Homework - Day 1: Data Exploration

Social Analysis and Simulation in R

Yui Nishimura

1 Problem

Instruction: Choose one observational data of your interest. Imagine that you are writing a paper using the data and an appendix for descriptive statistics. Based on your theoretical interest (and after explaining your research questions), provide a set of summaries of your data both numerically and visually. When visualizing the data using scatter plots, histogram, correlation plots, and etc, make sure that you do so by always following the theoretical range of your variables of interest (i.e., when plotting a variable denoting proportion, the plot must show its theoretical minimum and maximum, and thus, 0 and 1). Comment on several variables that you find interesting and discuss how much the empirical distributions of the values of these variables deviate from the theoretical distributions of them (or your expectations on them). If possible, visualize such expectation on the same plots as well.

1.1 Introduction

Why do states voluntarily participate to the Universal Periodic Review (UPR) even though party countries share political affinity? The UPR is the peer review system created by the United Nations (UN), which launched with the establishemnent the Human Rights Council in 2008. Given the cumulated critics against its predecessor, the Human Rights Commissions, member countries purported to avoid politicizing institutional settings and promote fair monitoring of human rights practices in the member countries.

According to the literature, peer review system across countries seemingly reflects relationships between them. Thus, my theoretical interest is whether recommendations submitted by the friendly countries are more latient on the states under the review. %>%

1.2 Variables and Data

The dataset is structured with recommendation as a unit of analysis. The outcome variable is *Severity of Recommendations*, which indicate the severity level of recommendation. The non-governmental organization, UPR Info, collected all the UPR data including the severity of recommendations based on wordings. According to their coding rules, the severity indicator provides 5 ordered category, where the larger number represents harsher contents, and vice versa. Recommendations numbered as small just suggest to report more information, or share their policy experiences. According to the coding rules, the details of categories are summarized as follows (https://www.upr-info.org/database/files/Database_Action_Category.pdf).

1. Recommendation directed at non-SuR states, or calling upon the SuR to request technical assistance, or share information

- Example of verbs: call on, seek, share

2. Recommendation emphasizing continuity
   - Example of verbs: continue, maintain, persevere, persist, pursue
3. Recommendation to consider change
   - Example of verbs: analyse, consider, envisage envision, examine, explore, reflect upon, revise, review, study
4. Recommendation of action that contains a general element
   - Example of verbs: accelerate, address, encourage, engage with, ensure, guarantee, intensify, promote, speed up, strengthen, take action, take measures or steps towards
5. Recommendation of specific action
   - Example of verbs: conduct, develop, eliminate, establish, investigate, undertake as well as legal verbs: abolish, accede, adopt, amend. implement, enforce, ratify

By using this scheme, each recommendation is coded within the range from 1 to 5, since wordings at different levels are used at one time. In such cases, the coders take averages of severity ranks based on frequency of each word.

Variables of interest to explain the variations across recommendation should capture how closely the two countries are related or the extent of shared strategic interests. Here I employ *alliance*, *foreign aid*, and *voting behavior at the UN* as ones of typical political partnership.

For the sake of this purpose, I employed datasets developed by Levoic and Voeten (2017), which include the necessary variables I argue here. This dataset is available from https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10087.

1.3 Preparation

Now I start to explore the relationship between variables of interest. First, I clean up the environment.

```
rm(list=ls()) # rm() cleans up your R
gc();gc()     # gc() cleans up your memory
```

There are packages the following excircise use.

```
library("haven")
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```
library("PerformanceAnalytics")
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
##
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```
library("xtable")
```

Here I load the datasets.

```
# For mac, I need to specify the encoding as below since this dataset includes the string va
df1 <- read_dta("data/merged_data.dta", encoding = "latin1")
head(df1)
```

As we can see on the above, this dataset originally include so many variables which unnecessary for the purpose here. Therefore, I extract only relevant variables as follows.

```
# the name of columns I need
keeps <- c("ccode1", "ccode2", "year", "action", "alliance", "SenderAidDonor", "TargetAidDon
df2 <- df1[keeps]

# check the new data
summary(df2)
```

```
##      ccode1          ccode2           year          action
##  Min.   :  2.0   Min.   :  2.0   Min.   :1817   Min.   :1.00
##  1st Qu.:110.0   1st Qu.:101.0   1st Qu.:1983   1st Qu.:3.00
##  Median :344.0   Median :325.0   Median :2007   Median :4.00
##  Mean   :358.6   Mean   :351.2   Mean   :1992   Mean   :3.76
##  3rd Qu.:600.0   3rd Qu.:572.0   3rd Qu.:2011   3rd Qu.:4.50
##  Max.   :990.0   Max.   :990.0   Max.   :2014   Max.   :5.00
##  NA's   :21298   NA's   :20317   NA's   :1      NA's   :202831
##     alliance     SenderAidDonor   TargetAidDonor      session
##  Min.   :0.000   Min.   :0.0      Min.   :0.00     Min.   : 1.00
##  1st Qu.:0.000   1st Qu.:0.0      1st Qu.:0.00     1st Qu.: 7.00
##  Median :0.000   Median :0.0      Median :0.00     Median :12.00
##  Mean   :0.022   Mean   :0.1      Mean   :0.09     Mean   :11.94
##  3rd Qu.:0.000   3rd Qu.:0.0      3rd Qu.:0.00     3rd Qu.:17.00
##  Max.   :1.000   Max.   :1.0      Max.   :1.00     Max.   :20.00
##  NA's   :23875   NA's   :158985   NA's   :158985   NA's   :182027
```

```
##   jointvotes3       response
##  Min.   : 1.00    Length:223093
##  1st Qu.:58.00    Class :character
##  Median :63.00    Mode  :character
##  Mean   :59.46
##  3rd Qu.:68.00
##  Max.   :77.00
##  NA's   :158584
```

