Assignment Bayesian Statistics

Elise van Wonderen (6027202)

5 January 2022

Introduction

For the current analysis I will make use of a dataset on bilingual children's language proficiency (van Wonderen & Unsworth, 2021). In the dataset, there are 51 Spanish-Dutch bilingual children between the ages of 3 and 9, who carried out a Sentence Repetition Task (SRT) as a measure of their general language proficiency in Dutch. In addition to their scores on the SRT (in percentages), we also know their age in years, and the level of paternal and maternal education measured on a 5-point scale (1 = primary education, 2 = secondary education, 3 = post-secondary non-tertiary education, 4 = bachelor or master's degree or equivalent, 5 = doctoral degree or equivalent). For the current analysis, I also computed level of parental education by averaging the paternal and maternal education levels. Descriptive statistics for all variables in the dataset can be found in Table 1.

The main research question that I will be answering is to what extent children's age and parental education predict children's language proficiency as measured by the SRT, and which of the two predictors is more important. Given the wide age range of the children in the current dataset and the limited variability in level of parental education, my hypothesis is that children's age will be a stronger predictor of their language proficiency than parental education.

In addition to the model specified above, I will also consider a model with age and paternal education only (model 2), age and maternal education only (model 3), and a model with age and age squared (to account for a potential quadratic relationship between age and language proficiency; model 4). I will investigate which of these models describe the data best. If there is a very strong non-linear relationship between age and SRT, and the effect of age is much stronger than the effect of parental education, then I expect the model 4 to describe the data best. If maternal education is a much stronger predictor than paternal education or vice versa, we may also see a difference in the model fits of model 2 and 3. If maternal and paternal education have additive effects, I expect the first model to be better than either model 2 or 3.

Table 1: Descriptive statistics of children's age and parental education and their score on the Sentence Repetition Task

	mean	sd	median	min	max
SRT score (%)	67.95	28.46	76.67	3.33	100.00
Parental education	4.05	0.48	4.00	2.50	5.00
Paternal education	4.08	0.56	4.00	2.00	5.00
Maternal education	4.02	0.62	4.00	2.00	5.00
Age (in years)	6.33	1.52	6.10	3.43	9.92

Methods

To run the regression analyses, I programmed a Gibbs sampler with an independent Metropolis-Hastings (M-H) step for the second predictor in the model. In each iteration, the Gibbs sampler samples parameter values from the (proportional) conditional posterior distributions, which are a function of the prior parameter values, the data and the current values of the other model parameters.

For the M-H step, a value is sampled from a proposal distribution. Then, the probability of acceptance (POA) is calculated by taking the ratio of the probability of the proposed value over that of the previous value for the conditional posterior density, times the ratio of the probability of the previous value over that of the proposed value for the proposal density. Finally, the POA is compared to a random value from a uniform distribution from 0 to 1. If the POA is greater than this value, the proposed value is accepted. Otherwise, the proposed value is rejected and the current value is retained. To prevent any of the probabilities to become very large, I have used log-probabilities instead (and thus I also used addition and subtraction in the calculations, rather than multiplication and division).

I specified normal priors for the model parameters β_0, β_1 and β_2, and an inverse-gamma prior for the variance. For the main analysis, I specified vague priors with a mean of zero and a variance of 1000 for parameters β_0, β_1 and β_2, and hyperparameters α and β of 0.001 for the variance. To see how the specification of the priors would influence the results, I also specified informative priors for β_2 (see "Results"). For the M-H step, I specified a normal proposal density for which I plugged in the maximum likelihood estimates for the mean and standard deviation. I used two chains with 10,000 iterations each, of which 2,000 were used for the burn-in period. Different but reasonable starting values were chosen for each chain. Predictor variables were mean-centered prior to estimation.

Results

I combined the posterior samples of both chains to compute the point estimates for Model 1 as presented in Table 2. Age has a clear effect on SRT score, with an increase of 12.68 percentage points for each year that children are older, 95% central credibility interval (CCI) = 8.82 - 16.47. In addition, for each one-level increase in parental education (which was measured on a scale from 1 to 5), children's scores on the SRT increased with 13.44 percentage points. However, the 95% CCI is much wider, ranging from 1.74 to 25.38. Therefore, we can be confident that there is some effect of parental education, but we are less confident on whether this is a large effect or not.

