Final exam - Statistics I for research in International Relations and Political Science

An example about how to apprehend a basic quantitative analysis

Rémi Viné

2022-12

Contents

0. Preambule

This document aims to replicate what could have been done and proposes some further tools that can be useful for further investigations. I insist on the importance of understanding the data structures before using it, and the variables used, before interpreting them. Most methods are simply the ones covered in class (Statistics 101), and the extra tools might come to incentivize you to further explore the possibilities to conduct a more specific quantitative analysis.

In this document, I purposefully run flawed models, where some key controls are omitted (for example the standard of living). The document starts with the raw data originally provided and hopes to be self-consistent in terms of packages called here.

The present analysis is not meant to give any insights on the topic: no literature is used. Rather, it tries to highlight the different steps and difficulties one can face in using a simple quantitative study with real-world data sets from different sources.

1. Data wrangling, research question and working hypotheses

The exercise only uses Rstudio and all the material is produced using Rmd (including the present .pdf). Many parts are purposefully kept raw to let the reader see the different steps before reaching conclusions, decisions regarding the data.

```
### Call packages
library(tidyverse)     # basic package for statistics
library(haven)         # necessary to upload Stata datasets
library(readxl)        # necessary to upload excel datasets
library(interactions)  # necessary to plot interactions
library(ggeffects)     # necessary for marginal effects
library(stargazer)     # useful for nice-looking tables
library(ggsci)         # color palette as in some journals
library(moments)       # for skewness and kurtosis (in descriptives)
library(ggcorrplot)    # for nice-looking correlation matrix
library(car)           # for VIF quick computations
library(lm.beta)       # for beta regression
library(lmtest)        # for the Breusch-Pagan test
library(nlme)          # necessary for gls
```

1.1. Upload the different data sets.

There are three main data sources: (i) Urban social disorder (USD) over 1960-2014 from Thomson et al. (2022), (ii) the city population data from the World urbanization prospects (WUP) of the United Nations, (iii) the World development indicators (WDI) from the World Bank. These data have different year ranges, and different observational units. It is important to be clear with each data to then merge them and conduct a quantitative analysis.

```
### Working directory
getwd()   # find the current directory (eg "/home/remi")
```

```
## [1] "/home/remi/Dropbox/Other_tasks/IHEID_STAT_I/FINAL/2022/Students_outputs"
```

```
          # change the current directory
setwd("/home/remi/Dropbox/Other_tasks/IHEID_STAT_I/FINAL/2022/EXAM_STAT_IRPS_2022")
### Import the data
cities <- read_excel("cities.xlsx")
events <- read_excel("events.xlsx")
wdi_selection <- read_dta("WDI_2019_selection.dta")
```

```
wup_pop <- read_dta("WUP_AnnualPopUrbanAggl.dta")
wup_sh_pop <- read_dta("WUP_PercTotPopUrbanAggl.dta")
```

1.2. Basic understanding of the data

```
### Grasp dimensions of dataframes
dim(cities)
```

```
## [1] 187  18
```

```
dim(events)
```

```
## [1] 13460    33
```

```
dim(wdi_selection)
```

```
## [1] 266  27
```

```
dim(wup_pop)
```

```
## [1] 2041    4
```

```
dim(wup_sh_pop)
```

```
## [1] 2041    4
```

The data sets “cities” and “events” do not have the same observational unit. In “events”, a city appears as many times as there was social unrest. Since the analysis is done at the city level using either city or country level variables (without time variation in WUP and WDI), then the refinement of having all events per city separated is not necessary here - rather, this advantage cannot be used. **As such, the “events” data does not appear to bring many insightful information in the context of the current analysis.**

Data sets must be merged, and it is important to use the appropriate variables to merge data sets. In particular, names (of countries, of cities) tend to differ across data sets and should therefore be avoided (unless a fastidious exercise of checking each manually).

Note: there exist some packages to merge strings that can differ. For example, you might have used the “fuzzyjoin” package. In the present data sets, this would have delivered a good merger, but clearly, this cannot always be satisfactory. Therefore, you are strongly advised to find the best variables to join data sets.

In the present context, one has:

- **Urban social disorder data:** “cities” data set has *ID_UNWUP* that gives the UN city code, and has *ISO3* that provides the country code
- **City population data:** “wup_pop” and “wup_sh_pop” also have such city code, named “ISO31661numericcode”
- **World development indicators data:** “wdi_selection” has *CountryCode* that can be merged with *ISO3* from the Urban Social Disorder data

1.3. Build a general dataframe

```
### Merge dataframes
cities_WUP1 <- merge(cities, wup_pop, by.x="ID_UNWUP", by.y="ISO31661numericcode")
cities_WUP2 <- merge(cities_WUP1, wup_sh_pop, by.x="ID_UNWUP", by.y="ISO31661numericcode")
overall_data <- merge(cities_WUP2, wdi_selection, by.x="ISO3", by.y="CountryCode")
dim(overall_data) # Dimension of the general merged data set
```

```
## [1] 168  50
```

```
ls(overall_data)  # List of all the variables
```

```
##  [1] "BX_KLT_DINV_WD_GD_ZS"   "BX_TRF_PWKR_CD_DT"      "CAPEND"
##  [4] "CAPITAL"                "CAPSTART"               "CITY"
##  [7] "CITY_ALT"               "CITY_ID"                "COUNTRY"
## [10] "CountryName"            "DEATHEVENTS"            "DT_ODA_ODAT_GN_ZS"
## [13] "EG_ELC_ACCS_ZS"         "FP_CPI_TOTL_ZG"         "FS_AST_CGOV_GD_ZS"
## [16] "GWNO"                   "ID_UNDAT"               "ID_UNWUP"
## [19] "ISO3"                   "LAT"                    "Location.x"
## [22] "Location.y"             "LONG"                   "MS_MIL_XPND_GD_ZS"
## [25] "NEVENTS"                "NODEATHEVENTS"          "NOTE"
## [28] "Note.x"                 "Note.y"                 "NY_GDP_MKTP_CD"
## [31] "NY_GDP_MKTP_PP_CD"      "population_city"        "REGION"
## [34] "SE_PRM_CUAT_ZS"         "SE_PRM_UNER_ZS"         "SE_TER_CUAT_BA_ZS"
## [37] "SH_DTH_COMM_ZS"         "SH_DTH_INJR_ZS"         "SH_DTH_NCOM_ZS"
## [40] "SH_XPD_CHEX_GD_ZS"      "share_pop_in_this_city" "SI_POV_GINI"
## [43] "SL_AGR_EMPL_ZS"         "SL_EMP_TOTL_SP_NE_ZS"   "SL_SRV_EMPL_ZS"
## [46] "SL_TLF_CACT_ZS"         "SL_UEM_TOTL_ZS"         "SP_DYN_AMRT_FE"
## [49] "SP_DYN_AMRT_MA"         "SP_POP_DPND"
```

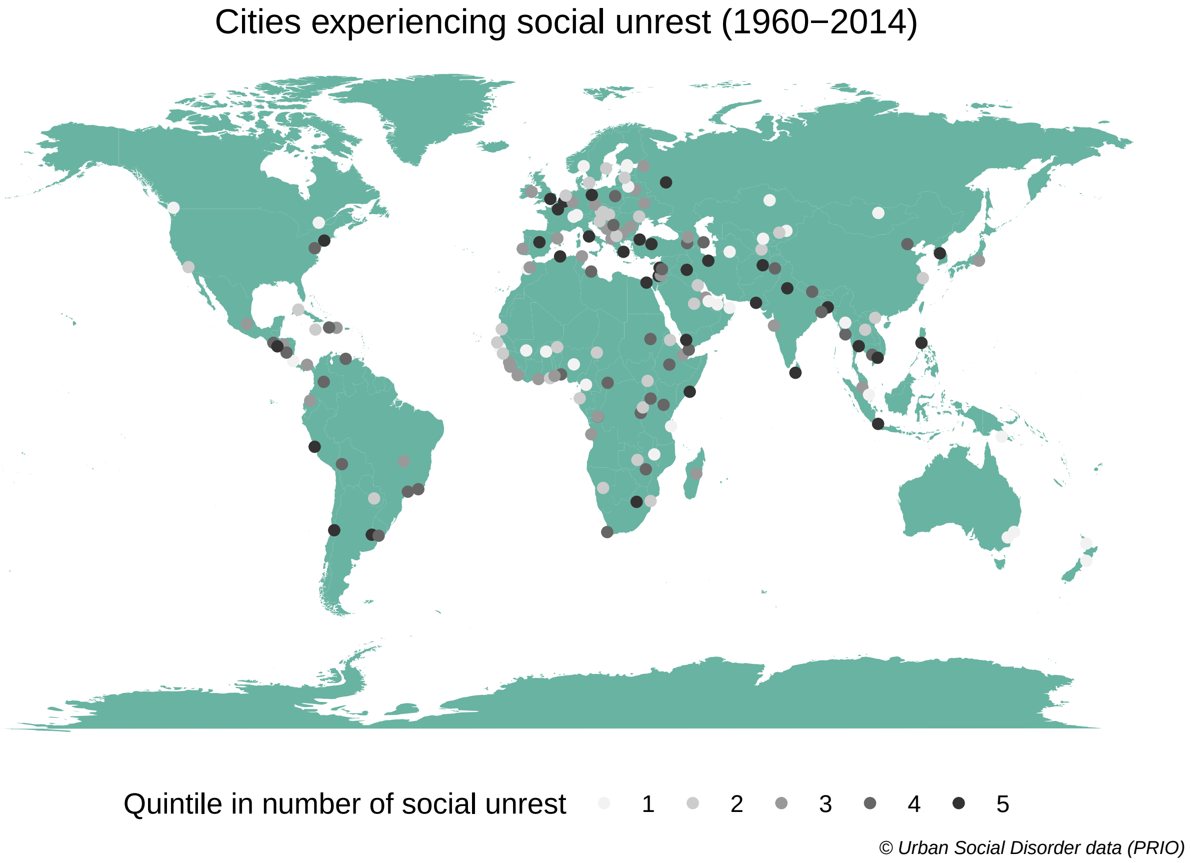
1.4. Build a map

Before going any further, a map can be useful to visualize some potential patterns. First, one might expect social unrest to happen in some countries/regions rather than in others. The map below does not show a noticeably clear pattern, apart from Oceania, even including Southeast Asia tends not to be well represented (or there is hardly any social unrest there). Second, the intensity of social unrest might vary across regions. This is confirmed, as Africa does not have many cities that belong to the 5th quintile in terms of number of social unrests (only 2 cities when excluding Middle East and North Africa (MENA) region while there are 4 in Latin America).

```
### Load a world map
world_map <- map_data("world")
### For readability of the map, let's build a quintile variable for NEVENTS
summary(overall_data$NEVENTS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   27.00   50.50   76.73  100.25  472.00
```

```
overall_data <- overall_data %>%
  mutate(quintile_NEVENTS = findInterval(NEVENTS, quantile(NEVENTS,
                                         probs = seq(0, 1, 0.2)),
                                       rightmost.closed = TRUE))
overall_data$quintile_NEVENTS <- as.factor(overall_data$quintile_NEVENTS)
                                # necessary for discrete color allocation
### Map
ggplot() +
  # Countries
  geom_polygon(data = world_map, aes(x=long, y=lat, group=group),
               fill="#69b3a2", col=NA) +
   # Cities
  geom_point(data= overall_data, aes(x = LONG, y = LAT,
                                     colour = quintile_NEVENTS)) +
  # Set colors
  scale_color_manual(name = "Quintile in number of social unrest",
                     breaks = c("1", "2", "3", "4", "5"),
                     values=c("gray95", "gray80", "gray60", "gray40", "gray20")) +
  # Labs
  labs(title = "Cities experiencing social unrest (1960-2014)", #subtitle = "",
       caption = "© Urban Social Disorder data (PRIO) ",
       fill = "Score") +
    # Theme
    theme_void() +
    theme(plot.caption = element_text(size = 7, face = "italic"),
          plot.title = element_text(hjust = 0.5),    # Center ggplot title
          legend.position = "bottom")
```


Cities experiencing social unrest (1960−2014)
Quintile in number of social unrest
● 1
● 2
● 3
● 4
● 5
© Urban Social Disorder data (PRIO)

1.5. Brief overview of the variables by theme

- *Explained variable (at the city level):* let us take the “NEVENTS” to keep things as simple as they can be. Note that a more elaborate analysis could have constructed a ratio of deadly events, of even an intensity index of the social unrest occurring in a city.
- *Explanatory variables at the city level:* these are the demographic variables taken from WUP
- *Explanatory variables at the country level:* these are the various variables provided by the WDI data.

