Happiness, Inequality, and Mental Health Policy Clustering and Regression

Bingyao Zou, Aurora Deng, Andrew Golden

Load in data

Collect the data we prepared in Python:

```r
link='https://github.com/auroraD-11/542Teamdata/raw/master/TEAMdata.RDS'
myFile=url(link)

#Reading in data:
fromPy=readRDS(file = myFile)

#Reset indexes to R format:
row.names(fromPy)=NULL
```

Clustering Analysis

Preparing data:

```r
#Subset the data frame:
dfClus=fromPy[,c(2,3)]

#Rename subset indexes:
row.names(dfClus)=fromPy$Country

#Set random seed:
set.seed(999)

#Decide distance method and compute distance matrix:
library(cluster)
dfClus_D=cluster::daisy(x=dfClus,metric="gower")
```

Density-based clustering:

We chose a density-based (DBSCAN) method for our research. Our research sample is based on the GINI Index and scores of happiness in all countries, as well as comparing those that do and do not publish their mental-health-related public policy.

Why we choose Density-based clustering?

The reason why we chose DBSCAN as the clustering method is to gather the country samples that have a high density of observations and mark out the outlier points, which are less dense with observations. Through dropping outlier points and exploring the clusters, we could determine the correlation between our variables of GINI Index, happiness score, and publication status of the mental-health policy on the country-level.

Our setting

At the beginning of the research, we set up the similarity distance equal to 0.05, with the minimal points equal to 4 to explore the correlations among three variables on the country-level. We got 5 clusters based on different density-levels and 26 outliers with the lowest density from 121 sample countries in total. We compared the GINI coefficient variables with happiness scores and with mental health policy publication status separately.

For both of the comparisons, we placed the GINI Index on the horizontal axis to facilitate a correlation between the happiness score and the publication status of the mental health policy. The GINI Index ranges from 0 to 100, with 0 representing perfect equality and 100 representing the perfect inequality.

```
library(dbscan)
# We set up minimal points equal to 4
minNeighs=4

#We set up the similarity distance equal to 0.05
distance=0.05
res.db = dbscan::dbscan(dfClus_D, eps=distance,
                     minPts=minNeighs)
```

```
#Identify the number of clusters and outliers, save result:
res.db
```

```
## DBSCAN clustering for 121 objects.
## Parameters: eps = 0.05, minPts = 4
## The clustering contains 5 cluster(s) and 26 noise points.
##
##  0  1  2  3  4  5
## 26 21  4 54 12  4
##
## Available fields: cluster, eps, minPts
```

```
fromPy$db=as.factor(res.db$cluster)
```

