A STATISTICAL ANALYSIS ON LANGUAGE ACQUISITION

Anthony YAZDANI - N°15 317 878

May 23, 2020 - GENEVA - SWITZERLAND

About the study

Linguists are interested in the features influencing the understanding and acquisition of relative clauses. One classical hypothesis is that object relatives are more difficult to comprehend and produce for children than subject relatives. Another feature that seem to have an effect on this acquisition is whether there is a match between certain characteristics of the subject and the object of the relative. The data comes from a research including 68 children of 5, 7, and 8 years old. The children perform a sentence repetition task. Each child was asked to repeat 32 sentences.

Executive summary

In this analysis, we found that all provided features have a statistically significant impact on the success of the child. More precisely, we found that the older the child is, the more likely he or she is to succeed. It appears that the probability of success increase when there is a match between certain characteristics of the subject and the object of the relative. We also found that the probability of success increase when there is a subject relative clause in contrast with an object relative clause.

INTRODUCTION

One of the main assumptions related to GLM theory is not valid in this context. Indeed, we can hypothesize that the data are correlated because they come from repeated measurements on the same individuals. In this context, we will make use of an appropriate theory named generalized mixed effects models (GLMM). GLMMs are appropriately considered to be subject-specific models, meaning that they model the expectation conditioned on the random effect. However, the conclusions drawn from a subject-specific model are generally very close to the population-specific (marginal) model. In order to conduct a detailed analysis, we will also mention and comment the population-specific modeling. To fulfill our goal, we will start with an exploratory data analysis. In a second section, we will develop several models to finally propose a parsimonious model that fits well our data. Finally, we will end up interpreting our model to determine which features influence the understanding and acquisition of relative clauses.

I - EXPLORATORY DATA ANALYSIS

This section will be dedicated to an exploratory analysis of our data. We will identify each child through their `Id` and their age will be indicated with the variable `Age`. All sentences are indicated as containing a subject relative when `Structure = SR`; an object relative when `Structure = OR`; a match when `Match_Mismatch = Ma` and a mismatch when `Match_Mismatch = Mis`. Finally, we will indicate if the child successfully repeated the sentence when `Success = 1` and `Success = 0` otherwise.

Data quality

The following table shows that there are no missing values in the data and that the variable classes are well specified.

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
Id	0	0.00	0	0	0	0	integer	68
Age	0	0.00	0	0	0	0	factor	3
Structure	0	0.00	0	0	0	0	factor	2
Match_Mismatch	0	0.00	0	0	0	0	factor	2
Success	1456	66.91	0	0	0	0	integer	2

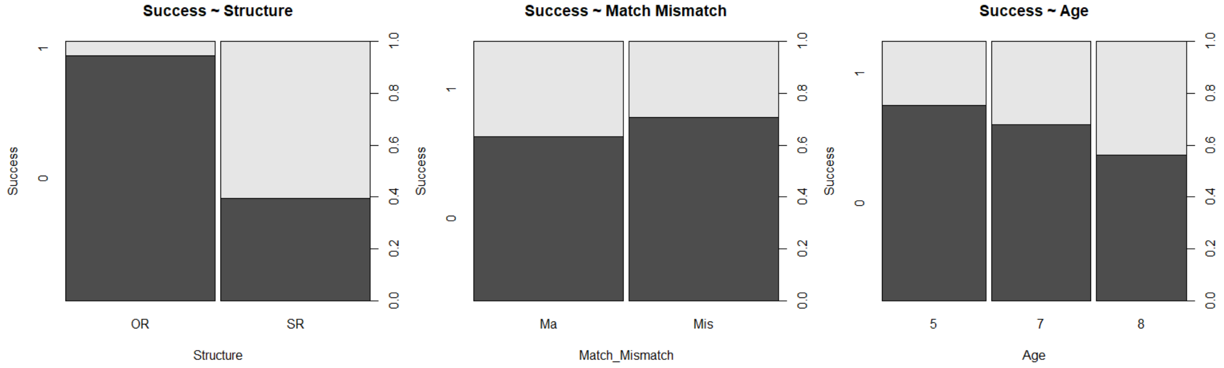
When working with factors, we may want to check whether we have quasi-separation/separation. One can show that if there is no overlapping, the estimate for the concerned coefficient diverges to infinity ($\hat{\beta}_{no-overlap} \to +\infty$). The following table shows which combination of response and covariates occurs less than ten times. Note that the rows are defined as Success.Age.Structure.Match_Mismatch.

