《国际关系定量分析基础》2020 秋季

第二次小组作业 (共计 100 分)

陈道想　　黄卓尔　　潘明花　　杨霖　　赵佳鹏

截止时间：**2020 年 11 月 2 日 11：59 am**

面板数据（panel data）也称为横截面时间序列数据（cross-sectional time-series data)，是国际关系中最常用的一种数据格式。本次作业的目标是练习如何利用既有数据库建立一个整洁的包含世界所有国家 1990-2016 年的面板数据，并添加国家特征相关的其他数据和变量。既有数据包括世界银行（world bank）、Pen World Table 以及 terrorism 数据库等。

注意事项：

- 小组作业截止时间：**2020 年 11 月 2 日 11：59 am**
- 请将文件解压缩后，直接在 R Markdown 文件中完成本次作业
- 作业在网络学堂提交，每个小组仅需提交一份
- 提交作业的文件名需以 `HW-2-Team-X.Rmd`，`HW-2-Team-X.pdf` 或者 `HW-2-Team-X.html`, 请将 `X` 替换为小组编号，如 `HW-2-Team-A.Rmd`、`HW-2-Team-A.pdf` 或 `HW-2-Team-A.html`。(若 R Markdown 出现无法 `knit` 为 `pdf` 情况，则使用 `bookdown::html_document2`: 会生成为 `html`)
- 请显示每道题的 R Code 于 pdf 中，注重 Code 的整洁性和可读性，可参考Google's R Style Guide
- 本次作业所需 R Packages 已经提供。本次作业需要的数据已经提供，请将数据与 `HW-2-Team-X.Rmd` 放在同一工作路径的文件夹内，通过 `load("globalterrorism.RData")` 加载。

```
load("globalterrorism.RData")
load("wdi.RData")
```

创建基本数据框（共 20 分）

1.(**20 分**) 请用 R 中的 `states`、`countrycode` 和 `dplyr` 三个 packages, 创建一个 1990-2016 年包括全世界所有国家的，以年为单位的面板数据，并将数据框命名为 `base_df`。**注意：**创建数据应使用基于 COW 的国家代码 (country code)，数据应该删除台湾省（台湾的 COW code 是 713），最终的数据只保留 `cowcode`, `country_name` 和 `year` 三个变量（参考表-1 的输出结果）。结合 `base_df` 数据，请回答这个数据的 “分析单元”（unit-of-analysis）是什么？

```
# 载入所需的包
library(states)
library(countrycode)
library(dplyr)
base_df <- state_panel("1990-01-01", "2016-12-31", by = "year",
                       partial = "any", useGW = FALSE) %>%
# 删除台湾省信息
          filter(., cowcode != 713) %>%
# 基于 cowcode 变量产生 country_name 变量
          mutate(country_name =
                 countrycode(sourcevar = .$cowcode,
                             "cown", "country.name"),
# 将 date 变量修饰为 year 变量
                 year = as.numeric(format(as.Date(date),"%Y")))  %>%
# 当 country_name 为 NA 值时，去除之
          filter(!is.na(country_name)) %>%
# 调整数据框，除去原始数据日期列
          select(cowcode, country_name, year)
# 使用 knitr:kable 产生一个新的表格
kable(head(base_df), caption = " 基本数据框")
```

我们观察发现，`base_df` 数据的 “分析单元” 是 country-year，即国家 • 年。具体而言，即 1990 年至 2016 年间某一年某一国家。

处理 WDI 数据（共 20 分）

世界银行发展指标（world bank development indicators）是社会科学研究中常用的关于国家层次的政治经济和社会数据。其中，R 中的 `WDI` 是一个基于世界银行数据的软件包，记录了世界银行及其开发指标的相关数据。利用 `WDI` 软件可包获取 1990-2016 年所有国家的包括 “军费开支占 GDP 比

表 1: 基本数据框

cowcode	country_name	year
2	United States	1990
2	United States	1991
2	United States	1992
2	United States	1993
2	United States	1994
2	United States	1995