Visualize the clustering result

Based on Gini Index and happiness scores for all countries

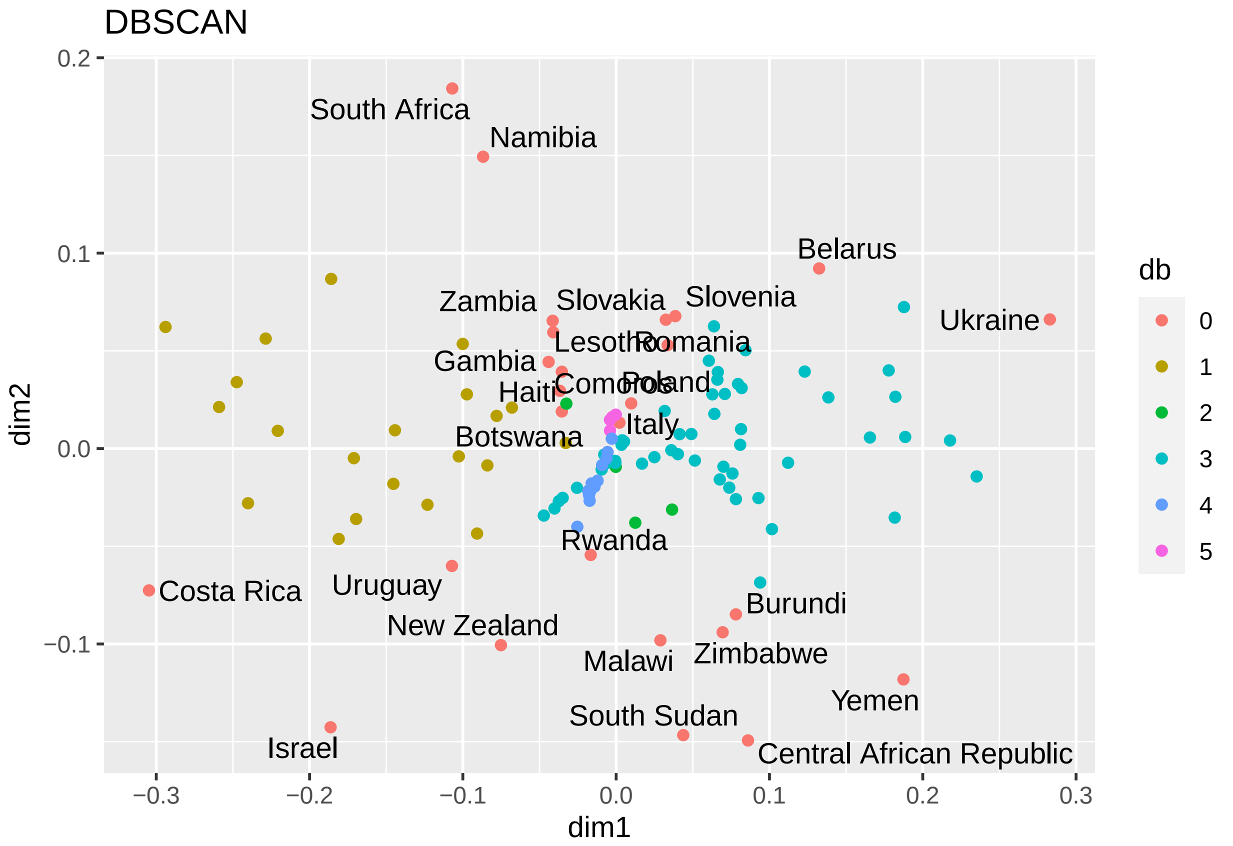
```
library(ggplot2)
library(ggrepel)

#Prepare a bidimensional map:
projectedData = cmdscale(dfClus_D, k=4)
fromPy$dim1 = projectedData[,2]
fromPy$dim2 = projectedData[,3]

#Create the fundamental map
base= ggplot(data=fromPy,
             aes(x=dim1, y=dim2,
                 label=Country))

#Plot results from DBSCAN
dbPlot= base + labs(title = "DBSCAN") + geom_point(aes(color=db),
                                               show.legend = T)
#Annotating:
LABEL=ifelse(fromPy$db==0,fromPy$Country,"") # annotating outliers

dbPlot + geom_text_repel(aes(label=LABEL))
```