Note that this selection offers a large range of potential narratives:

- Standard of living and inequality: Gross Domestic Product (GDP) in Purchasing Power Parity (or not), inflation rate, and Gini

- Cross-border capital inflows: Foreign Direct Investment (FDI), Official Development Assistance (ODA), Remittances,

- Labor “market”: share of employed people per sector, (Women) labor force participation, unemployment

- Human capital: education-related variables (and health: expenditures, mortality, and causes of death)

- Infrastructures: access to electricity

- Public sector information: share of public entities in the production, share of military expenditures, share of health expenditures

The narrative’s construction depends on the strand of the literature considered the most relevant. From a more trivial perspective, variables that are largely missing in the data might be best avoided. The output below shows a simple count for all variables. For the ones missing frequently, one might avoid even considering them. Here, the Gini, the share of primary school age children out of school, the educational attainment (beyond bachelor and/or at least primary school) are more often missing than not missing. A variable somehow in between is the ODA variable: 110 observations out of 168 are non-missing.

```
### Build a function for appropriate count (without NAs)
my.summary <- function(x,...){
  c(
    count = format(round(sum(!is.na(x)), ...), 3))
}
sapply(wdi_selection, my.summary) # Only WDI has severe problems in terms of missing data
```

```
##          CountryName.count          CountryCode.count
##                      "266"                      "266"
## BX_KLT_DINV_WD_GD_ZS.count    BX_TRF_PWKR_CD_DT.count
##                      "237"                      "240"
##    DT_ODA_ODAT_GN_ZS.count       EG_ELC_ACCS_ZS.count
##                      "175"                      "264"
##       FP_CPI_TOTL_ZG.count    FS_AST_CGOV_GD_ZS.count
##                      "221"                      "207"
##    MS_MIL_XPND_GD_ZS.count       NY_GDP_MKTP_CD.count
##                      "195"                      "253"
##    NY_GDP_MKTP_PP_CD.count       SE_PRM_CUAT_ZS.count
##                      "240"                       "36"
##       SE_PRM_UNER_ZS.count    SE_TER_CUAT_BA_ZS.count
##                      "170"                       "38"
##       SH_DTH_COMM_ZS.count       SH_DTH_INJR_ZS.count
##                      "231"                      "231"
##       SH_DTH_NCOM_ZS.count    SH_XPD_CHEX_GD_ZS.count
##                      "231"                      "234"
##          SI_POV_GINI.count       SL_AGR_EMPL_ZS.count
##                       "59"                      "235"
## SL_EMP_TOTL_SP_NE_ZS.count       SL_SRV_EMPL_ZS.count
##                      "145"                      "235"
##       SL_TLF_CACT_ZS.count       SL_UEM_TOTL_ZS.count
##                      "235"                      "235"
##       SP_DYN_AMRT_FE.count       SP_DYN_AMRT_MA.count
##                      "222"                      "222"
##          SP_POP_DPND.count
##                      "241"
```

The list of countries below (there are duplicates because some countries have several cities with recorded urban social disorders). Most likely, these countries are: (i) those contributing to ODA but not those receiving, (ii) countries that are in conflict with the United States, so that most ODA is channeled away from them (Venezuela, Cuba). Therefore, The ODA variable might replace the NA by zeros without distorting the reality (apart maybe for South Sudan, Yemen, Eritrea - three countries with extremely serious political instabilities and where the social unrest might not be well studied anyway).

```
### Check the countries with missing information on ODA
overall_data_miss_ODA <- overall_data[(is.na(overall_data$DT_ODA_ODAT_GN_ZS)), ]
list(overall_data_miss_ODA$CountryName)
```

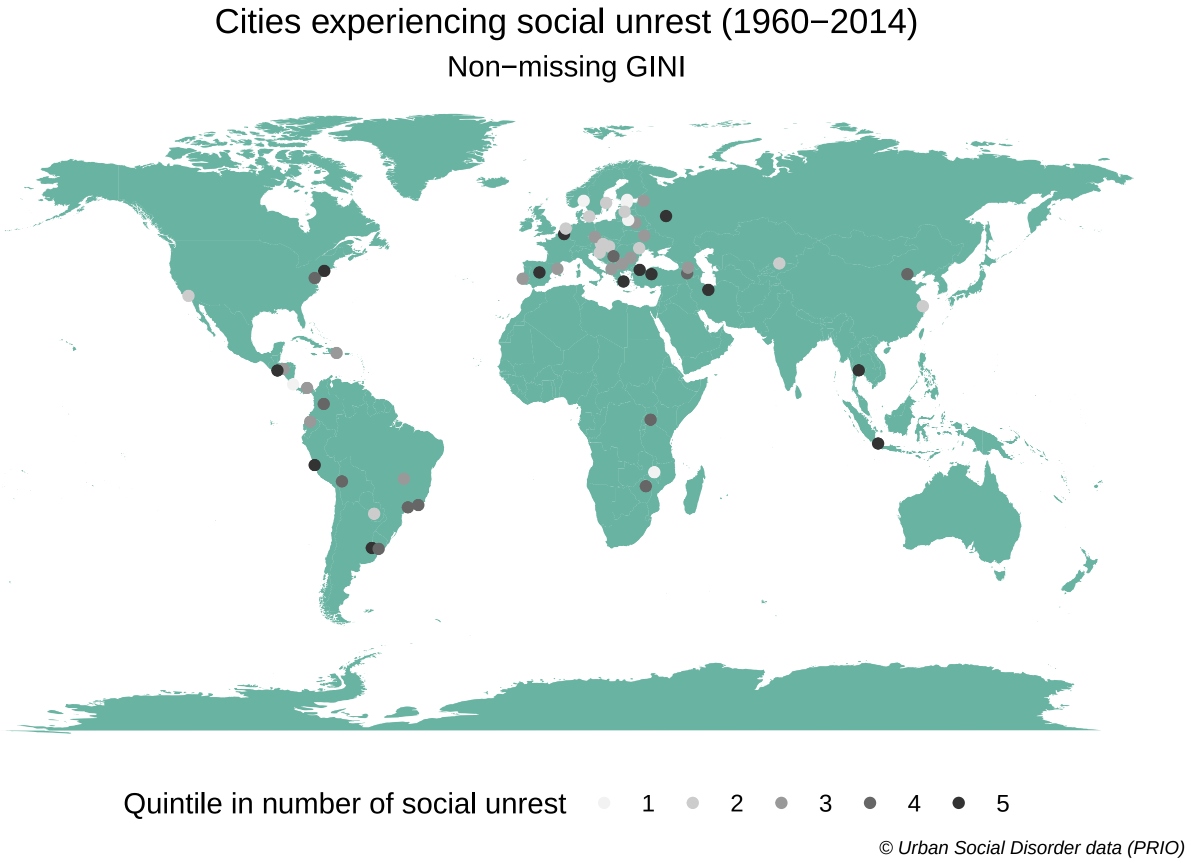
```
## [[1]]
##  [1] "United Arab Emirates" "Australia"            "Australia"
##  [4] "Austria"              "Belgium"              "Bulgaria"
##  [7] "Bahrain"              "Canada"               "Canada"
## [10] "Switzerland"          "Switzerland"          "Chile"
## [13] "Cuba"                 "Czech Republic"       "Germany"
## [16] "Germany"              "Denmark"              "Eritrea"
## [19] "Spain"                "Spain"                "Estonia"
## [22] "Finland"              "France"               "United Kingdom"
## [25] "Greece"               "Croatia"              "Hungary"
## [28] "Ireland"              "Israel"               "Italy"
## [31] "Japan"                "Korea, Rep."          "Kuwait"
## [34] "Lithuania"            "Latvia"               "Netherlands"
## [37] "Norway"               "New Zealand"          "New Zealand"
## [40] "Oman"                 "Poland"               "Portugal"
## [43] "Qatar"                "Romania"              "Russian Federation"
## [46] "Russian Federation"   "Saudi Arabia"         "Singapore"
## [49] "South Sudan"          "Slovak Republic"      "Sweden"
## [52] "Uruguay"              "United States"        "United States"
## [55] "United States"        "Venezuela, RB"        "Yemen, Rep."
## [58] "Yemen, Rep."
```

That being said, the point here is not to develop an entire strategy based on the ODA variable. A simple construction that is a binary equal to zero whenever the country does not receive any ODA and equal to one otherwise is easily constructed.

```
### Build a binary equal to zero when no (or zero) ODA is reported.
overall_data <- overall_data %>%
  dplyr::mutate(some_ODA = ifelse(is.na(DT_ODA_ODAT_GN_ZS), 0, DT_ODA_ODAT_GN_ZS),
                receive_ODA = ifelse((some_ODA > 0), 1, 0))
```

A similar exercise could have been done for the GINI variable. One would find that mostly developed countries have information on GINI (the map below shows this). Therefore, a study focusing on GINI here would in fact look at inequalities among developed countries and its association with social unrest.

```
### Define the subset without missing GINI information
overall_data_no_miss_GINI <- overall_data[!(is.na(overall_data$SI_POV_GINI)), ]
### Map (as before)
ggplot() +
  geom_polygon(data = world_map, aes(x = long, y = lat, group = group),
               fill="#69b3a2", col=NA) +
  geom_point(data= overall_data_no_miss_GINI, aes(x = LONG, y = LAT,
                                         colour = quintile_NEVENTS)) +
  scale_color_manual(name = "Quintile in number of social unrest",
                     breaks = c("1", "2", "3", "4", "5"),
                     values=c("gray95", "gray80", "gray60", "gray40", "gray20")) +
  labs(title = "Cities experiencing social unrest (1960-2014)",
       subtitle = "Non-missing GINI",
       caption = "© Urban Social Disorder data (PRIO) ",
       fill = "Score") +
  theme_void() +
  theme(plot.caption = element_text(size = 7, face = "italic"),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "bottom")
```



Cities experiencing social unrest (1960−2014)

Non−missing GINI

© Urban Social Disorder data (PRIO)

Apart from these variables, one can consider that all other variables can be used, whatever the narrative, to conduct over the analysis.

1.6. Research question: how is the propensity to have social disorders in a city related with the city population and labor market characteristics?

1.7. Different variables to be tested here (as an illustration), and there expected effects

Let us assume that we are interested in some aspects of society that characterize economic inclusion. We also want to control for the population of the city, as this can affect both the number and the magnitude of the social unrests (respectively the extensive margin and the intensive margin) and this can affect the labor market as well. Indeed, a large city shows more diversification in jobs, it tends to be more inclusive for women, migrants. In a nutshell, the size of the city can be correlated to both and not including it puts us at a high risk of endogeneity, where $E[\varepsilon|X] \neq 0$. **Note here that this point is mostly raised for the sake of the exercise; obviously, many variables are omitted, leading to biased results anyway (not even talking yet about the reserve causality issue in the present exercise).**

- **Social disorders:** This variable gives the count of social unrest that occurred in cities over the period of interest (1960-2014). These events are widely defined, from spontaneous demonstrations to General Warfare (see the codebook, Thomson et al. (2022). This large meaning leads to some concern in the interpretation. Can such different events all be considered as depending on the same types of variables? Certainly not. This would call for further investigation.

- **Receiving ODA or not:** So far, we have exposed the willingness to use the ODA variable, as a binary variable that can capture (i) the level of development, (ii) the solidity of the welfare state. A strong welfare state can be assumed to push to lower violence within the country, as inequalities are expected to be lower, the army is probably less powerful as compared to the police (and the police, relative to the army, tends to lower the stress in urban disorder). On the other hand, a strong welfare state might be assumed to allow more freedom to demonstrate, which can lead to important social disorders (e.g., Paris riots in 2005).

Note that this is a pure hypothesis for the sake of the present mock analysis. One would need to back such statement with literature for a sound analysis.