The original dataset consists of dyad-year as the unit of analysis. However, I do not examine the relations between dyad and recommendations. But, I am interested in how the contents of recommendation can be differensiated over the relationship of recommending countries and countries under review. Therefore, I have to reconstruct the data by making recommendation as the unit of analysis. To do so, I ommitted rows which include NA in the action variable, because all recommendations have been evaluated to indicate their level of requirement. In other words, if there is no evaluation for recommendation, recommendation does not exist for that dyad-year. And then, I confirm whether the new data is successfully created or not.

```
# extracting the targetted data
df3 <- subset(df2, !is.na(df1$action))
head(df3)
```

```
## # A tibble: 6 x 10
##   ccode1 ccode2  year action alliance SenderAidDonor TargetAidDonor session
##    <dbl>  <dbl> <dbl>  <dbl>    <dbl>          <dbl>          <dbl>   <dbl>
## 1      2     20  2010   5           1              0              0       9
## 2      2     40  2010   4.5         0              0              1       9
## 3      2     41  2010   4           1              0              1      NA
## 4      2     52  2010   3           1              0              1      NA
## 5      2     70  2010   4.60        1              0              1       9
## 6      2     90  2010   4.17        1              0              1      NA
## # ... with 2 more variables: jointvotes3 <dbl>, response <chr>
```

```
# creating id for row
df3$id <- NA
df3$id <- seq.int(nrow(df3))

# successfully capturing the time period - 2008 (the establishment of UPR) to 2014
summary(df3$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2008    2010    2012    2011    2013    2014
```

1.4 Numerical Descriptive Statistics

Now, I start to explore the variables of interests.

```r
summary(df3$action)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.758   4.500   5.000
```

```r
unique(df3$action)
```

```
##  [1] 5.000000 4.500000 4.000000 3.000000 4.600000 4.166667 4.736842
##  [8] 4.428571 4.583333 4.666667 4.750000 3.500000 4.090909 4.200000
## [15] 4.250000 4.125000 2.666667 3.600000 4.400000 4.833333 4.375000
## [22] 3.666667 2.000000 4.545455 4.333333 3.333333 4.285714 3.250000
## [29] 3.750000 4.625000 4.714286 3.400000 1.000000 3.166667 4.800000
## [36] 3.833333 2.500000 2.333333 1.250000 4.142857 1.500000 2.285714
## [43] 2.800000 1.333333 4.777778 2.600000 1.666667 3.200000 2.250000
## [50] 2.750000 3.857143 4.571429 3.800000 4.111111 4.857143 2.833333
## [57] 3.777778 4.909091 3.142857 4.444445 3.714286 3.909091 4.300000
## [64] 3.555556 2.400000 3.875000 1.750000 4.222222 3.285714 3.428571
## [71] 4.083333 3.888889 4.555555 4.307693 4.700000 3.571429 4.875000
## [78] 4.636364 4.357143 1.400000 4.727273 4.818182 4.363636 4.846154
## [85] 2.200000 4.272727 3.625000 4.100000 4.900000 4.888889 1.857143
## [92] 4.066667 2.375000 4.769231 3.538461 3.375000
```

We also can use `table()` to confirm the frequently observed combination of data by two variables. In this case, the outcome variable can take a lot of values, so it might not be sufficient. But if we are interested in some data which are theoretically observed more than others (in this case, the integer outcomes), it might be useful to look up once to confirm the expectation. Here I only extract first twenty observations as an example.

```r
table(head(df3$action, 20), head(df3$alliance, 20))
```

```
##
##                    0 1
##   3                0 1
##   3.5              0 1
##   4                0 3
##   4.16666650772095 0 1
##   4.42857122421265 0 1
##   4.5              1 3
##   4.58333349227905 0 1
##   4.59999990463257 0 1
##   4.66666650772095 2 0
##   4.73684215545654 0 1
##   4.75             0 2
##   5                0 2
```

To output the data itself in LaTeXformat, we can use 'xtable() function'. It is useful to save the table once and print it to add additional setting of the table. Here I only extract the first five observations.

```
library("xtable")
t1 <- xtable(head(df3, caption = "The Data Set", table.placement =""))
print(t1, scalebox=.8, caption.placement = "top")
```

% latex table generated in R 3.6.0 by xtable 1.8-4 package % Mon Aug 26 08:43:06 2019

	ccode1	ccode2	year	action	alliance	SenderAidDonor	TargetAidDonor	session	jointvotes3	response	id
1	2.00	20.00	2010.00	5.00	1.00	0.00	0.00	9.00	66.00	Accepted	1
2	2.00	40.00	2010.00	4.50	0.00	0.00	1.00	9.00	65.00	Accepted	2
3	2.00	41.00	2010.00	4.00	1.00	0.00	1.00		65.00		3
4	2.00	52.00	2010.00	3.00	1.00	0.00	1.00		66.00		4
5	2.00	70.00	2010.00	4.60	1.00	0.00	1.00	9.00	66.00	Accepted	5
6	2.00	90.00	2010.00	4.17	1.00	0.00	1.00		66.00		6

`xtable()` is also used for summary statistics of the dataset itself.[1] Here I extract some variables of interest.

```
t2 <- xtable(summary(df3[,c(3:5,7,9,10)],
                     caption = "Summary Statistics of the Data Set",
                     table.placement =""))
print(t2, scalebox=.8, caption.placement = "top")
```

% latex table generated in R 3.6.0 by xtable 1.8-4 package % Mon Aug 26 08:43:06 2019

	year	action	alliance	TargetAidDonor	jointvotes3	response
X	Min. :2008	Min. :1.000	Min. :0.000	Min. :0.0000	Min. : 1.00	Length:20262
X.1	1st Qu.:2010	1st Qu.:3.000	1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:61.00	Class :character
X.2	Median :2012	Median :4.000	Median :0.000	Median :0.0000	Median :65.00	Mode :character
X.3	Mean :2011	Mean :3.758	Mean :0.114	Mean :0.0901	Mean :62.86	
X.4	3rd Qu.:2013	3rd Qu.:4.500	3rd Qu.:0.000	3rd Qu.:0.0000	3rd Qu.:68.00	
X.5	Max. :2014	Max. :5.000	Max. :1.000	Max. :1.0000	Max. :77.00	
X.6				NA's :355	NA's :133	

