Income Inequality vs Index of Health and Social Problems : US

coop711

2016-10-07

Data Reading

```
library(knitr)
# library(xlsx)
# data.usa <- read.xlsx("../data/USA-inequality.xls", 1, stringsAsFactors = FALSE)
load("./Inequality_Index_HS_US.RData")
str(data.usa)
```

```
## 'data.frame':    50 obs. of  20 variables:
##  $ State                           : chr  "Alabama" "Alaska" "Arizona" "Arkansas"
 ...
##  $ State.Abbrev                    : chr  "AL" "AK" "AZ" "AR" ...
##  $ Income.Inequality               : num  0.475 0.402 0.45 0.458 0.475 ...
##  $ Trust                           : num  23 NA 47 29 43 46 49 NA 37 38 ...
##  $ Life.expectancy                 : num  74.6 76.7 77.5 75.1 78.3 ...
##  $ Infant.mortality                : num  9.1 5.5 6.4 8.3 5.5 ...
##  $ Obesity                         : num  32 30 28.5 31 31 21.5 26.5 27 27.5 30.5
 ...
##  $ Mental.health                   : num  3.3 2.8 2.2 3.2 3.3 ...
##  $ Maths.and.literacy.scores       : num  258 268 263 262 259 ...
##  $ Teenage.births                  : num  62.9 42.4 69.1 68.5 48.5 ...
##  $ Homicides                       : num  78.9 85.6 80.4 56.1 60.5 ...
##  $ Imprisonment                    : num  509 413 507 415 478 357 372 429 447 502
 ...
##  $ Index.of.health...social.problems: num  1.385 0.137 0.212 0.948 0.327 ...
##  $ Overweight.children             : num  35 31 30 33 30 22 27 35 32 32 ...
##  $ Child.wellbeing                 : num  8.5 4.4 4.9 9.3 -3.4 ...
##  $ Women.s.status                  : num  -0.932 0.74 -0.147 -1.318 0.969 ...
##  $ Juvenile.homicides              : num  12 8 7 6 10 4 4 0 NA 8 ...
##  $ High.school.drop.outs           : num  24.7 11.7 19 24.7 23.2 ...
##  $ Child.mental.illness            : num  11.5 8.2 8.7 11.8 7.5 ...
##  $ Pugnacity                       : num  41.8 NA 36.3 38.4 37.7 ...
```

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

```
data.usa.1 <- data.frame(Gini = data.usa$Income.Inequality,  HS.index = data.usa$Inde
x.of.health...social.problems)
str(data.usa.1)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ Gini    : num  0.475 0.402 0.45 0.458 0.475 ...
##  $ HS.index: num  1.385 0.137 0.212 0.948 0.327 ...
```

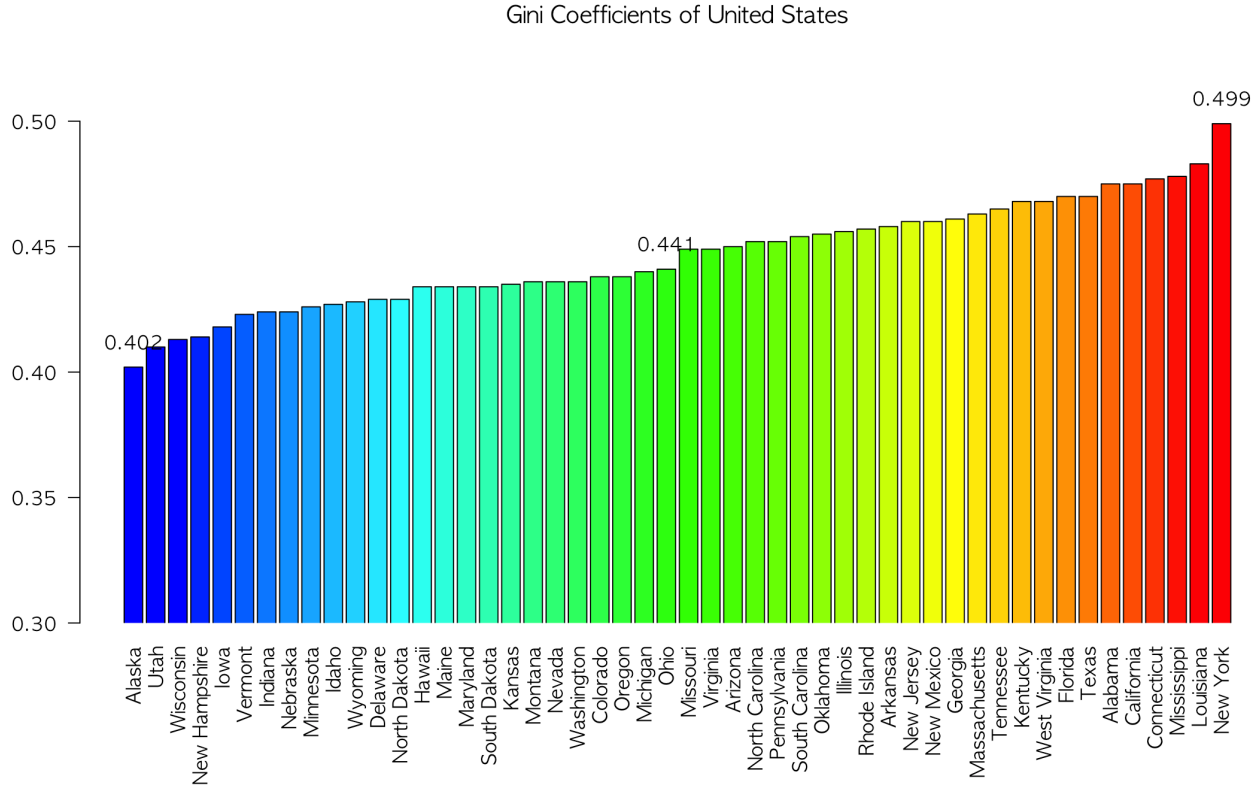
```
Gini <- data.usa.1$Gini
State <- data.usa$State
Abb <- data.usa$State.Abbrev
options(digits = 3)
kable(data.frame(State = State, State.Abb = Abb, data.usa.1))
```