- **City population:** The WUP database provides city population. One would expect more demonstrations and possibilities to have social unrests in big cities (even though one could argue that big cities are also better prepared to disorder and thus better able to *prevent* them). **Therefore, the first working hypothesis is the following:**

$$Population_{city} \uparrow \iff Social\ Unrest \uparrow$$

- Among the labor-oriented variables, we want to see the role played by unemployment. In developed countries, high unemployment might be a signal of a gloomy economy, which can in turn imply a higher degree of dissatisfaction. On the other hand, one might expect a tamed effect in developing countries, where the informal labor market captures a large share of the economic activity and therefore where the unemployed workers tend not to appear in the statistics as unemployed. Therefore, in developing countries, a large share of unemployed active people captures two opposite forces: (i) the one identical to developed countries (thus expected a positive relationship), (ii) an opposite effect due to a positive signal in terms of formal inclusion of economic activities (thus expecting a positive relationship). **All in all, one might expect the following second working hypothesis:**

$$Unemployment_{Rate} \uparrow \iff Social\ Unrest \uparrow |Developed\ Countries$$

$$Unemployment_{Rate} \uparrow \iff Social\ Unrest \updownarrow |Developing\ Countries$$

- Other variables are all meant to capture some of the variation that could occur in the $E[\varepsilon|Unemployment_{Rate}] \neq 0$. In order to capture some discrepancies across genders, one might be interested in the difference in mortality rate across genders. Here, we construct the difference using the women mortality rate as the reference. Hence, negative values indicate a higher men mortality and a large positive value a large women extra mortality rate relative to men.

The table below aims to provide details on the different variables

```r
### Rename and build necessary variables
overall_data <- overall_data %>% # rename some variables
  dplyr::rename(unemployment_rate = SL_UEM_TOTL_ZS,
               LFP = SL_TLF_CACT_ZS,
               Share_services  = SL_SRV_EMPL_ZS,
               Share_agriculture = SL_AGR_EMPL_ZS)
overall_data <- overall_data %>% # create the gender gap in mortality
  dplyr::mutate(gap_mortality_rate = SP_DYN_AMRT_FE - SP_DYN_AMRT_MA)
```

Variable	Codename	Datasource	Units of observation	Year(s)
Number of social disorders (count of disorders)	NEVENTS	USD	City	1960-2014
Whether the country receives ODA (binary)	receive_ODA	WDI	Country	2019
Population (in thousand)	population_city	WUP	City	2018
Unemployment rate (%)	unemployment_rate	WDI	Country	2019
Labor force participation (% of active agents over 14)	LFP	WDI	Country	2019
Gender gap in mortality rate (F-M, aged 15-60, rates in ‰)	gap_mortality_rate	WDI	Country	2019
Share of employments in services (% of total employment)	Share_services	WDI	Country	2019
Share of employments in agriculture (% of total employment)	Share_agriculture	WDI	Country	2019

Last, one might want to explore whether there is heterogeneity across regions in the relationships among the variables of interest. Although the map did not indicate a clear pattern of social disorders, exploring it statistically might deliver interesting results, as the labor market is different across regions: informality is larger in Sub Saharan Africa (SSA), Middle East and North Africa (MENA), Latin America and it tends to be lower in Europe (for example - **here again, one would need sources for a fully-fledged analysis**).

```
### Create region dummies
overall_data$Asia<-ifelse(overall_data$REGION=="Asia", 1, 0)
overall_data$Europe<-ifelse(overall_data$REGION=="Europe", 1, 0)
overall_data$LatAm<-ifelse(overall_data$REGION=="Latin America", 1, 0)
overall_data$MENA<-ifelse(overall_data$REGION=="Middle East and North Africa", 1, 0)
overall_data$NorAm<-ifelse(overall_data$REGION=="North America", 1, 0)
overall_data$Oceania<-ifelse(overall_data$REGION=="Oceania", 1, 0)
overall_data$SSA<-ifelse(overall_data$REGION=="Sub-Saharan Africa", 1, 0)
```

2. Methods and limitations

2.1. Ordinary Least Squares (OLS)

The present analysis is a simple quantitative analysis using linear regressions. Several linear regressions, trying to develop a narrative around the research question, will be studied. OLS estimations are convenient for their simplicity and their ease to interpret the results. When the data fulfills the basic Gauss-Markov assumptions (Linearity, no linear relationships across explanatory variables, spherical errors, strict exogeneity), OLS is the *Best Linear Unbiased Estimator (BLUE).*

2.2. Limitations

Although extremely appealing, assumptions for OLS to be BLUE are strong and hard to meet in reality. In particular, the last assumption is difficult to statistically test. In the context of the data used here, **reverse causality**, which is one key issue of endogeneity, is of large risk. Indeed, the explained variable is *prior* to the explanatory variables. Obviously, such an analysis cannot be seriously conducted from a theoretical perspective, but it cannot be from a statistical perspective either because of the reserve causality. Nevertheless, the macro variables used as explanatory variables might not be too volatile, so that the 2018-2019 variables might not have been very different 40 years ago (which is a very strong assumption).

Reverse causality is not the only problem for endogneity. **Measurement errors** and **omitted variable biases** can lead to a link between explanatory variables and errors ($E[\varepsilon|X] \neq 0$). Most aggregates in WDI and WUP are approximations leaving space for sizable measurement errors. Furthermore, given the small set of data provided (but anyway, given the research question), it would be extremely tedious to convince that there are no omitted variable biases. A very useful tool here is the use of **fixed effects**, but that is beyond the scope of the class (and fixed effects can virtually wipe out all variance in the model, making the model clean of endogeneity, but also clean of any substantial relationships).

In any case, even assuming that endogeneity is not an issue (although it surely is, given the data used), there is no reason to go beyond a correlation analysis to a *causal* analysis. Let us keep being cautious in **not conveying that the relationships established can be causal**.

Lastly, most explanatory variables are at the country level while the explained variable is expressed at the city level. This implies that, apart from the city population variable, two cities from the same country will be expected to have the exact same number of social disorders. This is a clear limitation: it leaves to a lack of variance that is not excessively worrying because one country in the data does not have too many cities included (Brazil has three cities included and this is the maximum number of occurrences).

```
### Count occurrences in countries
max(table(overall_data$CountryName))
```

```
## [1] 3
```

Because social disorders in a city are unlikely to affect the entire labor "market", it seems plausible to have an underlying plausible causal channel from the explanatory variables to the explained variable. Nevertheless, results here will not risk implying causality.

```
### Select the variables of interest
selected_data <- overall_data[, c("NEVENTS", "receive_ODA", "population_city",
                                  "unemployment_rate", "LFP" , "gap_mortality_rate",
                                  "Share_services", "Share_agriculture", "Asia",
                                  "Europe", "LatAm", "MENA", "NorAm",  "Oceania",
                                  "SSA", "CITY" )]
dim(selected_data)
```

```
## [1] 168  16
```

```
length(unique(overall_data$CountryName))
```

```
## [1] 145
```

Our final data set, used in the quantitative analysis, has 168 cities and 145 different countries. Although not all countries are including, the geographical spread tends to cover most parts of the world, although African countries might be underrepresented.

3. Descriptive Statistics

3.1. Basic descriptives

The table below displays basic descriptive statistics for the variables used in the present analysis. The number of events ranges from 0 to 472. Cities recording the largest amount of unrest events are Baghdad (472), London (359), Beirut (336), Paris (282), and Karachi (277). Interestingly, (extended) MENA and Western Europe seem to concentrate most social unrest. However, it is likely that the *type* (and the gravity) of unrests differ from war settings in the MEAN to demonstrations in Western Europe. Cities included in the Urban Social Disorder data set have on average 76.73 events recorded but the median value is lower (50.50), which indicates that the data is positively skewed. This is confirmed by the large coefficient of skewness (2.00). This large right skewness may justify the use of log-transformation. However, it is well-known that $ln(0) \rightarrow Undefined.$

In order to deal with this issue, some people use a problematic transformation: $log(x) \approx log(x+1)$. This might not change the data dramatically as long as the variable x is very large. Here, the range of $NEVENTS$ goes from 0 to 472 events, so that the range is not too large, and this modification might be problematic. That being said, one can easily notice that only one city records no social disorder (Astana) and thus the unrigorous modification should not be too problematic. Still, in order to encounter a new tool, let us use the *inverse hyperbolic sine transformation (IHS)*. Formally, it is written as: $\tilde{x} = arcsin(x) = ln(x + \sqrt{x^2+1})$. It is useful to approximate log-transformation whenever the variable contains zero-valued observations, as is the case in the present analysis. Now, the main question is about the interpretation. A nice property of log-transformation is also in its interpretation in terms of growth rates. Bellemare and Wichman (2020) indicate that for large enough values of the transformed variable (when used as explained variable), the interpretation of the log-transformation remains. For simplicity, we will assume here that $NEVENTS$ has large enough values so that we can use the IHS transformation and interpret it as a log-transformation. *This is a strong assumption, and the subsequent models, using log-transformations for some explanatory variables add to the complexity of the accurate interpretation. Nevertheless, we can consider that the imprecision is minor in the present context of the data.*

About 63% of all cities included in the data set are in ODA receiving countries. These cities tend to be large, with an average population of close to 5 million inhabitants. Given that the median (slightly more than 2 million inhabitants) is less than half the average population, there is a large positive skewness, driven by enormous cities. The largest city has more than 37 million inhabitants (Tokyo) with a 7 million inhabitants' gap with the second biggest city in the data set (Delhi). Therefore, implementing a log-transformation of this variable can be perceived as a promising step to have more normal-like distribution (which can be useful to implement the central limit theorem when interpreting the inference tools).

The rate of unemployment of the countries in the data is on average 6.93%. Note that all country-specific variables here are distorted because some countries appear several times, hence for Brazil, its unemployment rate is incorporated three times in the computation of the summary statistics. Rates of unemployment range from (almost) 0 in Qatar to 28% in South Africa. As for the previous variable, it is also positively skewed. By contrast, Labor Force Participation is not skewed and ranges from 32% of active individuals taking part of the labor market (in Djibouti) to 88% (in Qatar). The constructed variable *Gap mortality rate* has a negative average of -63.37‰ (in fact all observations are negative) indicating that mortality of men between 15 and 60 exceeds women mortality by 63.37‰. The last variables to include in the enriched model are the share in employment in services and in agriculture. The larger the share of services the more developed the economy. The share of services in total employment is on average 55.36% but ranges from 10% (in Burundi) to 84% (in Singapore). On the other side, the share of agriculture expresses the mirror image of the share in services. The range goes from Singapore with virtually no workers in agriculture to 86% in Burundi.

