Data Analysis 2 Term Project

Gyongyver Kamenar (2103380)

Introduction

This paper analyze the differences of English correctness between people with several different native languages. Nations have different linguistic characteristics, cultural and social norms, and education systems as well. Therefore, people with different nationalities or native languages have differences in English learning. That's why, the motivation of this study is to see, that whether there are differences in English correctness between English learners of different native languages controlled for other relevant factors. It is a relevant question for everyone, because if there are significant differences, scientist can further investigate the main reasons behind it. It might have policy implications on education systems or recommendation of learning practices, methods and habits for English learners.

Data

Origin of the data

The used dataset was orignially collected by Hartshorne et. al. (2018) who studied the critical period for second language acquisition. He collected data from 669,498 respondents through online English quizes. The ca. 10 minute-long English grammar quize was deliberately developed to enabled the researcher to measure the respondent's syntactic English knowledge and also provided several demographic variables. The critical items of the quize are the diagnostics of proficiency and there were additional items to distinguish between dialects. In my analysis I used the percentage of correct critical items as a proxy of English correctness.

- **English correctness** : percentage of critical quize items correct

Further information about the data is available (here)[https://osf.io/pyb8s/] and in Hartshorne et. al. (2018).

Explanatory variables

In this subsection I list the available variables in the dataset that I found relevant explanatory variable. In some cases I did not use a given variable in my analysis and I explain the reason.

- **Age** : age of subject (numeric)
- **Gender** : gender of the subject, either male, female or other (categorical)
- **Native language** : subject's native language(s) (categorical)
- **Native English** : English is the subjects native language (binary)
- **Primary language** : subject's primary language(s) now (categorical)
- **Primary English** : subject's primary language is English (binary)
- **Psychiatric** : subject repoted any psychiatric disorders (binary)
- **Starting age of English learning** : age at start of English learning (numerical)
- **Language status** : either monoeng (native speaker of English only), bileng (native speaker of English + at least one other lang), immersion learner(spending at least 90% of their life since age of first exposure in an English-speaking country), or non-immersion learner(spent at most 10% of post-exposure life in an English-speaking country and no more than 1 year in total) (categorical)
- **Years of English "learning"** : age - starting age of English learning (numerical)
- **Education** : Highest level of education (categorical)

Variables I found relevant but did not use:

- **Dyslexia** : subject reported difficulty with reading (binary)

Reason: there was no variation in this variable (noone with dyslexia)

- **Live with English** : subject lives with any native speakers (binary)

Reason: majority of the data were NULL values indicating missing observation and I cannot be sure what it means

- **Countries** : countries subject lived in (categorical)

Reason: too much distinct categories (thousands) with just few subjects, moreover most subject lived in 2 or more countries

- **Current country**: country currently lived in (categorical)

Reason: too much distinct values (hundreds) and NAs

- **English country years**: number of years living in English speaking countries

Reason: most observations are missing and a similar feature is captured by Language status variable

Scope

The data includes bilinguals and immersion learners as well, which is beyond the scope of my research question. The study's interests are subject with 1 native language (not bilingual) and in case it is not English, we want to investigate subject who are non-immersion learners (spent at most 10% of post-exposure life in an English-speaking country and no more than 1 year in total). Thus the language status is filtered for monoenglish or non-immersion.

Furthermore, I analyze subject with 1 native language and there are several different categories in the dataset. I decided to analyze just the top 10 native languages based on the number of subject in the dataset, namely; English, Finnish, Turkish, German, Hungarian,Russian, Dutch, Polish, Swedish and Spanish in the respective order. Even from the least frequent native language Spanish there were more then 9200 subjects in the filtered dataset.

Age is also a relevant factor in English correctness, so subjects at least 14 years old to make they are able to learn a language.

Descriptive statistics

Please see below the descriptive statistics of the numerical variables and the correlation matrix in Figure 1.