State	State.Abb	Gini	HS.index
Alabama	AL	0.475	1.385
Alaska	AK	0.402	0.137
Arizona	AZ	0.450	0.212
Arkansas	AR	0.458	0.948
California	CA	0.475	0.327
Colorado	CO	0.438	-0.507
Connecticut	CT	0.477	-0.660
Delaware	DE	0.429	0.133
Florida	FL	0.470	0.360
Georgia	GA	0.461	0.896
Hawaii	HI	0.434	-0.388
Idaho	ID	0.427	-0.429
Illinois	IL	0.456	0.206
Indiana	IN	0.424	0.370
Iowa	IA	0.418	-0.895
Kansas	KS	0.435	-0.442
Kentucky	KY	0.468	0.874
Louisiana	LA	0.483	1.595
Maine	ME	0.434	-0.769
Maryland	MD	0.434	0.187
Massachusetts	MA	0.463	-0.959
Michigan	MI	0.440	0.349
Minnesota	MN	0.426	-1.216
Mississippi	MS	0.478	1.692
Missouri	MO	0.449	0.392
Montana	MT	0.436	-0.906
Nebraska	NE	0.424	-0.583
Nevada	NV	0.436	0.803

State	State.Abb	Gini	HS.index
New Hampshire	NH	0.414	-1.242
New Jersey	NJ	0.460	-0.402
New Mexico	NM	0.460	0.564
New York	NY	0.499	-0.179
North Carolina	NC	0.452	0.494
North Dakota	ND	0.429	-1.145
Ohio	OH	0.441	0.058
Oklahoma	OK	0.455	0.494
Oregon	OR	0.438	-0.346
Pennsylvania	PA	0.452	-0.015
Rhode Island	RI	0.457	-0.389
South Carolina	SC	0.454	0.899
South Dakota	SD	0.434	-0.759
Tennessee	TN	0.465	0.788
Texas	TX	0.470	0.930
Utah	UT	0.410	-0.709
Vermont	VT	0.423	-1.183
Virginia	VA	0.449	-0.055
Washington	WA	0.436	-0.516
West Virginia	WV	0.468	0.482
Wisconsin	WI	0.413	-0.473
Wyoming	WY	0.428	-0.551