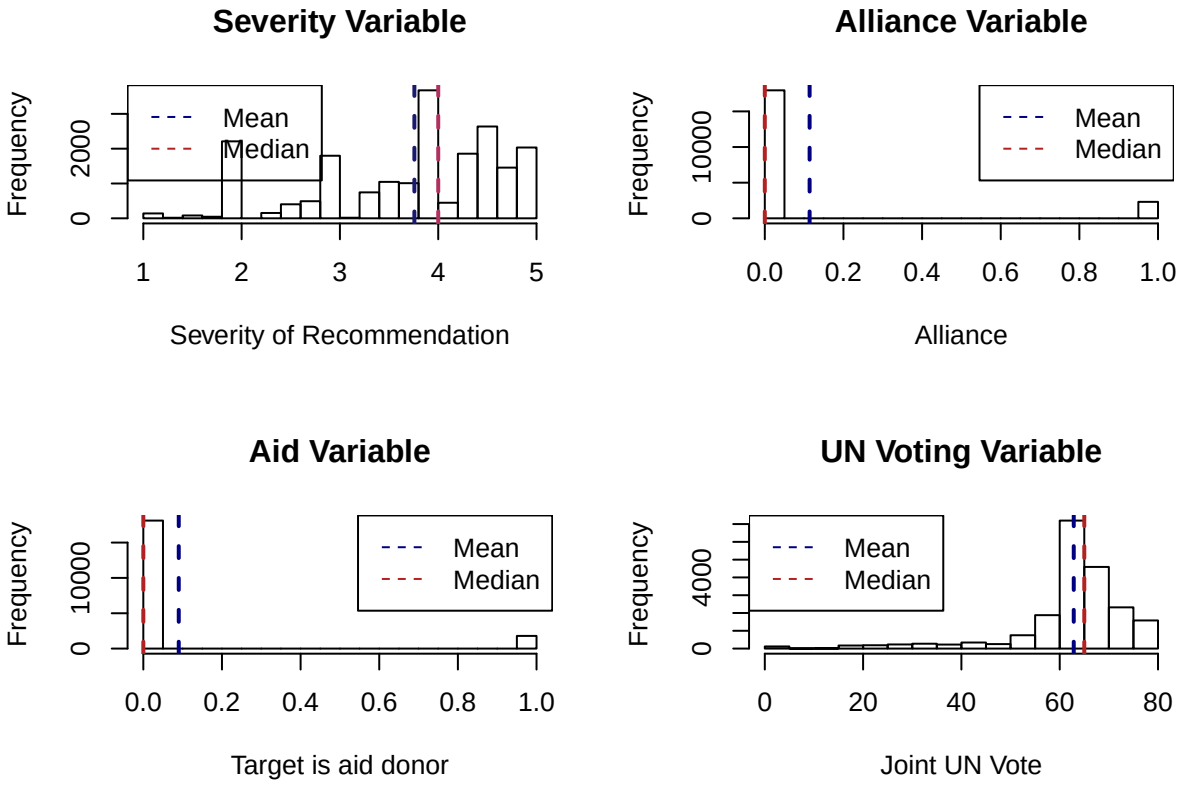
1.5 Visual Descriptive Statistics

Here I see the histogram of each variable first to see their distribution. To make the graph more informative, the mean and the median are calculated and added.

```
par(mfrow=c(2,2))
hist(df3$action, xlab="Severity of Recommendation", ylab="Frequency",
     main="Severity Variable")
abline(v=mean(df3$action, na.rm = TRUE),
       col="midnightblue", lwd="2", lty=2)  # Add a line
abline(v=median(df3$action, na.rm = TRUE),
       col="maroon", lwd="2", lty=2) # Add another line
legend("topleft", legend=c("Mean", "Median"),              # Add legend
```

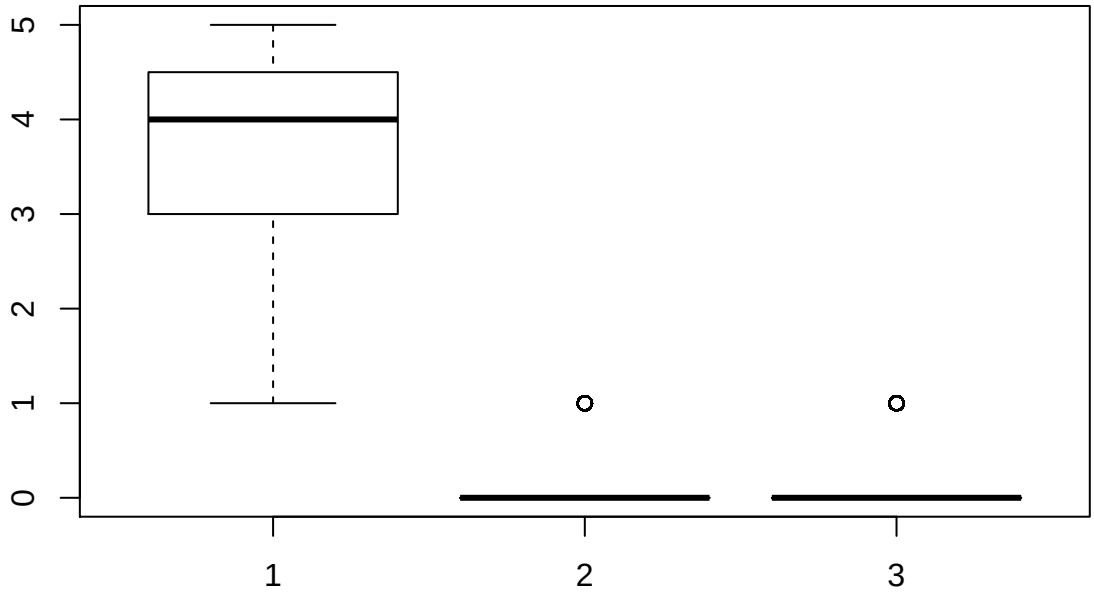
[1] The combination of 'xtable()' and 'summary()' is only available for dataframe. To see the descriptive statistics of a vector by using 'xtable', we have to use the vector directly to the 'xtable' or use 'data.frame()' within the 'summary()'.

```
        col=c("navy", "firebrick"), lty=c(2,2))

hist(df3$alliance, xlab="Alliance", ylab="Frequency", main="Alliance Variable")
abline(v=mean(df3$alliance, na.rm = TRUE), col="navy", lwd="2", lty=2)
abline(v=median(df3$alliance, na.rm = TRUE), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"),
        col=c("navy", "firebrick"), lty=c(2,2))

hist(df3$TargetAidDonor, xlab="Target is aid donor", ylab="Frequency", main="Aid Variable")
abline(v=mean(df3$TargetAidDonor, na.rm = TRUE), col="navy", lwd="2", lty=2)
abline(v=median(df3$TargetAidDonor, na.rm = TRUE), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"),
        col=c("navy", "firebrick"), lty=c(2,2))

hist(df3$jointvotes3, xlab="Joint UN Vote", ylab="Frequency", main="UN Voting Variable")
abline(v=mean(df3$jointvotes3, na.rm = TRUE), col="navy", lwd="2", lty=2)
abline(v=median(df3$jointvotes3, na.rm = TRUE), col="firebrick", lwd="2", lty=2)
legend("topleft", legend=c("Mean", "Median"),
        col=c("navy", "firebrick"), lty=c(2,2))
```

