Assignment: Lab Nº3 - ADV. ANALYTIC TECHNIQUES

Sebastian Urbina

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(ggplot2)
setwd("/Volumes/TOSHIBA EXT/0.1 Thesis/data/output")
df1 <- read.csv("mset.var_V_01.csv")
df1 <- df1 %>%
mutate(AGE = edad)

df1 <- df1 %>%
  mutate(education = ifelse(esc_nivel_1 == 99, NA, esc_nivel_1))

df1 <- df1 %>%
mutate(Percentage_k = percentage_k * 100)

df1 <- df1 %>%
mutate(gse_m = case_when(
    gse_t == "ABC1" ~ "ABC1-C2",
```

```
    gse_t == "C3" ~ "C3",
    gse_t == "D_E" ~ "DE",
    TRUE ~ NA
  ))

df1 <-  df1 %>%
  mutate(y_b = case_when(
    y == 0  ~ 1,
    y == 1  ~ 0,
    TRUE ~ NA
  ))


regions_to_exclude <- c('region de arica y parinacota', 'region de tarapaca',
                        'region de atacama', 'region de aysen del gral. carlos ibanez del
                        'region de magallanes y de la antartica chilena')

df2 <- df1 %>%
  filter(!(region_name %in% regions_to_exclude))
```

1. First run a naive OLS with clustered standard errors. Interpret it.

For the following analysis, I will use a Chilean survey dataset created by the CEP Institute. The universe of the study was all the individuals above 17 years old and residents of the country, achieving a sample of 1467 individuals. They interviewed each of them in their homes, and the sample method was stratified, random, and probabilistic in each of its three stages. I chose this dataset because it is among the most prestigious public opinion surveys, having high methodological standards.

In this study, we aim to investigate how individual political beliefs are influenced by factors such as sex, age, and presidential vote choice. Furthermore, we seek to examine how regional differences may also affect individual beliefs. To achieve this, we will employ a multilevel analysis approach.

The ideology variable is a scale ranging from 1 to 10, where 1 represents a left-leaning ideology and 10 represents a right-leaning ideology. The gender variable is coded as 1 for female and 0 for male. The presidential vote variable is coded as 1 for those who voted for the left-wing candidate in the 2021 election and 0 for those who voted for the right-wing candidate in the same election. The region variable identifies the region where each respondent resides. Some regions were discarded due to a low response rate in those areas.

```r
library(QMSS)
```

```
Loading required package: lme4

Loading required package: Matrix


Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack

Loading required package: lmtest

Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric

Loading required package: MASS


Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select

Loading required package: plm


Attaching package: 'plm'
```

```
The following objects are masked from 'package:dplyr':

    between, lag, lead

Loading required package: plyr

------------------------------------------------------------------------------

You have loaded plyr after dplyr - this is likely to cause problems.
If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
library(plyr); library(dplyr)

------------------------------------------------------------------------------


Attaching package: 'plyr'

The following objects are masked from 'package:dplyr':

    arrange, count, desc, failwith, id, mutate, rename, summarise,
    summarize

The following object is masked from 'package:purrr':

    compact

Loading required package: rdd

Loading required package: sandwich

Loading required package: AER

Loading required package: car

Loading required package: carData


Attaching package: 'car'
```

```
The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some

Loading required package: survival

Loading required package: Formula

Loading required package: VGAM

Loading required package: stats4

Loading required package: splines


Attaching package: 'VGAM'

The following object is masked from 'package:AER':

    tobit

The following object is masked from 'package:car':

    logit

The following object is masked from 'package:plm':

    has.intercept

The following object is masked from 'package:lmtest':

    lrtest
```

```r
df2 <- df2 %>%
  mutate(pol = ifelse(iden_pol_2 == 99, NA, iden_pol_2))

df2 <- df2 %>%
  mutate(pol = ifelse(pol == 88, NA, pol))


lm.pol <- lm(pol ~ y_b + female + AGE, data = df2)

clusterSE(fit = lm.pol, cluster.var = "region_name", data = df2)
```

```
t test of coefficients:

              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  6.6130872  0.3212810  20.5835   <2e-16 ***
y_b         -2.4290204  0.1887430 -12.8695   <2e-16 ***
female       0.0884901  0.1718117   0.5150   0.6067
AGE          0.0029532  0.0054636   0.5405   0.5890
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Controlling for other factors, individuals who voted for the left-wing candidate in the 2021 election exhibit ideological beliefs that are, on average, 2.4 points lower compared to those who voted for the right-wing candidate. This difference is statistically significant.

2. Then run an empty (random intercept) model. Interpret it.

```r
null_lm <- lmer(pol ~ (1 | region_name), data = df2, REML = FALSE)
summary(null_lm )
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: pol ~ (1 | region_name)
   Data: df2

     AIC      BIC   logLik deviance df.resid
  4290.8   4305.4  -2142.4   4284.8      963

Scaled residuals:
```

```
     Min       1Q   Median       3Q      Max
-2.16411 -0.23503 -0.08529  0.30956  2.18443

Random effects:
 Groups      Name        Variance Std.Dev.
 region_name (Intercept) 0.06021  0.2454
 Residual                4.90655  2.2151
Number of obs: 966, groups:  region_name, 11

Fixed effects:
            Estimate Std. Error t value
(Intercept)   5.3801     0.1124   47.88
```

```
rho(null_lm)
```

```
[1] 0.01212333
```

Rho means that 1.21% of the variance in the scale of ideology is between different regions.