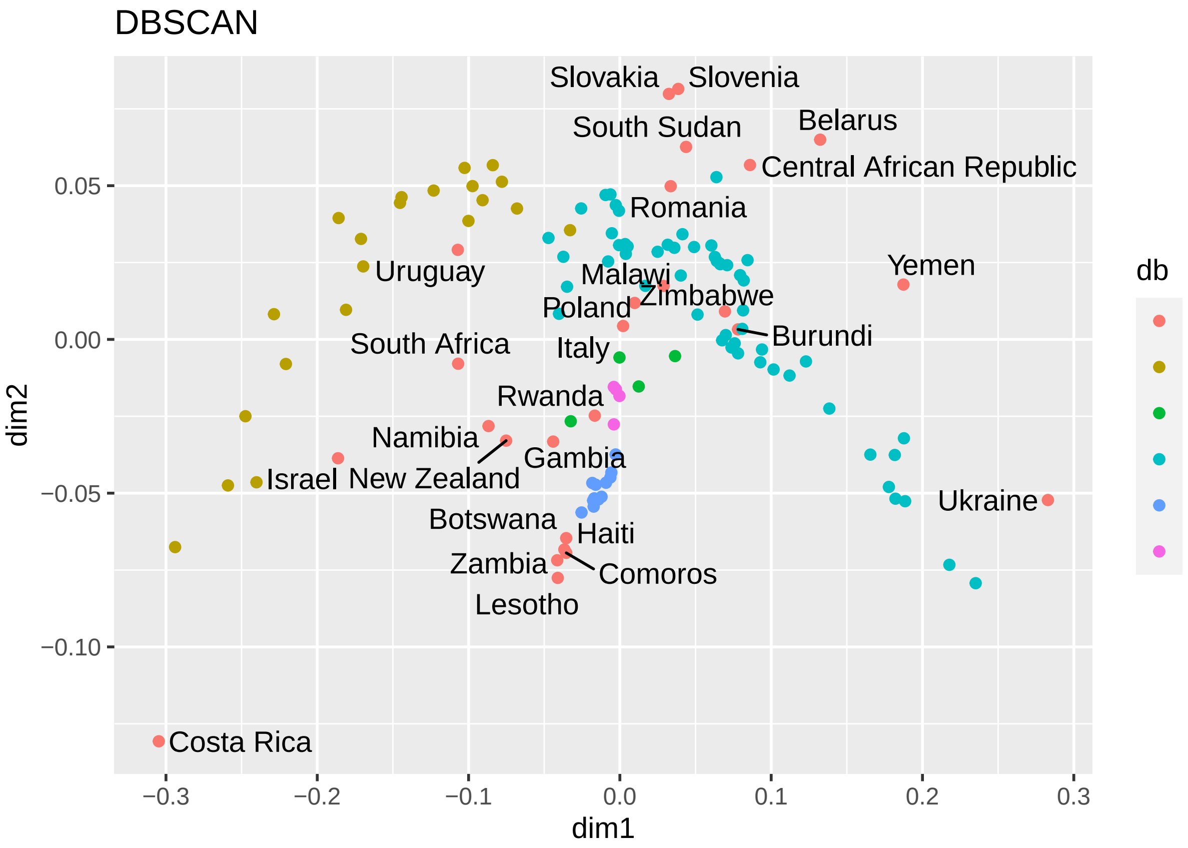
Based on Gini Index and the publication status of the mental-health public policy by country governments

```
#Prepare a bidimensional map:
projectedData = cmdscale(dfClus_D, k=4)
fromPy$dim1 = projectedData[,2]
fromPy$dim2 = projectedData[,4]

#Create the fundamental map
base1= ggplot(data=fromPy,
             aes(x=dim1, y=dim2,
                 label=Country))

#Plot results from DBSCAN
dbPlot= base1 + labs(title = "DBSCAN") + geom_point(aes(color=db),
                                               show.legend = T)
#Annotating:
LABEL1=ifelse(fromPy$db==0,fromPy$Country,"") # annotating outliers
```

```
dbPlot + geom_text_repel(aes(label=LABEL1))
```



Clustering findings:

1. In two DBSCAN visuals, the two higher-density clusters—Groups 4 and 5—were located in the central area of both graphs.
2. For the GINI Index/Happiness scores graph, we could see the highest density groups are the 0.5-GINI Index countries that were among the average happiness score overall.
3. In the GINI Index/Mental Health Publication Status graph, these two higher-density clusters within the same GINI Index are located below the zero of the vertical axis, which means in these two clusters, those 0.5 Gini coefficient countries, their governments did not publish their mental-health public policy.
4. The relatively lower-density groups—Groups 1, 2, and 3—are located from the higher GINI Index countries (more economic inequality) to the lower Gini Index countries (less economic inequality) separately, and all of those three clusters are situated within a similar range on the vertical axis.

Clustering Conclusion:

Through utilizing the DBSCAN clustering method, we did not find any substantial evidence to prove that there is a direct relationship between the GINI Index, happiness scores, and the government's mental health policy publication status.

Further Clustering Analysis

In some situations, the international comparison of GINI Indices is not straight forward, because different countries might use different equivalence scales. In further research, we might narrow down our locational variables into the region-level, state-level, or city-levels that are under the same equivalence scales and determine the relationship between income inequality and happiness.

Regression Analysis

Prepare data

The original mental policy data lists the specific mental-health related public policies of each country. We define a new binary variable based on the summary so that we can see directly whether a country's government publishes their mental policy or not

```r
# Turn policy plan variable into binary variable, 1 indicates the country has at least one mental polic
fromPy$PolicyPlan = factor(ifelse(fromPy$PolicyPlan == "Yes", 1, 0))
```

Regression Hypothesis and Reasoning

To gain insight into the relation between country-level happiness, economic inequality, and mental health, our team conducted a regression analysis using these variables. This linear regression model uses the happiness score as its dependent variable, the GINI Index as an independent variable, and the publication of mental health policy as a 0-1 indicator variable. The model analyzes these data from 121 countries.

