221105 Data Assignment 2

Jongrak

2022-11-08

Default Setting

Reset and clear the working environment

```
rm(list = ls())
```

Set the working directory

```
setwd("~/Library/Mobile Documents/com~apple~CloudDocs/Study/1_Univ/Lecture/2022-2/POLI223 POLITICAL METHODOLOGY/Data Assignment/Data Assingment 2")
```

Load the package

```
library(foreign)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Examine the data 'bes.dta'

Load the dataset

```
bes <- read.dta("bes.dta")
```

Overview of 'bes.dta'

```
dim(bes)
```

```
## [1] 30895     4
```

```
head(bes)
```

```
##                    vote       edlevel age leave
## 1 Stay/remain in the EU Undergraduate  26     0
```

```
## 2 Stay/remain in the EU Undergraduate  61     0
## 3 Stay/remain in the EU     GCSE A*-C  55     0
## 4 Stay/remain in the EU     GCSE A*-C  20     0
## 5 Stay/remain in the EU      Postgrad  66     0
## 6 Stay/remain in the EU Undergraduate  27     0
```

Q1. What percentage of people answer that they will vote to leave the EU (among those who expressed their intention to vote)? Do you think the survey results provided a clear prediction of the outcome of the referendum? Why, or why not?

48.82 % of people answer they will vote to leave the EU. We should remove NAs of 'leave' in that those mean the people who did not express their intention.

These results are statistically significant and can just provide the credible statistical estimation - still uncertain. In other words, these results can predict but still have uncertainty, hence, are not 'clear'.

1) By Law of Large Numbers, the mean value of samples converges to the one of population as the size of sample goes infinite or gets large enough. In this case, the size of sample is '28044'(without 'NA'), and it can be said to be large enough, hence, we can estimate the real outcome of the referendum with the mean value of this sample.

2) Also, by Central limit theorem, the distribution of specific samples' standardized sum converges to standard normal distribution. With this theorem, we can estimate the population mean and statistically evaluate the confidence. The 95% confidence interval of this sample mean is [48.22629, 49.42027].

This means just statistical significance and it can be different from the population or real value. In fact, the real mean value of the referendum is about '51.9%'.

https://www.bbc.co.uk/news/politics/eu_referendum/results

Mean value of 'leave' without 'NA'

```
mean_leave <- mean(bes$leave, na.rm=TRUE)
100 * mean_leave
```

```
## [1] 48.82328
```

Is this a clear prediction?

```
# Check the size of sample
bes1 <- select(bes, leave) %>%
  na.omit()

dim(bes1)
```

```
## [1] 28044     1
```

```
# Check the confidence interval(95%)
sd_leave <- sd(bes1$leave)/sqrt(28044)
```

```
interval0 <- mean_leave - (sd_leave * 2)
interval1 <- mean_leave + (sd_leave * 2)
interval_leave <- c(interval0, interval1)

100 * interval_leave

## [1] 48.22629 49.42027
```

Q2. What is the mean value of age? What is the proportion of people who received college education (undergraduate and postgraduate degree). How does that compare with the actual demographic pattern in the UK? Do you think the survey sample is representative of the population?

The mean value of age is about '50'. and the proportion of people who received college education is about '42.29 %' This value is similar to real value according to HESA's report.

"Labour Force Survey data published by the Office for National Statistics (ONS) in November 2017 shows how UK Higher Education contributes to the skills level of the nation. In July to September 2017, 42% of the UK population aged 21 to 64 had achieved higher education qualifications."

https://www.hesa.ac.uk/news/11-01-2018/sfr247-higher-education-student-statistics/qualifications

Of course, in the aspect of education level, this survey sample can be a representative of the population. However, there are other factors we need to consider such as region, age and wage. Therefore, we cannot easily say this sample is perfect representative of the population. In fact, the real average age of UK population in 2016 is about '40' according to statista's survery report.

https://www.statista.com/statistics/281288/median-age-of-the-population-of-the-uk/

The mean value of age

```
mean(bes$age) # There's no 'NA'

