```
---
title: "Income Inequality vs Index of Health and Social Problems"
author: "coop711"
date: "`r Sys.Date()`"
output: html_document
---

### Data Preparation

<!--
`xlsx` package는 Excel 자료를 다루는 데 매우 유용한데, `read.xlsx(filename, n)`의 구조로 되어 있으며, 여기서 `n`은 엑셀 시트의 번호이다.

```{r, data, message = FALSE}
# install.packages("xlsx", repos = "https://cran.rstudio.com")
library(xlsx)
```
-->

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,

```{r, data with GDP}
library(knitr)
# load("Inequality_Index_HS.rda")
data.full <- read.csv("../data/international-inequality_GDP.csv", stringsAsFactors = FALSE)
str(data.full)
```

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

```{r, which}
(country.na <- which(is.na(data.full$Index.of.health...social_problems)))
data.full$Country[country.na]
```

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

```{r, NA in Index.of.health...social_problems}
names(data.full)
which(names(data.full) == "Index.of.health...social_problems")
```

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

```{r, data for 21 countries}
options(digits = 2)
v.names <- c("Country", "Income.inequality", "Index.of.health...social_problems", "GDP_WB")
data.21 <- data.full[-c(11, 18), v.names]
names(data.21)[3] <- "Index.HS"
kable(data.21)
```

Plots

우선 소득불평등과 건강 및 사회문제 지표의 관계를 대략적으로 살펴보면,

```{r, base plot}
Index_inequality.df <- data.21[c("Income.inequality", "Index.HS")]
# str(Index_inequality.df)
plot(Index_inequality.df)
cor.1 <- cor(data.21["Income.inequality"], data.21["Index.HS"])
cor.1
```

매우 높은 양의 상관관계(r = `r cor.1`) 가 관찰됨을 알 수 있다. 자주 사용하는 `data.21[c("Income.inequality", "Index.HS")]`를 간단한 R 오브젝트로 assign하여 반복 사용하고 있다. `cor()`에도 data frame을 사용하면 어떻게 되는지 다음 결과와 비교해 보자.

```{r, cor data frame}
cor(Index_inequality.df)
```

각 점이 어느 나라를 나타내는지 표시하기 위하여 `text()` 를 활용하자. 동그라미 대신 까만 점으로 표시하고, 나라 이름을 올려보자.

```{r, text for countries, fig.width = 8, fig.height = 6}
(Country <- data.21[, "Country"])
(Country.2 <- data.21["Country"])
(Country.3 <- data.21["Country"]$Country)
str(Country)
str(Country.2)
str(Country.3)
plot(Index_inequality.df, pch = 20)
text(Index_inequality.df, labels = Country)
```

text label의 위치 기본값은 바로 점 위임을 알 수 있다. 위치 선정에 가능한 값들을 넣어보자.

```{r, pos = 1, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20)
text(Index_inequality.df, labels = Country, pos = 1)
```

```{r, pos = 2,  fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20)
text(Index_inequality.df, labels = Country, pos = 2)
```

```{r, pos = 3, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20)
text(Index_inequality.df, labels = Country, pos = 3)
```

```{r, pos = 4, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20)
text(Index_inequality.df, labels = Country, pos = 4)
```

우선 x-축과 y-축의 범위를 `xlim = c(3, 9), ylim = c(-1.5, 2.5)`로 하여 미국과 일본의 라벨이 도표 밖으로 나가지 않게 하자. `pos = 4`로 하고 `cex = 0.8`로 하여 글자 크기를 줄여보면,

```{r, xlim-ylim-cex, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20, xlim = c(3, 9), ylim = c(-1.5, 2.5))
text(Index_inequality.df, labels = Country, pos = 4, cex = 0.8)
```

오스트리아, 덴마크, 독일, 네덜란드의 라벨만 점 왼편에 위치시켜 보자. 각 인덱스를 찾아보면,

```{r, label left}
which(Country %in% c("Austria", "Denmark", "Germany", "Netherlands"))
text.left <- which(Country %in% c("Austria", "Denmark", "Germany", "Netherlands"))
text.left
text.right <- setdiff(1:nrow(data.21), text.left)
text.right
pos.text <- ifelse(1:nrow(data.21) %in% text.left, 2, 4)
```
```

```{r, plot labels adjusted, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20, col = "red", xlim = c(3, 9), ylim = c(-1.5, 2.5))
text(Index_inequality.df, labels = Country, pos = pos.text, cex = 0.8)
```