重”、“GDP 经济增长率”、“FDI 占 GDP 比重的数据”，其对应的变量分布为 `MS.MIL.XPND.GD.ZS`, `NY.GDP.MKTP.KD.ZG`, `BX.KLT.DINV.WD.GD.ZS`。通过如下代码或者 `load("wdi.RData")` 可以获得这一数据。

```
# Foreign direct investment, net inflows (% of GDP): BX.KLT.DINV.WD.GD.ZS
# Military expenditure (% of GDP): MS.MIL.XPND.GD.ZS
# GDP growth (annual %): NY.GDP.MKTP.KD.ZG
library(WDI)
wdi <- WDI(country = "all", indicator = c("MS.MIL.XPND.GD.ZS",
                                          "NY.GDP.MKTP.KD.ZG",
                                          "BX.KLT.DINV.WD.GD.ZS"),
           start = 1990, end = 2016)
```

2.(**10 分**) 请利用 `dplyr` 软件包,将数据 `wdi` 中以上三个变量重命名为 `mil_expen` (MS.MIL.XPND.GD.ZS), `gdp_growth` (NY.GDP.MKTP.KD.ZG), `fdi_percent` (BX.KLT.DINV.WD.GD.ZS)。同时，利用 R 中的 `countrycode` 包，创建一个名为 `ccode` 变量——该变量应当使用 COW 的国家编码。最后，基于 `wdi` 数据，只保留如下变量：`country`, `ccode`, `year`, `mil_expen`, `gdp_growth`, `fdi_percent`（参考表-2 的输出结果）。

```
load("wdi.RData")
wdi <- wdi %>%
# 进行变量重命名
      rename(mil_expen = MS.MIL.XPND.GD.ZS,
             gdp_growth = NY.GDP.MKTP.KD.ZG,
             fdi_percent = BX.KLT.DINV.WD.GD.ZS) %>%
# 基于 country 变量产生 ccode 变量
      mutate(ccode = countrycode(sourcevar = .$country,
                                 "country.name", "cown")) %>%
```

表 2: WDI 数据框

country	ccode	year	mil_expen	gdp_growth	fdi_percent
Andorra	232	1990	NA	3.781388	NA
Andorra	232	1991	NA	2.546003	NA
Andorra	232	1992	NA	0.929212	NA
Andorra	232	1993	NA	-1.031484	NA
Andorra	232	1994	NA	2.383188	NA
Andorra	232	1995	NA	2.757499	NA

```
# 当 ccode 为 NA 值（如，country 为地区）时，去除之
      filter(!is.na(ccode)) %>%
      select(country, ccode, year, mil_expen, gdp_growth, fdi_percent)
# 使用 knitr:kable 产生一个新的表格
kable(head(wdi), caption = "WDI 数据框")
```

注：如果基于 iso2c 变量产生 ccode 变量，在第 7 题统计 mil_expen，gdp_growth，fdi_percent 的 NA 值个数时将会比示例文件多 9 个。部分代码如下：

```
mutate(ccode = countrycode(sourcevar = .$iso2c,
                           "iso2c", "cown")) %>%
```

这是因为科索沃地区（Kosovo）的 iso2c 值 XK 无法被转化为 cown 值，因而被赋为 NA，并在下一行代码被 filter 命令去除。

3.(**10 分**) 基于 wdi 数据，利用 dplyr 包，以表格形式显示 2015 年全世界军费开支占 GDP 比重最高的 10 个国家（参考表-3 的输出结果）。**提示：**可利用 dplyr 包中 filter, slice_max、arrange 三个命令。

```
wdi %>%
# 筛选 2015 年数据
    dplyr::filter(year == 2015) %>%
# 选取 mil_expen 最高的 10 个国家
    slice_max(order_by = mil_expen, n=10) %>%
# 仿照样例进行降序排列
    arrange(desc(mil_expen)) -> max
# 使用 knitr:kable 产生一个新的表格
kable(max, caption = "2015 年全世界军费开支占 GDP 比重最高的 10 个国家")
```

表 3: 2015 年全世界军费开支占 GDP 比重最高的 10 个国家

country	ccode	year	mil_expen	gdp_growth	fdi_percent
Saudi Arabia	670	2015	13.325672	4.1064089	1.2442918
Oman	698	2015	10.930961	4.6776856	-3.1760577
South Sudan	626	2015	10.561462	-10.7933646	0.0012502
Algeria	615	2015	6.270243	3.7000000	-0.3240118
Israel	666	2015	5.496969	2.2900921	3.7810887
Azerbaijan	373	2015	5.464877	1.0495464	7.6263363
Iraq	645	2015	5.410251	2.4776646	-4.2671891
Kuwait	690	2015	5.007568	0.5930196	0.2484545
Russian Federation	365	2015	4.862758	-1.9727192	0.5026084
Bahrain	692	2015	4.632877	2.8630473	0.2084878

