

《国际关系定量分析基础》2020 秋季

第三次作业 (共计 100 分)

学生姓名 学生学号

截止时间: 2020 年 11 月 16 日 11: 59 am

注意事项:

- 作业在网络学堂提交
- 请将 Chunk 中的 `eval=FALSE` 改为 `eval=TRUE` 再 `knit`
- 请将文件解压缩后, 直接在 R Markdown 文件中完成本次作业
- 学生可以互相讨论作业, 但作业必须是自己本人独立完成
- 提交作业的文件名需以 `HW-3-YourName.Rmd`, `HW-3-YourName.pdf` 或者 `HW-3-YourName.html`, 请将 `YourName` 替换为你的姓名。(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

本次作业需要的数据已经提供, 请将数据与 `HW-3-YourName.Rmd` 放在同一工作路径的文件夹内

```
load("defense_spending.RData")
```

本次作业使用的数据 `defense_spending` 来自于 Matthew Fuhrmann 发表于《美国政治学科杂志》2020 年第 2 期的复制数据 (见 Matthew Fuhrmann, “When Do Leaders Free-Ride? Business Experience and Contributions to Collective Defense,” *American Journal of Political Science*, Vol. 64, No. 2, April 2020, pp. 416–431)。该文章检验了北约领导人的经商经历对于该国防务经费支出的影响。

其中部分变量如下:

- `ccode`: The Correlates of War (COW) country code
- `countryname`: The Correlates of War (COW) country name
- `year`: The year of the observation
- `leadername`: The leader's name

表 1: 变量的描述性统计

Statistic	N	Mean	Median	Max	Min	St. Dev.
defspend_ch_usa	917	2.776	0.391	113.962	-13.718	14.190
business	927	0.155	0	1	0	0.362
econ_finance	927	0.306	0	1	0	0.461
lnrgdpe	917	12.560	12.613	15.126	8.182	1.429
growth	907	3.730	3.754	24.568	-16.735	3.874
war	927	0.041	0	1	0	0.198
cpg_sw2014	884	2.891	3.000	4.000	0.093	0.884

- **defspend_ch_usa** (Δ Defense Expenditures): Annual **percentage change** in defense spending (is USA dollars). Source: SIPRI (2015).
- **business** (Business Experience): An indicator of the leader's executive-level business experience (1 = business experience; 0 = otherwise). Source: coded by the author.
- **econ_finance** (Economics and Finance Experience): An indicator of the leader's background in economics or finance (1 = with; 0 = without).
- **lnrgdpe** (Economic Capacity(ln)): Logged gross domestic product. Source: Feenstra et. al (2015).
- **growth** (Economic Growth): Annual change in gross domestic product (%). Source: Feenstra et. al (2015).
- **war**: An indicator of War in the year of the observation (1 = war; 0 = no war). Source: Reiter et. al (2016).
- **cpg_sw2014** (Government Ideology): Government ideology based on a 5-point scale with higher values indicating greater left-wing dominance. Source: Seki and Williams (2014).
- **nato**: NATO member = 1; 0 = otherwise. Source: https://www.nato.int/cps/en/natohq/topics_52044.htm.

表-1 统计了部分变量的统计分布特征。请利用 `defense_spending` 数据完成以下各题。

清理描述数据（共 40 分）

1.(10 分) `ggcorrplot` 是一款比较新颖的描述变量相关系数的 **R** 软件包。请利用 `ggcorrplot` 这一命令，选取图-1 中显示的变量，绘制它们之间的相关系数图。提示：部分代码已经给出。

请完成代码

```
library(ggcorrplot)
defense_spending %>%
  select(defspend_ch_usa, business, econ_finance,
         lnrgdpe, growth, war, cpg_sw2014) %>%
  na.omit() %>% # remove NAs
  cor() %>% # calculate correlations
  ggcorrplot(., hc.order = F, type = "lower",
             colors = c("blue", "white", "red"), digits = 2,
             lab = FALSE, sig.level = 0.05,
             ggtheme = ggplot2::theme_minimal)
```