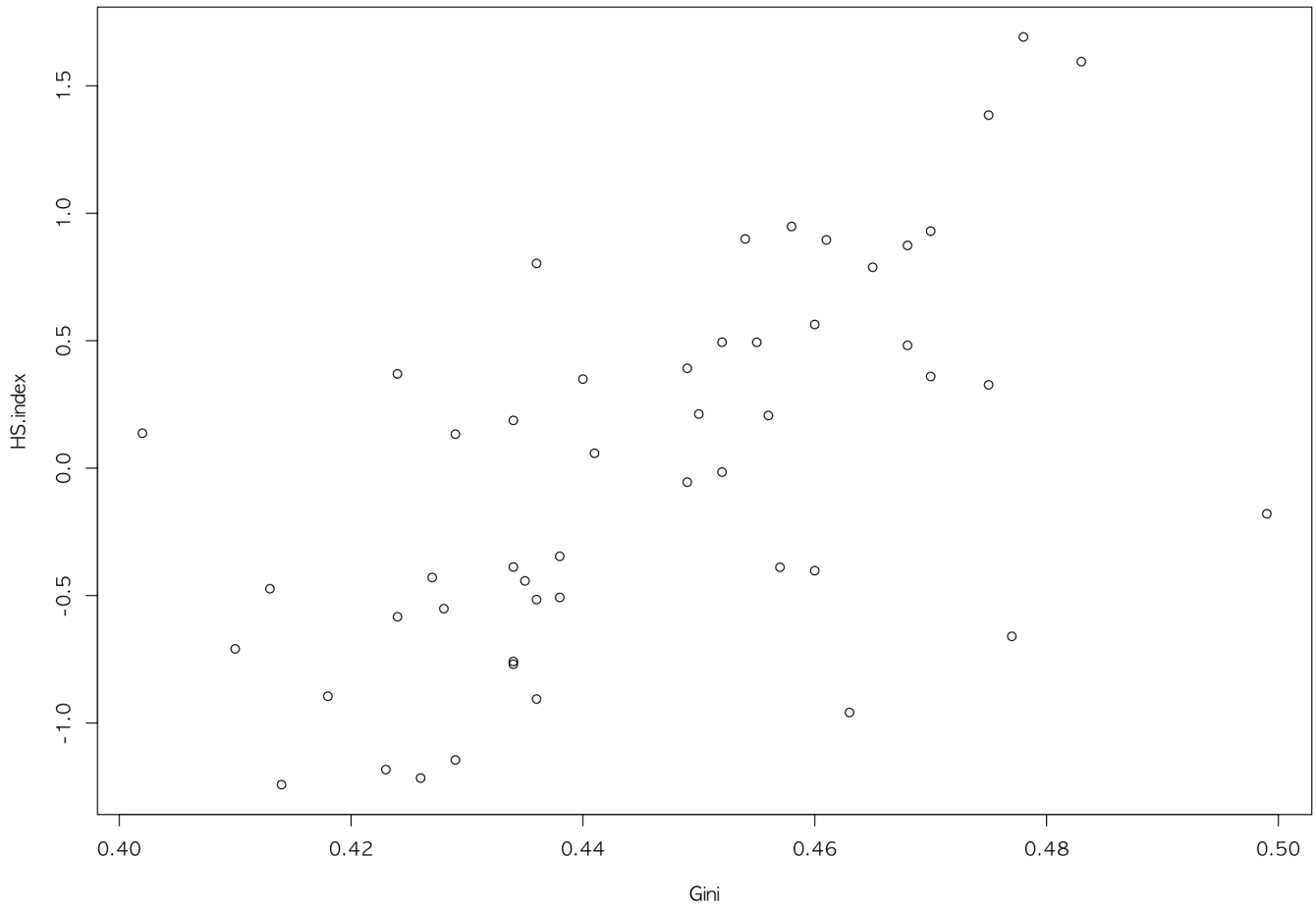
주별 Gini계수를 `barplot()` 으로 비교해 보자. 전부 0.4는 넘고 0.5는 넘지 않기 때문에 차이를 살피기 위해서 y축의 범위(`ylim =`)를 조정하였다. 이때 `xpd = FALSE` 가 어떤 역할을 하는지 잘 알아두자.

```
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, star
t = 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(G
ni[o.Gini][c(1, 25, 50)], digits = 3))
title(main = "Gini Coefficients of United States")
```



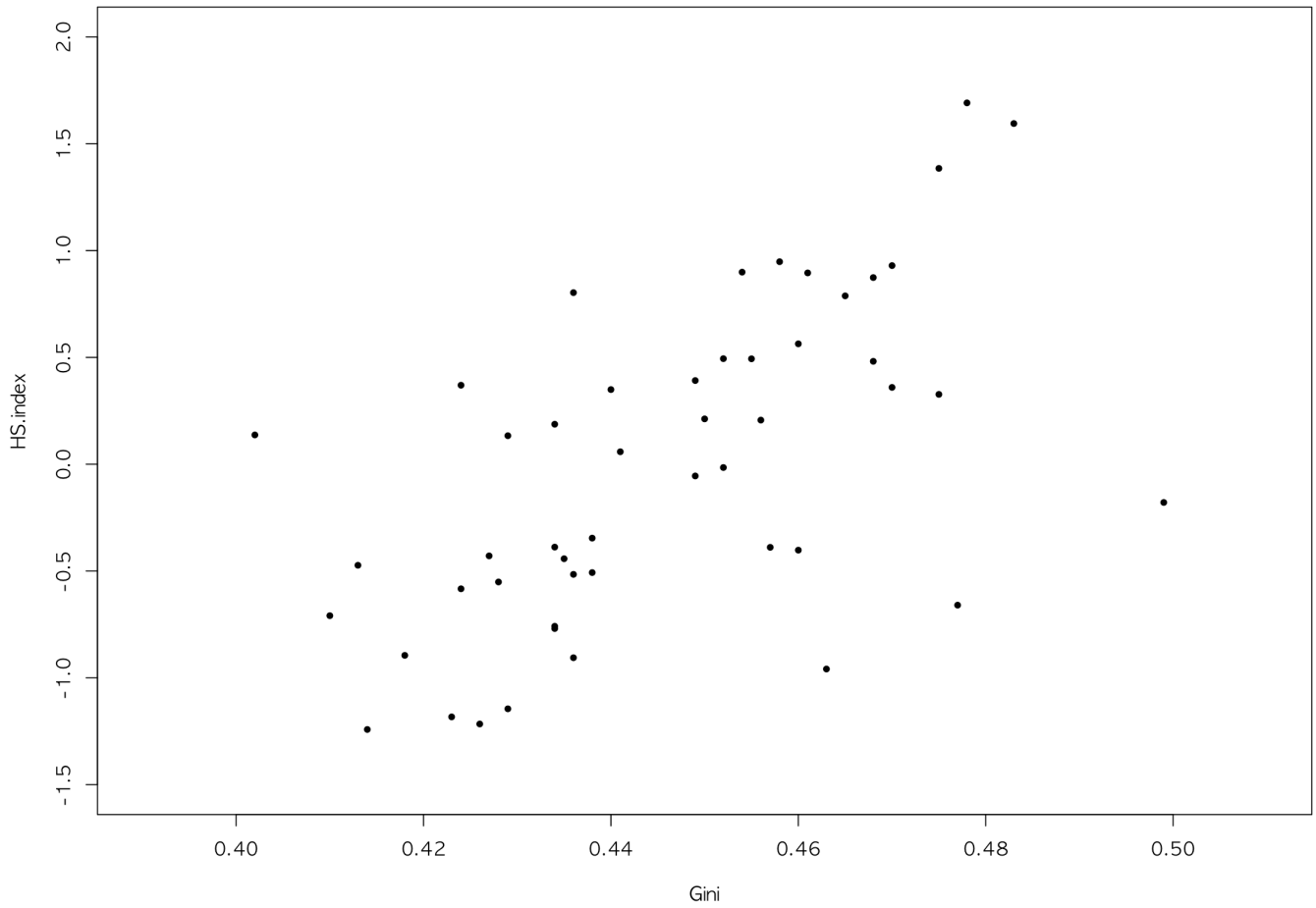
간단한 산점도를 그리고, 추가 작업을 생각한다.

```
plot(data.usa.1)
```