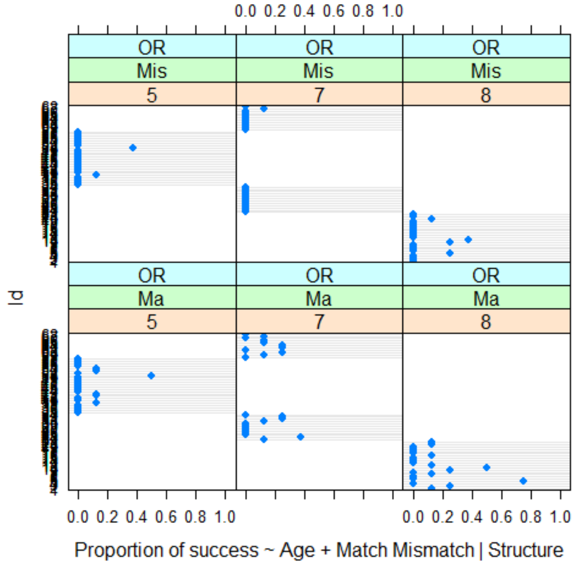
Combination	#
1.5.OR.Ma	9
1.5.OR.Mis	4
1.7.OR.Mis	1
1.8.OR.Mis	8

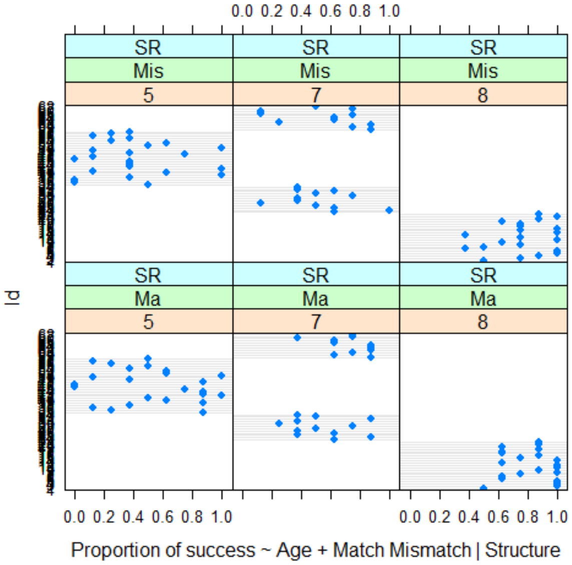
The dataset is large, divided in balanced clusters and has no missing values as well as overlap for any combinations of response and covariates. In fact, this is an ideal framework to make an analysis. Interestingly enough, the balanced clusters will lead to less numerical convergence problems when considering complex models.

Graphical overview

Because of the nature of the covariates (factors), we don't have a large number of tools at our disposal. However, we can still use mosaic plots and dotplots per `Id`.







As seen from the mosaic plots, all three covariates seem to have an impact on the response variable. As stated in the introduction, we should consider a random effect due to repeated measurements on the same individuals. In addition, we notice from the dotplots that some individuals perform better/worse on average compared to other similar individuals. The random effect should capture the overall difference between individuals letting the fixed effect coefficients explain the differences in performance due to the conditions of the experiment (Age, $Structure$, $MatchMismatch$). In other words, the random effect should capture the within cluster correlation. Second order interactions are hard to asses visually when factors only are available. However, due to the small number of covariates we can easily fit the three possible interactions and proceed to a classical variable selection based on AIC, BIC and statistical tests. The three possible second order interactions are (Age:$Structure$), (Age:$MatchMismatch$) and ($Structure$:$MatchMismatch$). Hence, we will suggest a method that consists of fitting a simple model without second order interactions, and after validating the latter, we will propose an extension with second order interactions.

II - MODEL BUILDING

To keep the log-likelihood available, we will use Laplace, Adaptive GQ and Gauss Hermite maximum likelihood estimation and disregard Penalized Quasi-Likelihood (PQL) estimation. Additionally, one can show that the PQL estimation can produce biased estimators (GLAM 2020). For the sake of clarity, all R outputs are not shown. We will use the AIC as a basis for choosing which method we will use at each stage of our model building and the chosen method will be implicitly shown through the commands. Furthermore, no interpretation will be made until we obtain a validated final model. For the ease of interpretation, only the canonical link function (logit link) is considered. Last but not least, one can show that the optimization problem between a Bernoulli and a Binomial model leads to the same estimates. However, the residuals are expected to be approximately normal in the Binomial case. This is why, in order to have better insights from the residuals, the data has been transformed. One can suggest that the deviance in a Binomial setting shall be used, however, deviance is badly approximate by the χ^2 distribution with a small number of trials.

