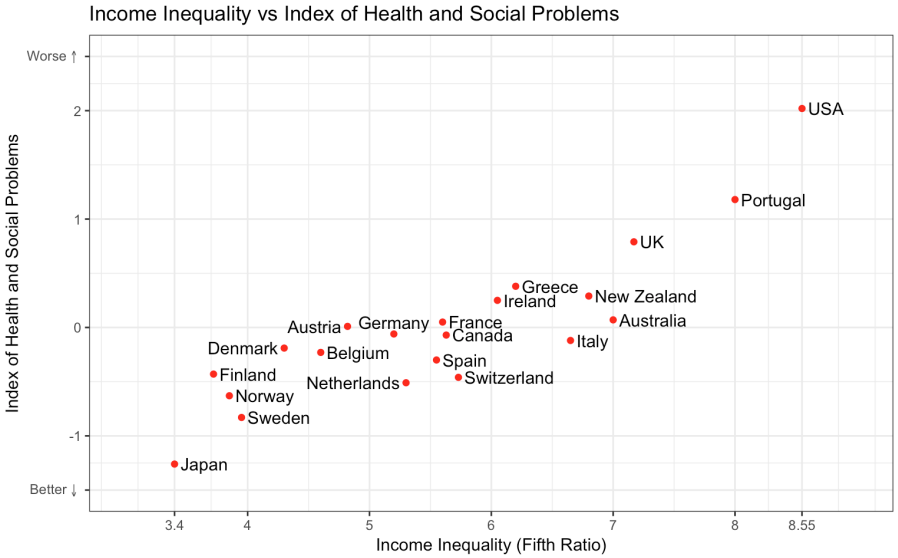

세부조정을 위해서 `nudge_x, nudge_y` 설정

```
nudge_y.text <- ifelse(vjust.text == "bottom", 0.05, 0)
nudge_x.text <- ifelse(hjust.text == "middle", 0, ifelse(hjust.text == "right",
-0.05, 0.05))
p6 <- p2 + geom_text(hjust = hjust.text, vjust = vjust.text, nudge_x = nudge_x.text,
nudge_y = nudge_y.text)
p6
```



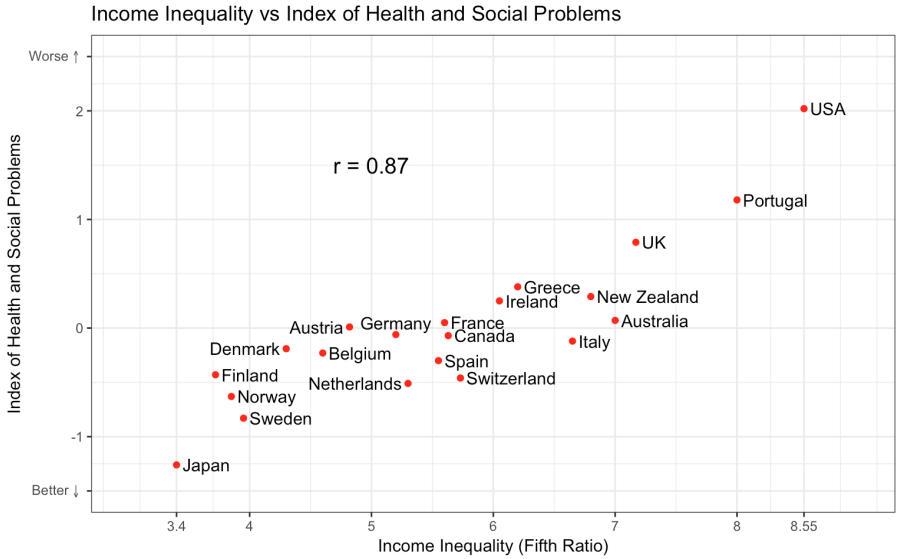
건강 및 사회문제 지표의 경우 어느 방향이 좋은지 알 수 없으므로 친절하게 도표의 주변에(margin)에 알려주고, 이제 조정된 text 외에 x-축과 y-축에 적절한 라벨과 메인 타이틀을 넣어보자.

```
main.title <- "Income Inequality vs Index of Health and Social Problems"
x.lab <- "Income Inequality (Fifth Ratio)"
y.lab <- "Index of Health and Social Problems"
lowest <- data.21$Income.inequality[Country == "Japan"]
highest <- data.21$Income.inequality[Country == "USA"]
p7 <- p6 +
  scale_x_continuous(name = x.lab, breaks = c(lowest, 4:8, highest), labels = c(lowes
t, 4:8, highest), limits = c(3, 9)) +
  scale_y_continuous(name = y.lab, breaks = c(-1.5, -1:2, 2.5), labels =
c(expression("Better" %down% ""), -1:2, expression("Worse" %up% "")), limits =
c(-1.5, 2.5)) +
  labs(title = main.title)
p7
```