Last, the region dummies are simply informative on the share of cities that are located in these regions, as such, African cities account for 24% of the sample, which is in fact quite large. Europe, as compared to Northern America but also as compared to its population share in the world population, is rather over-represented.

```r
### Simple descriptive statistics
  # Create a dataframe with only numerical data from descriptive statistics
selected_num_data <- selected_data[,!names(selected_data) %in% c("CITY")]
  # Function for summary, put NA out, and round accordingly
my.summary <- function(x,...){
  c(Mean= format(round(mean(x, na.rm = TRUE, ...), 2), nsmall = 2),
    SD= format(round(sd(x, na.rm = TRUE, ...), 2), nsmall = 2),
    Median= format(round(median(x, na.rm = TRUE, ...), 2), nsmall = 2),
    Min= format(round(min(x, na.rm = TRUE, ...), 0), nsmall = 0),
    Max= format(round(max(x, na.rm = TRUE,...), 0), nsmall = 0),
    Skewness = format(round(skewness(x, na.rm = TRUE,...), 2), nsmall = 2),
    Kurtosis = format(round(kurtosis(x, na.rm = TRUE,...), 2), nsmall = 2),
    Count = format(round(sum(!is.na(x)), ...), 1))
}
##" Output
  # t() is the transpose to get one line per variable
  # noquote() gets rid off the quotation marks
  # sapply is function applying a defined function
    # to the called vector (here a dataframe)
noquote(t(sapply(selected_num_data, my.summary)))
```

```
##                    Mean    SD      Median  Min  Max   Skewness Kurtosis Count
## NEVENTS            76.73   77.62   50.50   0    472   2.00     7.87     168
## receive_ODA        0.63    0.48    1.00    0    1     -0.54    1.29     168
## population_city    4740.92 6228.57 2144.00 320  37393 2.43     9.59     168
## unemployment_rate  6.93    5.41    5.06    0    28    1.65     5.94     168
## LFP                61.31   11.18   62.15   32   88    -0.10    2.85     168
## gap_mortality_rate -63.37  33.50   -56.06  -164 -9    -0.80    3.10     150
## Share_services     55.36   17.53   57.99   10   84    -0.44    2.25     168
## Share_agriculture  24.58   21.77   18.49   0    86    0.86     2.78     168
## Asia               0.21    0.41    0.00    0    1     1.39     2.94     168
## Europe             0.23    0.42    0.00    0    1     1.31     2.71     168
## LatAm              0.14    0.34    0.00    0    1     2.11     5.46     168
## MENA               0.12    0.33    0.00    0    1     2.27     6.14     168
## NorAm              0.03    0.17    0.00    0    1     5.53     31.63    168
## Oceania            0.02    0.15    0.00    0    1     6.25     40.02    168
## SSA                0.24    0.43    0.00    0    1     1.19     2.42     168
```

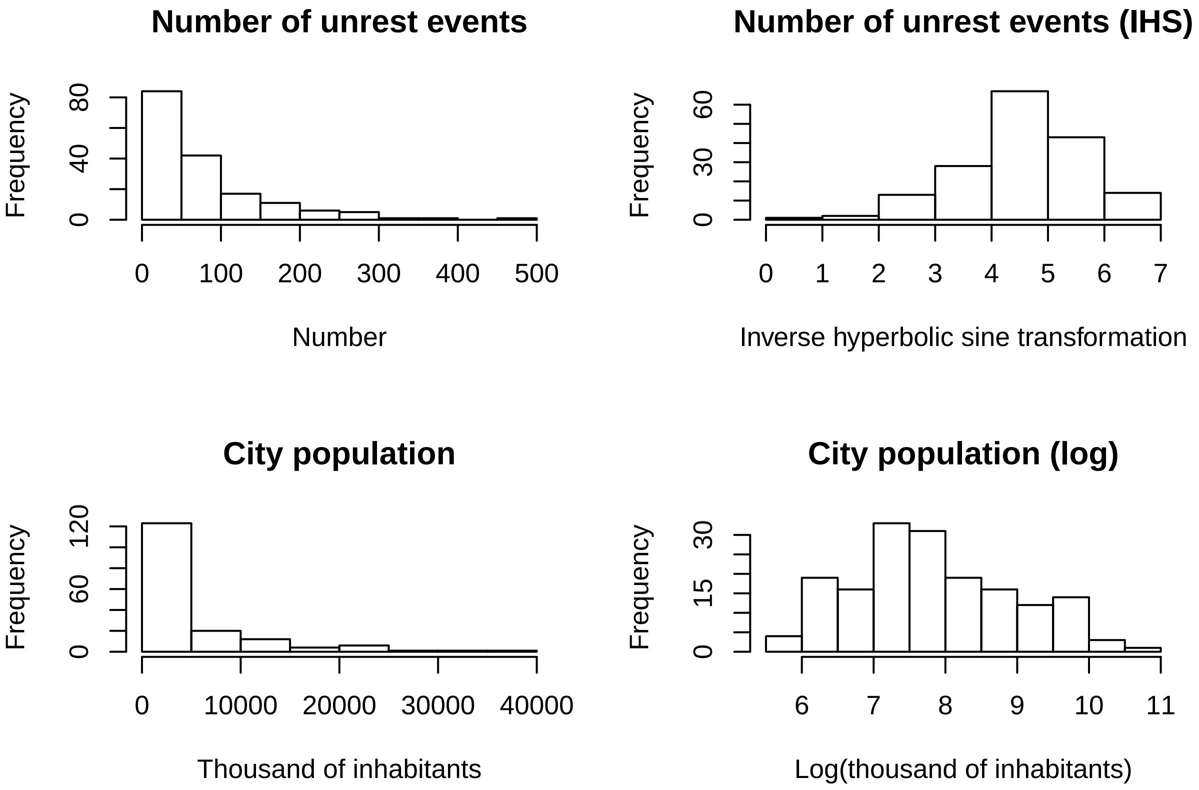
3.2. Transform the variables

The discussed variables' transformations are performed here:

```
### Inverse hyperbolic sine and log transformations
selected_data$NEVENTS_ihs <- asinh(selected_data$NEVENTS)
selected_data$NEVENTS_ihs <- log(selected_data$NEVENTS + sqrt(selected_data$NEVENTS^2+1))
selected_data$log_population_city <- log(selected_data$population_city)
```

One probably wants to see how does this change the distributions of the two transformed variables:

```
### Histogram of pre-post transformed variables
par(mfrow=c(2,2)) # display two by two
  # Number of unrest events
hist(selected_data$NEVENTS,
     main = "Number of unrest events",
     xlab = "Number")
hist(selected_data$NEVENTS_ihs,
     main = "Number of unrest events (IHS)",
     xlab = "Inverse hyperbolic sine transformation")
  # Population in cities
hist(selected_data$population_city,
     main = "City population",
     xlab = "Thousand of inhabitants")
hist(selected_data$log_population_city,
     main = "City population (log)",
     xlab = "Log(thousand of inhabitants)")
```



Clearly, the variables are more compact and less skewed. The descriptive statistics below confirm this. Nevertheless, it appears that the kurtosis of the IHS number of social unrests remain large, indicating that tails of the distribution are thick. This can be worrying, as the variance might not be finite in such cases (which is necessary to implement the central limit theorem). However, the standard deviation is equal to 1.1 IHS number of events, which does not seem to be extremely large in the present context of the data.

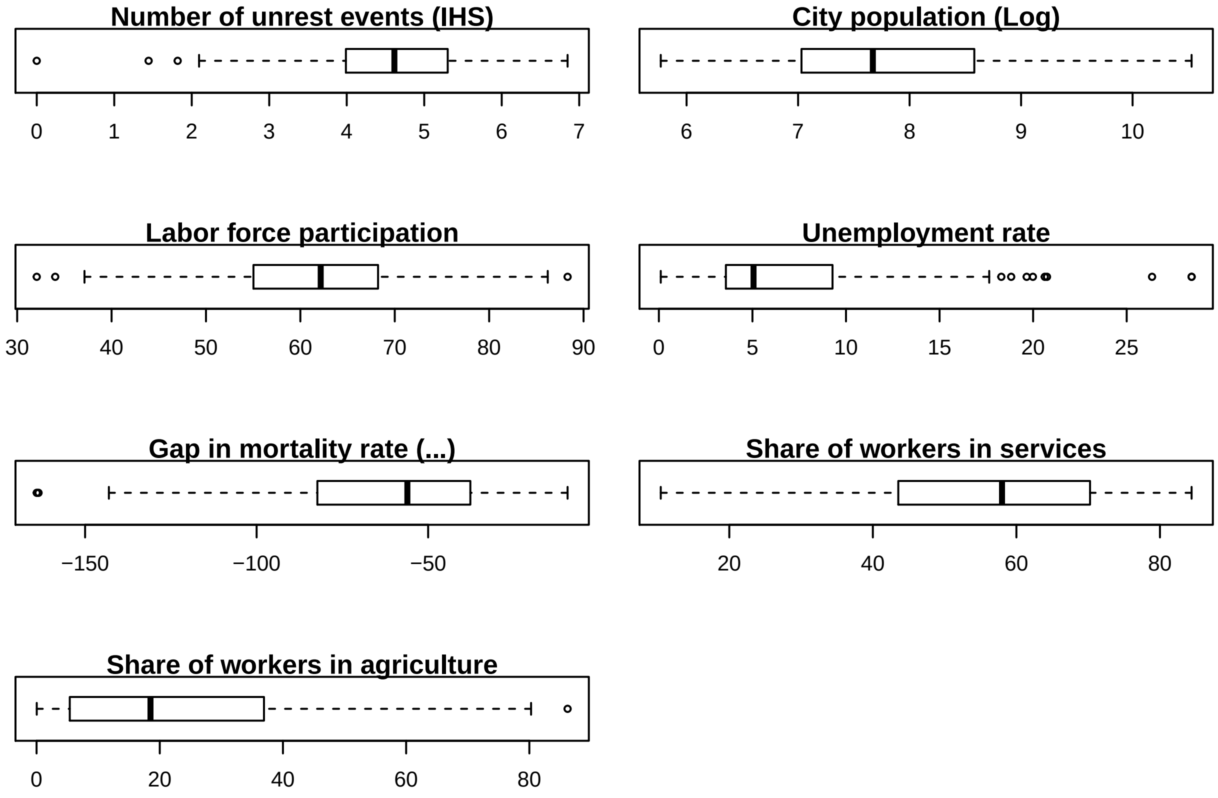
```
### Quick look at the transformed descriptive statistics
transformed_var <- selected_data[, c("NEVENTS_ihs", "log_population_city")]
noquote(t(sapply(transformed_var, my.summary)))
```

```
##                     Mean SD   Median Min Max Skewness Kurtosis Count
## NEVENTS_ihs         4.53 1.11 4.62   0   7   -0.71    4.07     168
## log_population_city 7.82 1.11 7.67   6   11  0.34     2.37     168
```

3.3. Check for outliers

Last, it is important to verify that there are no worrying outliers in each variable used. It appears that the transformation of the $NEVENTS$ variable is not sufficient in avoiding outliers, as three observations appear to be beneath the lower limit in the box plot. Many of the variables used have outliers: LFP has outliers on both sides of the distribution, unemployment rate, gap in mortality and the share of workers in agriculture have outliers on one side only. Technically, this should be discussed at greater lengths, as these outliers might be too influential in the forthcoming regressions. Nevertheless, although the outliers are indeed away from the upper/lower limits of the box plots, they do not appear to be extremely far. Maybe the case of Astana that has zero social unrest event and that might have in fact been dropped from the start - which would have prevented us the use of IHS transformation instead of log-transformation - is problematic, along which countries with extremely large unemployment rates (South Africa and Djibouti). Here, we take the stance to continue the analysis without dropping these variables or changing them, as they do not seem to be outliers due to measurement errors.

```
### Produce boxplots
#(not useful for the binary, given the descriptive statistics shown above)
par(mfrow = c(4,2), mar = c(5, 1, 1.0, 1)) # change margins to fit in
boxplot(selected_data$NEVENTS_ihs,
        main = "Number of unrest events (IHS)",
        horizontal=TRUE)
boxplot(selected_data$log_population_city,
        main = "City population (Log)",
        horizontal=TRUE)
boxplot(selected_data$LFP,
        main = "Labor force participation",
        horizontal=TRUE)
boxplot(selected_data$unemployment_rate,
        main = "Unemployment rate",
        horizontal=TRUE)
boxplot(selected_data$gap_mortality_rate,
        main = "Gap in mortality rate (%)",
        horizontal=TRUE)
boxplot(selected_data$Share_services,
        main = "Share of workers in services",
        horizontal=TRUE)
boxplot(selected_data$Share_agriculture,
        main = "Share of workers in agriculture",
        horizontal=TRUE)
```