Because of the uncertainty about the computation of the degrees of freedom, t-tests are controversial in the GLMM context. However, these tests could be referred as complementary tools. Knowing that we can face over/under dispersion in binomial modeling, we would like to refer to them based on a posterior estimation of the dispersion parameter ϕ. It turns out that we can easily estimate ϕ (GLAM 2020) and inflate or deflate the estimated standard errors. When considered appropriate, the P_{values} and the estimated ϕ will be displayed on the last two columns of R outputs ($Pr(>|t|)$ and Phi). More formally, we will proceed in the following way:

$$\hat{\phi} = \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{\hat{r}_{it}^2}{N-(p+1)}, \qquad \hat{sd}^{(\phi)}_{\hat{\beta}_i} = \sqrt{\hat{\phi}} \times \hat{sd}_{\hat{\beta}_i}, \qquad \mathrm{t} = \frac{\hat{\beta}_i - \beta_{H_0}}{\hat{sd}^{(\phi)}_{\hat{\beta}_i}} \sim t_{N-(p+1)}. \tag{1}$$

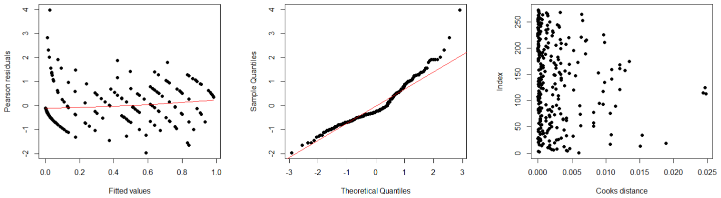
First model

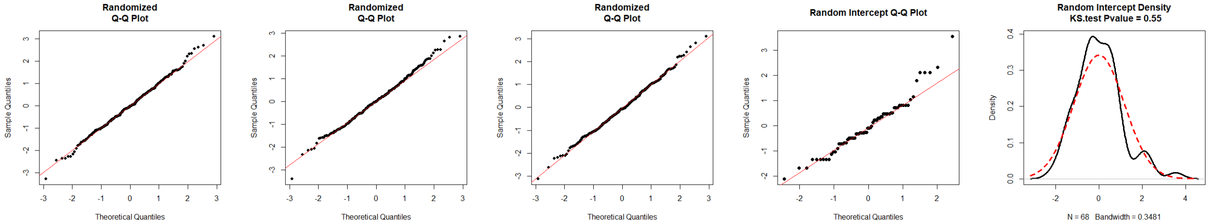
```
model_1_AdaptiveGQ <- glmer(cbind(Success, Failure) ~ 1 + Age + Structure + Match_Mismatch
                              + (1|Id), family=binomial, data=groupeddata, nAGQ = 24)
```

AIC	BIC	logLik	deviance	df.resid	|	RE		VAR	SD
376.8058	398.4406	-182.4029	364.8058	266	|	Id	(Intercept)	1.363229	1.167574

	Estimate	Std. Error	z value	Pr(>|z|)		Pr(>|t|)		Phi
(Intercept)	-4.0302580	0.3187422	-12.644256	0.0000000	***	0.0000000	***	0.68
Age7	0.6933248	0.3765561	1.841226	0.0655885	.	0.0259261	*	
Age8	1.7025431	0.3887542	4.379484	0.0000119	**	0.0000002	**	
StructureSR	4.1861021	0.1951993	21.445267	0.0000000	***	0.0000000	***	
Match_MismatchMis	-0.6745071	0.1304820	-5.169351	0.0000002	**	0.0000000	***	

First of all, we will plot the residuals and the randomized quantile residuals against theoretical quantiles to detect outlying/deviating residuals or patterns that may indicate a problem. Secondly, we will compute the cook distances to look for influential points. Finally, the distribution of the random effects are discussed. The function `qresiduals()` cannot be used for `glmer` models, however, we can simulate randomized quantile residuals on the uniform scale (`DHARMa` package) and invert the latter to the Gaussian scale.