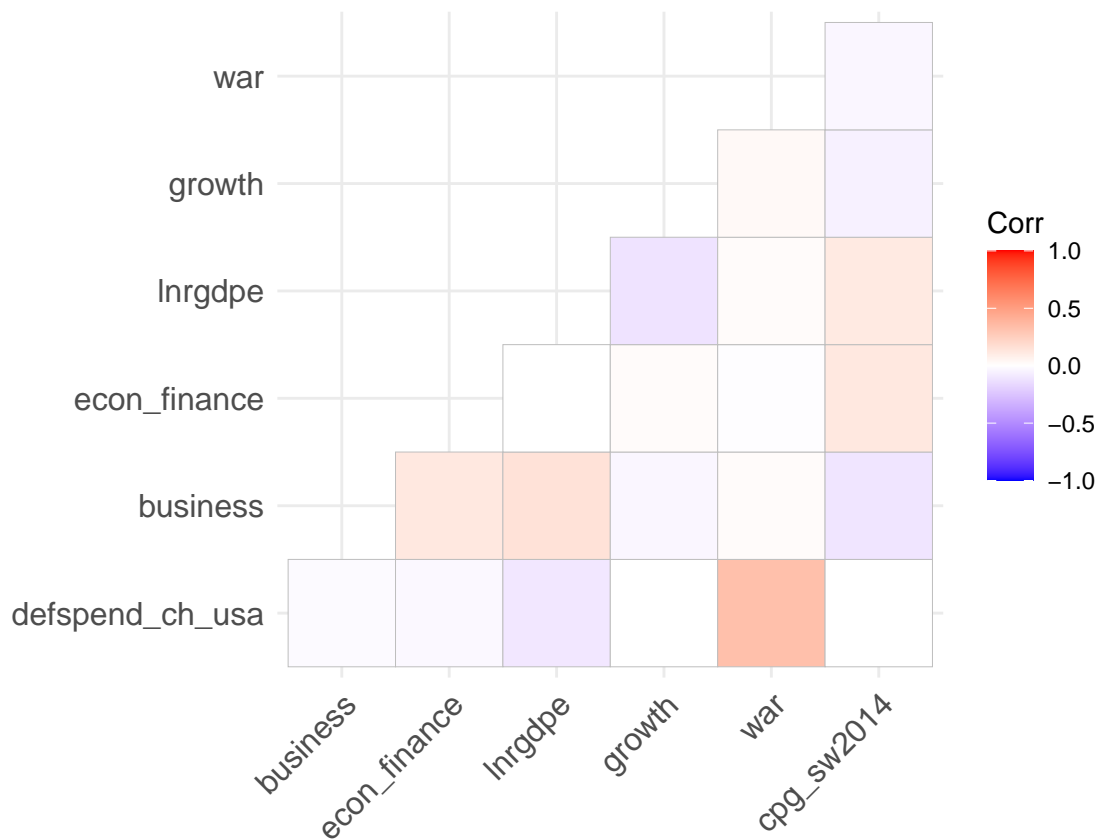


图 1: 变量间的相关系数

2.(10 分) 利用箱线图 (Box plots) 描述变量 `defspend_ch_usa` (Δ Defense Expenditures) 与 `business` (Business Experience) 的变化关系, 并简要描述数据分布有何特征或问题。

```
defense_spending %>%  
  filter(!is.na(business)) %>%  
  filter(!is.na(defspend_ch_usa)) %>%  
  ggplot(aes(y = defspend_ch_usa, x = factor(business))) +  
  geom_boxplot() +  
  scale_x_discrete(labels = c("No Business Experience", "Business Experience")) +  
  labs(x = "Business Experience",  
       y = "Change in Defense Spending")
```