Table 2: Estimates, standard deviations, MC errors and 95% central credibility intervals for the parameter values of Model 1

	Mean	SD	MC error	Lower	Upper
Intercept	67.38	2.94	0.02	61.62	73.21
Age	12.68	1.95	0.02	8.82	16.47
Parental education	13.44	6.03	0.05	1.74	25.38
Residual variance	442.45	94.44	0.75	295.04	662.32

Bayes factors

I used Bayes Factors (BFs) to (i) quantify the evidence that the effects of the predictors were meaningfully different from zero (defined as having a standardized effect of 0.2 or greater, conform the definition of a small effect by Cohen, 1992), and (ii) to investigate whether the standardized effect of age was greater than that of parental education. For calculating BFs, I used the R-package `bain` (Gu, Hoijtink, Mulder, & van Lissa, 2020). I computed all BFs twice, while setting a different seed, to ensure stability of the results. I tested the hypotheses "parental education > 0.2", "age > 0.2" and "age > parental education" against their complements, and obtained BFs of 1.76, 5217995373.02923 and 1475, respectively. Based on this, I conclude that there is overwhelming evidence for the hypothesis that the effect of age is meaningfully different from zero, and that the effect of age is greater than the effect of parental education. In contrast, there is no conclusive evidence that the effect of parental education is meaningfully different from zero.

Posterior predictive check

I tested the assumption of linearity with a posterior predictive check. If the assumption of linearity is met, the model's residuals should have the same mean value across the range of fitted values. However, if there is a curvilinear trend in the residuals, we may expect different mean residual values across the range of fitted values. Therefore, as a test statistic, I divided the fitted values into three quantiles, and computed the difference between the mean residual for the third versus the second quantile. If the assumption of linearity is met, this test statistic should be distributed around zero.

For the posterior predictive check, I generated new datasets by – for each iteration – sampling a value from a normal distribution with a mean and standard deviation based on the predictor variables and the sampled parameter values. For each replicated dataset, I computed a vector of residuals and then the test statistic. For the observed data, I computed a vector of residuals for each posterior sample and then computed the test statistic. As can be seen in Figure 1, the test statistic is – as expected – distributed around zero for the replicated datasets. However, for the observed data, the test statistic is distributed around -18.6 and the two distributions hardly overlap. This is also reflected in the posterior predictive p-value (the proportion of times that the absolute value of the test statistic is greater for the observed data than for the replicated data), which is 0.03. This is much smaller than 0.5 (i.e., the p-value if the medians of the two distributions are the same), which is why we conclude the assumption of linearity is not met. This is probably due to the non-linear relationship between age and SRT score: due to ceiling effects, the effect of age on SRT score levels off for older children. This may explain why we see more negative residuals for higher fitted values (the third quantile) than for fitted values in the middle range (the second quantile).

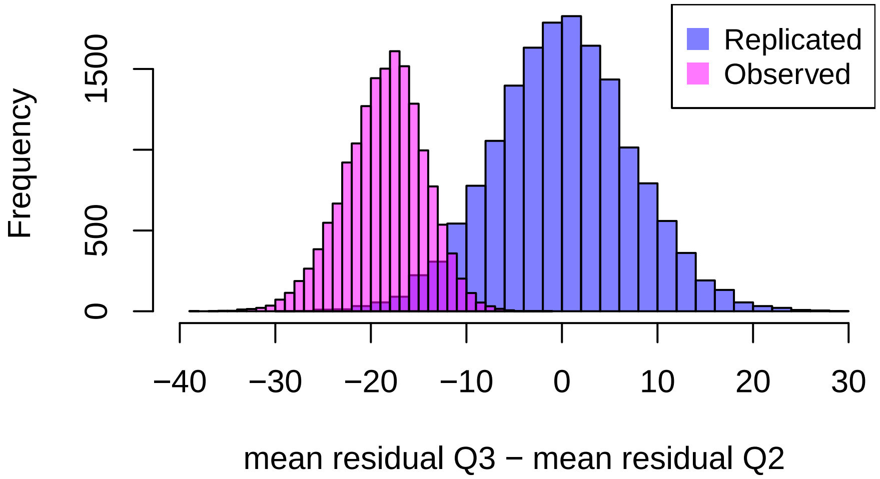

Figure 1: Distribution of the test statistic (mean residual Q1 - mean residual Q2) for the observed and replicated data