处理 Penn Word Table 9.1 数据（共 20 分）

4.(**10 分**) Penn Word Table 是经济学家常用的关于世界各国经济指标的数据，R 中的 `pwt9` 软件包已经收录了 1950-2017 年各国的相关数据。利用 `pwt9` 包，获取 1990-2016 年之间各国的数据，将数据框命名为 `pwt9`；同时利用 COW 的国家编码，创建一个 `ccode` 变量。**提示：**数据应该是 4450 行，48 列。

```
library(pwt9)
data("pwt9.0")
pwt9 <- pwt9.0 %>%
      filter(year >= 1990 & year <= 2016) %>%
# 基于 isocode 变量产生 ccode 变量
      mutate(ccode = countrycode(sourcevar = .$isocode, "iso3c", "cown"))
dim(pwt9)
```

```
## [1] 4550   48
```

```
# 使用下列命令可以实现类似于第 1、2 题的数据清洗效果
# 本命令将在下题中使用：filter(!is.na(ccode))
```

注：采用 `pwt9.0` 时，数据应该是 4550 行，48 列；若采用 `pwt9.1`，数据应该是 4914 行，53 列。

5.(**10 分**) 基于创建的 1990-2016 年的 `pwt9` 数据框，只保留各国 1990-2016 年间的 GDP 变量（对应的变量名为 `rgdpna`）、人口 (对应的变量名为 `pop`)。**提示：**最后保留的 `ccode`, `year`, `pop`, `rgdpna`

表 4: PWT9 数据

ccode	year	pop	rgdpna
540	1990	11.12787	65542.75
540	1991	11.47217	66192.51
540	1992	11.84897	62328.01
540	1993	12.24679	47379.62
540	1994	12.64848	48014.21
540	1995	13.04267	55216.34

变量和观察量如表-4 所示.

```
pwt9 <- pwt9 %>%
# 参见上题注释
        filter(!is.na(ccode)) %>%
        select(ccode, year, pop, rgdpna)
dim(pwt9)
```

```
## [1] 4225    4
```

```
# 使用 knitr:kable 产生一个新的表格
kable(head(pwt9), caption = "PWT9 数据")
```

处理 Terrorism 数据（共 20 分）

6.(**10 分**) `globalterrorism.RData` 记录了 1990-2016 年发生在世界各国的恐怖袭击事件。首先，描述这一数据的分析单位（unit-of-analysis）是什么？其次，根据 `globalterrorism.RData` 这一数据，利用 `dplyr` 包，将数据汇总到国家--年层次，以显示各国在每一年发生的恐怖袭击的次数和每年的伤亡人数总和。**提示：**需要使用 `group_by`, `summarise` 等命令。

```
load("globalterrorism.RData")
gtd <- globalterrorism %>%
# 基于国家名称产生 ccode 变量
       mutate(ccode = countrycode(sourcevar = .$country_txt,
                                  "country.name", "cown")) %>%
       filter(!is.na(ccode)) %>%
# 将数据汇总到国家 • 年层次
```

表 5: GTD 数据

iyear	ccode	sum_nkillter	sum_events
1990	2	0	32
1990	41	0	3
1990	42	0	2
1990	70	0	5
1990	90	0	83
1990	91	0	13

```
        group_by(ccode, iyear) %>%
# 统计各国各年发生的恐怖袭击的次数和每年的伤亡人数总和
        summarise(sum_nkillter = sum(nkillter, na.rm = TRUE),
                  sum_events = n()) %>%
        select(iyear, ccode, sum_nkillter, sum_events) %>%
# 仿照样例进行升序排列
        arrange(iyear, ccode)
dim(gtd)
```

```
## [1] 2372    4
```

```
# 使用 knitr:kable 产生一个新的表格
kable(head(gtd), caption = "GTD 数据")
```

我们观察发现，globalterrorism.RData 数据的 “分析单元” 是 event，即一次事件。而我们处理之后产生的数据汇总到了国家 • 年层次，因而其 “分析单元” 是国家 • 年。具体而言，即 1990 年至 2016 年间某一年某一国家。