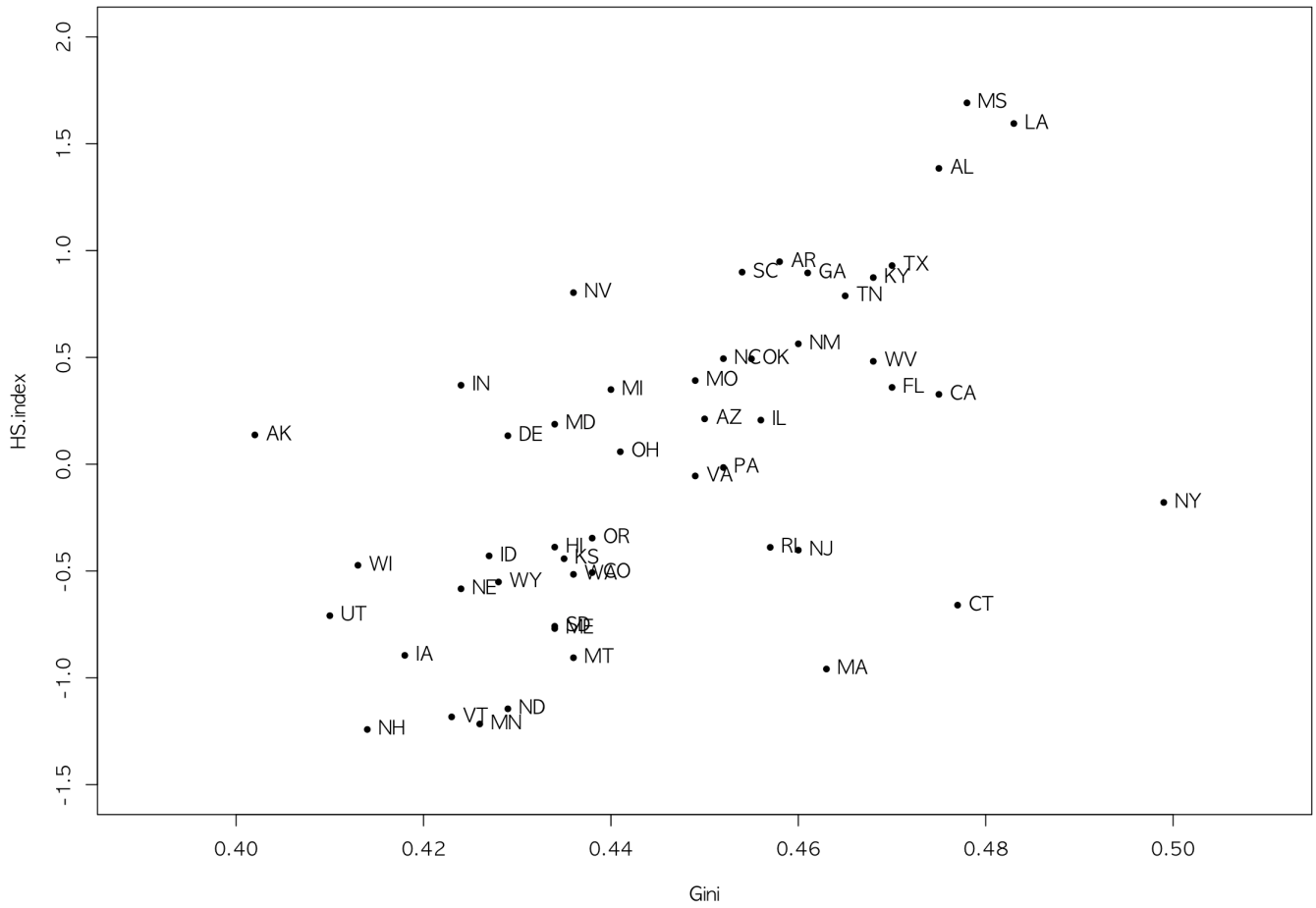
x-축과 y-축의 범위를 설정하고, `pch = 20` 으로 다시 그린다.

```
plot(data.usa.1, pch = 20, xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0))
```



각 주의 약칭을 새겨넣는다.

```
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.