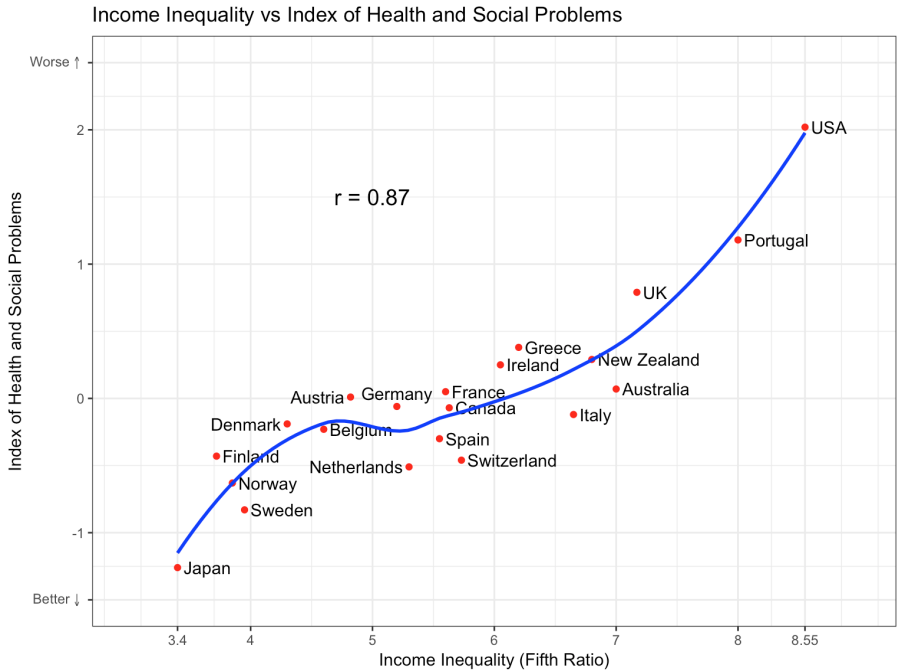

상관계수를 텍스트로 그림 안에 넣어주고 여기까지 작업한 내용을 별도의 파일로 저장해 놓으려면,

```
p8 <- p7 + annotate("text", x = 5, y = 1.5, label = paste("r =", round(cor.1, digits
= 2)), size = 5)
p8
```



선형회귀선을 추가하여 대체적인 추세를 보려면 `lm()` 을 이용하되, `x , y` 의 순서를 제대로 바꿔야 함에 유의.

```
lm.ineq <- lm(Index.HS ~ Income.inequality, data = Index_inequality.df)
# lm.ineq <- lm(Index_inequality.df[2:1])
# p9 <- p8 + geom_abline(intercept = lm.ineq$coef[1], slope = lm.ineq$coef[2], colour
 = "blue")
p9 <- p8 + geom_smooth(method = "loess", se = FALSE, colour = "blue")
p9
```




```
ggsave("../pics/Inequality_vs_HS_Index_ggplot.png", p9, width = 8, height = 6)
```

