Statistical analysis

Gonzalo García-Castro, Serene Siow, Nuria-Sebastian-Galles, and Kim Plunkett

10/28/2019

We will use **Logistic Growth Curve Analysis** to analyse our data (e.g. **???**; **???**). Growth Curve Analysis (GCA; **???**) allows us to model the probability of target fixation across the time course of trials. Traditional analyses of the Visual World Paradigm involve aggregating the number/proportion of fixation of each trial to single value to be predicted by certain variables of interest. By using GCA, we will we will minimise the need to aggregate data across trials, thus increasing the power of our analysis. GCA is a generalised case of regression models where multiple polynomials of the time course of the trial are included as predictors. This means that our model will be able to accomodate non-linear patterns of target fixations within a given trial. Since our response variable is binary (i.e. follows a binomial distribution), we use logistic regression, following Barr (2008) guidelines. We will include three predictors of interest to model the probability of target fixation at each time point: prime-target relatedness (related/unrelated, $Relatedness$), prime cognateness (cognate/non-cognate, $Cognateness$), linguistic profile (Spanish-English, Spanish-Catalan, $Profile$). We will include participants (ID), and items ($Item$) as random effects to account for variability caused by individual differences across participants or intrinsic properties of the stimuli involved in each trial.

Response variable

We want to predict whether a given participant is fixating the target in each time point within a given trial. This response variable follows a binomial distribution: at each time point within the trial, a participant is either fixating the target or not. This technique requires the transformation of the in-principle binary outcome to either the log odds scale or the logit scale. We will transform our response variable to the empirical logit scale, given that the first tends to be problematic for near-zero values (Agresti, Booth, Hobert, & Caffo, 2000; McCullagh & Nelder, 1989; both cited by Barr, Levy, Scheepers, & Tily, 2013).

The presence of a target fixation in each time point will be coded as 1, and its absence will be coded as 0. We create time bins of 50 ms duration (6 samples, given 120 Hz sampling rate). At each time bin within the trial, we will aggregate fixations for each participant, item, and condition using the empirical logit transformation suggested by Barr (2008):

$$elogit(Y, N) = \log\left(\frac{Y + 0.5}{N - Y + 0.51}\right)$$

Where $elogit(Y, N)$ is the resulting empirical logit of target fixation for Y number of fixations across trials and N number of trials, and ϕ is the proportion of target fixations across a given set of trials. This formula includes a 0.5 adjustment factor that avoids unreliable logit scores near endpoints. This empirical logic, from now on named *elog* indicates the probability of target fixation at each time bin within the trial.

Missing data

We will use multilevel multiple imputation (**???**).

Model

To predict the probablity of target fixation, we will include several terms in our model. Some terms will correspond to variables of intrinsic interest, others will be included to account for individual differences across participants and items.

Fixed effects

Effects of interest

We expect three variables to exert an effect on the probability of target fixation across the time course of each trial: prime-target phonological relatedess ($Relatedness$), prime cognateness ($Cognateness$), and linguistic profile ($LangProfile$). We will implement these predictors by including the following terms in our model as fixed effects:

- Overall intercept (β_0)
- Time ($ot1$): time within trial, specified in seconds to avoid coefficients being too small (Barr, 2008). Following Barr (2008) guidelines, we will time-lock the time domain at the time point at which the grand mean of the probability of fixations shows the first tendency to rise. This will be taken as an indicator of signal-driven looking behaviour.
- Time2 ($ot2$): second-order (quadratic) polynomial of the time domain.
- Time3 ($ot3$): third-order (cubic) polynomial of the time domain.
- Prime-target relatedness ($Relatedness$): do prime and target words hare phonological onset (related/unrelated)
- Prime cognateness ($Cognateness$): is the prime word a cognate between the languages the two toddler is learning?
- Interaction effect of prime-target relatedness and prime cognateness ($Relatedness \times Cognateness$).