3. Then run a full random-intercept model. Interpret that.

```
lmer.pol <- lmer(pol ~ y_b + AGE + female +  (1 | region_name), data = df2, REML = FALSE)
summary(lmer.pol)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: pol ~ y_b + AGE + female + (1 | region_name)
   Data: df2

     AIC      BIC   logLik deviance df.resid
  2569.7   2596.2  -1278.9   2557.7      597

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9893 -0.7390  0.2028  0.4002  2.8441

Random effects:
 Groups      Name        Variance Std.Dev.
 region_name (Intercept) 0.05616  0.237
 Residual                4.03503  2.009
```

```
Number of obs: 603, groups:  region_name, 11

Fixed effects:
             Estimate Std. Error t value
(Intercept)  6.716785   0.305875  21.959
y_b         -2.411474   0.169937 -14.190
AGE          0.002704   0.004779   0.566
female       0.070487   0.166278   0.424

Correlation of Fixed Effects:
       (Intr) y_b    AGE
y_b    -0.391
AGE    -0.805  0.095
female -0.350 -0.057  0.075
```

```
rho(lmer.pol)
```

```
[1] 0.01372649
```

In the random intercept model, the coefficients are virtually the same as those in the OLS model. The coefficient for voting for the left candidate (y_b) was -2.43 and is now -2.41. The age coefficient was 0.002 and remains unchanged at 0.002. The female coefficient was 0.09 and is now 0.07.

Now, the value of rho implies that only 1.37% of the variance in the ideology scale is due to unobserved differences between communities.

On the other hand, the standard deviation of the region is practically the same at 0.237, meaning that there is a significant amount of variation between states. Moreover, the constant has a mean of 6.716785 and a standard deviation of 0.254.

4. Then, lastly, run a random-intercept, random-slope model. Interpret that.

```
lmer.pol <- lmer(pol ~ y_b + AGE + female +  ( y_b| region_name), data = df2, REML = FALSE
summary(lmer.pol)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: pol ~ y_b + AGE + female + (y_b | region_name)
   Data: df2
```

```
     AIC      BIC   logLik deviance df.resid
  2573.3   2608.5  -1278.7   2557.3      595

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0104 -0.7442  0.2207  0.3930  2.8416

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 region_name (Intercept) 0.16046  0.4006
             y_b         0.09929  0.3151   -0.89
 Residual                4.01497  2.0037
Number of obs: 603, groups:  region_name, 11

Fixed effects:
             Estimate Std. Error t value
(Intercept)  6.679304   0.328039  20.361
y_b         -2.395074   0.209502 -11.432
AGE          0.002736   0.004771   0.573
female       0.073012   0.166564   0.438

Correlation of Fixed Effects:
       (Intr) y_b    AGE
y_b    -0.528
AGE    -0.748  0.076
female -0.338 -0.018  0.072
```

For a 0-year-old man who voted for the right-wing candidate, on average, they have a score of 6.68 on the ideology scale. However, there is substantial variation around that mean, with a standard deviation of 0.4006.

On the other hand, on average, people who voted for the left candidate in the 2021 election have a score that is 2.34 points lower on the ideology scale (compared to right-wing voters). Nevertheless, there is substantial variation around that average slope, with a standard deviation of 0.3151.

There is a very high negative correlation (= -0.89) between the constant and the slope for each community.

In conclusion, regions with above-average scores on the ideology scale (in their constant) have below-average values on the slope for voting for the left candidate in the 2021 election – and vice versa. This could be explained by the notion that in some regions, political ideologies may have deeper roots in the territory, which also affects the voting patterns.

5. Feel free to add in cross-level interactions too.

```
lmer.pol <- lmer(pol ~ y_b*female + AGE  +  (y_b| region_name), data = df2, REML = FALSE)
summary(lmer.pol)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: pol ~ y_b * female + AGE + (y_b | region_name)
   Data: df2

     AIC      BIC   logLik deviance df.resid
  2575.2   2614.8  -1278.6   2557.2      594

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9943 -0.7517  0.2140  0.4048  2.8300

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 region_name (Intercept) 0.15841  0.3980
             y_b         0.08792  0.2965   -0.89
 Residual                4.01487  2.0037
Number of obs: 603, groups:  region_name, 11

Fixed effects:
            Estimate Std. Error t value
(Intercept)  6.725439   0.346196  19.427
y_b         -2.470324   0.279231  -8.847
female      -0.003309   0.269493  -0.012
AGE          0.002773   0.004772   0.581
y_b:female   0.123079   0.341629   0.360

Correlation of Fixed Effects:
           (Intr) y_b    female AGE
y_b        -0.585
female     -0.453  0.523
AGE        -0.701  0.043  0.028
y_b:female  0.324 -0.677 -0.786  0.022
```

For women who voted for the left candidate in the 2021 election, their score on the ideology scale is 0.12 points higher compared to men who voted for the left candidate, holding all other variables constant.