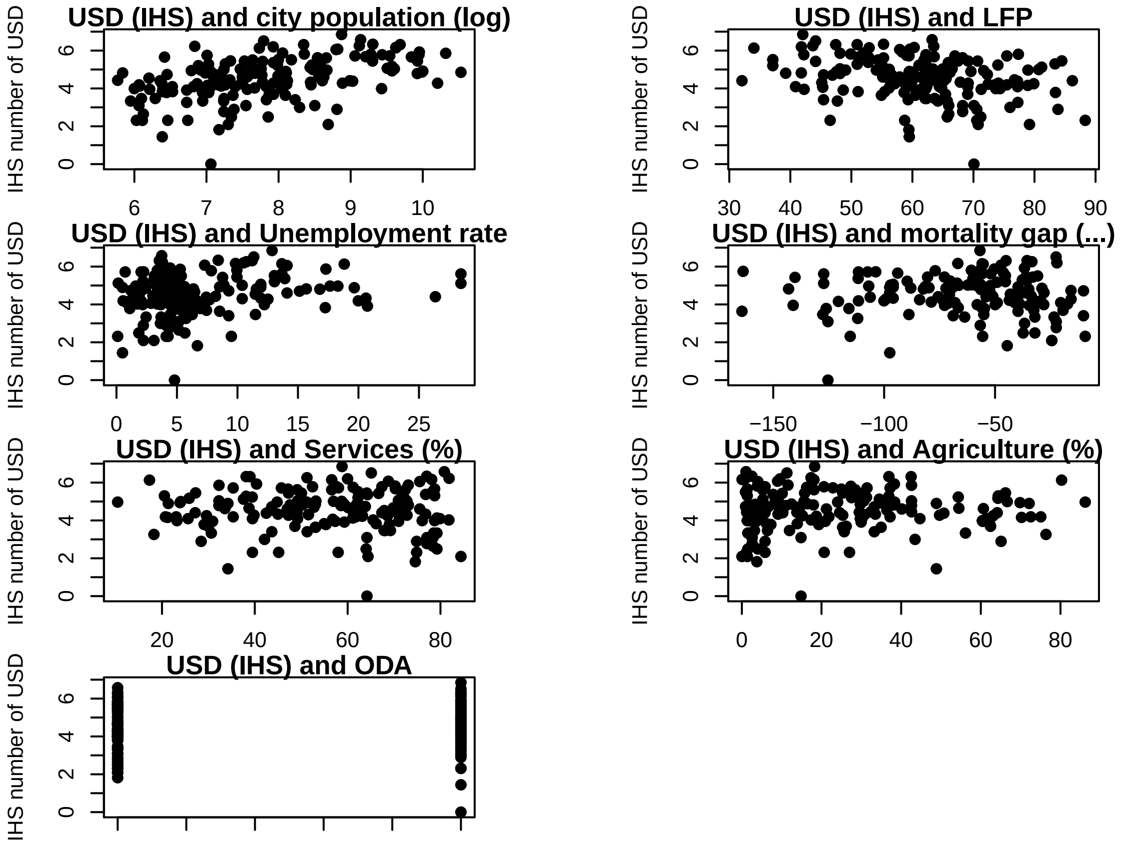


3.4. Joint distributions

Now, we can consider that we are well-aware of the variable distributions and characteristics. Now, we might be interested in knowing more about the *joint distributions*. This is important for two reasons: (i) check the associations between explanatory variables and the explained variable, (ii) check the associations among explanatory variables.

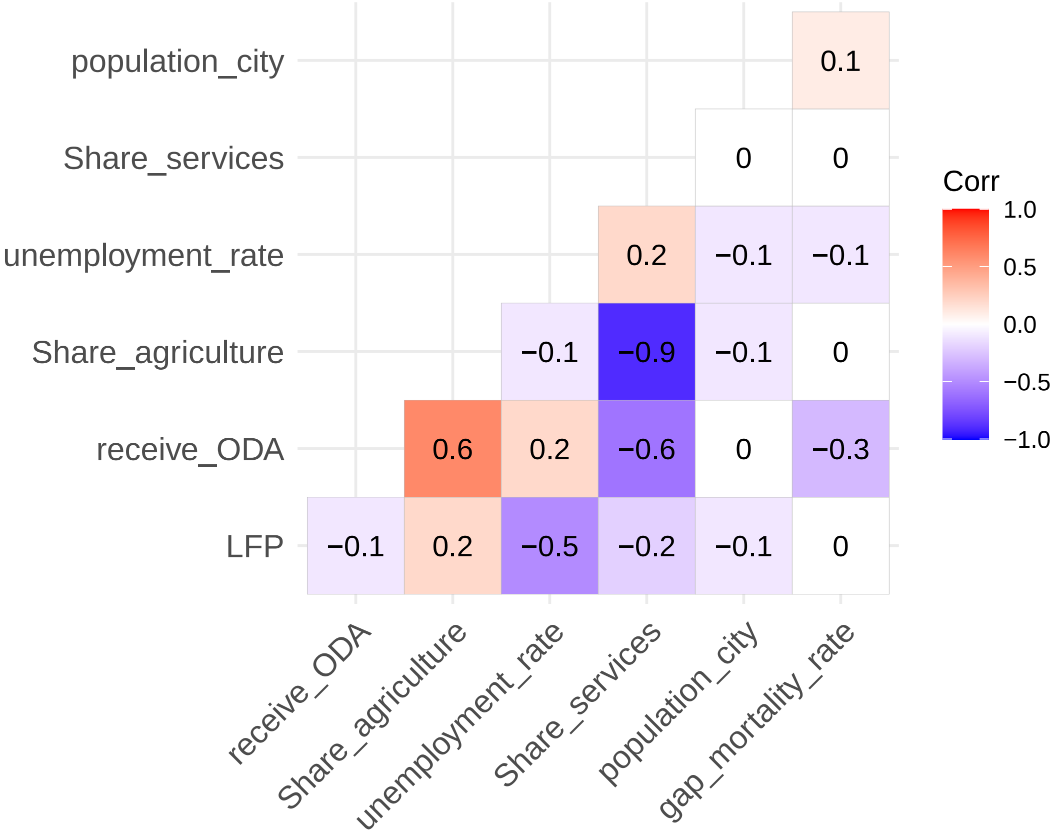
```r
### Joint distributions Y versus X variables
par(mfrow = c(4,2), mar = c(2, 5, 1.0, 5))
plot(selected_data$log_population_city, selected_data$NEVENTS_ihs,
     pch = 19, col = "black",
     main = "USD (IHS) and city population (log)",
     ylab = "IHS number of USD",
     xlab = "City population (log)")
plot(selected_data$LFP, selected_data$NEVENTS_ihs,
     pch = 19, col = "black",
     main = "USD (IHS) and LFP",
     ylab = "IHS number of USD",
     xlab = "Labor force participation")
plot(selected_data$unemployment_rate, selected_data$NEVENTS_ihs,
     pch = 19, col = "black",
     main = "USD (IHS) and Unemployment rate",
     ylab = "IHS number of USD",
     xlab = "Unemployment rate")
plot(selected_data$gap_mortality_rate, selected_data$NEVENTS_ihs,
     pch = 19, col = "black",
     main = "USD (IHS) and mortality gap (%)",
     ylab = "IHS number of USD",
     xlab = "Mortality gap (Women - Men), in %")
```

```
plot(selected_data$Share_services, selected_data$NEVENTS_ihs,
     pch = 19, col = "black",
     main = "USD (IHS) and Services (%)",
     ylab = "IHS number of USD",
     xlab = "Share of workers in services")
plot(selected_data$Share_agriculture, selected_data$NEVENTS_ihs,
     pch = 19, col = "black",
     main = "USD (IHS) and Agriculture (%)",
     ylab = "IHS number of USD",
     xlab = "Share of workers in agriculture")
plot(selected_data$receive_ODA, selected_data$NEVENTS_ihs,
     pch = 19, col = "black",
     main = "USD (IHS) and ODA",
     ylab = "IHS number of USD",
     xlab = "The country receives ODA (binary)")
```



Scatter plots show a positive linear relationship between (IHS) number of USD and (Log) city population: as expected (first working hypothesis), the larger the city the larger the number of USD. The relationship between the explained variable and unemployment rate is less striking but seems to display a positive relationship. Given the joint distribution, it might be a good idea to log-transform the unemployment rate variable. This might be even more necessary because of the large skewness of the unemployment rate variable exposed earlier. A rigorous analysis would probably use the log-transformation in such a case. The last explanatory variable that shows a clear pattern in its joint distribution with the (IHS) number of USD is the labor force participation: the higher the LFP, the lower the (IHS) number of USD. This is intuitive: unemployment rates and LFP might indeed be inversely associated with different variables. Regarding other variables, the joint distributions do not show striking patterns, these variables the explained variable seem to be close to independent. Scatter plots with binary variables seldom are a promising idea, as it is hard to read. From such a plot, one cannot conclude on an association or an absence of association.

```r
### Build correlation table across explanatory variables
  # X variable dataframe
selected_X_var <- selected_data[, c("receive_ODA", "population_city",
                                    "unemployment_rate", "LFP" ,
                                    "gap_mortality_rate", "Share_services",
                                    "Share_agriculture")]
  # Compute the correlations using all available observations
corr <- round(cor(selected_X_var, use = "pairwise.complete.obs"), 1)
  # display the table
ggcorrplot(corr,
           hc.order = TRUE,
           type = "lower",
           lab = TRUE)
```



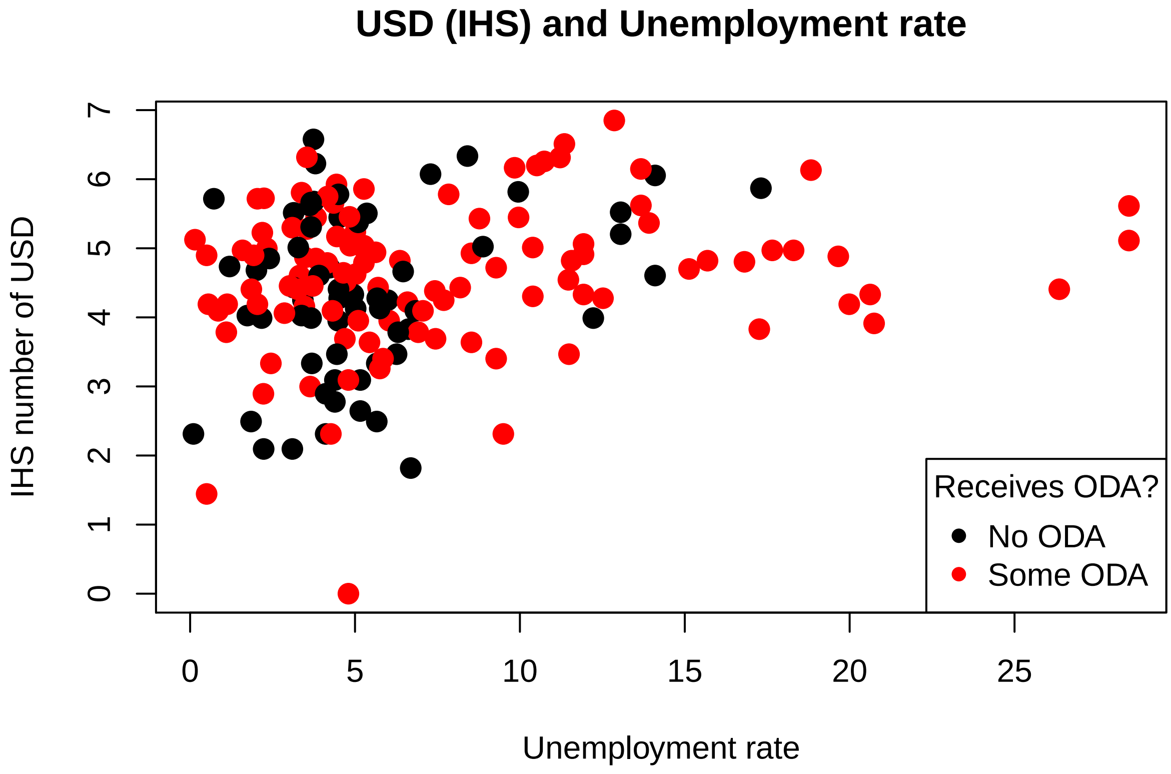
It is also important to verify that linear correlations across explanatory variables are not too high. Otherwise, the rank of the matrix composed of the explanatory variables might not be full because of the linear relationships among them. This would prevent the ordinary least squares from being run. Looking at the table produced above, the shares of workers in agriculture and services are very closely (linearly and inversely) related: the Pearson correlation coefficient is equal to -0.9. As mentioned in the descriptive statistics, this is obvious, and caution will be required when using these two variables as controls. As such, it even seems clear that one needs to choose to include only one of these two shares. Whether the country receives ODA or not is also closely related to the types of economic activities in the country: the more rural the country, the more likely it is to receive ODA, as the transition toward more industrial products and then to services has not been achieved yet. As mentioned earlier, the rate of unemployed and LFP are closely (and inversely) related as well. When running a model including all variables, it will therefore be extremely important to check for multicollinearity.

So far, descriptive statistics have allowed us to bring to an initial visual confirmation that the first working hypothesis (the larger the city the more social unrests) but this basic analysis cannot bring to conclusions, yet, on the role of receiving ODA.