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P25	P75	P95
Age	30.48	27.00	11.63	14.00	89.00	17.00	22.00	36.00	55.00
Starting age of English learning	4.02	0.00	5.42	0.00	69.00	0.00	0.00	9.00	13.00
Psychiatric disorder	0.03	0.00	0.18	0	1	0.00	0.00	0.00	0.00
English correctness	0.94	0.96	0.06	0.31	1.00	0.82	0.93	0.98	1.00
Years of English learning	26.46	23.00	13.06	1.00	89.00	10.00	18.00	32.00	54.00
Log(English correctness)	3.13	3.01	1.02	−0.81	5.25	1.50	2.47	3.62	5.25

The number of observations is 415616 for all of our key variables. The age of subjects is filtered to be at least 14 years, because below that they might not able to learn a language properly. There were some errors in the English correctness variable, because some native English people had correctness below 0.5. There is hardly and correctness score below 0.5 even for non-native English people, so these English people most probably did not make the effort to properly do the quize and that is why I removed them.

We can see, that the SD of English correctness variable is really small, the mean is 0.94 and the median is 0.96 so the distribution is left skewed. See the distributions Figure 3 .

As the focus is English correctness for different native languages, the Figure 2 shows the boxplot for each native language. The differences in English correctness are visible on the boxplot.

DESCRIPTION OF THE FIGURE. WHAT DOES IT TELS US?

