Exercise 1: Pooling

Henrik Singmann

January 2019 (updated: 2019-05-19)

Skovgaard-Olsen et al. (2016): The Relevance Effect and Conditionals

- Conditional = *if-then* statement; e.g., If global warning continues, London will be flooded.
- Bayesian reasoning often assumes 'the Equation': $P(\text{if } A \text{ then } B) = P(B|A)$
- Our question: Does the Equation hold?
- 94 (of 348) participants recruited via `crowdflower.com` worked on 12 items.
- Participant first saw a vignette:

  > Sophia's scenario: Sophia wishes to find a nice present for her 13-year-old son, Tim, for Christmas. She is running on a tight budget, but she knows that Tim loves participating in live role-playing in the forest and she is really skilled at sewing the orc costumes he needs. Unfortunately, she will not be able to afford the leather parts that such costumes usually have, but she will still be able to make them look nice.

- Then we asked participant for their rating for the conditional probability $P(B|A)$ on the probability scale from 0% to 100%:

  > Suppose Sophia makes Tim an orc costume. Under this assumption, how probable is it that the following sentence is true:
  > Tim will be excited about his present.

- On the next page, we asked participant for their rating of the probability of the conditional $P(\text{if } A \text{ then } B)$ on the probability scale from 0% to 100%:

  > Could you please rate the probability that the following sentence is true:
  > IF Sophia makes Tim an orc costume, THEN he will be excited about his present.

Design

Research question: Does the Equation (i.e., $P(\text{if } A \text{ then } B) = P(B|A)$) hold?

For each item, participants provide idiosyncratic estimates of $P(\text{if } A \text{ then } B)$ (`if_A_then_B`) and $P(B|A)$ (`B_given_A`).
Each participant worked on 12 items, that is each participant provided 12 estimates of $P(\text{if } A \text{ then } B)$ (`if_A_then_B`) and $P(B|A)$ (`B_given_A`).

Data prepared for this exercise is available in the folder (full data available at: https://osf.io/j4swp/)

Exercise 1: Analyse the data using the no-pooling approach

- Calculate the regression between $P(\text{if } A \text{ then } B)$ and $P(B|A)$ separately for each participant.
- Does this analysis suggest that there is an overall association between the two variables? If so, how strong is this relationship?

Getting started:

Don't forget to **restart R**: `Session -> Restart R`

Some package we might need.

```
library("tidyverse")
theme_set(theme_bw())
library("broom") # not automatically loaded
```

I have already downloaded the data from the OSF and prepared it according to the descriptions found there. The prepared data is in `dat`.

```
# Run complete chunk: Ctrl+Shift+Enter

# You might need to set the correct working directory via the menu:
# Session -> Set Working Directory -> To Source File Location

afex::set_sum_contrasts() # just in case we set orthogonal contrasts

load("ssk16_dat_prepared_ex1.rda") # data preapred in 'prepare_data.R'
glimpse(dat1)
```

```
## Observations: 376
## Variables: 7
## $ p_id          <fct> "36_P(if,then)", "36_P(if,then)", "36_P(if,then)...
## $ i_id          <fct> 1, 7, 10, 12, 2, 4, 5, 8, 1, 3, 9, 11, 5, 6, 10,...
## $ B_given_A     <dbl> 52, 79, 79, 77, 98, 98, 97, 90, 28, 100, 100, 10...
## $ B_given_A_c   <dbl> 0.02, 0.29, 0.29, 0.27, 0.48, 0.48, 0.47, 0.40, ...
## $ if_A_then_B   <dbl> 52, 84, 94, 81, 95, 97, 90, 95, 28, 100, 100, 10...
## $ if_A_then_B_c <dbl> 0.02, 0.34, 0.44, 0.31, 0.45, 0.47, 0.40, 0.45, ...
## $ rel_cond      <fct> PO, PO, PO, PO, PO, PO, PO, PO, PO, PO, PO, PO, ...
```

Variables in the data:

- `p_id`: participant identifier
- `i_id`: item identifier (i.e., id of vignette)
- `B_given_A`: original $P(B|A)$
- `B_given_A_c`: $P(B|A)$ centered at midpoint of scale (as used in paper)
- `if_A_then_B`: original $P(\text{if } A \text{ then } B)$
- `if_A_then_B_c`: $P(\text{if } A \text{ then } B)$ centered at midpoint of scale (as used in paper)
- `rel_cond`: relevance condition. Has only one level here, can be ignored.