合并三个数据（共 10 分）

7.(**10 分**) 分别将 wdi, pwt9 以及 gtd 合并到 base_df 中，产生一个包含 11 个变量，5094 个观测量的数据。注意需要将 sum_nkillter 与 sum_events 中的 NA 重新赋值为 0. **提示:** 使用 left_join 这一命令（参考如下最终结果）。

```
base_df <- left_join(base_df, wdi, by = c("cowcode" = "ccode",
                                          "year" = "year")) %>%
           left_join(., pwt9, by = c("cowcode" = "ccode",
```

```
                                    "year" = "year")) %>%
          left_join(., gtd, by = c("cowcode" = "ccode",
                                   "year" = "iyear")) %>%
# 将指定变量中的 NA 重新赋值为 0
          mutate(sum_nkillter = ifelse(is.na(sum_nkillter),
                                       0, sum_nkillter)) %>%
          mutate(sum_events = ifelse(is.na(sum_events), 0, sum_events))
summary(base_df)
```

```
##     cowcode      country_name            year        country
##  Min.   :  2.0   Length:5094        Min.   :1990   Length:5094
##  1st Qu.:255.0   Class :character   1st Qu.:1997   Class :character
##  Median :450.0   Mode  :character   Median :2003   Mode  :character
##  Mean   :463.6                      Mean   :2003
##  3rd Qu.:670.0                      3rd Qu.:2010
##  Max.   :990.0                      Max.   :2016
##
##    mil_expen         gdp_growth       fdi_percent             pop
##  Min.   :  0.000   Min.   :-64.047   Min.   : -58.3229   Min.   :   0.0408
##  1st Qu.:  1.094   1st Qu.:  1.494   1st Qu.:   0.8143   1st Qu.:   2.4391
##  Median :  1.653   Median :  3.800   Median :   2.3952   Median :   8.1120
##  Mean   :  2.319   Mean   :  3.665   Mean   :   6.6952   Mean   :  36.9909
##  3rd Qu.:  2.688   3rd Qu.:  6.095   3rd Qu.:   5.3144   3rd Qu.:  24.6399
##  Max.   :117.350   Max.   :149.973   Max.   :1282.6326   Max.   :1369.4357
##  NA's   :1186      NA's   :254       NA's   :395         NA's   :939
##      rgdpna          sum_nkillter        sum_events
##  Min.   :     275   Min.   :   0.000   Min.   :   0.00
##  1st Qu.:   13358   1st Qu.:   0.000   1st Qu.:   0.00
##  Median :   45244   Median :   0.000   Median :   0.00
##  Mean   :  417980   Mean   :   9.391   Mean   :  24.98
##  3rd Qu.:  265327   3rd Qu.:   0.000   3rd Qu.:   4.00
##  Max.   :17150538   Max.   :3949.000   Max.   :3926.00
##  NA's   :939
```

可视化数据（10 分）

8.(**5 分**) 利用 `stargazer` 包，基于合并的 `base_df` 数据，制作变量间的描述统计表，产生关于如下变量的描述性统计表格（见表-6 所示）。**提示**: 可以通过 `help(stargazer)` 选择对应的统计量(`summary.stat`)。

```
library(stargazer)
# 使用 stargazer 命令产生描述性表格
stargazer(base_df, summary.stat = c("n", "mean", "median", "max", "min", "sd"),
          header=FALSE, type='latex',
          title = " 变量的描述性统计", digit.separator = "")
```

表 6: 变量的描述性统计

Statistic	N	Mean	Median	Max	Min	St. Dev.
cowcode	5094	463.620	450	990	2	261.128
year	5094	2003.225	2003	2016	1990	7.735
mil_expen	3908	2.319	1.653	117.350	0.000	3.070
gdp_growth	4840	3.665	3.800	149.973	−64.047	6.485
fdi_percent	4699	6.695	2.395	1282.633	−58.323	40.883
pop	4155	36.991	8.112	1369.436	0.041	133.754
rgdpna	4155	417980.100	45243.920	17150538.000	275.172	1374661.000
sum_nkillter	5094	9.391	0	3949	0	113.735
sum_events	5094	24.977	0	3926	0	144.180

9.(**5 分**) 利用 `GGally` 包，基于合并的 `base_df` 数据，绘制包括如下变量的（图-1）相关系数图。

```
library(GGally)
ggpairs(base_df[, c("mil_expen", "gdp_growth", "fdi_percent", "pop",
                    "rgdpna", "sum_nkillter", "sum_events")],
# 去除坐标轴标注以使图形更为简洁
        axisLabels = "none")
```