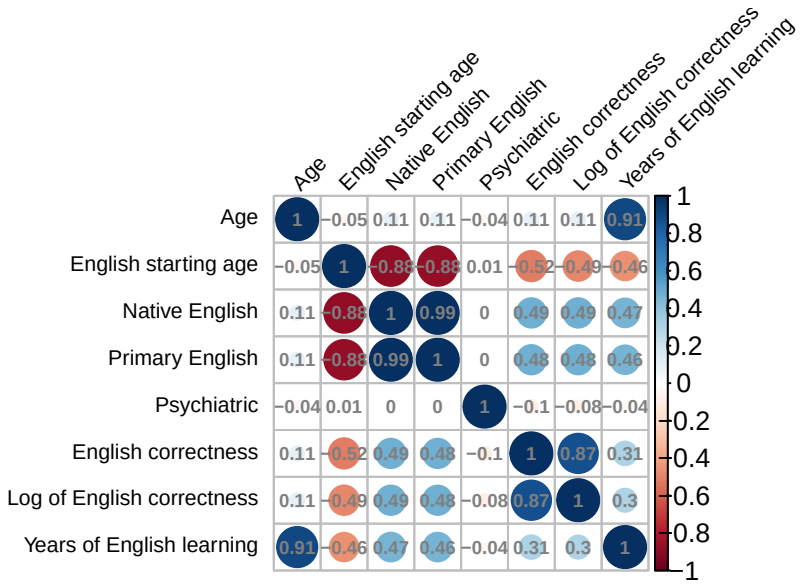

Figure 1: Correlation matrix

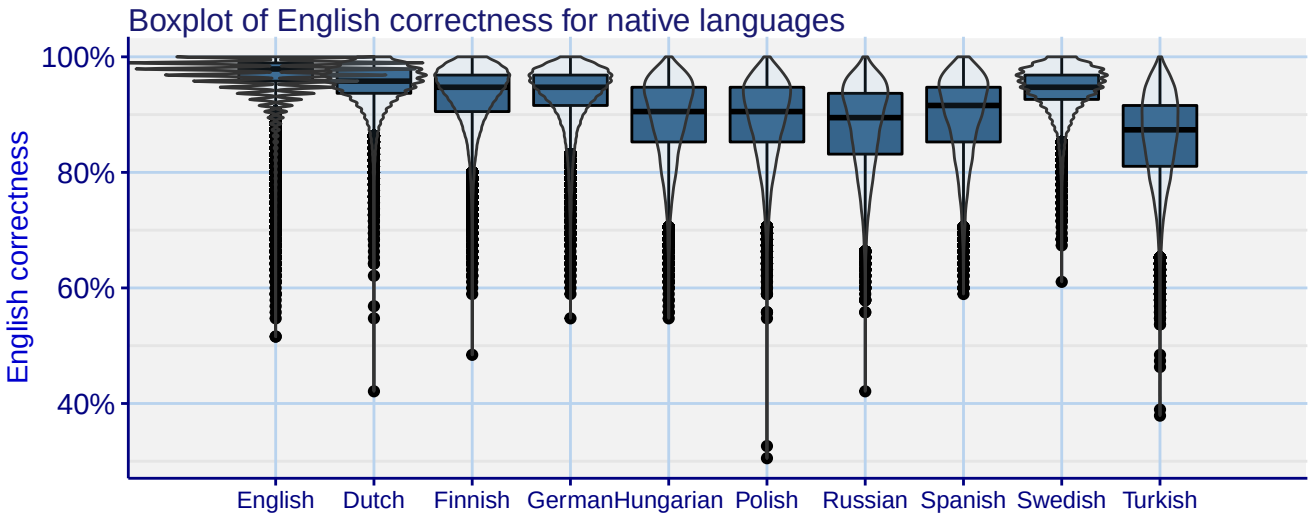

Figure 2: Boxplot of English correctness

Exploratory data analysis

The key pattern of association is:

How will you include this in your model?

Short description on the other variables: 2-10 sentence depends on the amount of variables you have. You should reference your decisions on the graphs/analysis which are located in the appendix.

Models

My preferred model is:

score = 0.95 0 $(student/teacher < 18)$ -0.01 $(student/teacher \geq 18) + \delta Z$

where Z are standing for the controls, which includes controlling for english language, lunch, other special characteristics and wealth measures. From this model we can infer:

- when every covariates are zero, students expected to have grade score of 0.95
- when the student to teacher is one unit larger, but below the value of 18, we see students to have on average 0 smaller grades.
- when the student to teacher is one unit larger, with the value above or equal to 18, we see students to have on average 0.01 smaller grades.

However, based on the heteroskedastic robust standard errors, these results are statistically non different from zero. To show that, I have run a two-sided hypothesis test:

$$H_0 := \beta_1 = 0$$

$$H_A := \beta_1 \neq 0$$

I have the t-statistic as 5.15 and the p-value as 0, which confirms my conclusion.

We compare multiple models to learn about the stability of the parameters. Bla-bla:

Table 2: Estimated models

	Simple	Multiple	Multiple with starting age	Splines	Splines
Intercept	0.862***	0.953***	0.953***	0.954***	0.955***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Native Dutch	0.087***	0.016***	0.016***	0.088***	0.112***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Native Finnish	0.072***	−0.002***	−0.002***	0.071***	0.095***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Native German	0.073***	0.002***	0.002***	0.074***	0.098***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Native Hungarian	0.031***	−0.037***	−0.037***	0.035***	