We notice from the first three plots that residuals are centered around zero, look approximately Gaussian and that there are no influential points. In fact, all cook distances are lower than 0.026. After simulating three randomized Q-Q plots, we can say that we have no outlying residuals and more generally a good fit to our data. Even if the distribution of the random intercept is right skewed, the Kolmogorov–Smirnov test provide a $P_{\text{value}} \approx 0.55$ meaning that we cannot reject the null hypothesis that the predictions of the random effects are drawn from a Gaussian distribution. It turns out that the P_{value} of the LRT test between this model and the one without the random intercept is smaller than 2^{-16}, meaning that the random intercept is significant. All fixed effect coefficients are significant according to the t-tests and z-tests.

None of the assumptions related to GLMMs seem to be violated, and no potential problems in the residuals, in the cook distances, or in the amplitude (absolute values) of the estimates are observed.

Second model

For this second model, the three possible interactions are added : (Age:$Structure$, Age:$MatchMismatch$ and $Structure$:$MatchMismatch$).

It turns out that the AIC of this model is equal to 370.25 ($< AIC_{model\ 1\ AdaptiveGQ} = 376.80$) and that the interactions (Age:$Structure$) and (Age:$MatchMismatch$) are not significant.

To obtain a more parsimonious model, a stepwise method based on AIC, BIC and LRT tests, is used.

	Removed.Interaction	AIC	BIC	LRT.pvalue	
Step 1	none	370.25	409.91		
	Age:Structure	370.69	403.14	0.10878	
	Age:Match_Mismatch	367.33	399.78	0.58251	<-
	Structure:Match_Mismatch	378.61	414.66	0.00564	
Step 2	none	367.33	399.78		
	Age:Structure	369.05	394.29	0.05733	<-
	Structure:Match_Mismatch	374.84	403.69	0.00859	
Step 3	none	369.05	394.29		
	Structure:Match_Mismatch	376.81	398.45	0.00758	

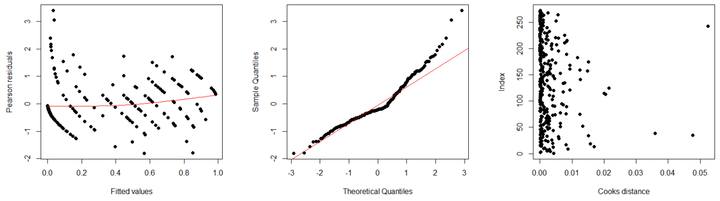
In the first step, all three criterions lead us to remove the (Age:$MatchMismatch$) interaction. In the second step, only BIC and LRT tests lead us to remove the (Age:$Structure$) interaction. However, the difference in AIC is small (< 2) and the coefficients is not significant. Hence, this interaction has been removed. Finally, the third step shows that the ($Structure$:$MatchMismatch$) interaction is needed. In fact, this interaction is highly significant.

