The general linear model – Additional exercises

Jan Vanhove

https://janhove.github.io

Ghent, 15–17 July 2024

Exercise 1: An experiment

From the abstract of Vanhove (2016):

> "This article investigates whether learners are able to quickly discover simple, systematic graphemic correspondence rules between their L1 and an unknown but closely related language in a setting of receptive multilingualism.
>
> (. . .)
>
> Eighty L1 German speakers participated in a translation task with written Dutch words, most of which had a German cognate. In the first part of the translation task, participants were shown 48 Dutch words, among which either 10 cognates containing the digraph ‹oe› (always corresponding to a German word with ‹u›) or 10 cognates with the digraph ‹ij› (corresponding to German ‹ei›). During this part, participants were given feedback in the form of the correct translation. In the second (feedback-free) part of the task, participants were shown another 150 Dutch words, among which 21 cognates with ‹oe› and 21 cognates with ‹ij›."

What I wanted to know was whether the exposure and feedback to ten words containing a specific interlingual orthographic correspondence (i.e., ‹oe›–‹u› or ‹ei›–‹ij›) was sufficient for the participants to pick up on this correspondence and use their knowledge in translating new words containing the correspondence.

First, we read in the data and compute, for each participant, the proportion of words in the second (feedback-free) part of the task which they translated correctly in each category (cognates containing ‹oe›, cognates containing ‹ij›, other cognates, non-cognates).

```r
library(tidyverse)
library(here)
```

```
d <- read_csv(here("data", "correspondencerules.csv"))
d_perParticipant <- d |>
  filter(Block != "training") |>
  group_by(Subject, LearningCondition, Category, WSTRight) |>
  summarise(ProportionCorrect = mean(Correct == "yes"),
            .groups = "drop")
```

I've also retained a measure of the participants' L1 (German) vocabulary knowledge, namely the `WSTRight` variable. Use `View(d_perParticipant)` to inspect the structure of the cleaned-up and restructured dataset.

Analyse this dataset to answer the research question.

Hints (also for other exercises):

- There are several acceptable ways of analysing these data.
- All of them involve plotting the data :)
- Note that `LearningCondition` varies *between* participants whereas `Category` varies *within* participants.
- First try to work out on a piece of paper what it is you want to do. You can always ask someone to help you implement your idea in R.
- Ask for help and feedback! :)

Exercise 2: An exercise in plotting

Pestana et al. (2017) measured the Portuguese reading skills of Portuguese children in Portugal, French-speaking Switzerland, and German-speaking Switzerland (`LanguageGroup`) at three points in time (`Time`).

Let's read in their data. We'll just retain the reading data in Portuguese.

```
skills <- read_csv(here("data", "helascot_skills.csv"))
background <- read_csv(here("data", "helascot_background.csv"))
d <- skills |>
  left_join(background, by = "Subject") |>
  filter(LanguageTested == "Portuguese") |>
  filter(!is.na(Reading))
```

Since the `Time` variable is just a number, we'll recode it as a factor:

```r
d$Time <- factor(d$Time)
```

Your task: Draw a plot showing how the reading scores differ between the different data collections (`Time`) and the language groups.

Exercise 3

From the abstract of Hicks (2021):

> "This study explores whether middle-school students can exploit explicitly addressed crosslinguistic lexical similarities between German and English to learn vocabulary more efficiently. Across six weeks, 260 Swiss German learners of English as a foreign language (17 classes) completed three vocabulary learning tests (T1, T2 and T3). Additionally, 7 of these 17 classes attended a 90-minute intervention between the first and second test: During a 45-minute introductory lesson students discovered four systematic orthographic correspondence rules (e.g. <p> to <f> as in ship and Schiff), followed by three 15-minute sessions to consolidate their knowledge."

She was interested in whether the children in the experimental condition were able to learn vocabulary more efficiently than the children in the control condition.

The dataset contains lots of variables that aren't of interest for this exercise, so let's discard those:

```r
d <- read_csv(here("data", "hicks2021.csv")) |>
  select(ID, Class, Group, T1cog, T3cog)
```

`ID` contains the participants' study-internal identification; `class` identifies their school class; `Group` the experimental condition to which they were assigned; `T1cog` is the participants' pretest performance; `T3cog` is the participants' posttest performance.

Task:

1. Assume the participants were randomly, and individually, assigned to the experimental conditions. Analyse the data.

2. In actual fact, all participants in the same class were assigned to the same condition. In other words, the participants were not assigned to the conditions individually. What kind of problems might this cause for the analysis?

3. Moreover, the participants weren't randomly assigned to the conditions. Those classes whose English teachers agreed to be part of the intervention group were assigned to the

experimental group; those whose English teachers did not agree to do so were assigned to the control group. What kind of problems might this have caused?

Exercise 4: Meaning of model parameters

We're given a data set `d` with the outcome `CorrectSpoken` and four predictors:

- `n.Sex`: 0.5 for men, -0.5 for women (sum coding).
- `NrLang`: Number of languages spoken for each participant (varies from 1 to 5).
- `DS.Span`: Score on the backward digit span, a working memory test (varies from 2 to 8, more is better).
- `Raven.Right`: Score on a test of fluid intelligence (varies from 0 to 35, more is better).

We fit the following model:

```
mod.lm <- lm(CorrectSpoken ~ NrLang + DS.Span + n.Sex*Raven.Right,
             data = d)
summary(mod.lm)$coefficients

                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.554      1.532    5.58  1.0e-07
NrLang               1.115      0.353    3.16  1.9e-03
DS.Span             -0.097      0.330   -0.29  7.7e-01
n.Sex               -5.017      1.716   -2.92  4.0e-03
Raven.Right          0.286      0.048    6.01  1.3e-08
n.Sex:Raven.Right    0.201      0.088    2.28  2.4e-02
```

1. What does the estimate of -5.02 for the `n.Sex` parameter refer to? That is, what does it mean literally.
2. Let's say we wanted to fit a similar model but one that has an intercept with a more useful interpretation. Explain how we could achieve this. Also explain how we can interpret the intercept in this new model.
3. Claim: We can glean from the model that, at least in this data set, participants with a high working memory capacity do not outperform participants with a low working memory capacity in terms of the `CorrectSpoken` variable.
   Is this claim correct? (Yes or no.) Justify your answer.
4. What gets computed here?

```
8.554 + 1.115*3 - 0.097*4 + 0.286*25

[1] 19
```

5. How large, according to the model, is the expected average difference in the `CorrectSpoken` variable between women with a `Raven.Right` score of 20 and women with a `Raven.Right` score of 30, keeping `NrLang` and `DS.Span` constant. It suffices to provide the computation without evaluating it.

References

Hicks, Nina Selina. 2021. Exploring systematic orthographic crosslinguistic similarities to enhance foreign language vocabulary learning. *Language Teaching Research* doi:10.1177/13621688211047353.

Pestana, Carlos, Amelia Lambelet & Jan Vanhove. 2017. Reading comprehension in Portuguese heritage speakers in Switzerland (HELASCOT project). In Raphael Berthele & Amelia Lambelet (eds.), *Heritage and school language literacy development in migrant children: Interdependence or independence?*, 58–82. Bristol: Multilingual Matters. doi:10.21832/9781783099054-005.

Vanhove, Jan. 2016. The early learning of interlingual correspondence rules in receptive multilingualism. *International Journal of Bilingualism* 20(5). 580–593. doi:10.1177/1367006915573338.