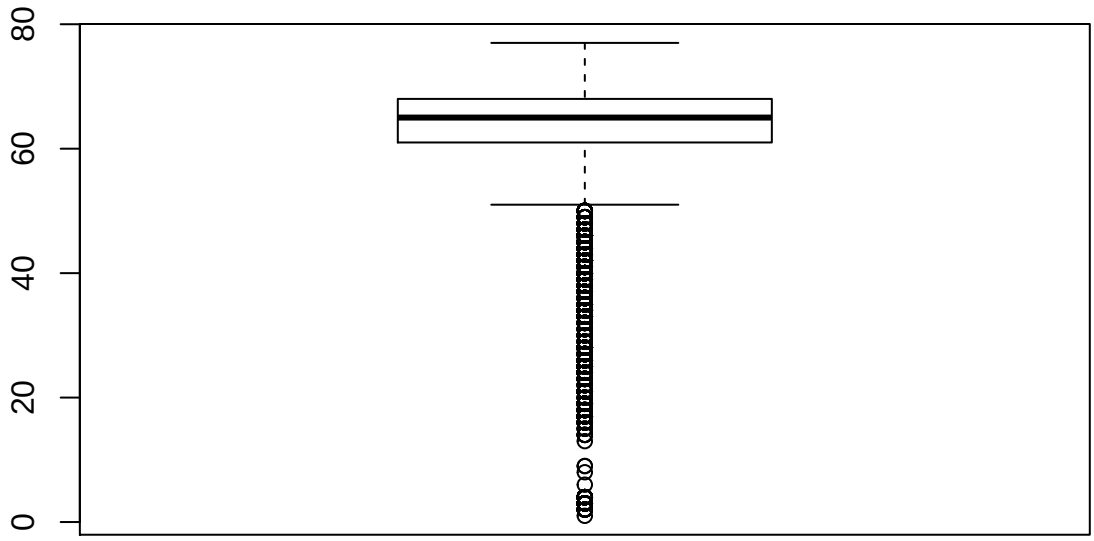


Here I use boxplot and the violine plot. Since the possible range of variable is largely different for joint UN voting vairable, I make the sole figure for it.

```r
boxplot(df3$action, df3$alliance, df3$TargetAidDonor)
```




```r
boxplot(df3$jointvotes3)
```



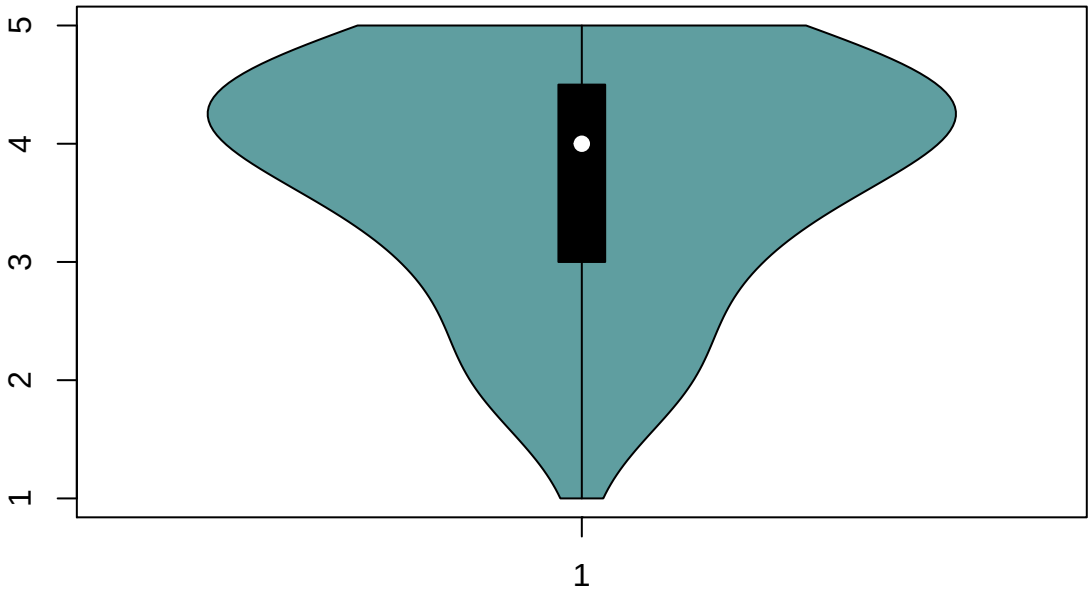

```r
head(df3)
```

```
## # A tibble: 6 x 11
##   ccode1 ccode2  year action alliance SenderAidDonor TargetAidDonor session
##    <dbl>  <dbl> <dbl>  <dbl>    <dbl>          <dbl>          <dbl>   <dbl>
## 1      2     20  2010   5           1              0              0       9
## 2      2     40  2010   4.5         0              0              1       9
## 3      2     41  2010   4           1              0              1      NA
## 4      2     52  2010   3           1              0              1      NA
## 5      2     70  2010   4.60        1              0              1       9
```

```
## 6       2     90  2010    4.17         1                0                1       NA
## # ... with 3 more variables: jointvotes3 <dbl>, response <chr>, id <int>
```

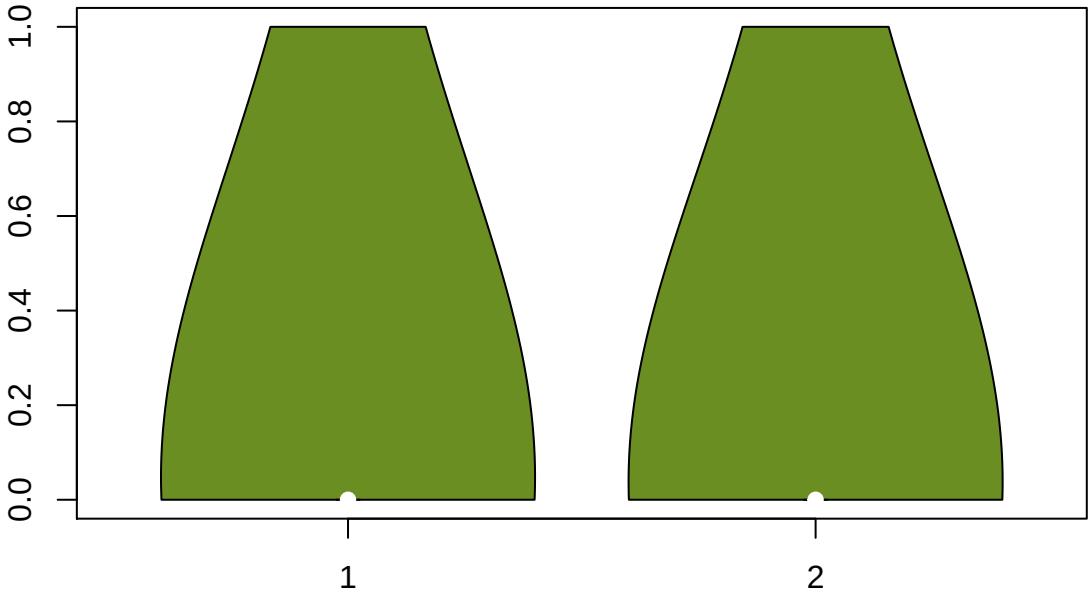
```
library("vioplot")
vioplot(df3$action, col="cadetblue")
```

```
## [1] 1 5
```