0.059***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Native Polish	0.028***	−0.042***	−0.042***	0.030***	0.054***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Native Russian	0.019***	−0.053***	−0.053***	0.019***	0.043***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Native Spanish	0.036***	−0.041***	−0.041***	0.032***	0.056***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Native Swedish	0.081***	0.006***	0.006***	0.080***	0.104***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Age		0.000***	−0.003***	0.000***	0.000***
		(0.000)	(0.000)	(0.000)	(0.000)
Male		−0.006***	−0.006***	−0.006***	−0.006***
		(0.000)	(0.000)	(0.000)	(0.000)
Other gender		−0.008***	−0.008***	−0.008***	−0.008***
		(0.001)	(0.001)	(0.001)	(0.001)
Native Turkish		−0.067***	−0.067***	0.005**	0.029***
		(0.001)	(0.001)	(0.002)	(0.002)
Graduate Degree		0.020***	0.020***	0.020***	0.020***
		(0.000)	(0.000)	(0.000)	(0.000)
HighSchool Degree (12-13years ed)		0.007***	0.007***	0.007***	0.007***
		(0.000)	(0.000)	(0.000)	(0.000)
Some Graduate School		0.010***	0.010***	0.011***	0.011***
		(0.000)	(0.000)	(0.000)	(0.000)
Some Undergrad (highered)		0.013***	0.013***	0.013***	0.012***
		(0.000)	(0.000)	(0.000)	(0.000)
Undergraduate Degree (3-5years highered)		0.017***	0.017***	0.017***	0.017***
		(0.000)	(0.000)	(0.000)	(0.000)
Psychiatric		−0.029***	−0.029***	−0.029***	−0.028***
		(0.001)	(0.001)	(0.001)	(0.001)
Starting age of English learning		−0.003***			
		(0.000)			
Years of English learning			0.003***		
			(0.000)		
Starting age of English learning (<2)				−0.042***	
				(0.001)	
Starting age of English learning (>=2)				−0.003***	
				(0.000)	
lspline(Eng_start, 1)1					−0.106***
					(0.002)
lspline(Eng_start, 1)2					−0.003***
					(0.000)
Num.Obs.	168 561	415 616	415 616	415 616	415 616
R2	0.200	0.390	0.390	0.397	0.401
R2 Within					
R2 Pseudo					
BIC	−459 206.0	−1 388 367.5	−1 388 367.5	−1 392 725.0	−1 395 634.4
Log.Lik.	229 657.167	694 313.116	694 313.116	696 498.343	697 953.066
Std.Errors	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust

** p < 0.05, *** p < 0.01

Robustness check / 'Heterogeneity analysis'

Task: calculate and report t-tests for each countries.

Conclusion

HERE COMES WHAT WE HAVE LEARNED AND WHAT WOULD STRENGHTEN AND WEAKEN OUR ANALYSIS.

Appendix

Here comes all the results which are referenced and not essential for understanding the MAIN results.

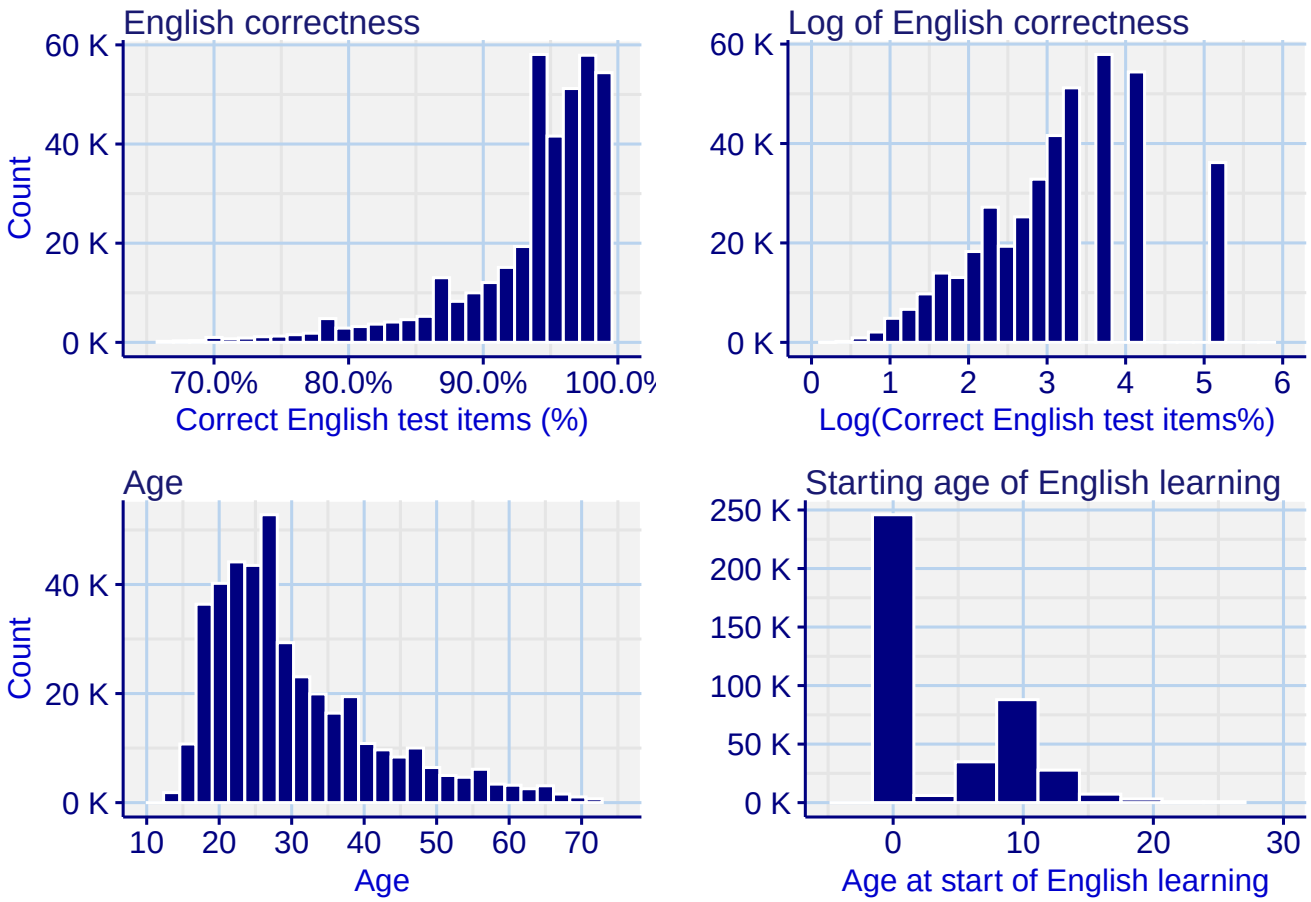

Figure 3: Distributions

```
f1
```

```
f2
```

Bibliography

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. Cognition, 177, 263-277.