```
which(Abb %in% c("VT", "ME", "NE", "WA", "VA", "HI", "RI", "SC", "AR", "NC", "GA", "K
Y"))
```

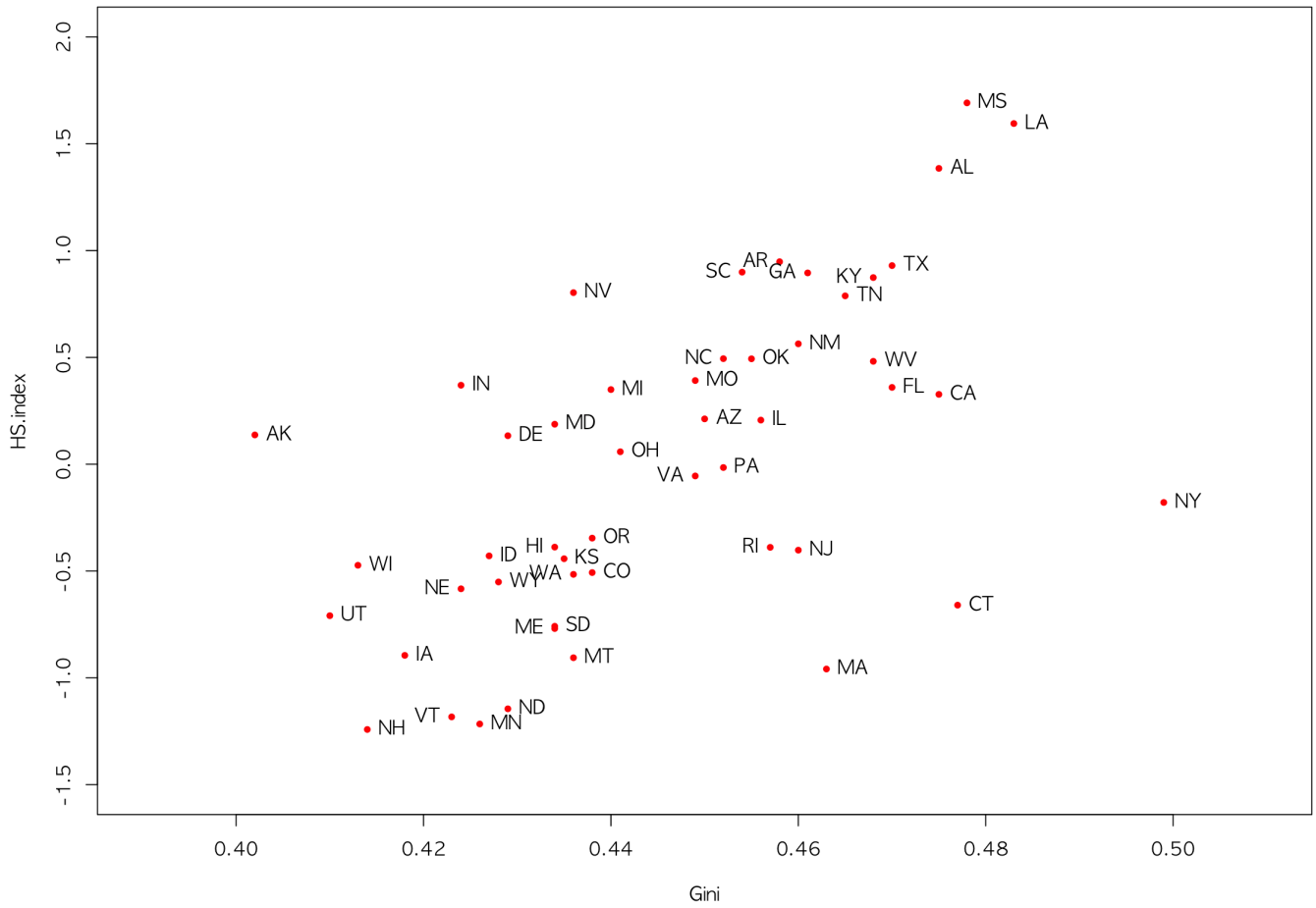
```
##  [1]  4 10 11 17 19 27 33 39 40 45 46 47
```

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