We expected to see a relationship between lower levels of inequality and happiness, as the holding of more wealth in a smaller number of people in countries would deprive the majority of the populace the money and resources they need. However, this could be affected by some countries with low GINI Indices being lower-wealth to begin with—for example former Soviet bloc states—confounding the hypothesis. Additionally, we expected publication of mental health policy to cause an increase in happiness, as the publication would indicate a sense of attention to public health and wellbeing, as well as being an indicator of governmental transparency and openness.

First Hypothesis:

```r
#State hypothese:
hypo=formula(scoreofhappiness~ GINI + PolicyPlan)

#Save columns needed:
colsNeeded=c('scoreofhappiness', 'GINI','PolicyPlan')

#Create subset:
DataRegGauss=fromPy[,colsNeeded]

#Rename indexes by country:
row.names(DataRegGauss)=fromPy$Country

#Compute regression moedels:
gauss=glm(hypo,data = DataRegGauss,family = 'gaussian')

#See result:
summary(gauss)
```

```
##
```

```
## Call:
## glm(formula = hypo, family = "gaussian", data = DataRegGauss)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.17255  -0.84255  -0.07923   0.73887   2.18745
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.30847    0.45743  15.977  < 2e-16 ***
## GINI        -0.05406    0.01088  -4.971 2.29e-06 ***
## PolicyPlan1  0.29281    0.24009   1.220    0.225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.096321)
##
##     Null deviance: 156.97  on 120  degrees of freedom
## Residual deviance: 129.37  on 118  degrees of freedom
## AIC: 359.47
##
## Number of Fisher Scoring iterations: 2
```

Second Hypothesis:

```r
gauss2=lm(hypo,data = DataRegGauss)
summary(gauss2)
```

```
##
## Call:
## lm(formula = hypo, data = DataRegGauss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17255 -0.84255 -0.07923  0.73887  2.18745
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.30847    0.45743  15.977  < 2e-16 ***
## GINI        -0.05406    0.01088  -4.971 2.29e-06 ***
## PolicyPlan1  0.29281    0.24009   1.220    0.225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.047 on 118 degrees of freedom
## Multiple R-squared:  0.1759, Adjusted R-squared:  0.1619
## F-statistic: 12.59 on 2 and 118 DF,  p-value: 1.105e-05
```

Search for better model

```r
with(summary(gauss), 1 - deviance/null.deviance)
```

```
## [1] 0.1758792
```

```
anova(gauss,test="Chisq")
```