# GDP와 건강 및 사회문제 지수

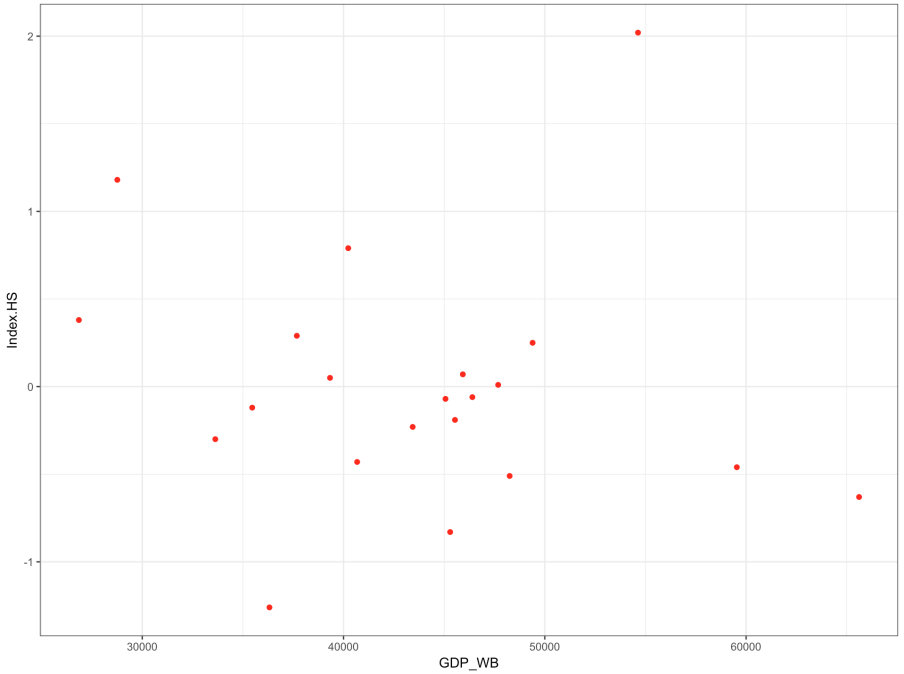
## Scatter Diagram

```
(Index_GDP.df <- data.21[c("Country", "GDP_WB", "Index.HS")])
```

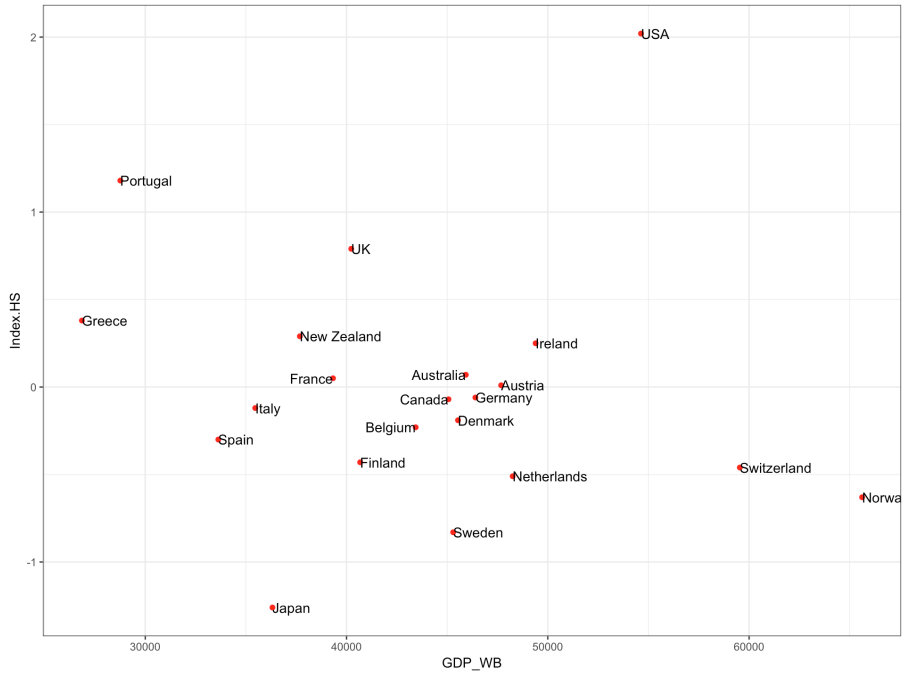
```
##         Country GDP_WB Index.HS
## 1     Australia  45926     0.07
## 2       Austria  47682     0.01
## 3       Belgium  43435    -0.23
## 4        Canada  45066    -0.07
## 5       Denmark  45537    -0.19
## 6       Finland  40676    -0.43
## 7        France  39328     0.05
## 8       Germany  46401    -0.06
## 9        Greece  26851     0.38
## 10      Ireland  49393     0.25
## 12        Italy  35463    -0.12
## 13        Japan  36319    -1.26
## 14  Netherlands  48253    -0.51
## 15  New Zealand  37679     0.29
## 16       Norway  65615    -0.63
## 17     Portugal  28760     1.18
## 19        Spain  33629    -0.30
## 20       Sweden  45297    -0.83
## 21  Switzerland  59540    -0.46
## 22           UK  40233     0.79
## 23          USA  54630     2.02
```

```
cor.2 <- cor(data.21["GDP_WB"], data.21["Index.HS"])
text.left.2 <- which(Country %in% c("Australia", "Belgium", "Canada", "France"))
text.right.2 <- setdiff(1:nrow(data.21), c(text.left.2))
hjust.text.2 <- ifelse(1:nrow(data.21) %in% text.left.2, "right", "left")
nudge_x.text.2 <- ifelse(hjust.text.2 == "right", -250, 250)
```