```
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.

```
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.

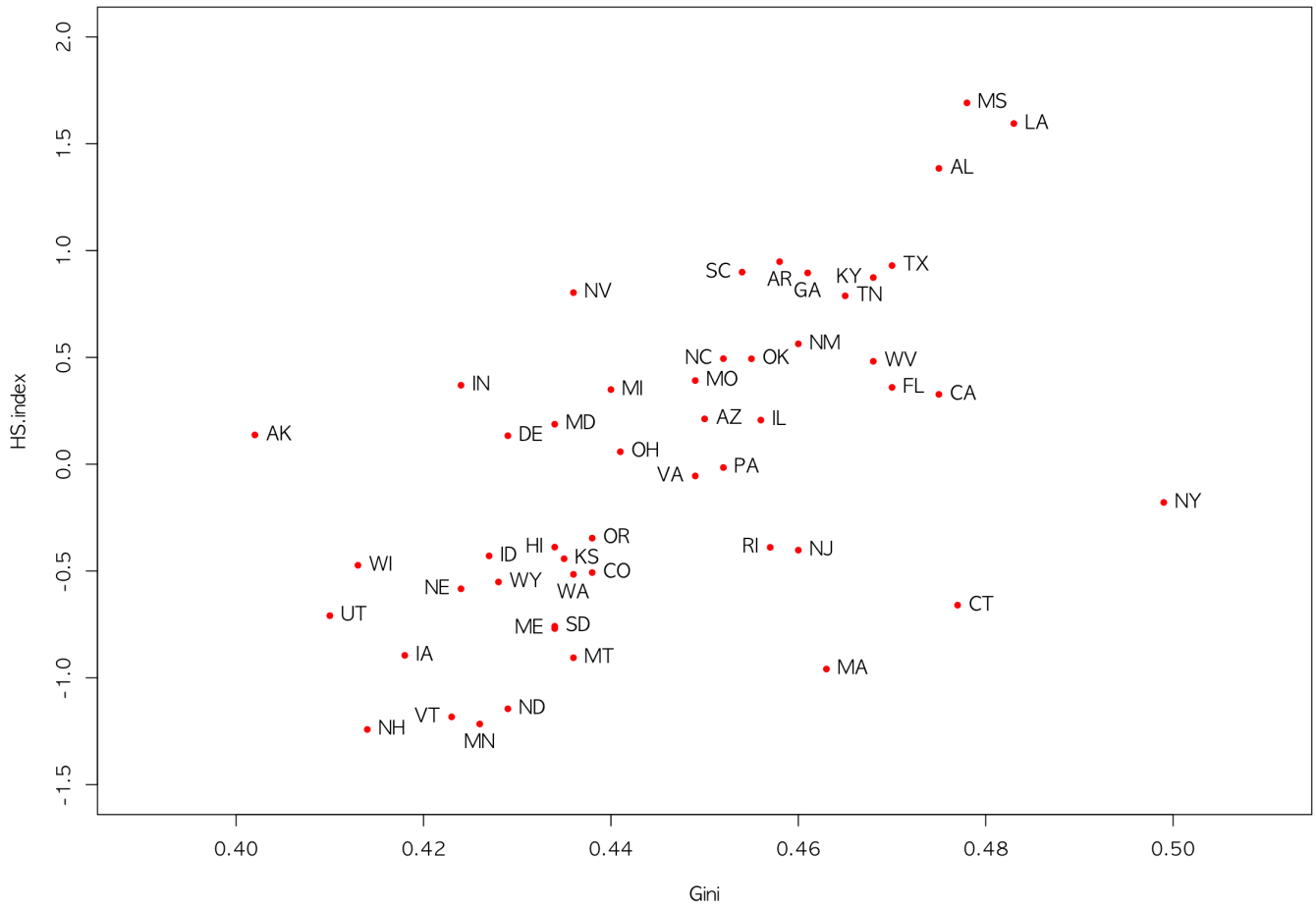
```
text.down.us <- which(Abb %in% c("WA", "AR", "GA", "MN"))
which(text.left.us %in% text.down.us)
```

```
## [1]  1  2 12
```

```
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.us
a.1) %in% text.left.us, 2, 4))
```

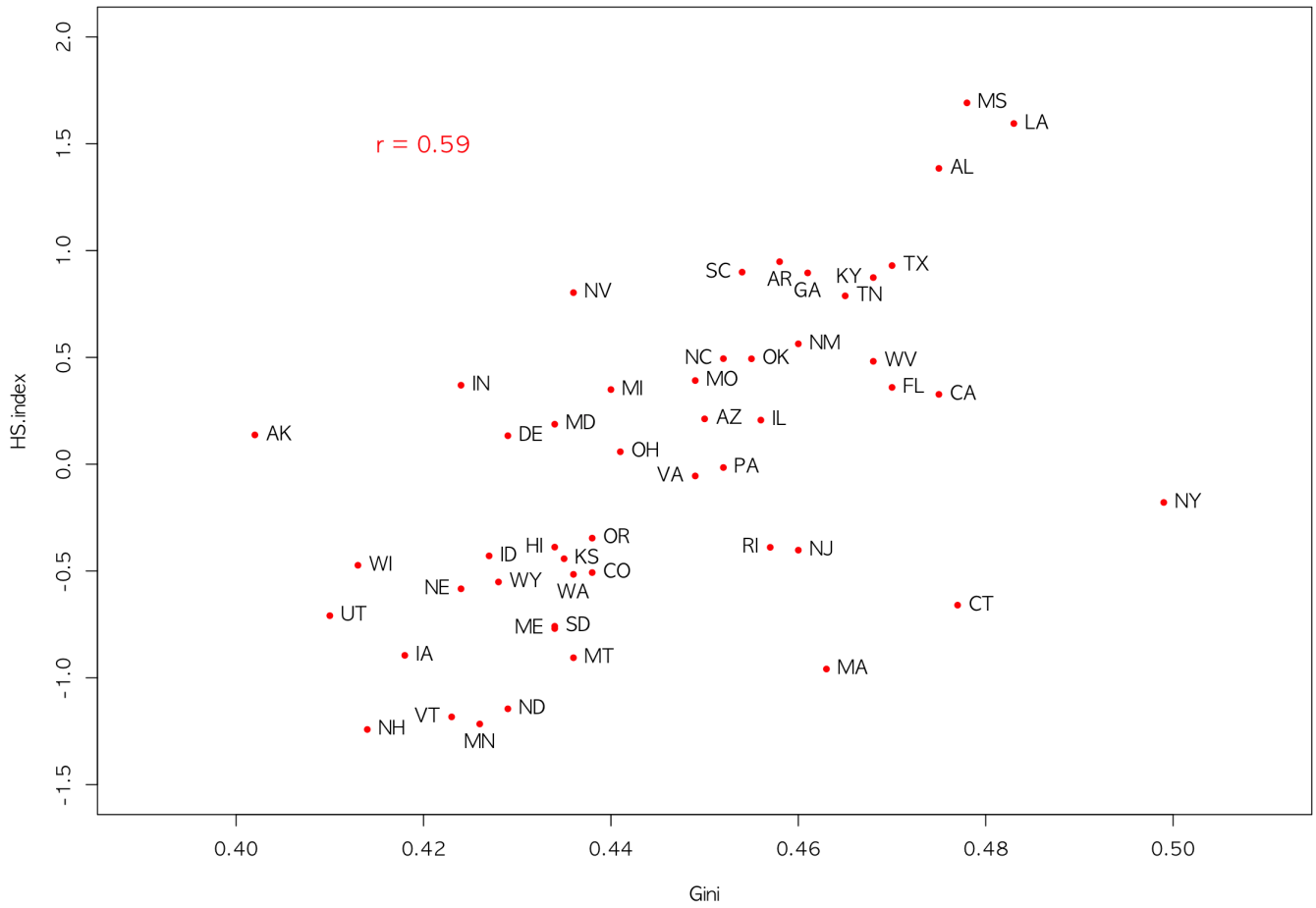
약칭 위치를 아래로 조정한 산점도를 다시 그린다.