Complete-Pooling Approach

```
m0 <- lm(if_A_then_B ~ B_given_A, dat1)
summary(m0)
```

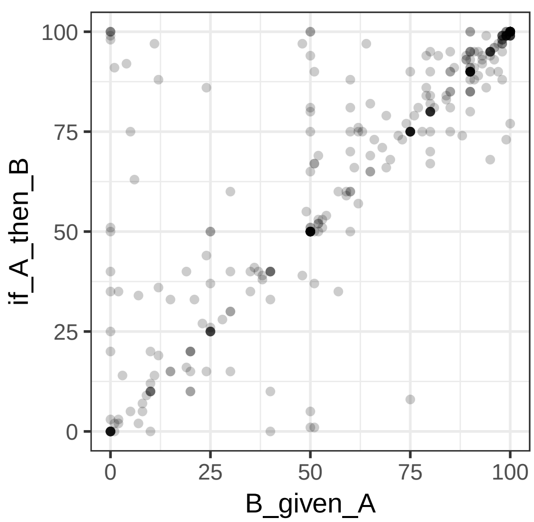
```
##
## Call:
## lm(formula = if_A_then_B ~ B_given_A, data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -69.513  -7.824   0.811   2.798  81.554
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.44616    1.96690   9.378   <2e-16 ***
## B_given_A    0.78756    0.02659  29.624   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.65 on 374 degrees of freedom
## Multiple R-squared:  0.7012, Adjusted R-squared:  0.7004
## F-statistic: 877.6 on 1 and 374 DF,  p-value: < 2.2e-16
```

When we completely ignore the dependencies in the data, we can see a clear relationship between the IV and DV. The regression parameter estimate for `B_given_A` is clearly significant (i.e., different from 0) but also not too far away from 1.0. If it were 1.0, this would mean that $P(\text{if } A \text{ then } B) = P(B|A)$ would hold exactly.

Before the next step, let's take a look at the data. It suggests indeed that the relationship between IV and DV. But does it hold when looking at the data of individual participants?

```
ggplot(data = dat1) +
  geom_point(mapping = aes(x = B_given_A, y = if_A_then_B), alpha = 0.2, pch = 16) +
  coord_fixed()
```



Full Instructions

- Your task is to calculate the regression parameter (i.e., slope, potentially also the intercept) for each participant (i.e., relationship of `if_A_then_B` and `B_given_A` for each `p_id`).
- Then investigate the distribution of resulting regression parameters. Perform this investigation in a graphical way and also statistically (i.e., using `lm`).
- The goal of this exercise is to combine your knowledge of the `tidyverse` and use it to solve the aforementioned task.
- In case you need some inspiration for `dplyr` and `broom`, you might want to take a look at chapter 25 (especially 25.2.1, 25.2.2, 25.2.3) of Wickham and Grolemund (2017) see: http://r4ds.had.co.nz/many-models.html

```
# go
```

Exercise 2: Analysing more conditions using complete pooling and no-pooling approach

The study of Skovgaard-Olsen et al. contained a further manipulation not considered so far. These additional data, `dat2`, can be found in file `ssk16_dat_prepared_ex2.rda`.

```
load("ssk16_dat_prepared_ex2.rda")
str(dat2)
```

```
## 'data.frame':    752 obs. of  7 variables:
##  $ p_id         : Factor w/ 94 levels "102_P(if,then)",..: 63 63 63 63 63 63 63 63 64 64 ...
##  $ i_id         : Factor w/ 12 levels "1","2","3","4",..: 1 2 7 8 9 10 11 12 2 4 ...
##  $ B_given_A    : num  52 60 79 0 51 79 80 77 98 98 ...
##  $ B_given_A_c  : num  0.02 0.1 0.29 -0.5 0.01 0.29 0.3 0.27 0.48 0.48 ...
##  $ if_A_then_B  : num  52 1 84 0 51 94 1 81 95 97 ...
##  $ if_A_then_B_c: num  0.02 -0.49 0.34 -0.5 0.01 0.44 -0.49 0.31 0.45 0.47 ...
##  $ rel_cond     : Factor w/ 2 levels "PO","IR": 1 2 1 2 2 1 2 1 1 1 ...
```

As discussed before, the initial research question was if the Equation holds (i.e., $P(\text{if } A \text{ then } B) = P(B|A)$). Furthermore, we were interested whether or not the Equation holds even if there is no apparent relationship between A and B? To this end we manipulated the relevance of A for B in the within-subjects variable `rel_cond`:

- positive relevance (PO): A is a reason for B (IF Sophia buys an orc costume for Tim, THEN Tim will be excited about his present.)
- irrelevance (IR): A and B have no apparent relationship (IF Sophia regularly wears shoes, THEN Tim will be excited about his present.)

Complete Pooling

- Your task is to calculate the regression parameter (and potentially also the intercept) for within-subject condition (i.e., relationship of `dv` and `if_A_then_B` and `B_given_A` for each level of `rel_cond`).
- There are different ways how to interpret complete pooling. Either one ignores individual differences or one aggregates across them. Can you find the different ways for implementing no pooling here?

```
# go
```

No Pooling

- Your task is to calculate the regression parameter (and potentially also the intercept) for each participant and within-subject condition (i.e., relationship of `dv` and `if_A_then_B` and `B_given_A` for each `p_id` and level of `rel_cond`).
- Then compare the individual regression parameters across conditions (i.e., for each level of `rel_cond`). Do this comparison in a graphical way and also statistically (i.e., ANOVA using `afex`).

```
# go
```

References

- Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition*, 150, 26-36. https://doi.org/10.1016/j.cognition.2015.12.017
- Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* Sebastopol CA: O'Reilly.