## [1] 50.75025
```

The proportion of people who received college education (undergraduate and postgraduate degree)

```
bes2 <- select(bes, edlevel) %>%
  na.omit()

edlevel <- prop.table(table(bes2$edlevel))
edlevel <- 100 * edlevel

edlevel_college <- edlevel[5] + edlevel[6]
names(edlevel_college) <- c("College education")
edlevel_college <- as.table(edlevel_college)

edlevel
```

```
##
## No qualifications          GCSE D-G         GCSE A*-C           A-level
##          7.840377          4.945410         22.376079         22.539658
##     Undergraduate          Postgrad
##          32.438087          9.860387

edlevel_college

## College education
##          42.29847
```

Q3. Next, examine the relationship between education and position on the EU referendum using the survey data (bes.dta). Let's create the binary variable (college_education) which is coded 1 for individuals who received undergraduate or postgraduate degree and 0 otehrwise, and examine the relationship between college education and position on the EU referendum. Which test would you use to examine the relationship between the two variables? onduct the test and interpret the results.

We can use the tabular analysis in that IV - Higher education or not - is categorical and DV - vote to leave or not - is categorical. By tabular chi-squared test, we can conclude that the education level and intention to vote to leave are related and it is statistically significant.

However, we should not easily conclude that the differences of education level 'cause' the differences of intention to vote to leave. This result means just 'correlation', and does not mean 'causal relation'. Therefore, we should examine the likelihood of reverse causality, and the causal mechanism. Also, It should be examined whether we controlled possible confounding variables or not.

First of all, it is seen that there is no likelihood of reverse causality in that we can expect the intention to vote to leave rarely affect on the education level. Economic or political knowledge can be a causal mechanism that links IV and DV. Education level affect on the difference of those knowledge and it can affect the attitude toward membership as EU. However, there can be confounding variables such as the difference of household income. This variable can affect on both of the education level and the intention to vote to leave.

Select and mutate the dataset

```
bes_college0 <- select(bes, edlevel, leave) %>%
  mutate(edlevel = ifelse(edlevel == "Postgrad" | edlevel == "Undergraduate",
1, 0)) %>%
  na.omit()
```

Label and make a table

```
bes_college1 <- mutate(bes_college0,
                       edlevel = ifelse(edlevel == 1, "HE", "Non-HE"),
                       leave = ifelse(leave == 1, "Yes", "No"))
bes_college_table <- prop.table(table(bes_college1))

bes_college_table
```

```
##         leave
## edlevel         No       Yes
##   HE     0.2871122 0.1440997
##   Non-HE 0.2323743 0.3364138
```

Tabular Chi-Squared Test

```
chisq.test(bes_college1$edlevel, bes_college1$leave)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  bes_college1$edlevel and bes_college1$leave
## X-squared = 1554.4, df = 1, p-value < 2.2e-16
```

Examine the data 'brexit_vote.dta'

Load the dataset

```
brexit <- read.dta("brexit_vote.dta")
```

Overview

```
dim(brexit)

## [1] 382  10

head(brexit)

##       region                 area leave_pct median_hourly_pay  eu_migrant
## 1 North East           Hartlepool     69.57              9.27 0.004085272
## 2 North East        Middlesbrough     65.48              8.58 0.006725742
## 3 North East Redcar and Cleveland     66.19             10.77 0.005361815
## 4 North East     Stockton-on-Tees     61.73             10.60 0.005722838
## 5 North East           Darlington     56.18              9.05 0.009607719
## 6 North West               Halton     57.42              9.35 0.004762791
##   non_eu_migrant eu_migrant_growth non_eu_migrant_growth unemployment_rate
## 1    0.011939827       0.005721637           0.005495932              10.1
## 2    0.032649883       0.010700382           0.030981424              12.1
## 3    0.012858293       0.002637783          -0.001487796               8.3
## 4    0.021843191       0.005588314           0.013794224               6.8
## 5    0.020574828       0.015781189           0.007962141               6.6
## 6    0.009559421       0.005422645           0.003933744               4.9
##   high_education
## 1      0.1140638
## 2      0.1282774
## 3      0.1246839
## 4      0.1517297
## 5      0.1645644
## 6      0.1131531
```

Q4. Let's now examine the relationship between education and support for the Brexit using the district-level data (brexit_vote.dta). We would like to examine the relationship between high_education and leave_pct. Which test would you use to examine the relationship between the two variables? Conduct the test and interpret the results.

Both of IV and DV are continuous variables, hence, we can examine the correlation coefficient. However, the regression model can be more useful in order to examine the relationship. Here, we would like to examine the relationship between two variables, we can use bivariate regression.

According to the result, the coefficient of 'high_education' is about '-1.12662' which means that it has negative correlation with the percentage of leaving voters. And it seems that it has statistically significance, hence, it has strong correlation. And we can expect or assume the causal mechanism that as the share of higher education population get higher, the share of high-skilled and high-wage people get higher. Those people may have more supportive attitude toward EU membership or free trade, and it may affect the percentage of leaving.

Refine and mutate the dataset