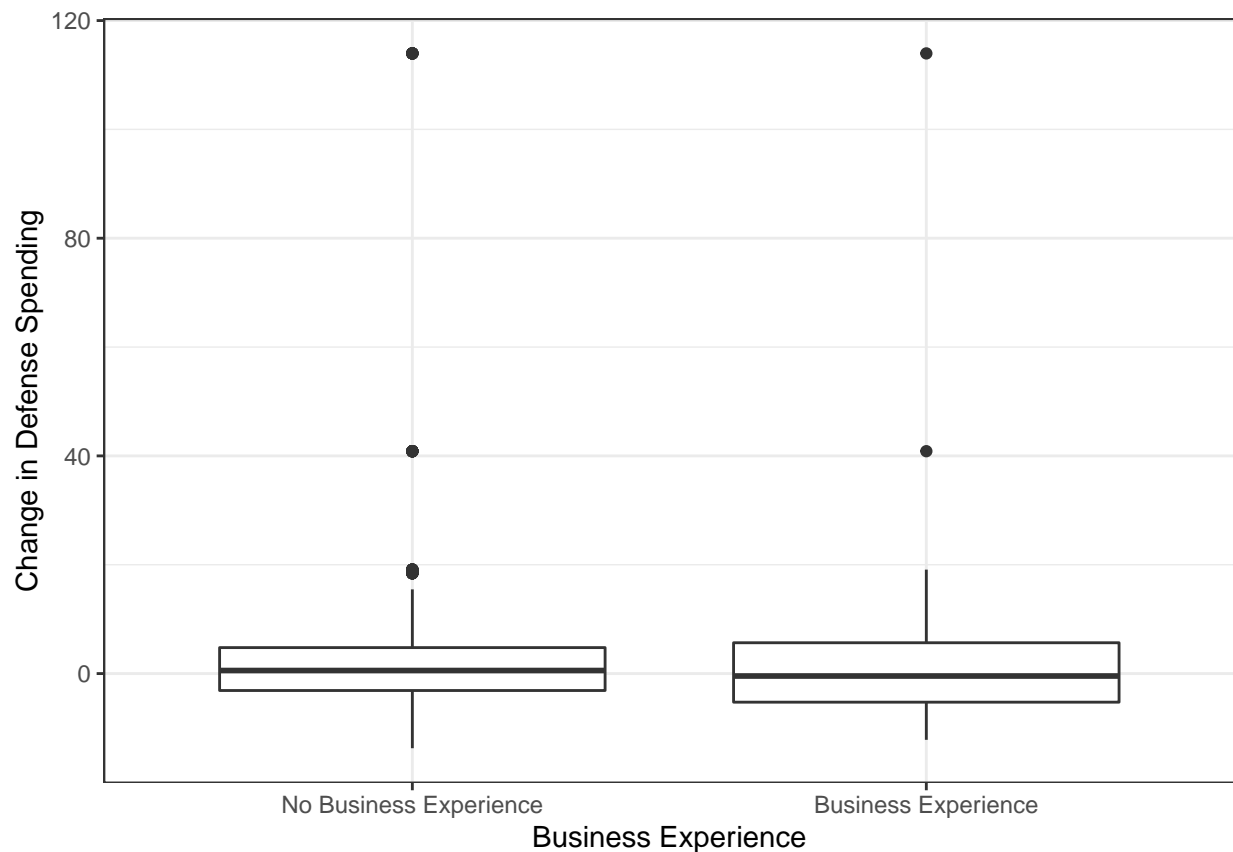


图 2: 经商经历与防务开支关系

- 数据存在奇异值，但是不能直接表明二者存在显著差别

3.(10 分) 利用 Welch Two Sample T-test 对于以下观点进行检验：拥有经商经历 (**business**) 的领导人在防务开支上 (**defspend_ch_usa**) 水平上存在不同。请说明零假设与备择假设分别是什么？根据 T-Test 结果，你得出什么结论？

- 以 μ 表示这两组防务开支的平均值，那么

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

```
t.test(defspend_ch_usa ~ business, data = defense_spending)

##
## Welch Two Sample t-test
##
## data:  defspend_ch_usa by business
## t = 1.0138, df = 228.26, p-value = 0.3117
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.082812  3.378087
## sample estimates:
## mean in group 0 mean in group 1
##      2.956654      1.809017
```

- p 值大于 0.1，不能拒绝零假设，即没有统计证据表明拥有经商经历与没有经商经历领导者在防务开支上存在显著的不同。

4.(10 分) 国际关系面板数据的定量分析中为了避免因变量与自变量存在“同时性偏差 (simultaneity bias)” (即, X_t 影响 Y_t , Y_t 也影响 X_t)，一般将自变量滞后一期 (lag one period)。如此，则变成了 X_{t-1} 影响 Y_t ，因为时间不可逆， Y_t 无法影响过去一期的 X_{t-1} 。利用 `dplyr` 中的 `lag` 命令，我们可以将部分变量滞后一年。请补充完成以下将变量滞后一年的代码，并重新提供一个如表-1 的描述性统计表格，简要说明观察量的变化。**注意：**新的描述性表格中的变量仅需包括新滞后一期的变量。

```
defense_spending <- defense_spending %>%
  group_by(ccode) %>%
  arrange(ccode, year) %>%
  mutate(business_lag1 = lag(business, n= 1L),
         econ_finance_lag1 = lag(econ_finance, n=1L),
         lnrgdpe_lag1 = lag(lnrgdpe, n = 1L),
         growth_lag1 = lag(growth, n = 1L),
         war_lag1 = lag(war, n = 1),
         cpg_sw2014_lag1 = lag(cpg_sw2014, n = 1L))

defense_spending %>%
  ungroup()%>%
  select(business_lag1, econ_finance_lag1, lnrgdpe_lag1,
```

```

      growth_lag1, war_lag1, cpg_sw2014_lag1) %>%
as.data.frame() %>%
stargazer(., type = "latex", header = FALSE, title = " 滞后一年变量的描述性统计",
          summary.stat = c("n", "mean", "median", "max", "min", "sd"),
          digit.separator = "")