Individual differences

The expect other variables not directly related to out hypothesis to also determine the probability of target fixation. Some of these variables will influence the outcome via individual differences across participants, while others will do so via intrinsic properties of the stimuli included in each trial. Regarding participant-level differences, we expect the effects of interest to be influenced by participants' receptive vocabulary size in their dominant language ($VocabSizeL1$), participants' receptive vocabulary in their non-dominant language ($VocabSizeL2$), and sex (Sex). Regarding stimulus-level differences, we expect our the effects of interest to be influenced by prime frequency, ($PrimeFreq$), and target frequency ($TargetFreq$). We will account for such differences among participants and items by including the following terms in our model as fixed effects:

- Receptive vocabulary size in dominant language ($VocabSizeL1$): number of words participants are able to understand in their dominant language (i.e. the language in which they were tested). Continous variable.
- Receptive vocabulary size in non-dominant language ($VocabSizeL2$): number of words participants are able to understand in their non-dominant language (i.e. the language in which they were not tested). Continuous variable.
- Sex (Sex): Sex of the participant. Categorical variable (female/male).
- Prime word frequency ($PrimeFreq$): Frequency of the label associated to the prime image, operationalised by the Zipf score of such word in SUBTLEX corpora. Continuous variable.
- Target word frequency ($PrimeFreq$): Frequency of the label associated to the target image, operationalised by the Zipf score of such word in SUBTLEX corpora. Continuous variable.

Random effects

We will include participant (ID) and item ($Item$) as crossed random effects (Barr, 2008):

- Random intercept by participant: The overall target fixation probability will be allowed to vary across participants.

- Random slope for main effect of prime-target relatedness by participant: The effect of prime-target relatedness will be allowed to vary across participants.
- Random slope for main effect of prime cognateness by participant: The effect of prime cognateness will be allowed to vary across participants.
- Random slope for interaction effect of prime-target relatedness and prime cognatess by participant: The effect of the interaction effect of prime-target relatedness and prime cognateness will be allowed to vary across participants.
- Random intercept by item: The overall target fixation probability will be allowed to vary across items (target-distractor pairs).
- Random slope for main effect of prime-target relatedness by item: The effect of prime-target relatedness will be allowed to vary across items.
- Random slope for main effect of prime cognateness by item: The effect of prime cognateness will be allowed to vary across items.
- Random slope for interaction effect of prime-target relatedness and prime-cognateness by item: The interaction effect between prime-target relateddness and prime cognateness will be allowed to very across items.
- Random slope for the main effect of Time, Time2 and Time3 by participant: The intercept of the target fixation probability will be allowed to vary across time points for each participant.
- Random slope for all combinations of interactions between the three time terms by participant.
- Random slope for the main effect of Time, Time2, and Time3 by item: The intercept of the target fixation probability will be allowed to vary across time points for each item.
- Random slope for all combinations of interactions between the three time terms by item.

Formal model

The maximal-random structure of our multilevel model can be summarised in the following set of formulas:

Level 1

The likelihood of target fixation, expressed as the epirical logic transformation of the originally binary outcome, is predicted for each participant, item and a linear, a quadratic, and a cubic transformation of the time domain, parsed into time bins of 50 ms.

$$\begin{aligned} elog_{fip} = &\pi_0 + \pi_1 Time_{fip} + \pi_2 Time2_{fip} + \pi_3 Time3_{fip} + \dots \\ &\pi_4 (Time1 \times Time2)_{fip} + \dots \\ &\pi_5 (Time1 \times Time3)_{fip} + \dots \\ &\pi_6 (Time2 \times Time3)_{fip} + \dots \\ &\pi_7 (Time1 \times Time2 \times Time3)_{fip} \end{aligned}$$

Where:

- $elog$ is the empirical logit of target fixation
- t is the time bin
- f indexes time bins
- i indexes items
- p indexes participants
- π_0 is the overal intercept for participant p, in item i, at time bin f
- π_1 is the regression coefficient (slope) of the linear time term ($Time$)
- π_2 is the regression coefficient (slope) of the quadratic time term ($Time2$)
- π_3 is the regression coefficient (slope) of the cubic time term ($Time3$)