3.5. Joint distributions by ODA

We initially hypothesized that the relationship between unemployment rate and the (IHS) number of USD depends on whether the country receives ODA (as a proxy for the level of development, the quality of the welfare state and of the institutions in the country). Replicating the joint distribution by whether the country receives ODA helps provide a first guess about the second working hypothesis formulated. Indeed, on can see the the positive relationship between unemployment and USD is stronger for developed countries, those that do not receive ODA; for ODA receivers, the relationship is in fact less striking: it might remain positive but that is not obvious visually.

```
### Scatter plots by ODA
plot(selected_data$unemployment_rate, selected_data$NEVENTS_ihs,
     pch = 16, cex = 1.5, col = factor(selected_data$receive_ODA),
     main = "USD (IHS) and Unemployment rate",
     ylab = "IHS number of USD",
     xlab = "Unemployment rate")
  # Legend
  legend("bottomright",
         title = "Receives ODA?",
         legend = c("No ODA", "Some ODA"),
         pch = 16,
         col = factor(levels(factor(selected_data$receive_ODA))))
```



Overall, the simple descriptive statistics came to confirm that the two working hypotheses formulated are worth investigating further. This is why we are now exploring univariate and multivariate analyses. The descriptive statistics section also raised many potential issues that one needs to keep in mind in the subsequent analysis.

4. Model construction

Notation convenience is important here: let us denote i the country and j the city. In theory, the number of social unrests (IHS) should be denoted $IHS(USD_{ij})$ because it varies across countries and cities. However, anytime a variable varies across cities, it mechanically varies across countries because a city cannot belong to different countries. Therefore, in such cases, we will only use the index of the mostly varying unit.

4.1. Univariate model (model 1)

The univariate model that helps us further investigate the first working hypothesis is (with e_j the error term):

$$IHS(USD_j) = a + b_1 Log(Population_j) + e_j$$

Since we simplify the interpretation and consider the IHS-transformed variables are equivalent to Log-transformed, we have here a log-log model. This implies that the slope coefficient is an **elasticity:** $b_1 = \frac{\delta USD_j}{\delta Population_j} \frac{Population_j}{USD_j}$. This means that the slope coefficient captures the corresponding change in % of the explained variable when the explanatory changes by 1%.

```
### Univariate model
lm1 <- lm(NEVENTS_ihs ~ log_population_city, data = selected_data)
summary(lm1)
```

```
##
## Call:
## lm(formula = NEVENTS_ihs ~ log_population_city, data = selected_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1714 -0.4968  0.1468  0.6793  2.1615
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.82045    0.53984   1.520     0.13
## log_population_city  0.47454    0.06831   6.946  8.1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9838 on 166 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2205
## F-statistic: 48.25 on 1 and 166 DF,  p-value: 8.099e-11
```

The univariate model is in line with the first working hypothesis: when the city population is larger by 1%, the corresponding number of USD events is larger by 0.47%. The intercept here (0.82) would indicate the (IHS) number of USD events in an empty city, but this will then not be a city at all so that the interpretation of the intercept is not meaningful here. The positive relationship is statistically significant, as the t-value is very large (almost 7) which implies that the rejection area of such a t-value is virtually non-existent ($p-value \approx 0$). This implies that the elasticity between the city population and the number of social disorders (the slope coefficient) is statistically significant.

The univariate model performs relatively well, as about 22.52% of all the variance compared to a model with only the intercept is captured (and the F-stat is 48.25, far above all conventional critical value at conventional levels of significance). In a univariate model, $\sqrt{R^2} = \rho$ so that the correlation between the explained and the explanatory variables is $\rho = 0.47$. It is positive because the sign of the slope coefficient is positive as well. The linear correlation between the (IHS) number of USD and the (Log) population is +0.47, rather a strong one.

4.2. Model with interaction (model 2)

In order to pursue the analysis and investigate the second working hypothesis, one needs to study the relationship between unemployment rate and the explained variable. This analysis will look at the role of receiving ODA, which implies interacting the unemployment rate with the ODA binary variable. This will allow us to decipher whether there is a discrepancy between the two groups of countries, as the joint scatter plot by ODA suggested. The model becomes:

$$IHS(USD_j) = a + b_1 Log(Population_j) + b_2 Unemployment\ Rate_i + b_3 ODA_i + b_4 Unemployment\ Rate_i \times ODA_i + e_j$$

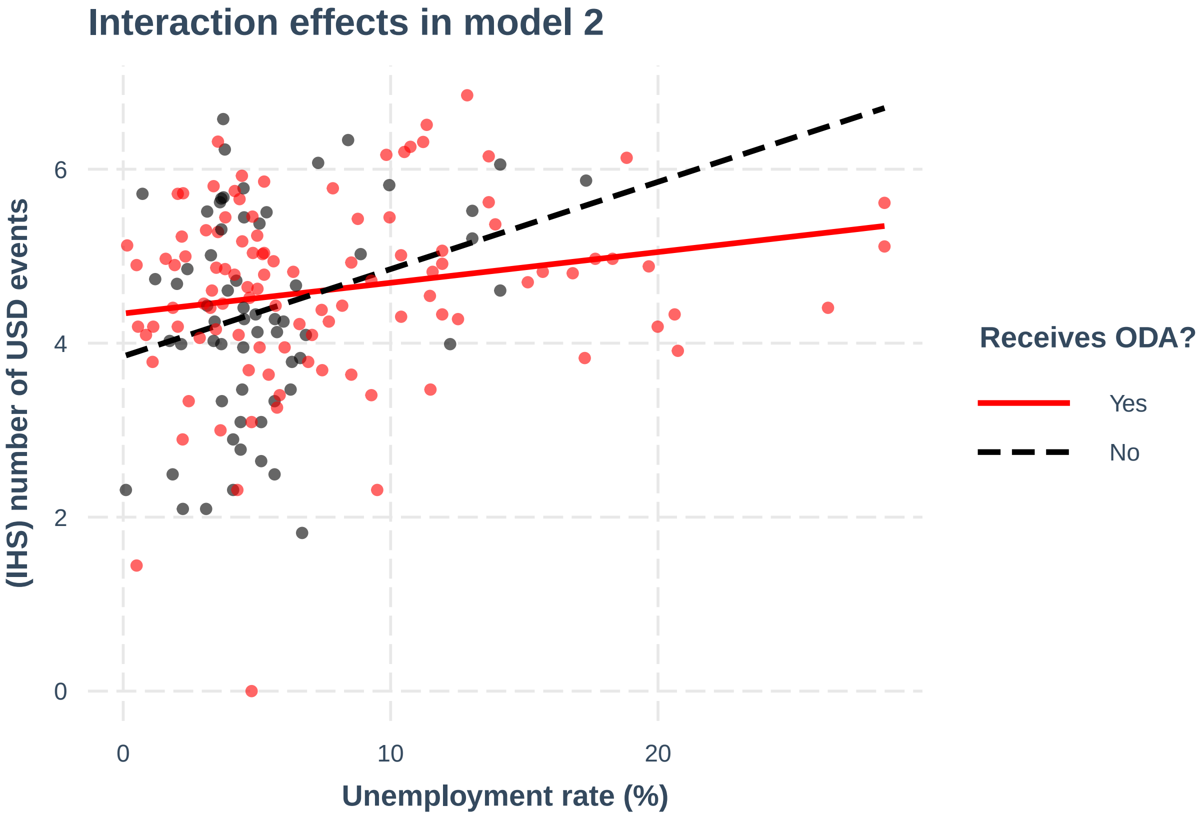
```
### Baseline interaction model
lm2 <- lm(NEVENTS_ihs ~ log_population_city + unemployment_rate * receive_ODA, data = selected_data)
summary(lm2)
```

```
##
## Call:
## lm(formula = NEVENTS_ihs ~ log_population_city + unemployment_rate *
##     receive_ODA, data = selected_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1474 -0.5271  0.1252  0.5720  2.4666
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.12017    0.55894   0.215  0.83005
## log_population_city            0.47660    0.06597   7.225 1.83e-11 ***
## unemployment_rate              0.10028    0.03555   2.821  0.00538 **
## receive_ODA                    0.49227    0.27128   1.815  0.07142 .
## unemployment_rate:receive_ODA -0.06495    0.03864  -1.681  0.09466 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.948 on 163 degrees of freedom
## Multiple R-squared:  0.2935, Adjusted R-squared:  0.2762
## F-statistic: 16.93 on 4 and 163 DF,  p-value: 1.254e-11
```

Adding the interaction of ODA and unemployment hardly changes the relationship between (Log) population and (IHS) number of USD events, both in terms of magnitudes and statistical significance ($b_1 = 0.4766$ with a $p - value \approx 0$). Receiving ODA seems to be positively related to the (IHS) number of USD events, as $b_3 = 0.4923$, so that receiving ODA (as compared to not receiving it) corresponds to a larger number of USD by 49.23% ($= 100 * 0.4923$), which is quite sizable.

Regarding the role of unemployment rate, we can partially confirm the second working hypothesis: for non-ODA receiving countries, an increase in unemployment rate by 1 percentage point corresponds to an increase in the number of USD by 10.03%, which is again quite sizable. On the other hand, an increase in unemployment rate by 1 percentage point in ODA-receiving countries corresponds to an increase in the number of USD by 3.53% ($\frac{\delta USD_j}{\delta Unemployment\ Rate_i}|_{ODA=1} = (b_2 - b_4) = 0.10028 - 0.06495$). In both cases, unemployment rate and (IHS) number of USD are positively related, but the relationship is largely tuned for ODA-receiving countries. The conflicting forces mentioned earlier might indeed be playing a role in this effect. Visually, this can be produced as:

```
### Visualize the interaction
interact_plot(lm2, pred = unemployment_rate, modx = receive_ODA,
              plot.points = TRUE,
              main.title = "Interaction effects in model 2",
              legend.main = "Receives ODA?",
              modx.labels = c("No", "Yes"),
              x.label = "Unemployment rate (%)",
              y.label = "(IHS) number of USD events",
                            colors = c("black", "red"))
```



Importantly, slope coefficients estimated are statistically significant at a level of significance $\alpha = 0.1$. Although this implies that type I errors are likely (occurring 10% of the time), it allows state that the relationships between these variables and the explained variables are statistically significant. Nevertheless, given that b_3 and b_4 are not statistically significant at lower values of α, this raises doubts on the statistical significance when all controls will be included.

Overall, the fit of this model is better than in the univariate case but not by much (the $adjusted - R^2$ is about 5 percentage point larger). It should be noted that the number of observations remains identical $(= 168)$, as the degrees of freedom are now 163 and 5 coefficients are estimated.

4.3. Model including controls

Now, in order to limit the omitted variable bias and the potential effects of confounding factors, we include the other explanatory variables in the model.

```
### Model with all controls
lm3 <- lm(NEVENTS_ihs ~ log_population_city + unemployment_rate * receive_ODA +
            LFP + gap_mortality_rate + Share_services + Share_agriculture
          , data = selected_data)
summary(lm3)
```

```
## 
## Call:
## lm(formula = NEVENTS_ihs ~ log_population_city + unemployment_rate *
##     receive_ODA + LFP + gap_mortality_rate + Share_services +
##     Share_agriculture, data = selected_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.8059 -0.4637  0.0873  0.6008  1.7271 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                   2.2642570  1.1351890   1.995  0.04801 *  
## log_population_city           0.4783988  0.0678548   7.050 7.32e-11 ***
## unemployment_rate             0.0811122  0.0453040   1.790  0.07554 .  
## receive_ODA                   0.2146987  0.3626344   0.592  0.55476    
## LFP                          -0.0221730  0.0077593  -2.858  0.00492 ** 
## gap_mortality_rate           -0.0012064  0.0023293  -0.518  0.60534    
## Share_services               -0.0123682  0.0129594  -0.954  0.34153    
## Share_agriculture             0.0004256  0.0106852   0.040  0.96828    
## unemployment_rate:receive_ODA -0.0522232  0.0467463  -1.117  0.26582    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9007 on 141 degrees of freedom
##   (18 observations deleted due to missingness)
## Multiple R-squared:  0.3541, Adjusted R-squared:  0.3174 
## F-statistic: 9.661 on 8 and 141 DF,  p-value: 1.265e-10
```

The whole model shows that the first hypothesis is robust to the inclusion of different control variables, but the second hypothesis is not. Indeed, there is no statistically different relationship between the ODA_receiving and the non-ODA-receiving countries (the $p-value$ of the interaction term is 0.2658, beyond any conventional levels of significance). Among all included controls, only the LFP appears to have a statistically significant impact and is negatively related to the number of USD events.