```

表 2: 滞后一年变量的描述性统计

Statistic	N	Mean	Median	Max	Min	St. Dev.
business_lag1	910	0.153	0.000	1.000	0.000	0.360
econ_finance_lag1	910	0.310	0.000	1.000	0.000	0.463
lnrgdpe_lag1	900	12.543	12.607	15.093	8.182	1.429
growth_lag1	890	3.759	3.786	24.568	-16.735	3.897
war_lag1	910	0.042	0.000	1.000	0.000	0.200
cpg_sw2014_lag1	868	2.894	3.000	4.000	0.093	0.884

推论统计 (共 60 分)

5.(20 分) 利用 `lm` 命令和第 4 题的新数据, 估计以下这个线性回归模型, 并利用回归表格 (`stargazer`) 报告回归结果, 并解释自变量的回归系数和模型的 R^2 。

$$Y_{\Delta \text{Defense Expenditures}} = \beta_0 + \beta_1 * \text{Business Experience}_{t-1} + \epsilon$$

```

# 估计模型
fit1 <- lm(defspend_ch_usa ~ business_lag1, data = defense_spending)
stargazer(fit1, header = FALSE, title = " 回归统计结果")

```

- 在其他变量不变的情况下, 拥有经商经历的领导人相比于没有经商经历的领导人而言, 在防务开支上平均少支出 -2.3363407 个百分点。
- 模型的 R^2 是: 0.0035476, 表明这个模型解释了因变量全部方差的 0.3547595 %。

6.(30 分) 利用 `lm` 命令和第 4 题的新数据, 估计以下多元线性回归模型, 绘制回归系数图 (`dotwhisker`) 并对各个自变量的回归系数进行统计学上意义的解读。

表 3: 回归统计结果

	<i>Dependent variable:</i>
	defspend_ch_usa
business_lag1	−2.336* (1.299)
Constant	3.050*** (0.508)
Observations	910
R ²	0.004
Adjusted R ²	0.002
Residual Std. Error	14.102 (df = 908)
F Statistic	3.233* (df = 1; 908)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

$$\begin{aligned}
 Y_{\Delta \text{Defense Expenditures}} = & \beta_0 + \beta_1 * \text{Business Experience}_{t-1} + \\
 & \beta_2 * \text{Economics and Finance Experience}_{t-1} + \\
 & \beta_3 * \text{Economic Capacity}_{t-1} + \\
 & \beta_4 * \text{Economic Growth}_{t-1} + \\
 & \beta_5 * \text{War}_{t-1} + \beta_6 * \text{Government Ideology}_{t-1} + \epsilon
 \end{aligned}$$

```

fit2 <- lm(defspend_ch_usa ~ business_lag1 + econ_finance_lag1 + lnrgdpe_lag1 +
           growth_lag1 + war_lag1 + cpg_sw2014_lag1, data = defense_spending)
library(dotwhisker)
dwplot(list(fit2), conf.level = .95, show_intercept = TRUE,
        model_name = "model 1") +
  theme_bw() +
  ggtitle("Coefficient Plot ")