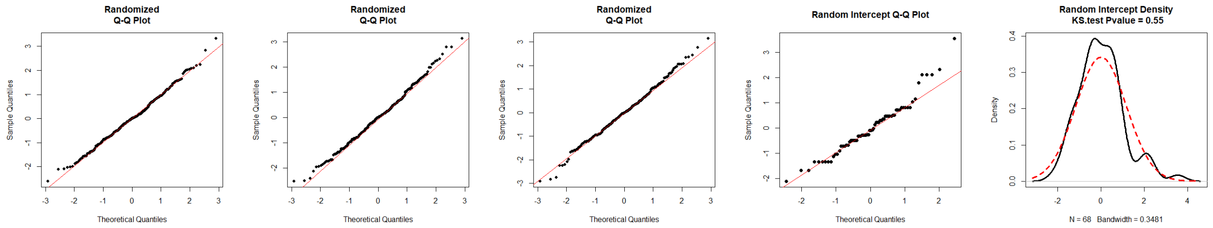
Final model

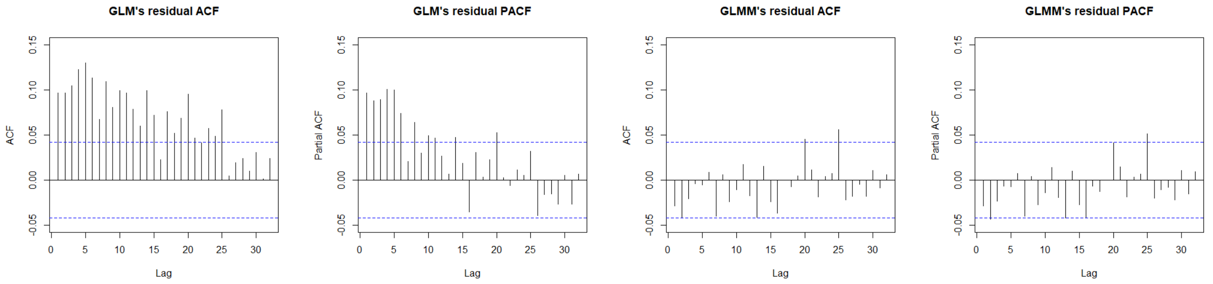
```
model_3_AdaptiveGQ <- glmer(cbind(Success, Failure) ~ 1 + Age + Structure + Match_Mismatch
+ Structure:Match_Mismatch + (1|Id), family=binomial, data=groupeddata, nAGQ = 24)
```

AIC	BIC	logLik	deviance	df.resid	|	RE		VAR	SD
369.0428	394.2834	-177.5214	355.0428	265	|	Id	(Intercept)	1.38331	1.176142

	Estimate	Std. Error	z value	Pr(>|z|)		Pr(>|t|)		Phi
(Intercept)	-3.7593063	0.3263782	-11.518253	0.0000000	***	0.0000000	***	0.63
Age7	0.6833405	0.3786861	1.804504	0.0711523	.	0.0242811	*	
Age8	1.7003960	0.3911719	4.346927	0.0000138	**	0.0000001	**	
StructureSR	3.8228408	0.2210853	17.291250	0.0000000	***	0.0000000	***	
Match_MismatchMis	-1.5826109	0.3428220	-4.616422	0.0000039	**	0.0000000	***	
SR:Mis	1.1007109	0.3713572	2.964022	0.0030365	**	0.0002419	**	







We see no outlying residuals and more generally a good fit to our data. There are no influential points and the distribution of the random intercept is still right skewed, but the Kolmogorov–Smirnov test provide a $P_{\text{value}} \approx 0.55$. The estimated dispersion parameter ($\hat{\phi}$) indicates underdispersion with respect to the Binomial distribution. However, one can show that the coefficient estimates are independent of ϕ in the optimization process. The z-tests and t-tests confirm the significance of all factors, however, LRT tests are more suitable in GLMM context and it allows to test the significance of the random effect ($H_0 : \sigma^2_\gamma = 0$). Before testing the random effect, we can see from the ACF and PACF plots (computed on 32 lags) that it captured the autocorrelation structure in the residuals. Indeed, by estimating a random intercept, there are no more autoregressive or moving average components apparent when compared to the GLM's residuals.

Equivalently to the LRT test, the latter can be inverted to obtain profile likelihood confidence intervals at a 95% level.

	2.5 %	97.5 %	|		2.5 %	97.5 %
.sig01	0.9418	1.4766	|	StructureSR	3.4027	4.2707
(Intercept)	-4.4258	-3.1355	|	Match_MismatchMis	-2.2943	-0.9410
Age7	-0.0651	1.4459	|	SR:Mis	0.3983	1.8624
Age8	0.9323	2.4921	|			

From the above table, we see that all variables are significant. As one cannot say that *Age*7 is statistically different from *Age*5, we see that *Age*8 is significantly different. Hence, we cannot remove this variable.

Population-specific model

The marginal coefficients from mixed models with nonlinear link functions can be computed using Monte Carlo integration (Hedeker et al. 2018). This can be done with the functions `mixed_model()` and `marginal_coefs()` from the `GLMMadaptive` package.

	Estimate	Std.Err	z-value	p-value	
(Intercept)	-3.1367540	0.2989948	-10.490999	0.0000000	***
Age7	0.5889956	0.3267298	1.802699	0.0714355	.
Age8	1.4585601	0.3401181	4.288394	0.0000180	**
StructureSR	3.1376498	0.2041833	15.366827	0.0000000	***
Match_MismatchMis	-1.4647129	0.3238467	-4.522859	0.0000061	**
SR:Mis	1.0814398	0.3442014	3.141881	0.0016787	**

The amplitude of the coefficients shrank down for all of them, and they are close to the subject-specific model. The conclusions made from the subject-specific model will not be largely different from the marginal model. Indeed, the signs of the estimates haven't changed meaning that a variable that influence positively the probability to succeed will still have a positive impact but with a small change in the amplitude.