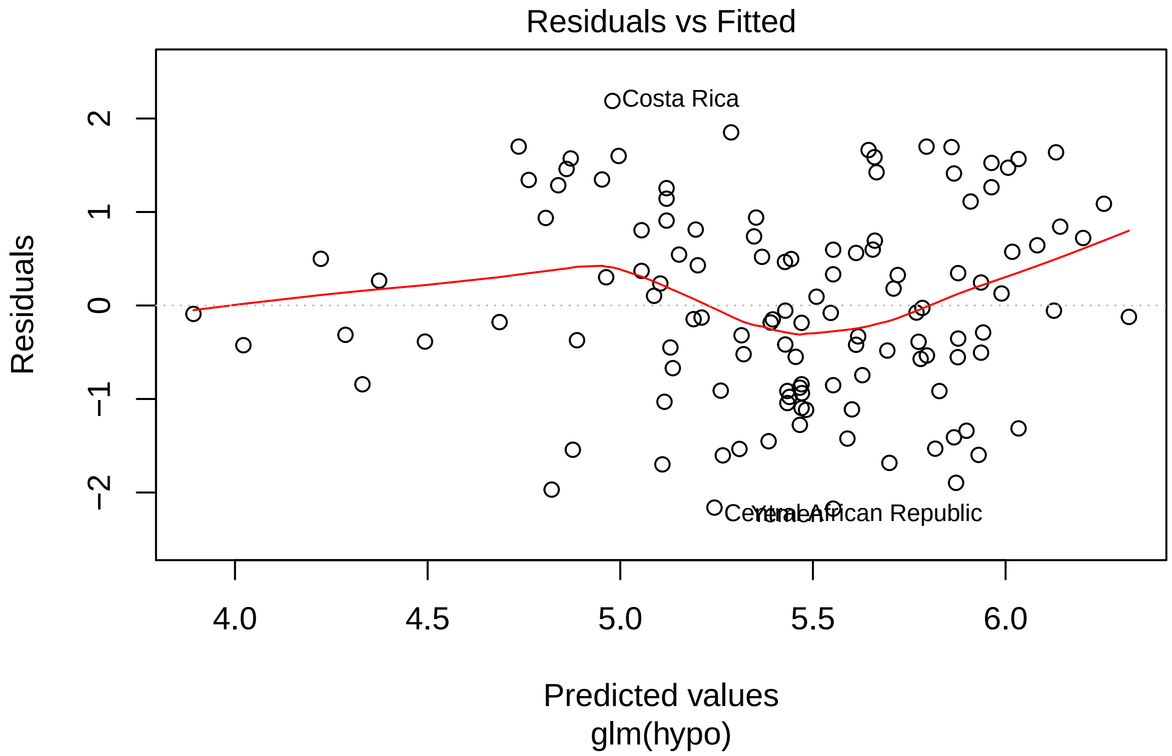
```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: scoreofhappiness
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         120     156.97
## GINI        1  25.9779       119     131.00 1.128e-06 ***
## PolicyPlan  1   1.6306       118     129.37    0.2226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(rsq)
rsq(gauss,adj=T) #Get the RSquared
```

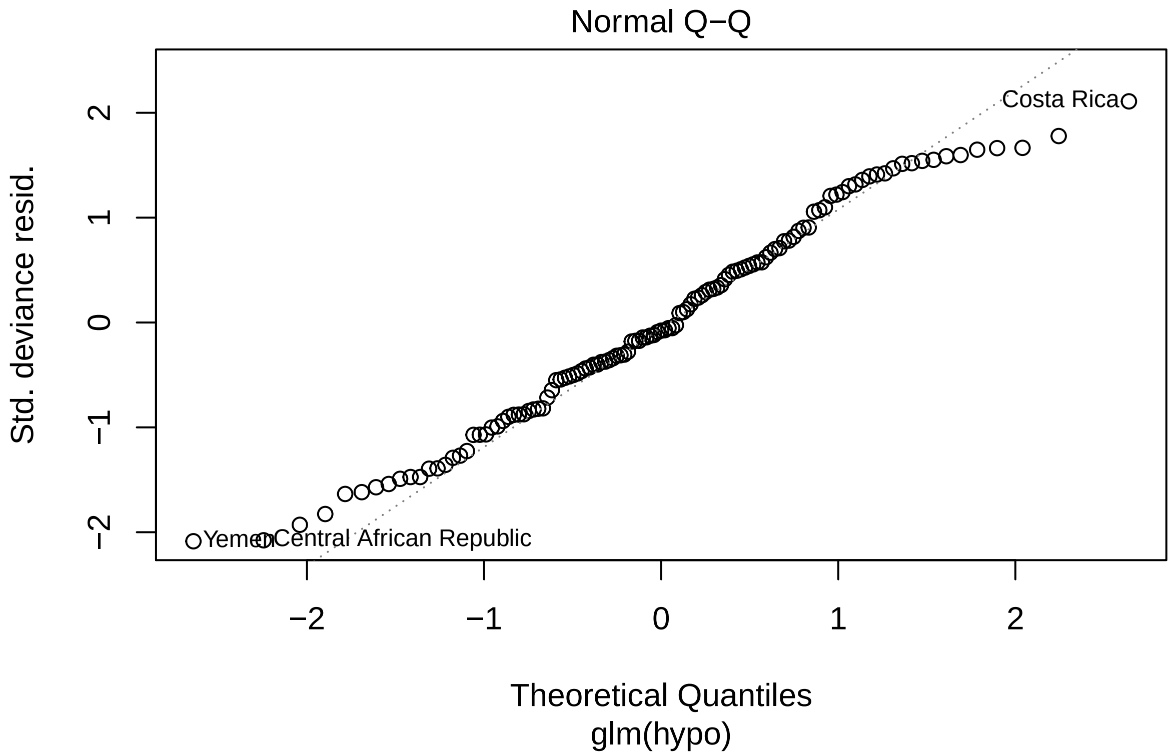
```
## [1] 0.161911
```

Verify the situation of chosen model

```
#Assume linear relationship
plot(gauss,1)
```