```
brexit0 <- mutate(brexit, eu_migrant = eu_migrant * 100,
                  non_eu_migrant = non_eu_migrant * 100,
                  eu_migrant_growth = eu_migrant_growth * 100,
                  non_eu_migrant_growth = non_eu_migrant_growth * 100,
                  high_education = high_education * 100)

brexit1 <- select(brexit0, leave_pct, high_education)
```

Bivariate Regression (The share of high-level educated and the percentage of 'leave')

```
brexit_lm1 <- lm(formula = brexit1$leave_pct ~ brexit1$high_education, data =
brexit1)

summary(brexit_lm1)

##
## Call:
## lm(formula = brexit1$leave_pct ~ brexit1$high_education, data = brexit1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.068  -2.162   1.284   3.791  17.396
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             75.03778    0.93945   79.87   <2e-16 ***
## brexit1$high_education -1.12662    0.04527  -24.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.42 on 378 degrees of freedom
##   (2 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.621,  Adjusted R-squared:   0.62
## F-statistic: 619.4 on 1 and 378 DF,  p-value: < 2.2e-16
```

Q5. Choose one continous variable that you think was an important factor on voters' decisions on the Brexit. Why do you think the factor was important? Which test would you use to examine the relationship between the two variables? Conduct the test and interpret the results.

'eu_migrant(EU migrant resident share 2001)' can be a important factor on voters' decisions in that it may voters' attitude toward international economic system by EU membership. Also, it may affect on the share of higher educated people.

We can examine the relationship with the multiple regression models. According to the results, 'high_education' factor has a negative coefficient, of course there are some changes of the magnitude, whenever other factors are added. It means that the share of resident population with higher education affects on the voters' decision and it is negative direction. And it has statistically significance.

We assumed the causal mechanism that as the share of higher education population get higher, the share of high-skilled and high-wage people get higher. However, when we consider the factor 'median_hourly_pay', it has a positive effect on the percentage of leaving vote. Therefore, we need to more precisely examine this factor or need to find out other causal mechanisms.

Considering the factor 'EU migrant resident growth', it has no statistically significance.

And although it is not a continuous variable, considering the factor region, Scotland or not, the factor 'high_education' still has negative coefficient. Interestingly, the coefficient of the factor 'eu_migrant' decreased as the factor 'Scotland or not' is added.

With EU migrant resident share (2001)

```
brexit2 <- select(brexit0, high_education, leave_pct, eu_migrant)

brexit_lm2 <- lm(formula = brexit2$leave_pct ~ brexit2$high_education + brexi
t2$eu_migrant)

summary(brexit_lm2)

##
## Call:
## lm(formula = brexit2$leave_pct ~ brexit2$high_education + brexit2$eu_migra
nt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.577  -2.434   1.215   4.122  18.786
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              77.30149    1.03700  74.543  < 2e-16 ***
## brexit2$high_education -1.39859    0.07336 -19.064  < 2e-16 ***
```

```
## brexit2$eu_migrant      2.29757    0.49539   4.638 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.253 on 377 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6415, Adjusted R-squared:  0.6396
## F-statistic: 337.3 on 2 and 377 DF,  p-value: < 2.2e-16
```

Add median hourly pay (2005)

```
brexit3 <- select(brexit0, high_education, leave_pct, eu_migrant, median_hour
ly_pay)

brexit_lm3 <- lm(formula = brexit3$leave_pct ~ brexit3$high_education + brexi
t3$eu_migrant + brexit3$median_hourly_pay)

summary(brexit_lm3)