Having validated our final model and shown the significance of the parameters, we will turn our attention to the interpretation. In the next section, the subject-specific model as well as the population-specific model - which are per se - the same model at different scales, will be discussed.

III - INTERPRETATION

Subject-specific model

Random Intercept

To interpret properly the model, it is essential to interpret the random effects. To do so, the extreme cases are taken as example. The child n°97 has the highest estimated random effect and the child n°92 the lowest. It turns out that it is coherent with the empirical proportion of successes of these children. The average proportion of successes is 33% while it is 72% for the child n°97 and 0% for the child n°92. The most relevant is that these two children were consistent in their performance. Indeed, regardless of the conditions of the experiment, child n°92 did not succeed in any sentence repetition task while child n°97 has consistently outperformed. In fact, the child n°97 is the 5 years old child who outperforms the other children in each type of exercise in the dotplots on page 3. In other words, the random effects seemed to reflect well the within cluster correlation.

Fixed effects

For the sake of clarity, the theoretical equation (with rounded estimates) is presented:

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = -3.76+0.68\iota[Age=7]_{it}+1.70\iota[Age=8]_{it}+3.82\iota[SR]_{it}-1.58\iota[Mis]_{it}+1.10\iota[SR,Mis]_{it}+\gamma_i$$

The latter can be rewritten on the odd scale as:

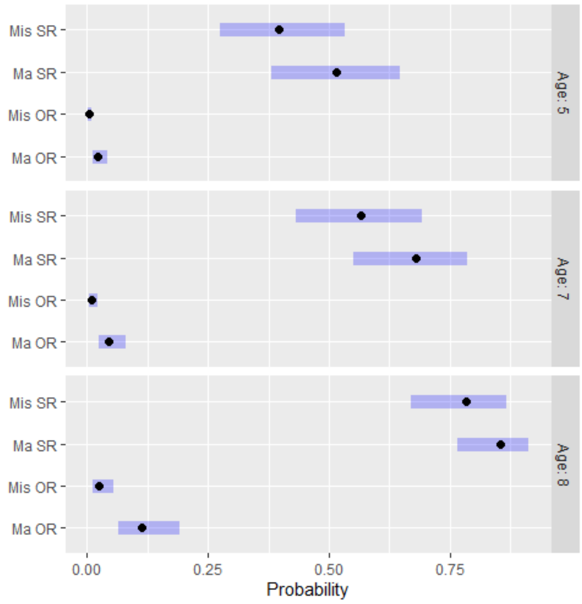
$$odd_{it} = e^{-3.76} \times e^{0.68\iota[Age=7]_{it}} \times e^{1.70\iota[Age=8]_{it}} \times e^{3.82\iota[SR]_{it}} \times e^{-1.58\iota[Mis]_{it}} \times e^{1.10\iota[SR,Mis]_{it}} \times e^{\gamma_i} \tag{2}$$

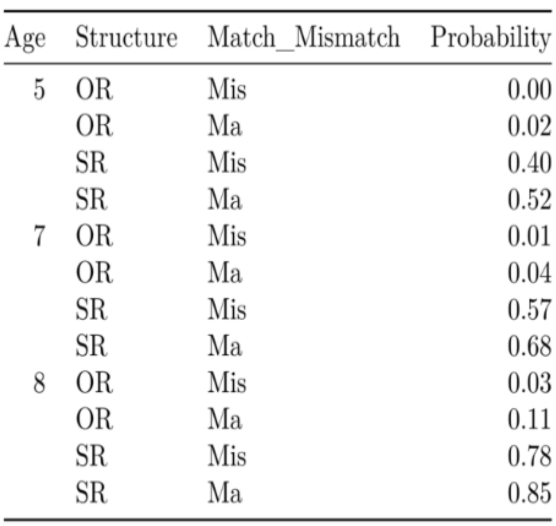
It is important to note that the contrast used is the default one in R i.e. the treatment contrast. If the condition (c) of an indicator function is satisfied it implies that $\beta\boldsymbol{\iota}[c] = \beta$. Consequently, when keeping all the other variables fixed, the odd is multiplied by e^{β} when $\boldsymbol{\iota}[c] = 1$. Note that when we keep all the other variables fixed, the random effect is included.