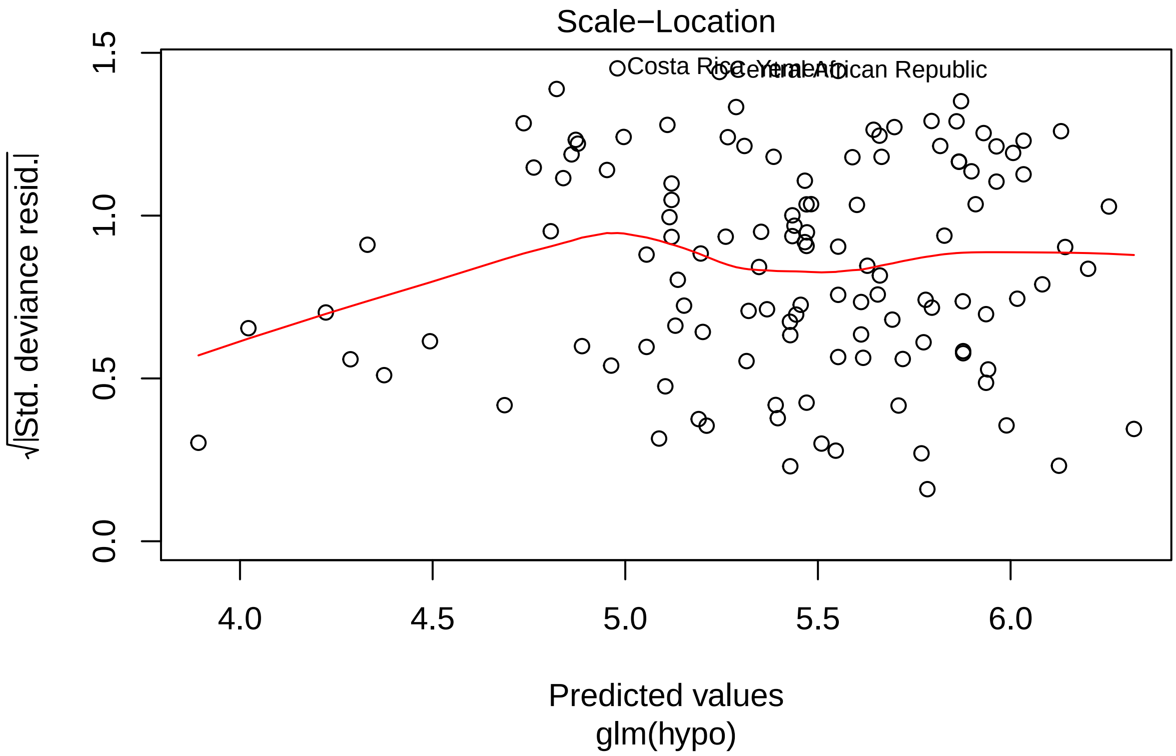
```
#Assume normality of residuals
plot(gauss,2) #visual
```



```
shapiro.test(gauss$residuals) #mathematical
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gauss$residuals
## W = 0.97962, p-value = 0.06354
```

```
#Assume homoscedasticity
plot(gauss, 3) #visual
```



```r
library(lmtest) #mathematical
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
bptest(gauss)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  gauss
## BP = 0.19823, df = 2, p-value = 0.9056
```

