Problem Set 2

Data Analysis and Visualization using R

RStudy2020

Due by 2020년 11월 2일 월요일 오전 12시 (한국시간)

Information & Instructions

출제자: 박상훈

데이터 파일

첫 번째 문제는 여기에서 2012년 미국 총선 시계열 설문조사 자료를 사용.

두 번째 문제는 여기의 'Quality of Governance의 시계열 자료를 사용.

제출

정해진 시간까지 정해진 문제에 대한 코드를 작성한 R 스크립트를 `sp23@email.sc.edu` 혹은 Dropbox의 `4_QnA or Discussion` 폴더에 업로드.

모든 코드는 다른 작업환경에서 열더라도 재생산가능하게 디렉토리 설정 등을 모두 고려한 결과물이어야 함.

Here starts the actual test

문제 1

`ANES 2012` 시계열 자료를 `anes`라고 하는 R의 객체로 저장하라.

```
library(ezpickr)
library(tidyverse)

anes <- pick("anes_timeseries_2012_stata12.dta")
```

A. 정당 일체감(party ID; `pid_x`) 변수를 민주당(매우 강한 민주당 지지, 약한 민주당 지지, 민주당 우호(leaning))일 경우 1, 공화당일 경우 0으로 나타내는 이항변수(binary variable)로 코딩하라. 이때, 무당파(independents)와 결칙치를 나타내는 값들(응답거부(refused), 무응답(no answer), 기타(etc))은 제외하라.

```
table(anes$pid_x)

anes <- anes %>%
  mutate(
    pid_x_dum = case_when(
      pid_x == -2L | pid_x == 4L ~ NA_integer_,
      # -2 무응답 및 응답거부, 4는 무당파(independents)
      pid_x < 4L ~ 1L,
      T ~ 0L
    )
  )
```

```
# 교차 확인
table(anes$pid_x, anes$pid_x_dum)
```

B. 교육수준을 나타내는 변수(dem_edu)를 숫자형(numerical) 변수로 재코딩하되, 모든 결측치와 other 카테고리를 포함하지 않도록 하라. 결과적으로 숫자형 변수로 재코딩된 변수는 1부터 16까지의 값을 가져야만 한다.

```
table(anes$dem_edu)
class(anes$dem_edu) # double, 즉 숫자형 변수임을 확인
anes <- anes %>%
  mutate(
    dem_edu_re =
      ifelse(anes$dem_edu == -9 | anes$dem_edu == 95, NA, anes$dem_edu)
    )
table(anes$dem_edu_re)
```

C. 민주당 지지자들과 공화당 지지자들 간의 평균 교육 수준의 차이가 존재하는지 여부를 평균 차이 분석을 통해 검증하라. 두 집단의 차이가 유의미하다고 할 수 있는가? 두 정당 지지 집단 간의 교육 수준의 평균 차이가 실질적으로 중요한 차이라고 생각되는지에 대해 자신의 의견을 서술하라.

```
anes %>%
     t.test(dem_edu_re ~ pid_x_dum, data = ., var.equal = TRUE)
```

t-검정의 결과는 p-값이 통상적인 유의수준의 기준인 0.05보다 작게 나타났으므로 우리는 민주당원과 공화당원 간 평균 교육수준에 차이가 없을 것이라는 영가설을 기각할 수 있는 경험적 근거를 가진다고 할 수 있습니다. 그러나 교육수준의 표준편차는 약 2보다 크고, 두 정당 일체감을 지는 유권자들의 평균적인 교육 수준의 차이는 -0.53이므로, 두 집단의 교육 수준의 차이는 실질적으로 큰 함의를 가지지 못하는 것으로 보입니다.

D. 새롭게 만든 교육 수준의 숫자형 변수의 측정 수준은 무엇인가 (e.g. 명목형, 순위형, 연속형 등)? 이 측정수준이 C에서 요구한 평균 차이를 검증하는 데 있어서 문제의 소지가 있는지 혹은 없다고 생각하는지 자신의 의견을 서술하라.

```
class(anes$dem_edu_re)
table(anes$dem_edu_re)
```

새롭게 만든 변수는 결측치와 무응답이 제거된 순위형 척도를 가지고 있으며, 숫자형 자료로 서로 다른 정당일체감을 가진 유권자들 간의 교육 수준의 평균적 차이가 존재하는지를 검증하는 데 사용되었습니다.

교육 수준의 변수를 연속형으로 사용하였을 때, 해석에 유의해야하는데 그 이유는 실제 교육 변수는 각 급간(interval)의 차이가 동등하지 않을 수 있기 때문입니다. 따라서 교육 수준의 한 단위 변화는 정확하게 민주당과 공화당의 서로 다른 정당일체감을 가진 집단의 교육 수준 차이를 정확하게 보여주지는 못한다는 사실을 인지해야 합니다.

E. 정당 일체감에 따른 평균 교육 수준의 차이가 존재하는지를 분석하기 위해 ANOVA를 시행하라. 이때, 정당일체감은 7점 척도의 변수를 사용하라(pid_x를 다시 사용하되, 결측치만 제외하고 일체감의 강도는 그대로 놔두어 분석하라).

```
anes <- anes %>%
  mutate(
    pid_x_con =
      ifelse(pid_x == -2 | pid_x == 4, NA, pid_x)
    )
table(anes$pid_x_con)

library(rstatix)

anes %>%
  anova_test(dem_edu_re ~ pid_x_con)
```

F. ANOVA의 결과를 어떻게 해석할 수 있는가? 이때, ANOVA의 영가설은 무엇인가? 분석 결과는 ANOVA의 영가설을 기각할만한 충분한 근거를 제시하는가? 집단 간의 차이를 보여주는 데 있어서 ANOVA가 가지는 한계는 무엇인가?