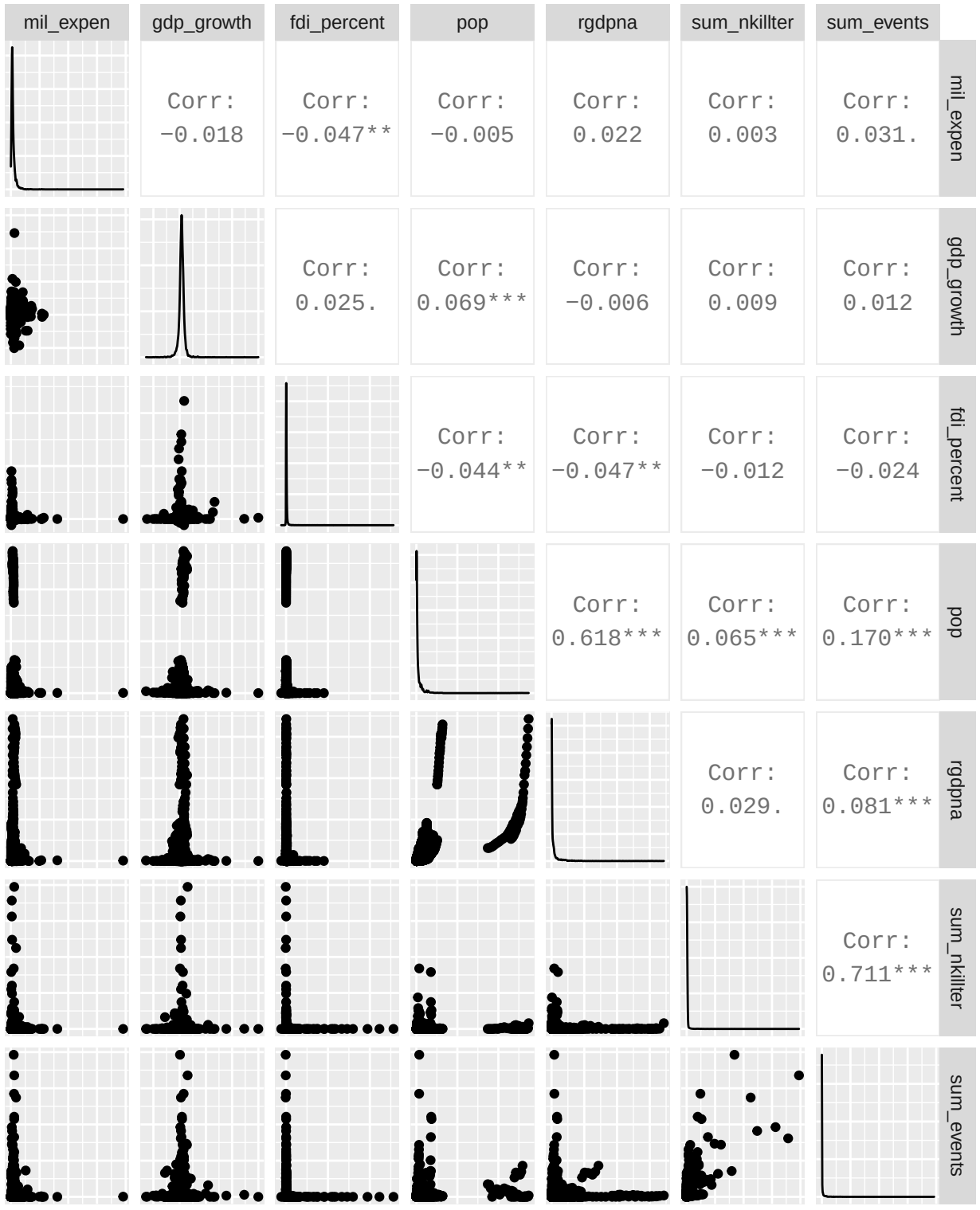
mil_expen
gdp_growth
fdi_percent
pop
rgdpna
sum_nkillter
sum_events
Corr: −0.018
Corr: −0.047**
Corr: −0.005
Corr: 0.022
Corr: 0.003
Corr: 0.031.
Corr: 0.025.
Corr: 0.069***
Corr: −0.006
Corr: 0.009
Corr: 0.012
Corr: −0.044**
Corr: −0.047**
Corr: −0.012
Corr: −0.024
Corr: 0.618***
Corr: 0.065***
Corr: 0.170***
Corr: 0.029.
Corr: 0.081***
Corr: 0.711***

图 1: 变量的相关关系