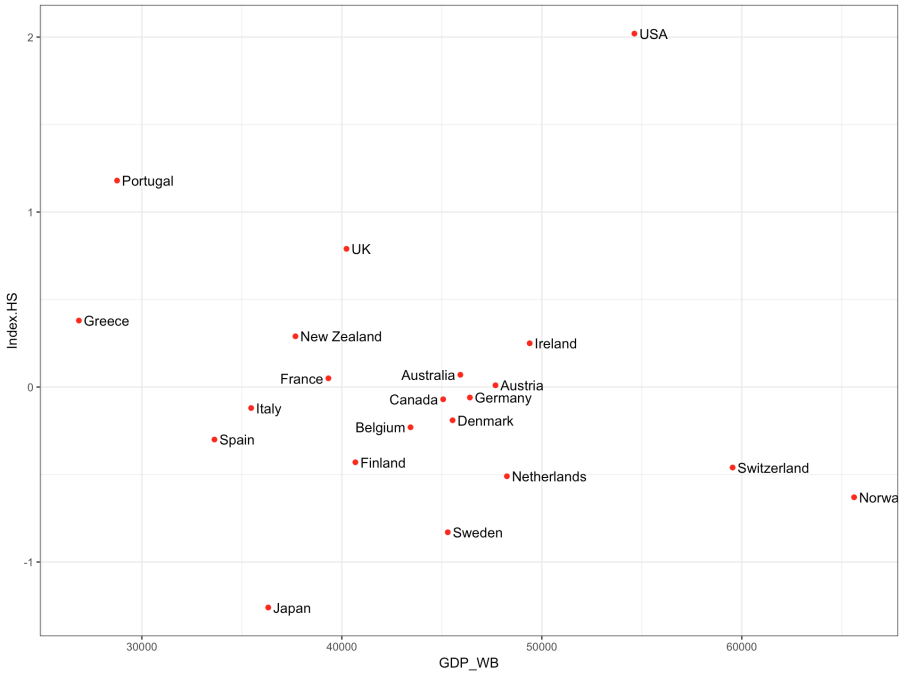
```
gd1 <- ggplot(data = Index_GDP.df, aes(x = GDP_WB, y = Index.HS, label = Country)) +
theme_bw()
gd2 <- gd1 +
  geom_point(colour = "red")
gd2
```



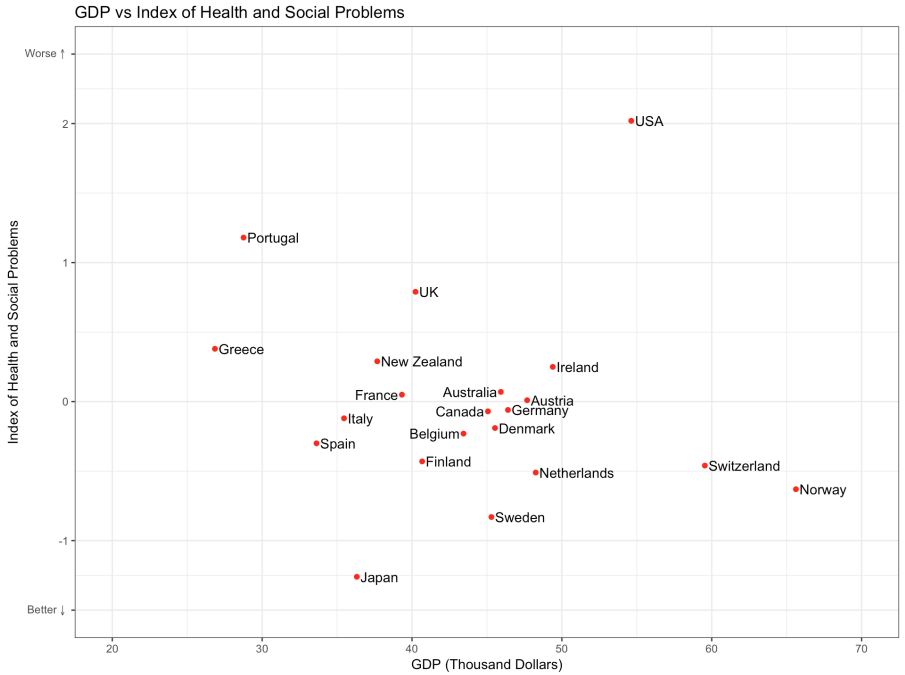

```
gd3 <- gd2 +
  geom_text(hjust = hjust.text.2)
gd3
```