독일의 라벨을 위로 붙이면 보기가 나아질 것으로 생각되므로,

```{r, points for germany}
which(Country %in% "Germany")
text.up <- which(Country %in% "Germany")
text.up
text.left <- setdiff(1:nrow(data.21), c(text.right, text.up))
text.left
pos.text <- ifelse(1:nrow(data.21) %in% text.up, 3, ifelse(1:nrow(data.21) %in%
text.left, 2, 4))
```

이제 조정된 text 외에 x-축과 y-축에 적절한 라벨과 메인 타이틀을 넣어보자.

```{r, labels and title, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20, col = "red", xlim = c(3, 9), ylim = c(-1.5, 2.5), ann
= FALSE)
text(Index_inequality.df, labels = Country, pos = pos.text, cex = 0.8)
main.title <- "Income Inequality vs Index of Health and Social Problems"
x.lab <- "Income Inequality (5th Ratio)"
y.lab <- "Index of Health and Social Problems"
title(main = main.title, xlab = x.lab, ylab = y.lab)
```

건강 및 사회문제 지표의 경우 어느 방향이 좋은지 알 수 없으므로 친절하게 도표의 주변에(margin)에 알려주려면,

```{r, mtext, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20, col = "red", xlim = c(3, 9), ylim = c(-1.5, 2.5), ann
= FALSE)
text(Index_inequality.df, labels = Country, pos = pos.text, cex = 0.8)
main.title <- "Income Inequality vs Index of Health and Social Problems"
x.lab <- "Income Inequality (5th Ratio)"
y.lab <- "Index of Health and Social Problems"
title(main = main.title, xlab = x.lab, ylab = y.lab)
mtext(c("Better", "Worse"), side = 2, at = c(-1.8, 2.8), las = 1)
```

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

```{r, correlation and separate file, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20, col = "red", xlim = c(3, 9), ylim = c(-1.5, 2.5), ann
= FALSE)
text(Index_inequality.df, labels = Country, pos = pos.text, cex = 0.8)
main.title <- "Income Inequality vs Index of Health and Social Problems"
x.lab <- "Income Inequality (5th Ratio)"
y.lab <- "Index of Health and Social Problems"
title(main = main.title, xlab = x.lab, ylab = y.lab)
mtext(c("Better", "Worse"), side = 2, at = c(-1.8, 2.8), las = 1)
text(x = 5, y = 1.5, labels = paste("r =", round(cor(Index_inequality.df[1],
Index_inequality.df[2]), digits = 2)))
# dev.copy(png, file = "../pics/inequality_health_social_en_72dpi.png", width = 640,
height = 480)
# dev.off()
```

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

```{r, lm to abline, fig.width = 8, fig.height = 6}
plot(Index_inequality.df, pch = 20, col = "red", xlim = c(3, 9), ylim = c(-1.5, 2.5), ann
= FALSE)
text(Index_inequality.df, labels = Country, pos = pos.text, cex = 0.8)
main.title <- "Income Inequality vs Index of Health and Social Problems"
x.lab <- "Income Inequality (5th Ratio)"
y.lab <- "Index of Health and Social Problems"
title(main = main.title, xlab = x.lab, ylab = y.lab)
mtext(c("Better", "Worse"), side = 2, at = c(-1.8, 2.8), las = 1)
text(x = 5, y = 1.5, labels = paste("r =", round(cor(Index_inequality.df[1],
Index_inequality.df[2]), digits = 2)))
lm.ineq <- lm(Index.HS ~ Income.inequality, data = Index_inequality.df)
# lm.ineq <- lm(Index_inequality.df[2:1])
abline(lm.ineq$coef, col = "blue")
```