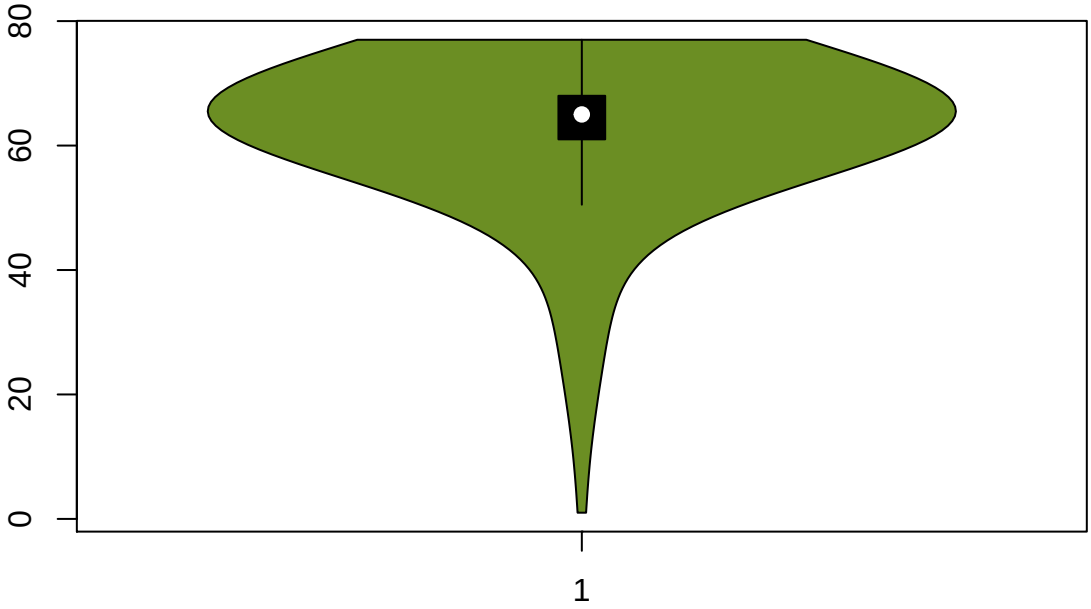

```
vioplot(df3$alliance, df3$TargetAidDonor, col="olivedrab")
```

```
## [1] 0 1
```



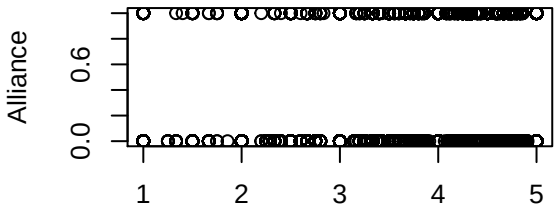

```
vioplot(df3$jointvotes3, col="olivedrab")
```

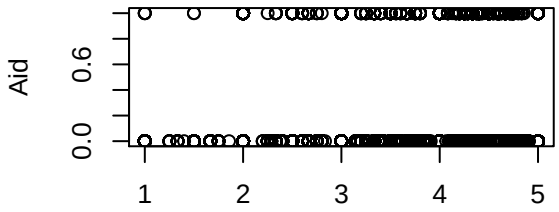
```
## [1]  1 77
```

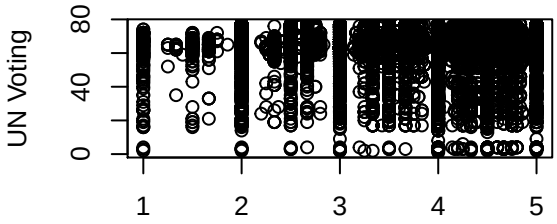


Plotting the relationships between the outcome and the independent variables.

```
par(mfrow=c(2,2))
plot(df3$action, df3$alliance, xlab = "Sevirity of Recommendation",
     ylab = "Alliance")
plot(df3$action, df3$TargetAidDonor, xlab = "Sevirity of Recommendation",
     ylab = "Aid")
plot(df3$action, df3$jointvotes3, xlab = "Sevirity of Recommendation",
     ylab = "UN Voting")
```







Looking at the bivariate relaitonship between variables, by using `corrplot()` and `chart.Corrrelation()`.