- π_4 is the regression coefficient (slope) of the interaction between the linear and the quadratic time terms ($Time1 \times Time2$)
- π_5 is the regression coefficient (slope) of the interaction between the linear and the cubic time terms ($Time1 \times Time3$)
- π_6 is the regression coefficient (slope) of the interaction between the cuadratic and the cubic time terms ($Time2 \times Time3$)
- π_7 is the regression coefficient (slope) of the interaction between the linear, the quadratic, and the cubic time terms ($Time1 \times Time2 \times Time3$)

Level 2

The regression coefficients previously used to predict the likelihood of target fixation are a function of the predictors of interest (i.e. fixed effects). They are computed for each participant in each item. fixed effects include prime-target relatedness, prime-cognateness, and their interaction.

$$\pi_0 = \beta_{00} + \beta_{01}Relatedness_{ip} + \beta_{02}Cognateness_{ip} + \beta_{03}(Relatedness \times Cognateness)_{ip} + r_0 + u0$$

Where:

- β_{00} is the grand mean of the intercept.
- β_{01} is the slope of the main effect of Relatedness on the overall intercept.
- β_{02} is the slope of the main effect of Cognateness on the overall intercept.
- β_{03} is the slope of the interaction effect between Relatedness and Cognateness on the overall intercept.
- r_0:

$$\pi_1 = \beta_{10} + \beta_{11}Relatedness_{ip} + \beta_{12}Cognateness_{ip} + \beta_{13}(Relatedness \times Cognateness)_{ip} + r_1 + u1$$

Where:

- β_{10} is the grand mean of the slope of the meain effect of the linear time term.
- β_{11} is the slope of the main effect of Relatedness on the overall slope of the main effect of the linear time term.
- β_{12} is the slope of the main effect of Cognateness on the overall slope of the main effect of the linear time term.
- β_{13} is the slope of the interaction effect of Relatedness and Cognateness on the overall slope of the main effect of the linear time term.
- r_1:

$$\pi_2 = \beta_{20} + \beta_{21}Relatedness_{ip} + \beta_{22}Cognateness_{ip} + \beta_{23}(Relatedness \times Cognateness)_{ip} + r_2 + u2$$

Where:

- β_{20} is the grand mean of the slope of the main effect of the quadratic time term.
- β_{21} is the slope of the main effect of Relatedness on overall slope of the main effect of the quadratic time term.
- β_{22} is the slope of the main effect of Cognateness on overall slope of the main effect of the quadratic time term.
- β_{23} is the slope of the interaction effect of Relatedness and Cognateness on overall slope of the main effect of the quadratic time term.
- r_2:

$$\pi_3 = \beta_{30} + \beta_{31} Relatedness_{ip} + \beta_{32} Cognateness_{ip} + \beta_{33} (Relatedness \times Cognateness)_{ip} + r_3 + u3$$

Where:

- β_{30} is the grand mean of the slope of the main effect of the cubic time term.
- β_{31} is the slope of the main effect of Relatedness on overall slope of the main effect of the cubic time term.
- β_{32} is the slope of the main effect of Cognateness on overall slope of the main effect of the cubic time term.
- β_{33} is the slope of the interaction effect of Relatedness and Cognateness on overall slope of the main effect of the cubic time term.
- r_3:

$$\pi_4 = \beta_{40} + \beta_{41} Relatedness_{ip} + \beta_{42} Cognateness_{ip} + \beta_{43} (Relatedness \times Cognateness)_{ip} + r_4 + u4$$

Where:

- β_{40} is the grand mean of the slope of the interaction effect of the linear and the quadratic time terms.
- β_{41} is the slope of the main effect of Relatedness on overall slope of the interaction effect of the linear and the quadratic time terms.
- β_{42} is the slope of the main effect of Cognateness on overall slope of the interaction effect of the linear and the quadratic time terms.
- β_{43} is the slope of the interaction effect of Relatedness and Cognateness on overall slope of the interaction effect of the linear and the quadratic time terms.
- r_4:

$$\pi_5 = \beta_{50} + \beta_{51} Relatedness_{ip} + \beta_{52} Cognateness_{ip} + \beta_{53} (Relatedness \times Cognateness)_{ip} + r_5 + u5$$