The inclusion of the different controls led to reducing the sample by about 10% as 18 observations are now missing. Since the variable does not seem to be related with the explained variable, we consider it better to drop this variable from the model and retrieve the full set of observations (this will be done in future models included in the summary table). Given the poor results of most control variables, the optimal model probably lies between the baseline interaction model (model 2) and this model (model 3).

5. Model choice and diagnostics

5.1. Initial check for multicollinearity

A good way to figure out what model this should be would consist in checking what variables cause multicollinearity. We have seen that this is a serious risk in the present study. We simply check the variance inflation factor in the previous model when the interaction term is excluded, as the Variance inflation factor (VIF) of the interacted term would mechanically be very large.

```
### Compute the VIF in the model without the interaction
lm3_bis <- lm(NEVENTS_ihs ~ log_population_city + unemployment_rate + receive_ODA +
            LFP + gap_mortality_rate + Share_services + Share_agriculture
          , data = selected_data)
vif(lm3_bis)
```

```
## log_population_city   unemployment_rate         receive_ODA                 LFP
##            1.049970            1.550806            1.812000            1.447559
##  gap_mortality_rate      Share_services   Share_agriculture
##            1.109670            8.708410            8.737602
```

As a result, the VIF of the share of services and the share of agriculture are both high, as expected. Although they do not reach the conventional threshold of $VIF > 10$, it is probably better to drop one on the two shares of workers in a sector. Here, let us drop the share of workers in services.

```
### Model with all controls except share in services
lm4 <- lm(NEVENTS_ihs ~ log_population_city + unemployment_rate * receive_ODA +
            LFP + Share_agriculture
          , data = selected_data)
summary(lm4)
```

```
##
## Call:
## lm(formula = NEVENTS_ihs ~ log_population_city + unemployment_rate *
##     receive_ODA + LFP + Share_agriculture, data = selected_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7816 -0.5160  0.0445  0.6177  2.4866
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.869074   0.773444   2.417  0.01679 *
## log_population_city            0.478397   0.064598   7.406  6.9e-12 ***
## unemployment_rate              0.058299   0.036564   1.594  0.11280
## receive_ODA                    0.103646   0.328801   0.315  0.75300
## LFP                           -0.025658   0.007674  -3.344  0.00103 **
## Share_agriculture              0.008358   0.004563   1.832  0.06884 .
## unemployment_rate:receive_ODA -0.036481   0.039367  -0.927  0.35546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.92 on 161 degrees of freedom
## Multiple R-squared:  0.3429, Adjusted R-squared:  0.3184
## F-statistic:    14 on 6 and 161 DF,  p-value: 8.68e-13
```

```
lm4_bis <- lm(NEVENTS_ihs ~ log_population_city + unemployment_rate + receive_ODA +
            LFP + Share_agriculture
            , data = selected_data)
vif(lm4_bis)
```

```
## log_population_city   unemployment_rate         receive_ODA                 LFP
##            1.019952            1.439956            1.757763            1.424591
##   Share_agriculture
##            1.773381
```

5.2. Add geographical heterogeneity

Now that the model is clean from multicollinearity, we may want to have a last layer of complexity by controlling for the region of the city.

$$IHS(USD_j) = a + b_1 Log(Population_j) + b_2 Unemployment\ Rate_i + b_3 ODA_i + b_4 Unemployment\ Rate_i \times ODA_i$$

$$+ \sum_{k=5}^{6} b_k X_{ik} + \sum_{l=7}^{13} b_l Region_l + e_j$$

The notation of this model is somehow more elaborate than the previous models but does not convey any more complex messages than a linear regression where 2 extra controls (X_{i5} being the LFP, X_{i6} being the share of workers in agriculture) and seven binary variables, one per region included (Asia, Europe, Latin America, MENA, Northern America, Oceania, Sub Saharan Africa). Since the interpretation of a multivariate analysis implies to comment on **partial correlations** between the explanatory variable of interest and the explained variable, the inclusion of region binary variables implies that the interpretations of each of the main explanatory variables (not the region dummies) will be an average effect of the relationships **within** regions. *In fact, by including region dummies, we proceeded as if we included* **region fixed effects** *in the model.* This is a refinement that is very common in order to control for omitted variable biases that could be region-specific in the present analysis.

```
### Model with regional dummies (fixed effects)
lm5_0 <- lm(NEVENTS_ihs ~ log_population_city + unemployment_rate * receive_ODA +
            LFP + Share_agriculture + Asia + Europe +
            LatAm + MENA + NorAm  + SSA + Oceania
          , data = selected_data)
summary(lm5_0)
```

```
##
## Call:
## lm(formula = NEVENTS_ihs ~ log_population_city + unemployment_rate *
##     receive_ODA + LFP + Share_agriculture + Asia + Europe + LatAm +
##     MENA + NorAm + SSA + Oceania, data = selected_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3254 -0.4176  0.0385  0.5533  2.4804
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.343056   0.887468   0.387 0.699615
## log_population_city             0.510891   0.065572   7.791 8.94e-13 ***
## unemployment_rate               0.034700   0.035448   0.979 0.329158
## receive_ODA                    -0.252065   0.333389  -0.756 0.450756
## LFP                            -0.021218   0.007867  -2.697 0.007766 **
## Share_agriculture               0.023576   0.005991   3.935 0.000125 ***
## Asia                            0.562717   0.503273   1.118 0.265248
## Europe                          1.260582   0.461729   2.730 0.007064 **
## LatAm                           1.337524   0.503610   2.656 0.008737 **
## MENA                            1.103770   0.494944   2.230 0.027178 *
## NorAm                           0.587180   0.591421   0.993 0.322341
## SSA                             0.218029   0.537008   0.406 0.685298
## Oceania                               NA         NA      NA       NA
## unemployment_rate:receive_ODA  0.007133   0.038771   0.184 0.854267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.867 on 155 degrees of freedom
## Multiple R-squared:  0.4382, Adjusted R-squared:  0.3947
## F-statistic: 10.07 on 12 and 155 DF,  p-value: 1.943e-14
```

```
lm5 <- lm(NEVENTS_ihs ~ log_population_city + unemployment_rate * receive_ODA +
            LFP + Share_agriculture + Asia + Europe +
            LatAm + MENA + NorAm  + SSA
          , data = selected_data)
```

Since the inclusion of all regions binary variables would lead the sum of these to be perfectly equal to 1, then the set of all dummies but 1 can lead us to retrieve the full set: $b_7+b_8+...+b_12+b_13 = 1 \Rightarrow 1-b_7-b_8-...-b_12 = b_13$. Hence, including all the regional dummies is a mistake because it would lead to perfect collinearity. Fortunately, *R* software is allowing us to run the model dropping one of the dummies to let the OLS find its unique solution. This is what is indicated by "Coefficients: (1 not defined because of singularities)" in the model output.

Model 5 continues to confirm that the elasticity between city population and USD events is about 0.5 and significant statistically spoken. Interestingly, when controlling by region, the unemployment rate relationship with USD event is no more tuned for ODA_receiving countries: the coefficient of the interaction term becomes positive (although not statistically significant at conventional levels of significance).

For several labor market characteristics and within each region, a higher city population by 1% corresponds to a higher number of social disorder events in the city by 51.1%. Compared to other regions, there are more events in Europe, Latin America and MENA: being in either of these regions corresponds to more than a doubling in the number of USD events.

One may want to be able to compare the magnitude of the unemployment rate and the LFP variables. To do so, a convenient technique is beta regression, that provides *standardized* coefficients. In our case, beta regression is not remarkably interesting because for the population - USD relationship, we can already interpret it as elasticity, which is quite convenient. Beta regressions for dummy variables are not especially useful either, as the interpretation is already very straightforward. Therefore, the key interest here lies in comparing LFP and unemployment, and both variables are expressed initially in percentage. Here, the beta coefficients are, respectively, -0.213 and 0.168. A higher LFP in a country by 1 standard deviation leads to a lower IHS number of USD by 0.213 standard deviations. A higher unemployment rate in a country by 1 standard deviation leads to a higher IHS number of USD by 0.168 standard deviations. Therefore, LFP's impact is larger than unemployment rate.

```
### Get standardized regression coefficients
lm.beta(lm5)
```

```
##
## Call:
## lm(formula = NEVENTS_ihs ~ log_population_city + unemployment_rate *
##     receive_ODA + LFP + Share_agriculture + Asia + Europe + LatAm +
##     MENA + NorAm + SSA, data = selected_data)
##
## Standardized Coefficients::
##                   (Intercept)           log_population_city
##                            NA                    0.51091933
##             unemployment_rate                   receive_ODA
##                    0.16847293                   -0.10948006
##                           LFP             Share_agriculture
##                   -0.21290104                    0.46068647
##                           Asia                         Europe
##                     0.20782734                     0.47468892
##                          LatAm                           MENA
##                     0.41383181                     0.32856523
##                          NorAm                            SSA
##                     0.08981000                     0.08429110
## unemployment_rate:receive_ODA
##                     0.03946436
```