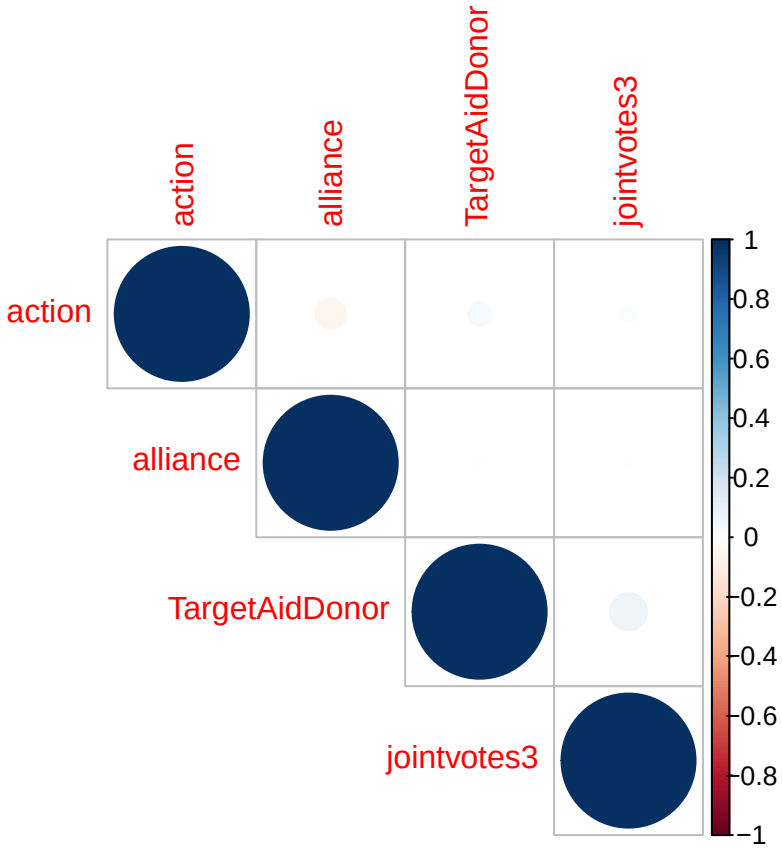
```
cor(df3[,c(4,5,7,9)], use = "complete.obs")    # correlation matrix without NA (only with co
```

```
##                    action     alliance TargetAidDonor  jointvotes3
## action          1.00000000 -0.050313201    0.030364254  0.015996332
## alliance       -0.05031320  1.000000000    0.004997134 -0.005176501
## TargetAidDonor  0.03036425  0.004997134    1.000000000  0.077017002
## jointvotes3     0.01599633 -0.005176501    0.077017002  1.000000000
```

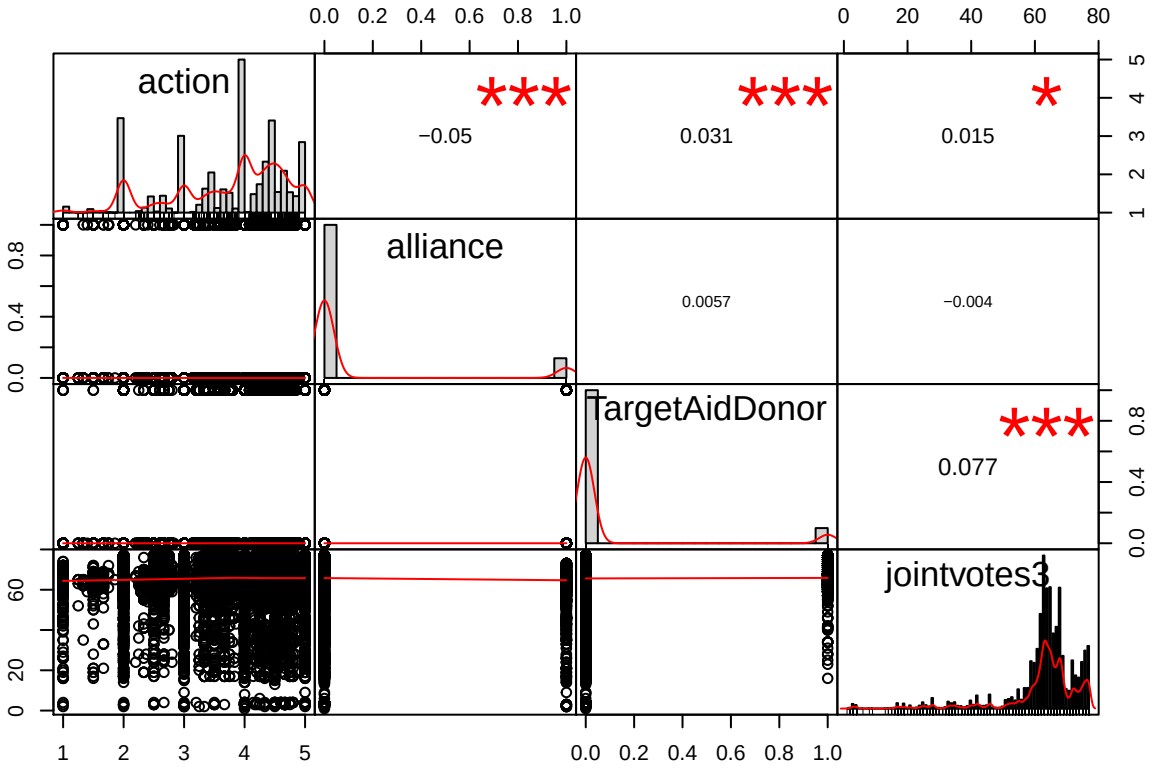
```
round(cor(df3[,c(4,5,7,9)], use = "complete.obs"), digit=2) # rounding numbers
```

```
##                 action alliance TargetAidDonor jointvotes3
## action            1.00    -0.05           0.03        0.02
## alliance         -0.05     1.00           0.00       -0.01
## TargetAidDonor    0.03     0.00           1.00        0.08
## jointvotes3       0.02    -0.01           0.08        1.00
```

```
# visualizing correlation
# library("corrplot")
co_mat <-  cor(df3[,c(4,5,7,9)], use = "complete.obs")     # ssing correlation matrix
corrplot(co_mat, type="upper")
```




```
# library("PerformanceAnalytics")
chart.Correlation(df3[,c(4,5,7,9)], histogram=TRUE, pch=19)
```



Interestingly, only the relation between the alliance and action is negative. This is in line with the theoretical expectation. Surprisingly, the other two show the opposite direction of coorelation. So, it is unintuitive but correlation take into account the all recommendations, regardless of time variation. Further investigations are important to think more about these variables.

2 Question

Instruction: Using the same dataset, explore a set of possible theoretical relationships between two variables of your interest. In other words, find several quantitatively predictive logical models that may account for the functional forms of Variable X explaining Variable Y. Remember (from the readings) that the relationship between the two variables may not be necessarily linear. While you can choose any functional form based on quadratic, square root, natural log, and other transformations, always provide a theoretical rationale of using such transformation of one or two variables. When you find the best functional form you can get, plot the functional form with the scatter plot of these variables. Add any "anchor points" and conceptually "forbidden areas" to the graph. If you have some grouping variables that you are interested in (e.g., majoritarian systems v. proportional representation systems), make sure that you visualize such distinction on the plot and discuss whether your best logical model fits both types of data. Finally, perform a Normal-linear model with the two variables and report a coefficient table and R^2 statistic. Comment on the extent to which "linear regression" is helpful to discover the theoretical functional form.