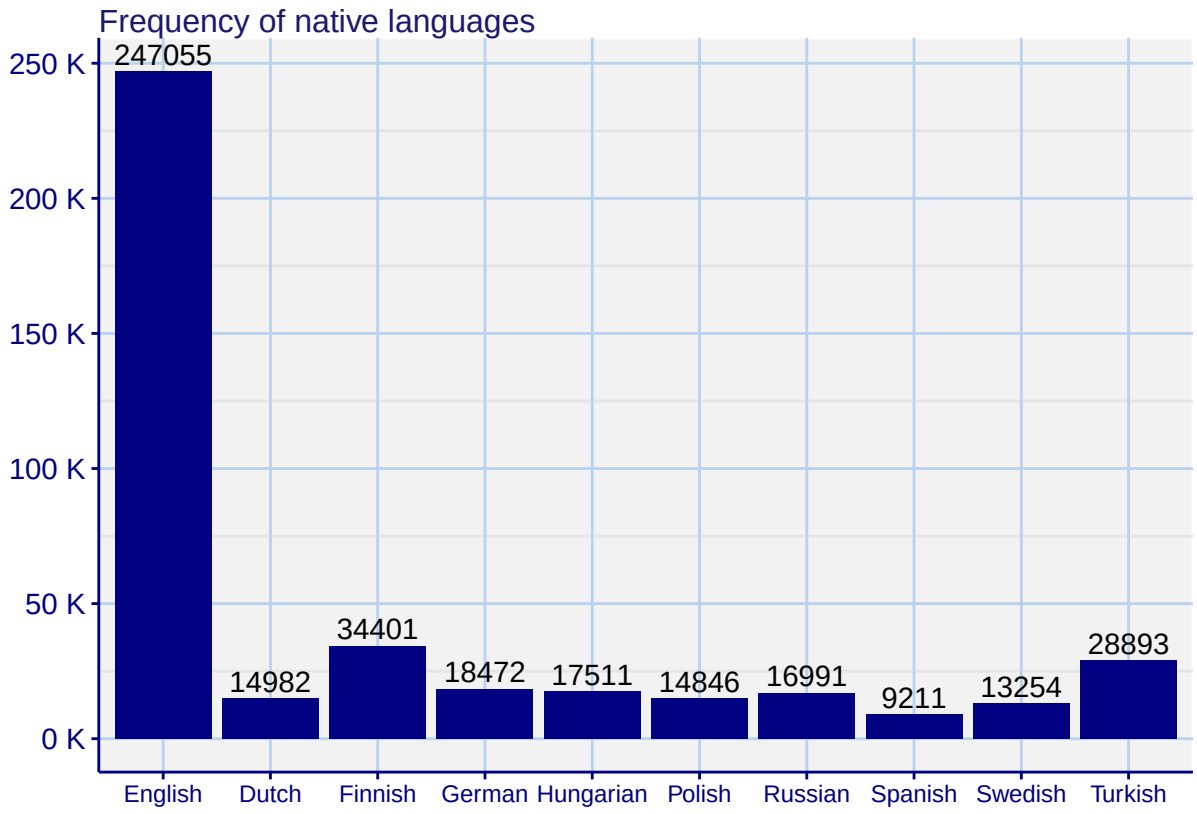

Figure 4: Frequency of native languages

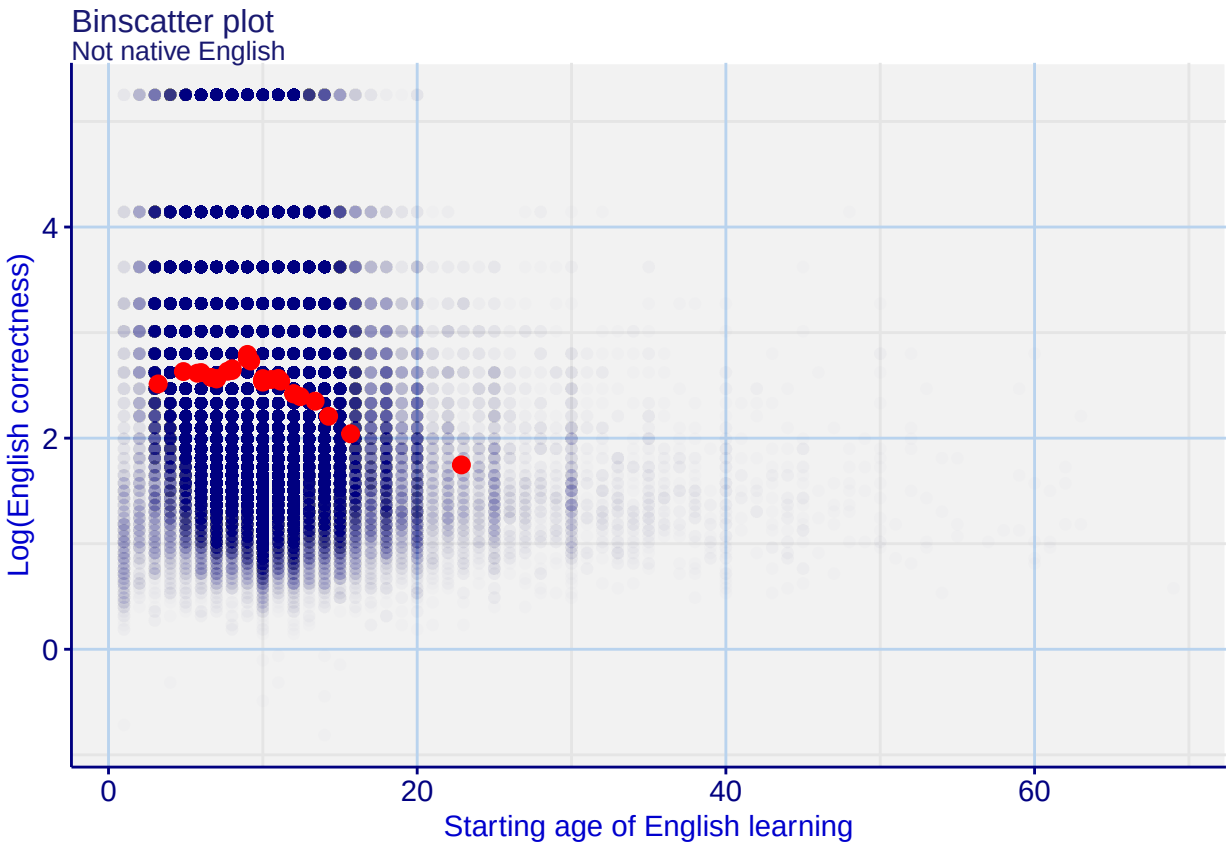