##
## Call:
## lm(formula = brexit3$leave_pct ~ brexit3$high_education + brexit3$eu_migra
nt +
##     brexit3$median_hourly_pay)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.844  -2.087   0.989   3.620  15.995
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                68.38710    1.97447  34.636  < 2e-16 ***
## brexit3$high_education    -1.61928    0.08248 -19.633  < 2e-16 ***
## brexit3$eu_migrant         1.99889    0.48224   4.145 4.20e-05 ***
## brexit3$median_hourly_pay  1.23996    0.23661   5.240 2.68e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.045 on 376 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6659, Adjusted R-squared:  0.6632
## F-statistic: 249.8 on 3 and 376 DF,  p-value: < 2.2e-16
```

Add EU migrant resident growth (2001-2011)

```
brexit4 <- select(brexit0, high_education, leave_pct, eu_migrant, median_hour
ly_pay, eu_migrant_growth)

brexit_lm4 <- lm(formula = brexit4$leave_pct ~ brexit4$high_education + brexi
t4$eu_migrant + brexit4$median_hourly_pay + brexit4$eu_migrant_growth)

summary(brexit_lm4)

##
## Call:
## lm(formula = brexit4$leave_pct ~ brexit4$high_education + brexit4$eu_migra
nt +
##     brexit4$median_hourly_pay + brexit4$eu_migrant_growth)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.761  -2.076   0.898   3.454  15.951
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                69.07183    2.01647  34.254  < 2e-16 ***
## brexit4$high_education     -1.61887    0.08231 -19.668  < 2e-16 ***
## brexit4$eu_migrant          2.34593    0.52800   4.443 1.17e-05 ***
## brexit4$median_hourly_pay   1.18912    0.23826   4.991 9.22e-07 ***
## brexit4$eu_migrant_growth -0.26930    0.16854  -1.598    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.032 on 375 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6681, Adjusted R-squared:  0.6646
## F-statistic: 188.7 on 4 and 375 DF,  p-value: < 2.2e-16
```

ANOVA

```
anova(brexit_lm4, brexit_lm3)

## Warning in anova.lmlist(object, ...): models with response '"brexit3$leave
_pct"'
## removed because response differs from model 1

## Analysis of Variance Table
##
## Response: brexit4$leave_pct
##                             Df  Sum Sq Mean Sq  F value    Pr(>F)
## brexit4$high_education       1 25533.6 25533.6 701.7469 < 2.2e-16 ***
## brexit4$eu_migrant           1   841.1   841.1  23.1154 2.212e-06 ***
## brexit4$median_hourly_pay    1  1003.3  1003.3  27.5753 2.537e-07 ***
## brexit4$eu_migrant_growth    1    92.9    92.9   2.5529    0.1109
## Residuals                  375 13644.7    36.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What about region...? Especillay Scotland or not.

```
brexit5 <- select(brexit0, high_education, leave_pct, eu_migrant, median_hour
ly_pay, region) %>%
  mutate(region = ifelse(region == "Scotland", 1, 0))

brexit_lm5 <- lm(formula = brexit5$leave_pct ~ brexit5$high_education + brexi
t5$eu_migrant + brexit5$median_hourly_pay + brexit5$region)

summary(brexit_lm5)

##
## Call:
## lm(formula = brexit5$leave_pct ~ brexit5$high_education + brexit5$eu_migra
nt +
##     brexit5$median_hourly_pay + brexit5$region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.9213  -2.4291   0.3045   2.7833  15.6933
```

```
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 71.47460    1.50622  47.453  < 2e-16 ***
## brexit5$high_education      -1.37941    0.06407 -21.531  < 2e-16 ***
## brexit5$eu_migrant           0.66539    0.37368   1.781   0.0758 .
## brexit5$median_hourly_pay    0.80611    0.18101   4.454 1.12e-05 ***
## brexit5$region             -14.70531    0.87735 -16.761  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.576 on 375 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.809,  Adjusted R-squared:  0.8069
## F-statistic:   397 on 4 and 375 DF,  p-value: < 2.2e-16
```