2.1 Two Variables of Interest

According to the above, it is interesting to see the relationship between alliance and the level of recommendaiton further. Theoretically, it is expected that the alliance formation hinders members to make a statement criticizing the other member(s), since they share the strategic interest and members are interdependent each other, so damaging their relationship is risky for future cooperation. Based on this expectation, the data generation process is:

$$\text{Severity Level}_i \sim (\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \times \text{alliance}_i$$

Here, the main indenpendent variable, alliance, is dichotomous variable, which only can take 0 or 1. On the other hand, the severity level vairable can take values from 0 to 5. Considering expectedly negative relationship between alliance and the severity level, β is greater than 0. Since the outcome variable should be larger than 0, then α should be within the range between 1 and 5.

2.1.1 Visualization of the Functional Form

2.1.2 Performing Normal-Linear Model

```
fit <- lm(action ~ alliance, data = df3)
summary(fit)
```

```
##
## Call:
```

```
## lm(formula = action ~ alliance, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7745 -0.7745  0.2255  0.7255  1.3734
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.774468   0.007017 537.866  < 2e-16 ***
## alliance    -0.147819   0.020788  -7.111 1.19e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9403 on 20260 degrees of freedom
## Multiple R-squared:  0.00249,    Adjusted R-squared:  0.00244
## F-statistic: 50.56 on 1 and 20260 DF,  p-value: 1.192e-12
```

To produce the LaTeXformat table, I use `stargazer()`. In R markdown, the chunk option need to be setted as `results = 'asis'`. In default, R^2 is included in stargazer output.

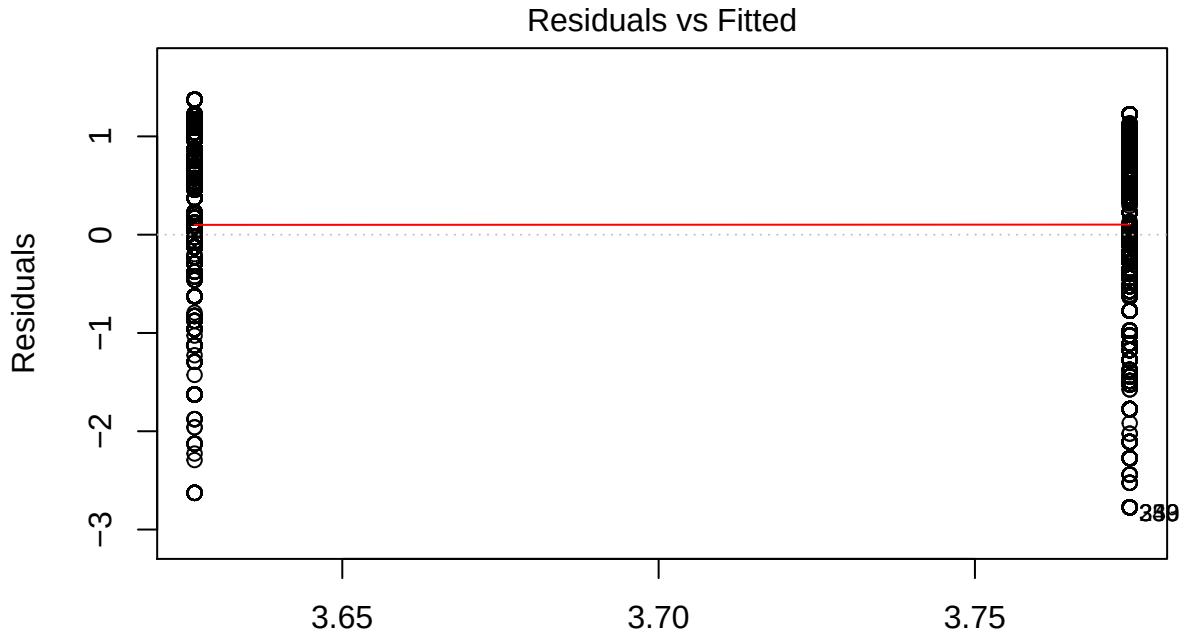
```
library("stargazer")
stargazer(fit)
```

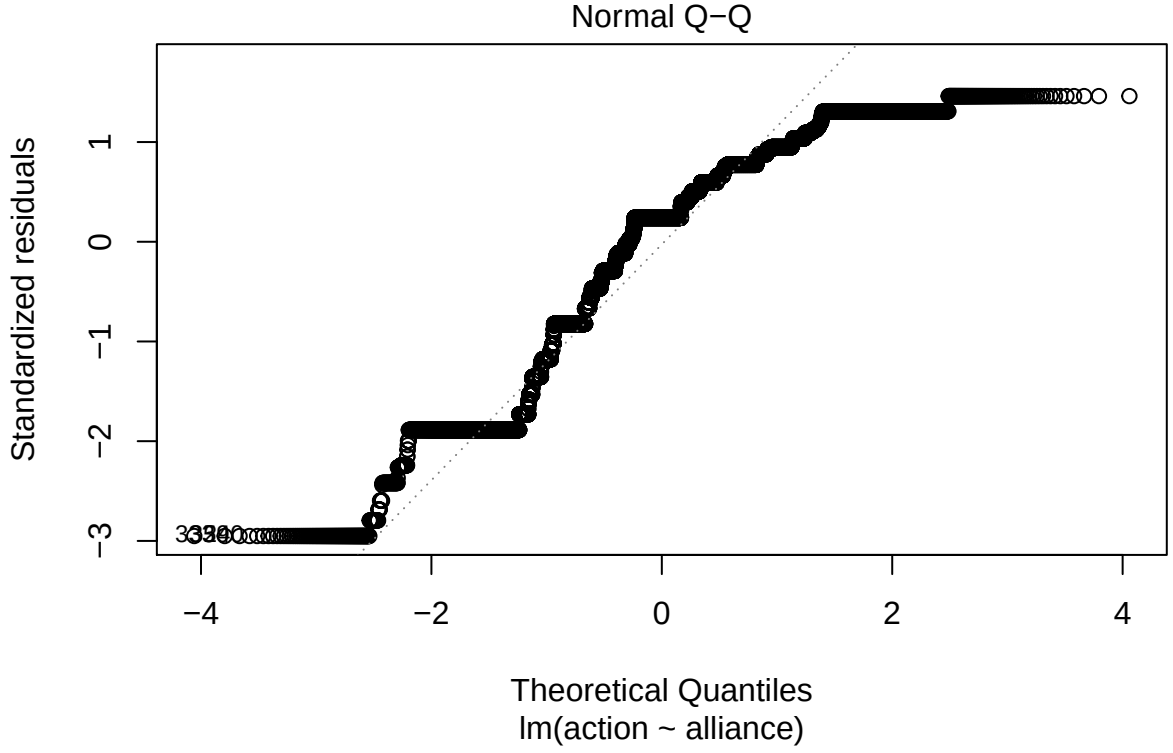
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Mon, Aug 26, 2019 - 08:43:25