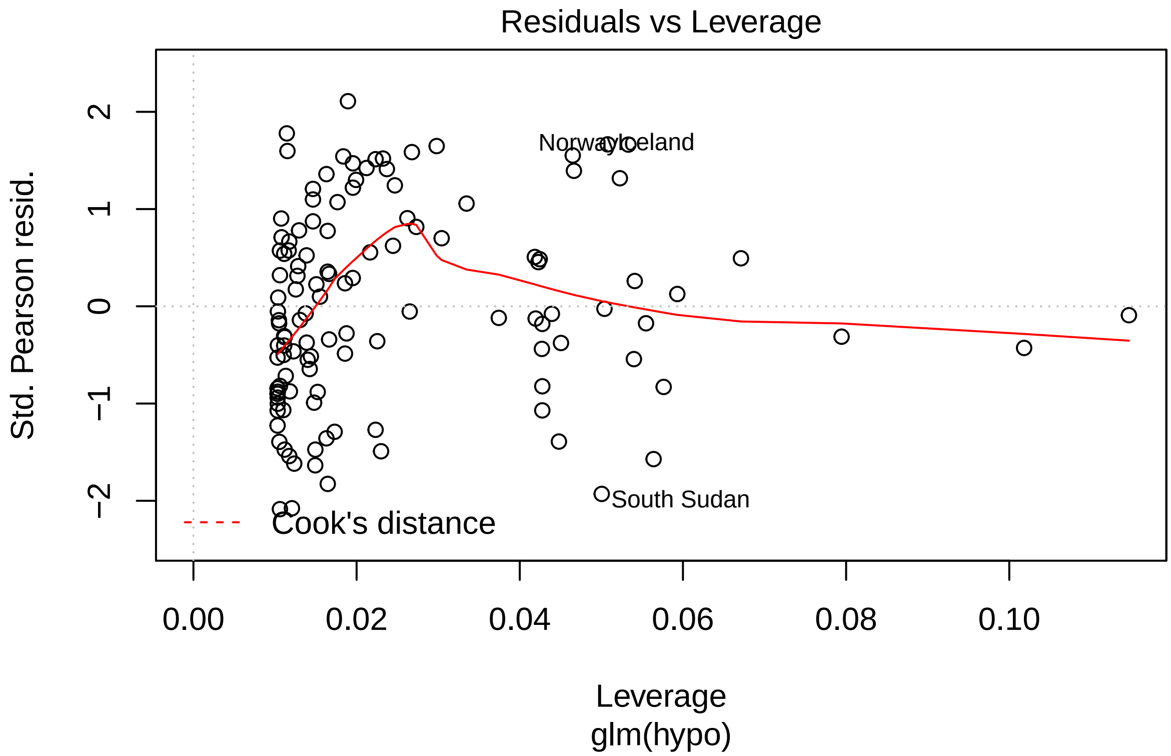
```r
#Assume no colinearity
library(car)
```

```
## Loading required package: carData
```

```
vif(gauss)
```

```
##       GINI PolicyPlan
##   1.011631   1.011631
```

```
#Analize the effect of atypical values
plot(gauss,5) #visual
```