```
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)
```



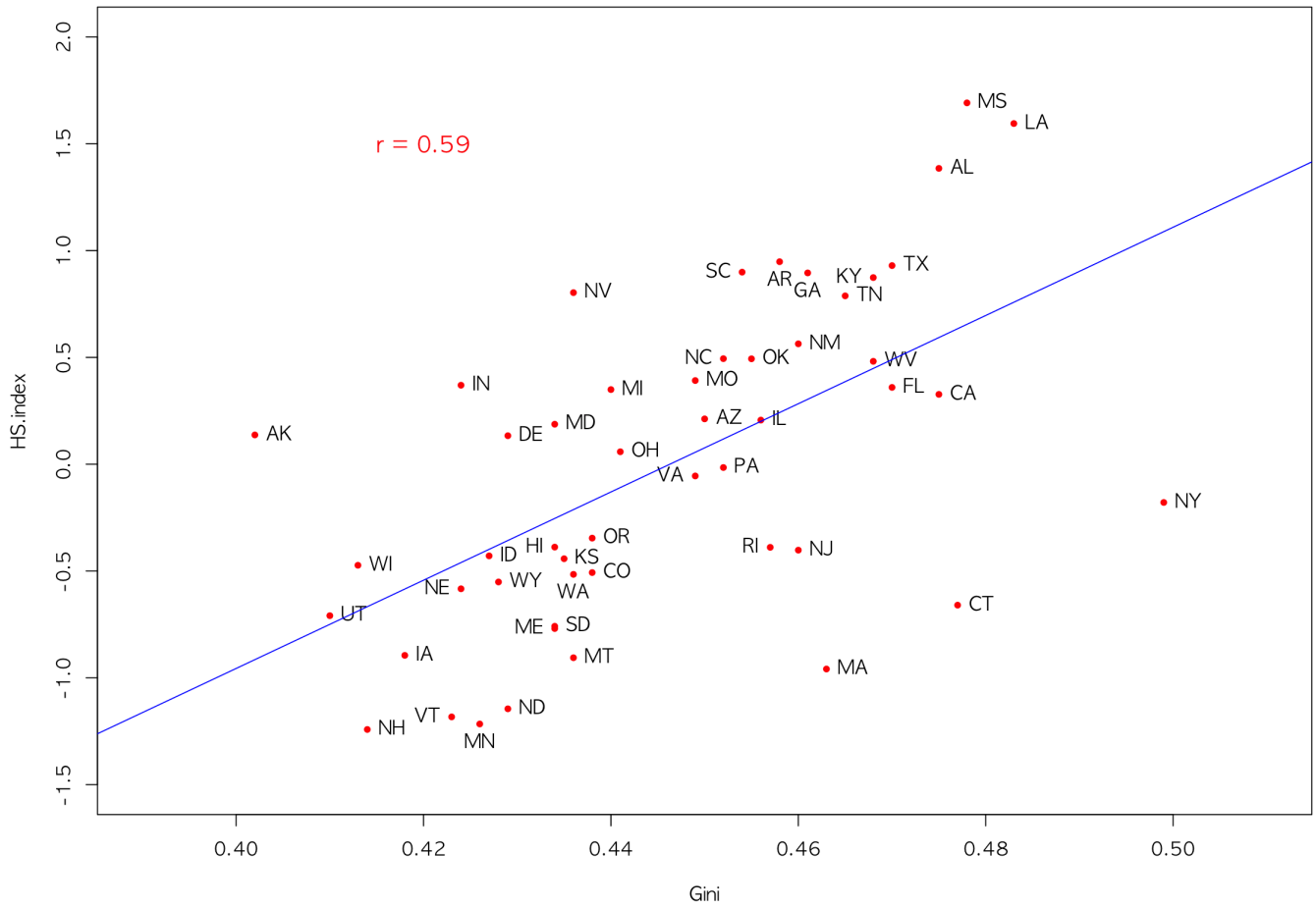
상관계수를 추가한다.