5.3. Summary table of OLS models

```
### This super useful function is taken from:
  # https://stackoverflow.com/questions/43245920/how-to-resize-tables-
  # generated-by-stargazer-in-r-markdown
# it allows resizing the table à la \resizebox in LaTeX
resizebox.stargazer = function(..., tab.width = "!", tab.height = "!"
){
  #Activate str_which() function:
  require(stringr)
  #Extract the code returned from stargazer()
  res = capture.output(
    stargazer::stargazer(...)
  )
  #Render the arguments:
  tab.width = tab.width
  tab.height = tab.height
  #Attach "}" between \end{tabular} and \end{table}
  res =
    prepend(res, "}", before = length(res))
  #Input \resizebox before \begin{tabular}
  res =
    c(res[1:str_which(res, "^\\\\begin\\{tabular\\}.*")-1],
      paste0("\\resizebox{",tab.width,"}{",tab.height,"}{%"),
      res[str_which(res, "^\\\\begin\\{tabular\\}.*"):length(res)]
    )
  #Produce the whole strings
  cat(res, sep = "\n")
}
### get labels
labels <- c(
  'Log City population',
  'Unemployment rate',
  'Receive ODA',
  'Labor force participation',
  'Gap in mortality (F - M)',
  'Share of workers in services',
  'Share of workers in agriculture',
  'Asia',
  'Europe',
  'Latin American',
  'MENA',
  'Northern America',
  'SSA',
  'Unemployment * ODA'
```

```
)
### Produce table
resizebox.stargazer(lm1, lm2, lm3, lm4, lm5,
                    title = "OLS and IHS number of urban social disorders (1960-2014)",
                    dep.var.labels = "IHS number of USD",
                    column.labels = c("Model 1", "Model 2", "Model 3", "Model 4", "Model 5"),
                    covariate.labels = labels,
                    #notes = ,
                    type = "latex",
                    tab.width = "0.95\\textwidth")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: vie, dic 30, 2022 - 15:29:54

Table 2: OLS and IHS number of urban social disorders (1960-2014)

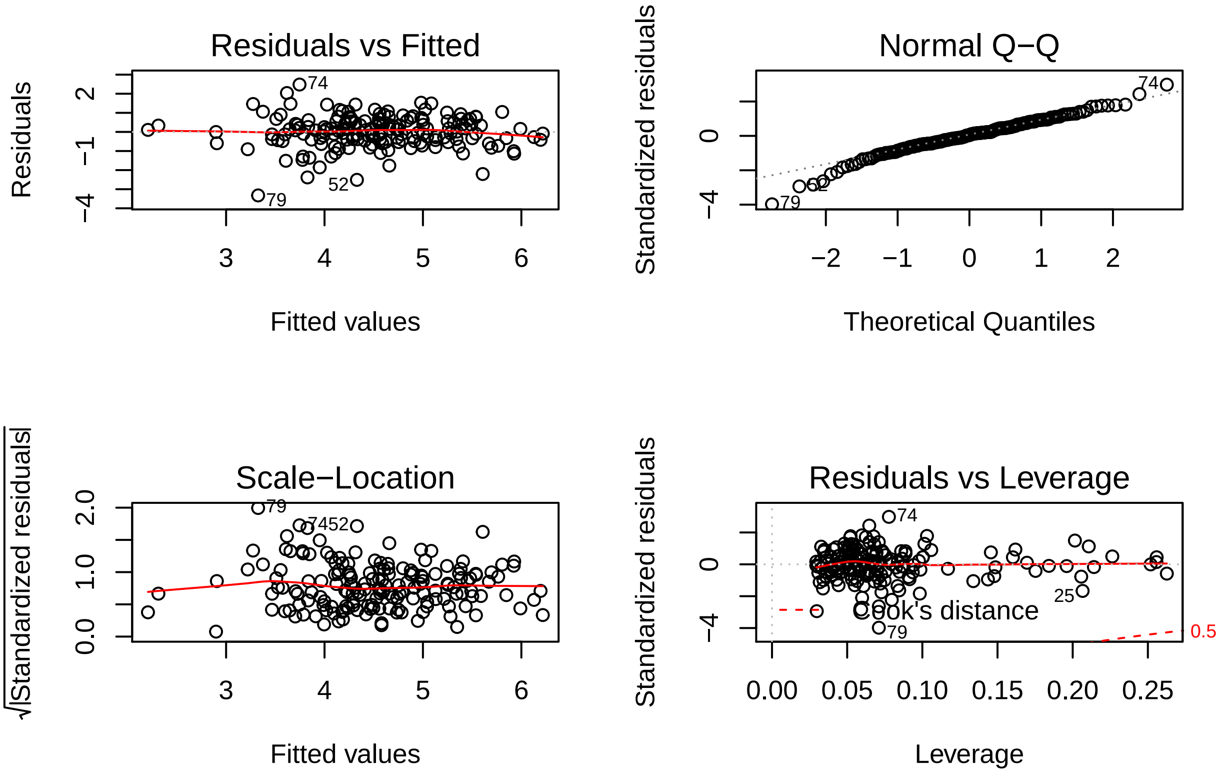
	Dependent variable:				
			IHS number of USD		
	Model 1	Model 2	Model 3	Model 4	Model 5
	(1)	(2)	(3)	(4)	(5)
Log City population	0.475*** (0.068)	0.477*** (0.066)	0.478*** (0.068)	0.478*** (0.065)	0.511*** (0.066)
Unemployment rate		0.100*** (0.036)	0.081* (0.045)	0.058 (0.037)	0.035 (0.035)
Receive ODA		0.492* (0.271)	0.215 (0.363)	0.104 (0.329)	−0.252 (0.333)
Labor force participation			−0.022*** (0.008)	−0.026*** (0.008)	−0.021*** (0.008)
Gap in mortality (F - M)			−0.001 (0.002)		
Share of workers in services			−0.012 (0.013)		
Share of workers in agriculture			0.0004 (0.011)	0.008* (0.005)	0.024*** (0.006)
Asia					0.563 (0.503)
Europe					1.261*** (0.462)
Latin American					1.338*** (0.504)
MENA					1.104** (0.495)
Northern America					0.587 (0.591)
SSA					0.218 (0.537)
Unemployment * ODA		−0.065* (0.039)	−0.052 (0.047)	−0.036 (0.039)	0.007 (0.039)
Constant	0.820 (0.540)	0.120 (0.559)	2.264** (1.135)	1.869** (0.773)	0.343 (0.887)
Observations	168	168	150	168	168
R^2	0.225	0.294	0.354	0.343	0.438
Adjusted R^2	0.221	0.276	0.317	0.318	0.395
Residual Std. Error	0.984 (df = 166)	0.948 (df = 163)	0.901 (df = 141)	0.920 (df = 161)	0.867 (df = 155)
F Statistic	48.253*** (df = 1; 166)	16.930*** (df = 4; 163)	9.661*** (df = 8; 141)	14.000*** (df = 6; 161)	10.074*** (df = 12; 155)
Note:					*p<0.1; **p<0.05; ***p<0.01

The table gathers the different models constructed. It comes with no surprise that we tend to prefer the last model, as this allows to interpret the result *within* regions (compared to all models 1 to 4), and as it comprises all observations in the sample, and does not bear multicollinearity risks (compared with model 3). To pursue diagnostics, we will use model 5.

5.4. Diagnostics

Traditional diagnostics aim to verify whether the Gauss-Markov assumptions are realistic. Here, we will briefly discuss linearity, spherical errors (here simply the heteroskedasticity), and the applicability of the central limit theorem (normality of errors and outliers).

```
### Some diagnostics
par(mfrow = c(2,2))
plot(lm5)
```



The four graphs above aim to help us address these questions. The top left plot shows that linearity might not be too strong of an assumption as residuals tend to be well scattered around the horizontal zero residual line. On the other hand, this same plot rings a bell toward a potential heteroskedasticity. Indeed, it seems that $Vr(\varepsilon|X) \neq constant$. The variance of errors seems to be larger for lower fitted values. The Residual *versus* fitted plot shows a somehow funnel-looking pattern, which casts doubts on the spherical error assumption. Note also that observations 52 (Libreville), 74 (Jerusalem), 79 (Astana - now Nur-Sultan) might be outliers in the model (this is hardly a surprise for Astana, which has no USD event).

In order to accurately conclude on this issue, we can run a Breusch-Pagan test, which has $H_0 : Equal\ Variance\ (homoskedasticity)$ and $H_a : not\ equal\ variance\ (heteroskedasticity)$. The test is based on a χ^2 distribution. The squared residuals of the regression are used as the dependent variable with the explanatory variables initially used. A measure of the goodness of fit of this second regression is then compared with a critical value taken from the χ^2 distribution. Intuitively, we test the variance of the residuals with the initial component introduced to verify whether this variance depends on the initial explanatory variables. The better the fit in the second regression the larger the computed statistics and thus the more likely the odds to reject H_0 (thus rejecting homoskedasticity). Here, the Breusch-Pagan test gives a critical value of 25.032 for 12 degrees of freedom, the corresponding p-value is 0.015, so that we would reject H_0 for significance level beyond 1.5%. Heteroskedasticity is therefore likely. Many tools exist to address this, among which the **generalized least squares**, which basically correct the residual matrix to recover constant residual over the explanatory variables. To refine the analysis, GLS should be implemented.

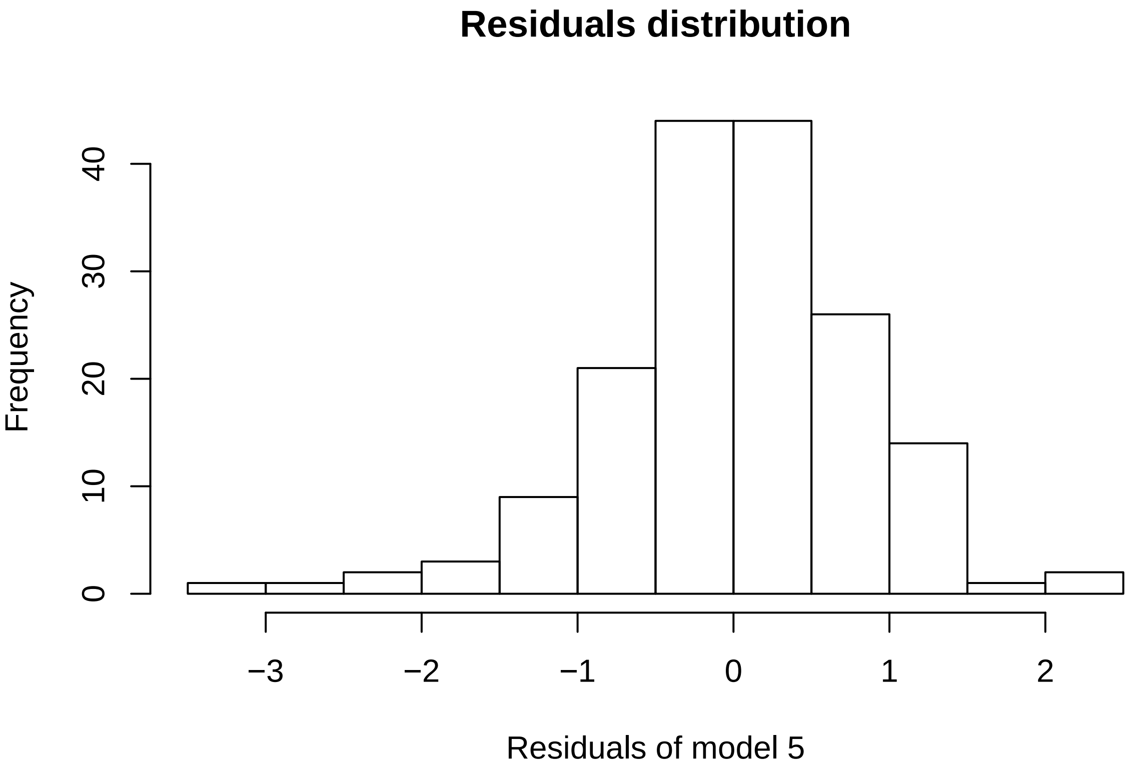
```
### Breusch-Pagan test
bptest(lm5)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm5
## BP = 25.032, df = 12, p-value = 0.01467
```

The top right plot informs on the normality of residuals, comparing empirical values with theoretical values of residuals, would they be perfectly normally distributed. The smallest empirical quintiles of residuals are somehow distant from their theoretical quintiles. This is coherent with the fact that some residuals show some distance, being far away under the regression line. The question is now whether this is sufficiently serious to question the normality of residuals. The histogram below confirms the observation of the quintile-quintile plot: lower values might not be normal.

In order to get to a definite conclusion, one can use a statistical test. The Kolmogorov-Smirnov test (non-parametric) allows to verify whether an empirical distribution is close enough to a theoretical distribution. Here we want to test whether the distribution of the model's residuals is close enough to a normal distribution (it compares cumulative distributions). Other tests exist (like the Shapiro-Wilk test) but they all have pros and cons. To keep things simple, we restrict ourselves to the Kolmogorov-Smirnov test. H_0 is that the empirical cumulative distribution is like the theoretical cumulative distribution of interest (when the distance between these cumulative density functions is not too large). Here, the resulting $p - value$ is 0.0993 so that H_0 cannot be rejected, unless the risk of type I error is 10% and beyond. It seems therefore acceptable to accept that residuals are normally distributed.

```
### Residual histogram
lm5_resid <- resid(lm5) # get the residuals
hist(lm5_resid, main='Residuals distribution',
     xlab = "Residuals of model 5")
```



```
### Kolmogorov-Smirnov  test
ks.test(lm5_resid, 'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  lm5_resid
## D = 0.094533, p-value = 0.0993
## alternative hypothesis: two-sided
```

Linearity and normality do not appear to be severely questioned and heteroskedasticity can be easily addressed using generalized least squares. Furthermore, multicollinearity has been addressed. Therefore, core hypotheses underlying linear least squares models do not seem to be unrealistic.

That being said, the most important hypothesis, exogeneity is surely not met: (i) reverse causality, (ii) omitted variables, (iii) measurement errors, are all critical issues that cannot be addressed in the present study. Not only can no causality be raised, but results are biased, so that magnitudes and even the signs might be wrong (not to say anything about the inference that is also probably wrong).

6. Wrap-up

Assuming that the constraints on the data and the endogeneity issues do not fully prevent any conclusions, we found that city population and the number of social events is strongly linked, even when controlling for many labor market conditions. The elasticity is equal to about 0.5, so that an increase in the city population by 1% corresponds to an increase in the number of urban social disorder events by 0.5%. The first working hypothesis is therefore verified.

On the other hand, the second working hypothesis is only verified when controls are not included. With the inclusion of controls, the link between unemployment rate and the number of urban social disorder events is not different according to whether the countries received official development assistance. In fact, it appears the the labor force participation plays a more important role than the rate of unemployment in the linear model constructed in the present analysis.

References

Bellemare, Marc F, and Casey J Wichman. 2020. “Elasticities and the Inverse Hyperbolic Sine Transformation.” *Oxford Bulletin of Economics and Statistics* 82 (1): 50–61.

Thomson, Henry, Karim Bahgat, Henrik Urdal, and Halvard Buhaug. 2022. “Urban Social Disorder 3.0: A Global, City-Level Event Dataset of Political Mobilization and Disorder.” *Journal of Peace Research*, 00223433221082991.