F-Answer. ANOVA의 영가설은 정당일체감이 서로 다른 유권자 집단들 간의 교육 수준의 차이가 존재하지 않는다는 것입니다. 즉, 모든 집단의 평균이 동일하다는 것입니다. 그러나 ANOVA 결과는 최소 하나 이상의 집단이 교육 수준의 평균에 있어서 통계적으로 유의미한 차이를 가지고 있다는 것을 보여줍니다.

ANOVA는 영가설을 기각할만한 충분한 근거를 제시하지만 어떤 그룹이 차이가 있는지를 특정하지는 못한다는 점에서 구체적인 정보를 제공하지 않는다는 한계가 있습니다.

문제 2.

`Quality of Governance` 시계열 자료를 `qog`라고 하는 R의 객체로 저장하라.

```
qog <- pick("http://www.qogdata.pol.gu.se/data/qog_std_ts_jan20.dta")
```

A. 가장 최근의 `qog` 데이터를 사용하여 각 지역(`ht_region`), 각 연도, 국가간 분쟁 경험 횟수 (`UCDP-PRIO`의 inter-state armed conflict 변수를 사용할 것)를 요약하여 보여주는 데이터셋, `qog_agg`을 만들어라. 하지만 단순하게 분쟁 경험 횟수만을 측정하는 변수가 아니라 무력 분쟁 변수를 3개의 항목을 가진 분류형 변수로 재구성하라: (1) 0: 국가간 분쟁을 경험한 횟수가 없는 경우, (2) 1: 국가 간 분쟁을 한 번 경험한 경우, (3) 2: 2번 이상의 국가간 분쟁 경험이 있는 경우.

```
qog_small <- qog %>%
  select(ccode, cname, year, ucdp_type2, ht_region)

table(qog_small$ht_region)

qog_agg <- qog %>%
  group_by(ht_region, year) %>%
  summarize(
    sum.intlconflict = sum(ucdp_type2, na.rm = T)
  ) %>% mutate(
    intl.conflict = case_when(
      sum.intlconflict == 0L ~ 1L,
      sum.intlconflict == 1L ~ 2L,
      T ~ 3L
    )
  )

head(qog_agg)
glimpse(qog_agg)
```

B. 새롭게 구축한 데이터셋의 분석 단위(level of analysis)는 무엇인가? 그리고 새롭게 구축한 무력분쟁 변수의 측정 수준은 무엇인가?

B-Answer. 새롭게 만든 데이터셋, `qog_agg`의 분석수준은 체계 수준(지역)이라고 할 수 있습니다. 또한 분석단위는 지역-연도(region-year)입니다. 그리고 무력분쟁(armed conflict) 변수의 측정수준은 순위형입니다.

C. 지역과 무력분쟁 변수 간의 관계를 보여주는 교차표를 만들어라. 두 변수를 대상으로 카이스퀘어(χ^2) 검정을 수행하고, 각 변수의 관측치들이 서로 독립적이라는 카이스퀘어 검정을 충족시키는지를 확인하라.

```
table(qog_agg$intl.conflict, qog_agg$ht_region)
```

```
model.region.conflict <-
  xtabs( ~ ht_region + intl.conflict, data = qog_agg)

Chisq <- chisq.test(model.region.conflict)
Chisq
```

D. 카이스퀘어 검정 결과를 설명하라. 카이스퀘어 검정의 영가설은 무엇인가? C의 결과는 카이스퀘어 검정의 영가설을 기각할 수 있는가? 카이스퀘어 검정의 한계는 무엇인가?

D-Answer. 카이스퀘어 검정의 영가설은 두 변수 간 관계가 없다(독립적이다)는 것입니다. 카이스퀘어 검정 결과, p-값이 통상적인 유의수준 기준인 0.05보다 작으므로 영가설을 기각하기에 충분한 경험적 근거를 갖추었다고 할 수 있습니다.

카이스퀘어 검정을 통해 우리는 관계의 존재유무와 강도에 대해서 확인할 수 있습니다. 하지만 카이스퀘어 검정은 그 관계의 방향성에 대해서는 알려주지 않습니다.

E. 만약 지역별로 얼마나 많은 국가간 분쟁이 발생하였는지를 알고 싶다고 할 때, 우리가 재조작하여 만든 분쟁 변수가 가질 수 있는 문제점은 무엇인가?

E-Answer. 국가간 분쟁 변수는 국가들이 얼마나 많은 분쟁을 경험하였는지, 그리고 그 분쟁이 지역별로 어떻게 분포되었는지를 보여줍니다. 이 변수의 문제는 동일한 지역에서 서로 다른 두 국가가 분쟁을 경험할 경우와 서로 다른 지역에 속한 국가간 분쟁이 있었을 경우 두 사례를 제대로 구분하지 못한다는 것에 있습니다.

만약 분쟁을 경험한 두 국가가 한 지역에 속해있다면, 해당 지역에 대한 국가간 분쟁 변수는 2회로 셀 것입니다. 그러나 만약 두 국가가 서로 다른 지역에 속해 있다면, 국가간 분쟁 변수는 각기 다른 지역에 대해 분쟁을 1회씩 계산할 것입니다. 문제는 전자의 경우 한 지역의 분쟁은 그 해에 동일한 지역 내의 분쟁이므로 1회로 계산해야하지만 2회로 계산되어 국가간 분쟁이 과다하게 측정될 수 있다는 문제가 존재합니다.

또한 이 변수는 국가간 분쟁에 관여한 국가의 수가 3개국 이상일 경우, 분쟁의 변량(variations)을 제대로 포착하지 못한다는 한계를 지닙니다.