GDP와 건강 및 사회문제 지수

```{r, GDP vs Index.HS, fig.width = 10, fig.height = 7.5}
Index_GDP.df <- data.21[c("GDP_WB", "Index.HS")]
text.left.2 <- which(Country %in% c("Canada", "Belgium", "Australia"))
text.right.2 <- setdiff(1:nrow(data.21), c(text.left.2))
pos.text.2 <- ifelse(1:nrow(data.21) %in% text.left.2, 2, 4)
plot(Index_GDP.df, pch = 20, col = "red", xlim = c(25000, 70000), ylim = c(-1.5, 2.5),
xaxt = "n", ann = FALSE)
axis(side = 1, at = seq(30000, 70000, by = 10000), labels = paste(3:7, "만", sep = ""))
text(Index_GDP.df, labels = Country, pos = pos.text.2, cex = 0.8)
cor.2 <- cor(Index_GDP.df["GDP_WB"], Index_GDP.df["Index.HS"])
text(x = 40000, y = 2, labels = paste("r =", round(cor.2, digits = 2)), cex = 1.2)
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"
title(main = main.title.2, xlab = x.lab.2, ylab = y.lab.2)
mtext(c("Better", "Worse"), side = 2, at = c(-1.8, 2.8), las = 1)
# dev.copy(png, file = "../pics/GDP_health_social_en_72dpi.png", width = 640, height =
480)
# dev.off()
```

한글화

국가명을 한글로 만들어 `Country.kr`로 저장하자.

```{r, Korean}
Country.kr<-c("호주", "오스트리아", "벨기에", "캐나다", "덴마크",
"핀란드", "프랑스", "독일", "그리스", "아일랜드", "이탈리아",
"일본", "네덜란드", "뉴질랜드", "노르웨이", "포르투갈",
"스페인", "스웨덴", "스위스", "영국", "미국")
```

```{r, Korean names, fig.width = 8, fig.height = 6}
# library(extrafont)
# par(family = "HCR Dotum LVT")
plot(Index_inequality.df, pch = 20, col = "red", xlim = c(3, 9), ylim = c(-1.5, 2.5), ann
= FALSE)
text(Index_inequality.df[text.right, ], labels = Country.kr[text.right], pos = 4, cex =
0.8)
text(Index_inequality.df[text.left, ], labels = Country.kr[text.left], pos = 2, cex =
0.8)
text(Index_inequality.df[text.up, ], labels = Country.kr[text.up], pos = 3, cex = 0.8)
main.title.kr <- "소득불평등과 건강 및 사회문제 지수"
x.lab.kr <- "소득불평등(소득5분위계수)"
y.lab.kr <- "건강 및 사회문제 지수"
title(main = main.title.kr, xlab = x.lab.kr, ylab = y.lab.kr)
mtext(c("좋음", "나쁨"), side = 2, at = c(-1.8, 2.8), las = 1)
```

상관계수 r = `r cor.1` 를 도표 안에 표시하고 별도의 파일로 출력하려면,

```{r fig.width=8, fig.height=6}
# par(family = "HCR Dotum LVT")
plot(Index_inequality.df, pch = 20, col = "red", xlim = c(3, 9), ylim = c(-1.5, 2.5), ann
= FALSE)
text(Index_inequality.df, labels = Country.kr, pos = pos.text, cex = 0.8)
main.title.kr <- "소득불평등과 건강 및 사회문제 지수"
```

```
x.lab.kr <- "소득불평등(소득5분위계수)"
y.lab.kr <- "건강 및 사회문제 지수"
title(main = main.title.kr, xlab = x.lab.kr, ylab = y.lab.kr)
mtext(c("좋음", "나쁨"), side = 2, at = c(-1.8, 2.8), las = 1)
text(x = 5, y = 1.5, labels = paste("r =", round(cor(Index_inequality.df[1],
Index_inequality.df[2]), digits = 2)))
dev.copy(png, file = "../pics/inequality_health_social_72dpi.png", width = 640, height =
480)
# dev.off()
```

선형회귀선을 이번에는 `lsfit`을 이용하여 삽입

```{r, simple regression line, fig.width = 8, fig.height = 6}
# par(family = "HCR Dotum LVT")
plot(Index_inequality.df, pch = 20, col = "red", xlim = c(3, 9), ylim = c(-1.5, 2.5), ann
= FALSE)
text(Index_inequality.df, labels = Country.kr, pos = pos.text, cex = 0.8)
main.title.kr <- "소득불평등과 건강 및 사회문제 지수"
x.lab.kr <- "소득불평등(소득5분위계수)"
y.lab.kr <- "건강 및 사회문제 지수"
title(main = main.title.kr, xlab = x.lab.kr, ylab = y.lab.kr)
mtext(c("좋음", "나쁨"), side = 2, at = c(-1.8, 2.8), las = 1)
text(x = 5, y = 1.5, labels = paste("r =", round(cor(Index_inequality.df[1],
Index_inequality.df[2]), digits = 2)))
lsfit.ineq <- lsfit(x = Index_inequality.df[, 1], y = Index_inequality.df[, 2])
abline(lsfit.ineq$coefficients, col = "blue")
```