This multiplicative effect corresponds to the shift with respect to the reference. Indeed, when the condition (c) is not satisfied, the odd corresponds to the reference level because when $\boldsymbol{\iota}[c] = 0$ it implies that $e^{\beta\boldsymbol{\iota}[c]} = 1$. The interest of the logit link function is the interpretability it confers on the odd scale. Indeed, odd_{it} is a strictly monotonous and p_{it}-increasing function on the interval [0,1].

- The reference levels are $Age = 5$, $Structure = OR$ and $Match\ Mismatch = Ma$. Hence, $e^{-3.76}$ correspond to the odd of a 5 years old child with an object relative and a match.
- Keeping all other variables fixed, the odd is multiplied by $e^{0.68}$ if the child is 7 years old or by $e^{1.70}$ if the child is 8 years old. This means that the older the child is, the more likely he or she is to succeed.
- When considering the variable $Structure$ and $Match\ Mismatch$, one need to take into account the interaction between them. Keeping all other variables fixed, the odd is multiplied by $e^{-1.58}$ with an object relative (OR) and a mismatch (Mis), is multiplied by ($e^{3.82} \times e^{-1.58} \times e^{1.10}$) $= e^{3.34}$ with a subject relative (SR) and a mismatch (Mis) and is multiplied by $e^{3.82}$ with a subject relative (SR) and a match (Ma). This means that the worse to the best conditions to succeed according to our model are the combination OR and Mis ($e^{-1.58}$), than OR and Ma (e^{0}), than SR and Mis ($e^{3.34}$) and finally SR and Ma ($e^{3.82}$).

Informally, the same conclusions can be drawn from the plot and the table below. Indeed, we see that the older the child is, the more likely he or she is to succeed. Also, it appears that the probability of success increase when there is a match and when there is a subject relative clause. Moreover, it allows us to understand which variable influence the most the probability to succeed (i.e. it allows us to get a feeling of what is a high value of a coefficient or a small value). In fact, we observe that the most influential variable is the structure.





Age	Structure	Match_Mismatch	Probability
5	OR	Mis	0.00
	OR	Ma	0.02
	SR	Mis	0.40
	SR	Ma	0.52
7	OR	Mis	0.01
	OR	Ma	0.04
	SR	Mis	0.57
	SR	Ma	0.68
8	OR	Mis	0.03
	OR	Ma	0.11
	SR	Mis	0.78
	SR	Ma	0.85

Now that we have interpreted our model, the importance of the random effect can be emphasized through an example. By selecting a child whose random effect is large to compute the probabilities of success (5 years old child n°97) and, a child whose random effect is equal to 0 (average 5 years old child), the following probabilities were obtained.

	Probability Average 5YO child (RE = 0)	Probability 5YO child n°97 (RE = 3.61)
OR Mis	0.00	0.15
OR Ma	0.02	0.46
SR Mis	0.40	0.96
SR Ma	0.52	0.98

Population-specific model

When it comes to the population-specific model, the interpretation is done in the same way. In fact, it is even simpler because there is no more random effect. One can easily compute the probabilities for all combinations of the covariates.

	Probability 5YO child	Probability 7YO child	Probability 8YO child
OR Mis	0.01	0.02	0.04
OR Ma	0.04	0.07	0.16
SR Mis	0.41	0.55	0.75
SR Ma	0.50	0.64	0.81

CONCLUSION

In this analysis, the features influencing the understanding and acquisition of relative clauses were determined. The first classical hypothesis states that object relatives are more difficult to comprehend and produce for children than subject relatives. The second one states that this acquisition is also impacted when there is a match between certain characteristics of the subject and the object of the relative. In this analysis, these assumptions were verified and proven true. More formally, the provided features had a statistically significant impact on the probability of success. Indeed, we have found that age has a positive effect on the chances of success, that the probability of success increase when there is a subject relative clause in contrast with an object relative clause and that the probability of success increase when there is a match between certain characteristics of the subject and the object of the relative.

In addition, by conducting an analysis both at the subject and population scale, our results shown the same conclusions with minor changes in the amplitude of the coefficients. However, one may propose an alternative method to investigate this data set. Indeed, another suitable type of model exist. They are based on generalized estimating equations (GEE).

References

GLAM. 2020. "GLAM 2020 Course." - *UNIGE - GSEM*.

Hedeker, Donald, Stephen HC du Toit, Hakan Demirtas, and Robert D Gibbons. 2018. "A Note on Marginalization of Regression Parameters from Mixed Models of Binary Outcomes." *Biometrics* 74 (1): 354–61.