Model comparison using the DIC

I compared the four models mentioned in the introduction by means of the DIC, which is a measure of model fit while penalizing for model complexity. It is given by

$$DIC = -2 \ \log f \ (y \mid \bar{\theta_y}) + 2p_D,$$

where $\bar{\theta}_y$ is the posterior mean, and p_D is the number of effective parameters in the model which is calculated by taking the average likelihood over the posterior distribution of θ minus the likelihood evaluated at the posterior mean of θ, i.e.,

$$p_D = \frac{1}{Q}\sum_{q=1}^{Q} -2\log f(y \mid \theta^q) \; + \; 2\log f\;(y \mid \frac{1}{Q}\sum_{q=1}^{Q}\theta^q),$$

where Q is the number of Gibbs samples.

Lower values for the DIC indicate a better fitting model. Based on the DIC, Model 4 with only age and age squared is the best fitting model (Table 3). It is substantially better than Models 1 and 2, and definitely better than Model 3. The differences between the other three models are not large, but the model with paternal education seems to be slightly better than the model with maternal education.

Table 3: DIC per model

Model	DIC
1. Age + parental educ	458
2. Age + paternal educ	456
3. Age + maternal educ	462
4. Age + age squared	450

Comparison of Bayesian and frequentist approaches

For the main analysis with vague priors, the Bayesian regression analysis leads to almost identical parameter estimates and credible intervals as the frequentist approach using the `lm` function, so from a practical point of view these analyses are similar. However, technically you would not be able to conclude from the frequentist confidence intervals that those are the 95% most probable values for the parameters, as they do not represent a probability distribution and only have an interpretation in the long-run (i.e., if you repeat the study an indefinite number of times, the true population value will be within the confidence interval 95% of the time). In addition, if based on previous literature we have good reason to specify an informative prior, we can obtain more precise estimates with the Bayesian approach as it then takes this prior information into account.

In addition, rather than using *p*-values to assess the significance of a predictor, we can make use of the Bayes Factor which allows us to directly test the evidence in favor of a hypothesis, in contrast to the p-value, which only allows us to (not) reject the null hypothesis that the predictor has an effect of zero (which is a rather unlikely hypothesis to start with, since we already know that age and parental education most likely have at least some effect on children's language proficiency). Based on the *p*-values obtained with the `lm` function we would reject the null hypothesis. However, with the Bayes Factor we were able to specify a more sensible hypothesis, namely that the standardized effects are larger than 0.2 (i.e., they minimally represent a small effect). We saw that there was overwhelming evidence for this hypothesis for age, but for parental education we concluded that there was no conclusive evidence for this hypothesis. It is true that the *p*-value for age ($p < .001$) is also much smaller than that for parental education ($p = 0.3$), but the Bayes Factor is more intuitive in that it can directly quantify evidence in favor of a hypothesis (rather than quantifying the probability of the data if the null hypothesis were true), and it is much more flexible (we could have also specified a medium or a large effect, for example, and we also tested whether age was a stronger predictor than parental education).

With regards to testing model assumptions, we saw in Figure 1 that even under the assumed model, we may obtain a test statistic that deviates from zero, so one test statistic in isolation cannot immediately prove a model assumption was violated. However, because we could compare the distributions of the test statistic for the observed and replicated data, we saw that the distributions were so different that we concluded the assumption of linearity was most likely violated.

References

Cohen, J. (1992). A power primer. *Psychological bulletin, 112*(1), 155.

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. J. (2020). *bain: Bayes Factors for Informative Hypotheses. R package version 0.2.4.* https://CRAN.R-project.org/package=bain

van Wonderen, E., & Unsworth, S. (2020). Testing the validity of the Cross-Linguistic Lexical Task as a measure of language proficiency in bilingual children. *Journal of Child Language, 48*(6), 1101-1125. https://doi.org/10.1017/S030500092000063X