GDP와의 관계

```{r, GDP vs Index, fig.width = 12, fig.height = 9}
# par(family = "HCR Dotum LVT")
Index_GDP.df <- data.21[c("GDP_WB", "Index.HS")]
text.left.2 <- which(Country %in% c("Canada", "Belgium", "Australia"))
text.right.2 <- setdiff(1:nrow(data.21), c(text.left.2))
pos.text.2 <- ifelse(1:nrow(data.21) %in% text.left.2, 2, 4)
plot(Index_GDP.df, pch = 20, col = "red", xlim = c(25000, 70000), ylim = c(-1.5, 2.5),
xaxt = "n", ann = FALSE)
axis(side = 1, at = seq(30000, 70000, by = 10000), labels = paste(3:7, "만", sep = ""))
text(Index_GDP.df, labels = Country.kr, pos = pos.text.2, cex = 0.8)
text(x = 40000, y = 2, labels = paste("r =", round(cor(Index_GDP.df[1], Index_GDP.df[2]),
digits = 2)), cex = 1.2)
main.title.2.kr <- "GDP와 건강 및 사회문제 지수"
x.lab.2.kr <- "GDP(달러)"
y.lab.2.kr <- "건강 및 사회문제 지수"
title(main = main.title.2.kr, xlab = x.lab.2.kr, ylab = y.lab.2.kr)
mtext(c("좋음", "나쁨"), side = 2, at = c(-1.8, 2.8), las = 1)
dev.copy(png, file = "../pics/GDP_health_social_72dpi.png", width = 640, height = 480)
# dev.off()
```

미국의 경우

`xlsx` 패키지를 이용하여 자료를 읽어들인다.

```{r, data_US}
data.usa <- read.xlsx("../data/USA-inequality.xls", 1, stringsAsFactors = FALSE)
str(data.usa)
```

당장 필요한 변수들만 모아서 data frame으로 재구성한다. 변수명 설정에 유의한다.

```{r, variables}
data.usa.1 <- data.frame(Gini = data.usa$Income.Inequality,  HS.index =
data.usa$Index.of.health...social.problems)
str(data.usa.1)
Gini <- data.usa.1$Gini
State <- data.usa$State
Abb <- data.usa$State.Abbrev
options(digits = 3)
kable(data.frame(State = State, State.Abb = Abb, data.usa.1))
```

주별 Gini계수를 `barplot()`으로 비교해 보자. 전부 0.4는 넘고 0.5는 넘지 않기 때문에 차이를 살피기 위해서 y축의 범위(`ylim = `)를 조정하였다. 이때 `xpd = FALSE`가 어떤 역할을 하는지 잘 알아두자.

```{r, barplot of Gini, fig.width = 12, fig.height = 8}
par(mai = c(2.0, 0.8, 0.8, 0.4) + 0.2)
o.Gini <- order(Gini)
b.Gini <- barplot(Gini[o.Gini], names.arg = State[o.Gini], col = rev(rainbow(50, start =
0, end = 4/6)), ylim = c(0.3, 0.52), xpd = FALSE, las = 2)
text(x = b.Gini[c(1, 25, 50)], y = Gini[o.Gini][c(1, 25, 50)] + 0.01, labels =
format(Gini[o.Gini][c(1, 25, 50)], digits = 3))
title(main = "Gini Coefficients of United States")
```

간단한 산점도를 그리고, 추가 작업을 생각한다.

```{r, first plot,fig.width = 12, fig.height = 9}
plot(data.usa.1)
```

x-축과 y-축의 범위를 설정하고, `pch = 20`으로 다시 그린다.

```{r, pch-xlim-ylim, fig.width = 12, fig.height = 9}
plot(data.usa.1, pch = 20, xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0))
```