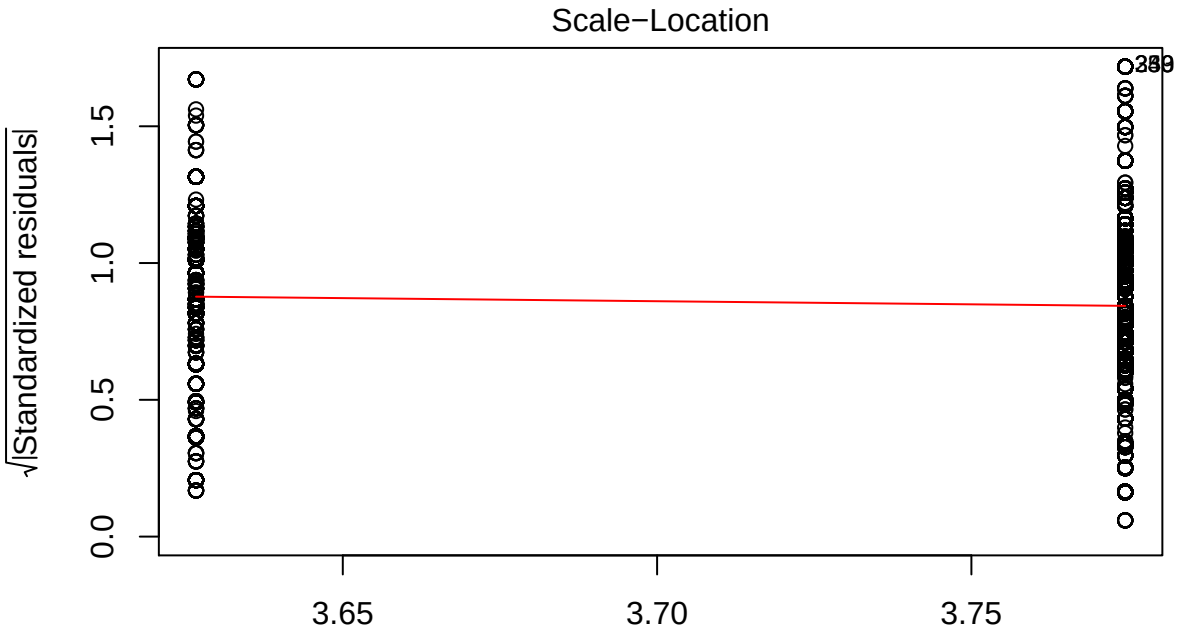
Table 1:

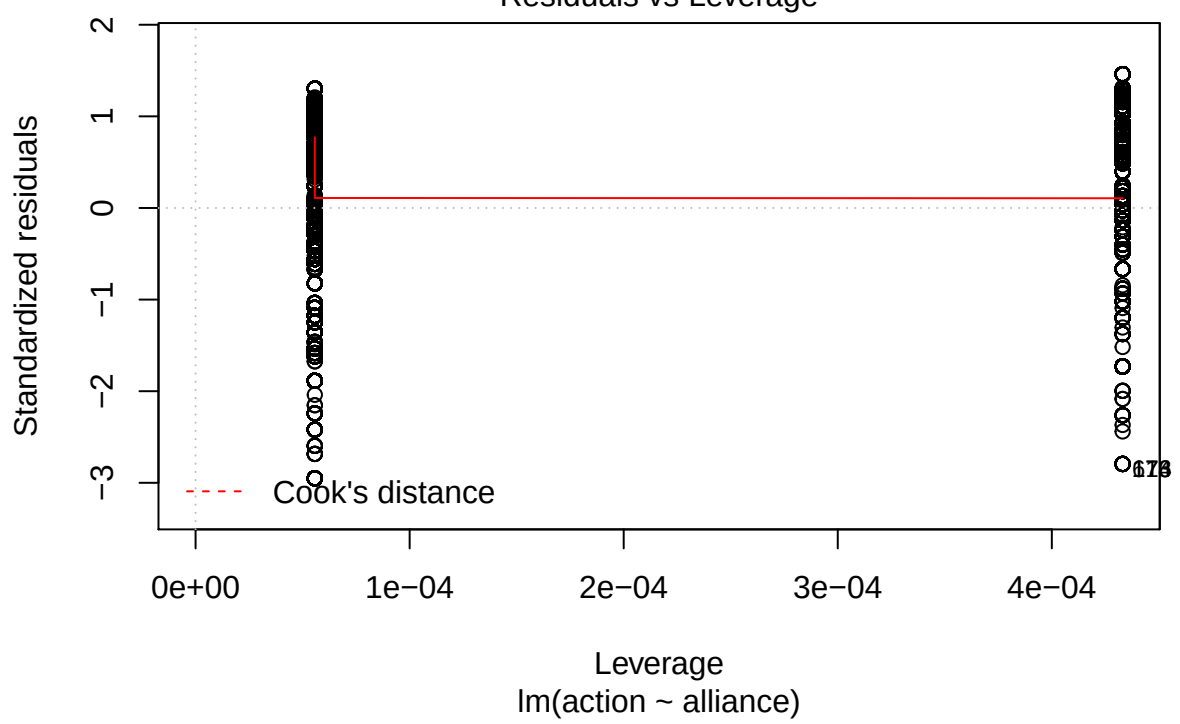
	Dependent variable:
	action
alliance	-0.148^{***}
	(0.021)
Constant	3.774^{***}
	(0.007)
Observations	20,262
R^2	0.002
Adjusted R^2	0.002
Residual Std. Error	0.940 (df = 20260)
F Statistic	50.564^{***} (df = 1; 20260)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
plot(fit)
```


Residuals vs Fitted
Residuals
1
0
−1
−2
−3
3.65
3.70
3.75
Fitted values
lm(action ~ alliance)


Normal Q−Q
Standardized residuals
1
0
−1
−2
−3
−4
−2
0
2
4
Theoretical Quantiles
lm(action ~ alliance)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3.65
3.70
3.75
Fitted values
lm(action ~ alliance)


Residuals vs Leverage
Standardized residuals
2
1
0
−1
−2
−3
Cook's distance
0e+00
1e−04
2e−04
3e−04
4e−04
Leverage
lm(action ~ alliance)