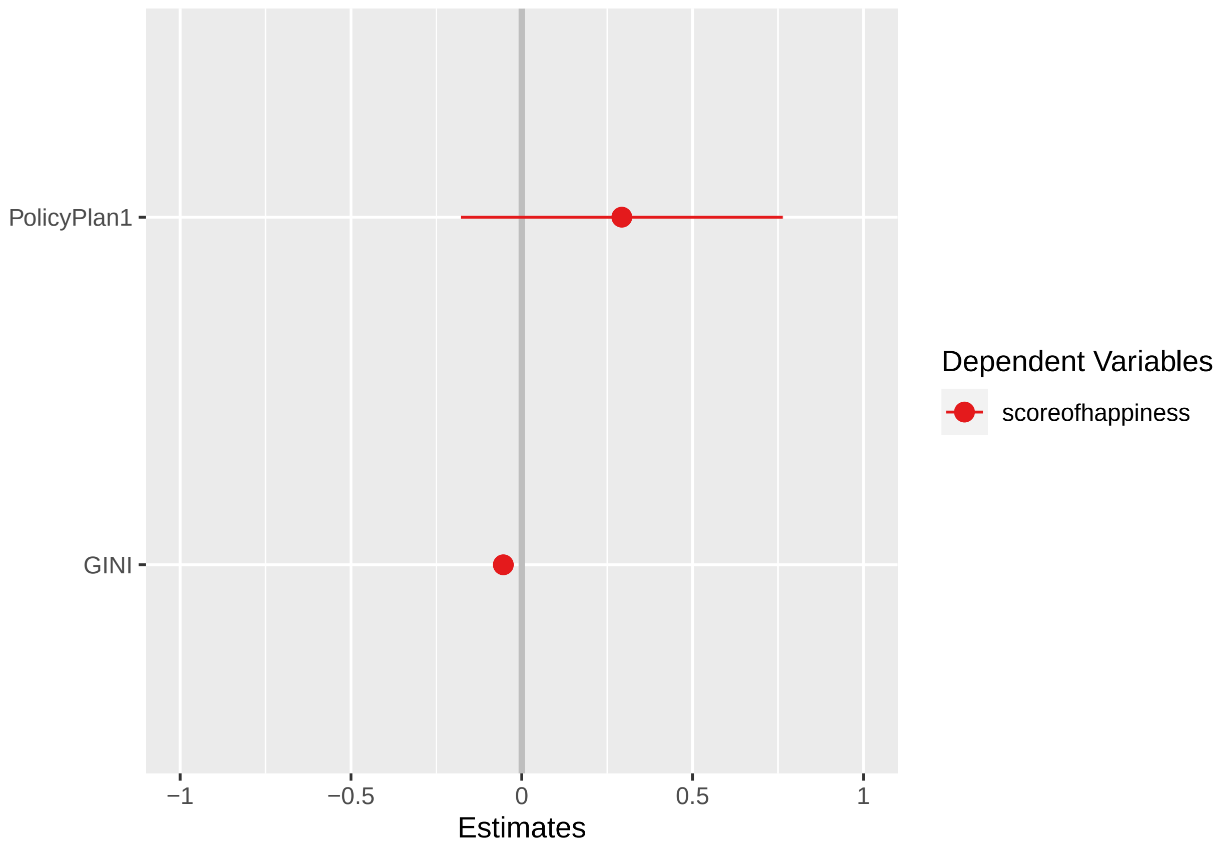


```
#Summary plot
library(sjPlot)
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

```
plot_models(gauss,vline.color = "grey")
```



Regression Result

Score of Happiness = 0.293(Policy Plan) – 0.054(GINI) + 7.31 * GINI statistically significant at the p<0.001 level * The intercept estimate is also statistically significant at the p<0.001 level * The PolicyPlan coefficient is not statistically significant * Adjusted R2 = 0.162

Our regression analysis shows statistically significant results for the impact of the GINI Index on happiness on happiness levels (p<0.001), but not for the publication of mental health policy. The intercept is also statistically significant at the p<0.001 level. Further, the Adjusted R-squared value for the regression is 0.162, indicating that this model does not predict the happiness score very well, which we would expect given the model has two variables, only one of which is statistically significant

The model estimates a country's happiness level will be equal to 7.31 minus 0.054 times the GINI Index (which ranges from 0 to 100) plus 0.293 if the country publishes their policy plan. For context regarding the GINI Index, the lowest GINI score in the data set is 22.7 (Faroe Islands) and the highest is 63.2 (Lesotho). The United States' GINI score is 45.0, while most of the Nordic countries—which received among the highest happiness scores in the Gallup World Happiness Report—have GINI scores in the mid-20s.

Further Regression Analysis

Further analysis could be conducted to better explore the relationship between various economic and social indicators and happiness. Other variables to include and control for could be per capita GDP, region of the world, political system, presence of conflict or instability, life expectancy, maternal mortality rate, literacy rate, and more. However, the Happiness Score as calculated by Gallup includes a number of these variables in their analysis. Thus, we would recommend delineating these factors from the score, and use only aggregated self-reported happiness scores by country as the measure of happiness, and then utilize these variables in a regression analysis. This would help to better understand factors impacting happiness without collinearity caused by double-counting variables on each side of the regression "equation."

Additional analysis could assess the impact of these various indices and variables on happiness at the state and local levels. Gallup or another such organization could conduct happiness surveys by state, county, metropolitan area, and/or city. This analysis could follow the above suggestion, using aggregated self-reported happiness scores as the dependent variable and the economic and social measures as independent variables.

Lastly, our data were not standardized by year. Additional analysis would require utilizing data all from the same year, rather than GINI Index scores varied by year, happiness scores from 2019, and publication status of mental health programs also varied by year.