각 주의 약칭을 새겨넣는다.

```{r, Abb, fig.width = 12, fig.height = 9}
plot(data.usa.1, pch = 20, xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0))
text(data.usa.1, labels = Abb, pos = 4)
```

겹쳐보이는 주의 약칭들로부터 인덱스를 추출한다.

```{r, index extraction}
which(Abb %in% c("VT", "ME", "NE", "WA", "VA", "HI", "RI", "SC", "AR", "NC", "GA", "KY"))
```

점 왼쪽에 약칭을 넣을 주들의 인덱스를 저장한다. 나머지 인덱스는 오른쪽에 넣을 것으로 따로 저장한다.

```{r, left to the point}
text.left.us <- which(Abb %in% c("VT", "ME", "NE", "WA", "VA", "HI", "RI", "SC", "AR",
"NC", "GA", "KY"))
text.right.us <- setdiff(1:nrow(data.usa.1), text.left.us)
pos.text.us <- ifelse(1:nrow(data.usa.1) %in% text.left.us, 2, 4)
```

왼쪽, 오른쪽 위치를 조정한 주 약칭을 다시 넣는다.

```{r, right or left, fig.width = 12, fig.height = 9}
plot(data.usa.1, pch = 20, col = "red", xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0))
text(data.usa.1, labels = Abb, pos = pos.text.us)
```

점 아래에 약칭을 넣을 주들의 인덱스를 찾는다. 왼쪽 인덱스, 오른쪽 인덱스에서 조정한다.

```{r, Abb under dots}
text.down.us <- which(Abb %in% c("WA", "AR", "GA", "MN"))
which(text.left.us %in% text.down.us)
text.left.us <- setdiff(text.left.us, text.down.us)
text.right.us <- setdiff(text.right.us, text.down.us)
pos.text.us <- ifelse(1:nrow(data.usa.1) %in% text.down.us, 1, ifelse(1:nrow(data.usa.1)
```

```
%in% text.left.us, 2, 4))
```

약칭 위치를 아래로 조정한 산점도를 다시 그린다.

```{r, point under, fig.width = 12, fig.height = 9}
plot(data.usa.1, pch = 20, col = "red", xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0))
text(data.usa.1, labels = Abb, pos = pos.text.us)
```

상관계수를 추가한다.

```{r, correlation, fig.width = 12, fig.height = 9}
plot(data.usa.1, pch = 20, col = "red", xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0))
text(data.usa.1, labels = Abb, pos = pos.text.us)
cor.us <- cor(data.usa.1$HS.index, data.usa.1$Gini)
text(x = 0.42, y = 1.5, labels = paste("r =", round(cor.us, digits = 2)), col = "red",
cex = 1.2)
```

단순회귀선을 추가한다.

```{r fig.width=12, fig.height=9}
plot(data.usa.1, pch = 20, col = "red", xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0))
text(data.usa.1, labels = Abb, pos = pos.text.us)
text(x = 0.42, y = 1.5, labels = paste("r =", round(cor.us, digits = 2)), col = "red",
cex = 1.2)
# lm.ineq.us <- lm(HS.index ~ Gini, data = data.usa.1)
lm.ineq.us <- lm(data.usa.1[2:1])
abline(lm.ineq.us$coef, col = "blue")
# abline(lm(HS.index ~ Gini, data = data.usa.1)$coef)
```

주제목을 추가하고, `xlab`, `ylab`을 수정한다. 수직축의 의미를 명확히 한다.

```{r, labs and title, fig.width = 12, fig.height = 9}
plot(data.usa.1, pch = 20, col = "red", xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0), ann =
FALSE)
text(data.usa.1, labels = Abb, pos = pos.text.us)
text(x = 0.42, y = 1.5, labels = paste("r =", round(cor.us, digits = 2)), col = "red",
cex = 1.2)
abline(lm.ineq.us$coef, col = "blue")
mtext(c("Better", "Worse"), side = 2, at = c(-1.8, 2.3), las = 1)
main.title.us <- "Income Inequality vs Health and Social Index (USA)"
x.lab.us <- "Gini Coefficients"
y.lab.us <- "Index of Health and Social Problems"
title(main = main.title.us, xlab = x.lab.us, ylab = y.lab.us)
```

<!--
```{r, save}
save.image(file = "Inequality_Index_HS.RData")
```
-->