Where:

- β_{50} is the grand mean of the slope of the interaction effect of the linear and the cubic time terms.
- β_{51} is the slope of the main effect of Relatedness on overall slope of the interaction effect of the linear and the cubic time terms.
- β_{52} is the slope of the main effect of Cognateness on overall slope of the interaction effect of the linear and the cubic time terms.
- β_{53} is the slope of the interaction effect of Relatedness and Cognateness on overall slope of the interaction effect of the linear and the cubic time terms.
- r_5:

$$\pi_6 = \beta_{60} + \beta_{61} Relatedness_{ip} + \beta_{62} Cognateness_{ip} + \beta_{63} (Relatedness \times Cognateness)_{ip} + r_6 + u6$$

Where:

- β_{60} is the grand mean of the slope of the interaction effect of the quadratic and the cubic time terms.
- β_{61} is the slope of the main effect of Relatedness on overall slope of the interaction effect of the quadratic and the cubic time terms.
- β_{62} is the slope of the main effect of Cognateness on overall slope of the interaction effect of the quadratic and the cubic time terms.

- β_{63} is the slope of the interaction effect of Relatedness and Cognateness on overall slope of the interaction effect of the quadratic and the cubic time terms.
- r_6:

$$\pi_7 = \beta_{70} + \beta_{71} Relatedness_{ip} + \beta_{72} Cognateness_{ip} + \beta_{73} (Relatedness \times Cognateness)_{ip} + r_7 + u7$$

Where:

- β_{70} is the grand mean of the slope of the interaction effect of the linear, the quadratic and the cubic time terms.
- β_{71} is the slope of the main effect of Relatedness on overall slope of the interaction effect of the linear, the quadratic and the cubic time terms.
- β_{72} is the slope of the main effect of Cognateness on overall slope of the interaction effect of the quadratic and the cubic time terms.
- β_{73} is the slope of the interaction effect of Relatedness and Cognateness on overall slope of the interaction effect of the linear, the quadratic and the cubic time terms.
- r_7:

Model implementation

We will use the `lme4` R package (**???**) to fit the model. We will use population-average parameter estimation (Fitzmaurice, Laird, & Ware, 2004; cited by Barr, 2008) with Restricted Maximum Likelihood criterion (REML). We will use the "bobyqa" optimiser from the `minqa` R package (**???**), and will set the maximum number of iterations at 1000 (**???**).

Following Mirman (2016), we will adjust the importance of observations based on their reliability. We want more reliable observations to be more determining on how coefficients are estimated to fit the model to the data. Empirical logit scores resulting from aggregating fixations across a higher amount of trials will be weighted higher that those resulting from aggregating less trials. Weights will be calculated using the following formula:

$$w(Y, N) = \frac{1}{Y + 0.5} + \frac{1}{N - Y + 0.5}$$

The model will be implemented in R using the following code:

```
lme4::lmer(
  # response variable
  elog ~                                          # empirical logic of target fixation
    # fixed effects
    (ot1+ot2+ot3) +                               # f eff of time, time^2 and time^3
    Relatedness +                                 # f eff of prime-target relatedness
    Cognateness +                                 # f eff of prime cognateness (C)
    LangProfile +                                 # f eff of language profile (LP)
    Relatedness*Cognateness +                     # f int of R and C
    Relatedness*LangProfile +                     # f int of R and LP
    Cognateness*LangProfile +                     # f int of C and LP
    Relatedness*Cognateness*LangProfile +         # f int of C and LP
    VocabSizeL1 +                                 # f eff of receptive L1 vocab
    VocabSizeL2 +                                 # f eff of receptive L2 vocab
    PrimeFreq +                                   # f eff of prime label freq
    TargetFreq +                                  # f eff of target label freq
```

```
    # random effects
    (1|ID) +                                        # r int by ID
    (1+ot1+ot+ot3|ID) +                             # r int and slope of time by ID
    (1+Relatedness|ID) +                            # r int and slope of R by ID
    (1+Cognatedness|ID) +                           # r int of of of C by ID
    (1+LangProfile|ID) +                            # r int of of of LP by ID
    (1+Relatedness*Cognatedness|ID) +               # r int and slope of RxC by ID
    (1+Relatedness*LangProfile|ID) +                # r int and slope of RxLP by ID
    (1+Cognateness*LangProfile|ID) +                # r int and slope of CxLP by ID
    (1+Relatedness*Cognateness*LangProfile|ID) +    # r int and slope of RxCxLP by ID
    (1|Item) +                                      # r int by item
    (1+ot1+ot+ot3|Item) +                           # r int and slope of time by item
    (1+Relatedness|Item) +                          # r int and slope of of R by item
    (1+Cognateness|Item) +                          # r int and slope of of C by item
    (1+LangProfile|Item) +                          # r int and slope of of R by item
    (1+Relatedness*Cognateness|Item) +              # r int and slope of RxC by item
    (1+Relatedness*LangProfile|Item) +              # r int and slope of RxLP by item
    (1+Cognateness*LangProfile|Item) +              # r int and slope of CxLP by item
    (1+Relatedness*Cognateness*LangProfile|Item), # r int and slope of RxCxLP by item

  # other settings
  data = CognatePriming,                            # data set
  REML = TRUE,                                      # use Restricted Maximum Likelihood
  control = lmerControl(
    maxit = 1000,                                   # number of max iterations to 1000
    optimizer = "bobyqa"                            # define optimiser
  ),
  verbose = TRUE,                                   # for more information during fitting
  weights = 1/weights                               # weight of each observation
)
```

Dealing with convergence issues

Due to the large number of parameters to be estimated in our model, and the limited amount of data we can collect, it is posible that we run into convergence issues. Logistic regression models are more likely than linear regression models to be subject of convergence issues as the random effects structure gets more complex. Lack of convergence means that some of the estimated coefficients are unreliable, as other values of the same coefficients could provide a similar fit to the data. We will handle this issue as follows:

1. We will change the optimiser until we reach convergence: first to "optimx" from the `optimx` R package (**???**), second to "nloptwrap" from the `nloptr` R package (**???**). If lack of convergence persist we will skip to step 2.
2. We will start a pruning process to simplify our model. Terms will be dropped from the model according to their scientific interest of for theoretical reasons. The order of pruning will be as follows:
    1) Random slopes of the interaction effect between `Relatedness` and `Cognateness` by item.
    2) Random slopes of main effect of `Relatedness` and `Cognateness` by item.
    3) Random slopes of the interaction effect between `Relatedness` and `Cognateness` by participant.
    4) Random slopes of the main effects of `Relatedness` and `Cognateness` by participant.
    5) Main (fixed) effect of the 3rd order (cubic) polynomial of time (`ot3`), exluding this term from the random by-participant and by-item effect.

If after taking the previous steps, convergence issues persist, we consider that dropping more terms would lead to the oversimplification of the model. This model would no longer capture the sources of variation that we consider intrincis to our hypotheses. For this reason, at this point we will change the approach and will perform Bayesian modelling.

Inference criteria

We will use a frequentist approach toward statistical inference. Extracting p-values in mixed-effect models is not straightforawrd due to the difficulty to estimate the degrees of freedom of the distributions from which parameters are drawn. We will use the Satterthwaite's approximation to degrees of freedom (**???**) to extract p-values using the `lmerTest` R package (**???**). If we run into convergence issues that persist even after model simplification, we will shift to a fully Bayesian approach toward statistical inference (see previous section).

Data exclusion

How will you determine which data points or samples if any to exclude from your analyses? How will outliers be handled? Will you use any awareness check?

References

Agresti, A., Booth, J. G., Hobert, J. P., & Caffo, B. (2000). Random-Effects Modeling of Categorical Response Data. *Sociological Methodology*, *30*(1), 27–80. https://doi.org/10.1111/0081-1750.t01-1-00075

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. CRC Press.

Mirman, D. (2016). *Growth Curve Analysis and Visualization Using R*. CRC Press.