```
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.

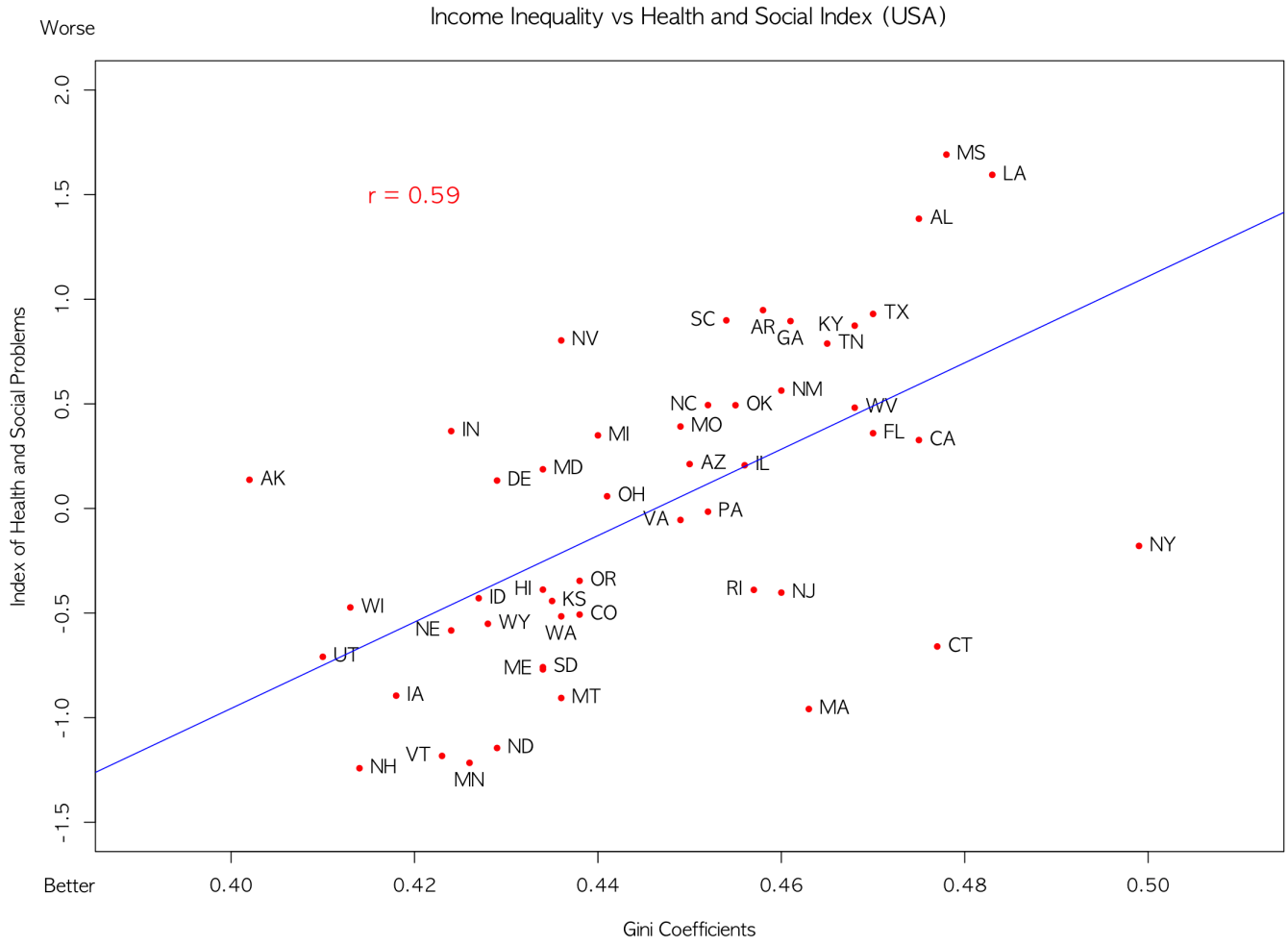
```
plot(data.usa.1, pch = 20, col = "red", xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0))
text(data.usa.1, labels = Abb, pos = pos.text.us)
text(x = 0.42, y = 1.5, labels = paste("r =", round(cor.us, digits = 2)), col =
"red", cex = 1.2)
# lm.ineq.us <- lm(HS.index ~ Gini, data = data.usa.1)
lm.ineq.us <- 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.

```
plot(data.usa.1, pch = 20, col = "red", xlim = c(0.39, 0.51), ylim = c(-1.5, 2.0), an
n = 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)
```


Income Inequality vs Health and Social Index (USA)
Worse
Better
Index of Health and Social Problems
Gini Coefficients
r = 0.59
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.40
0.42
0.44
0.46
0.48
0.50
MS
LA
AL
SC
AR
GA
KY
TN
TX
NV
NC
OK
NM
WV
MO
IN
MI
FL
CA
AZ
IL
AK
DE
MD
OH
VA
PA
NY
OR
HI
ID
KS
CO
RI
NJ
WI
NE
WY
WA
UT
CT
ME
SD
IA
MT
MA
ND
VT
NH
MN