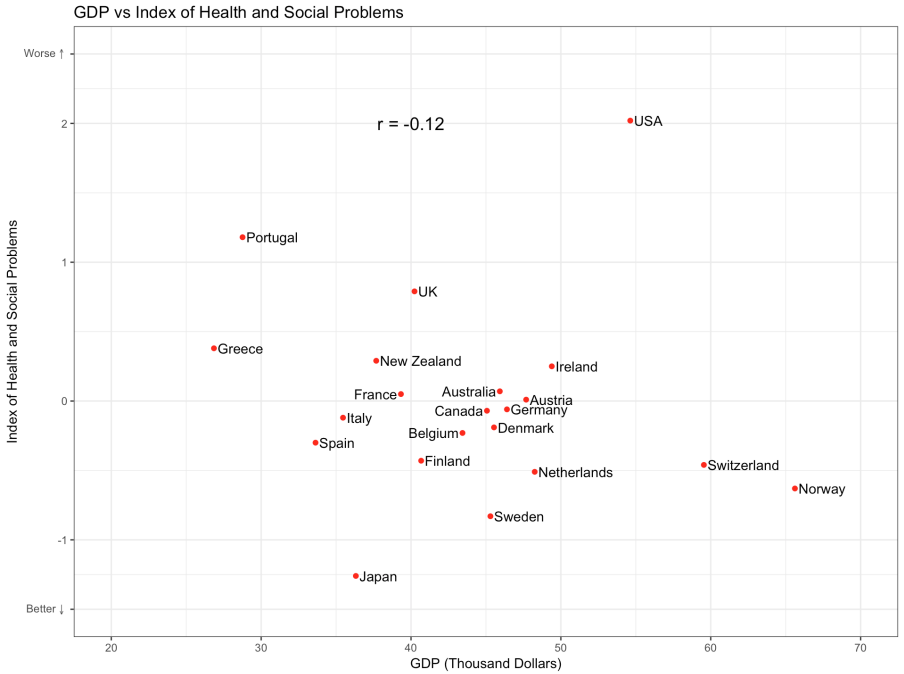
```
gd4 <- gd2 +
  geom_text(hjust = hjust.text.2, nudge_x = nudge_x.text.2)
gd4
```




```
main.title.2 <- "GDP vs Index of Health and Social Problems"
x.lab.2 <- "GDP (Thousand Dollars)"
y.lab.2 <- "Index of Health and Social Problems"
gd5 <- gd4 +
  scale_x_continuous(name = x.lab.2, breaks = seq(20000, 70000, by = 10000), labels =
 seq(20, 70, by = 10), limits = c(20000, 70000)) +
  scale_y_continuous(name = y.lab.2, breaks = c(-1.5, -1:2, 2.5), labels = c(expressi
on("Better" %down% ""), -1:2, expression("Worse" %up% "")), limits = c(-1.5, 2.5)) +
  labs(title = main.title.2)
gd5
```



```
gd6 <- gd5 + annotate("text", x = 40000, y = 2, label = paste("r =", round(cor.2, dig
its = 2)), size = 5)
gd6
```




```
gd7 <- gd6 + geom_smooth(colour = "blue", se = FALSE)
gd7
```

```
## `geom_smooth()` using method = 'loess'
```