```

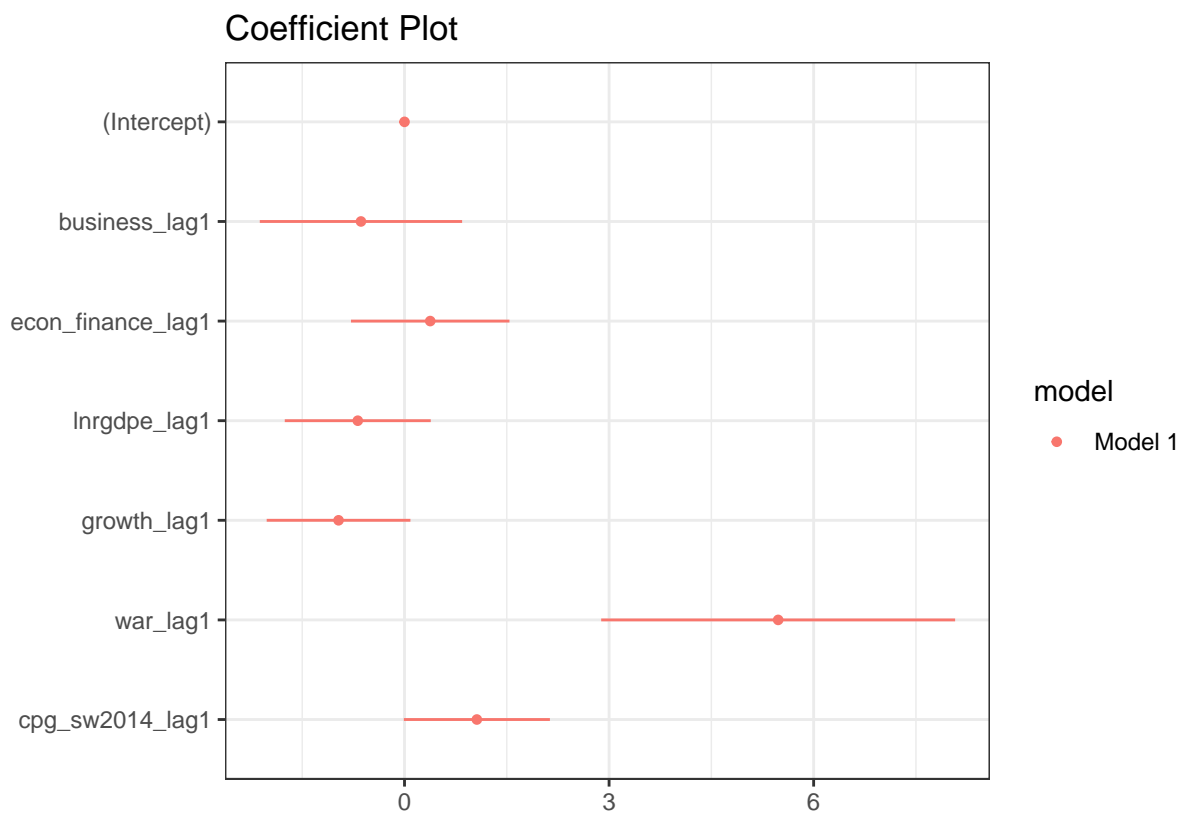


图 3: 多元回归模型结果

- 统计学解读: 变量 X 的回归系数为 β (正/负), 并且在 1%/5%/10% 的置信水平是统计上显著的,

- 这意味着在其他变量不变的情况下, X 每增加一个单位 1, 平均来说其国防务开支年增长率将增加/减少 β %。(这一部分不要就求)

7.(10 分) 根据第 4 题的数据和第 6 题的回归模型结果, 请预测当某北约国家的各自变量取为第 4 题表格中自变量的中位数 (median) 时, 该国预期的防务开支。**提示:** 利用第 4 题新表格中变量对应的中位数带入第 6 题中的回归模型, 算出预测值及其 95% 置信区间。

```
newdata <- defense_spending %>%
  ungroup()%>%
  summarise_at(c('business_lag1', 'econ_finance_lag1', 'lnrgdpe_lag1',
                 'growth_lag1', 'war_lag1', 'cpg_sw2014_lag1'),
               median, na.rm = TRUE)
#The confidence interval reflects the uncertainty around the mean predictions.
predict(fit2, newdata = newdata, interval = "confidence", level = .95)
```

```
##           fit           lwr           upr
## 1 1.270281 0.6068481 1.933713
```

```
## alternative
newdata <- defense_spending %>%
  ungroup()%>%
  summarise(across(ends_with("_lag1"), median, na.rm = TRUE))
#The prediction interval gives uncertainty around a single value.
predict(fit2, newdata = newdata, interval = "predict")
```

```
##           fit           lwr           upr
## 1 1.270281 -13.95798 16.49854
```

A prediction interval reflects the uncertainty around a single value, while a confidence interval reflects the uncertainty around the mean prediction values. Thus, a prediction interval will be generally much wider than a confidence interval for the same value.