Figure 5: Binned Scatterplots

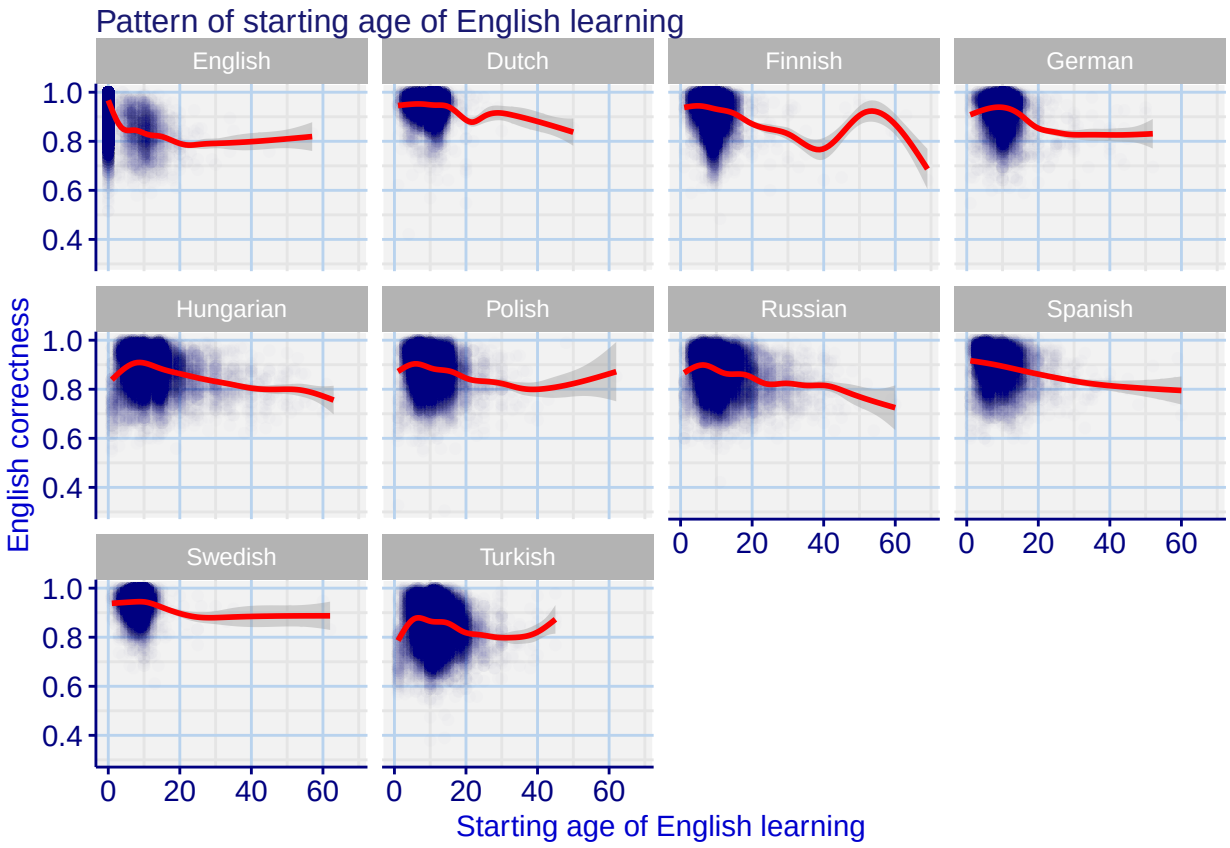

Figure 6: Smoothed patter

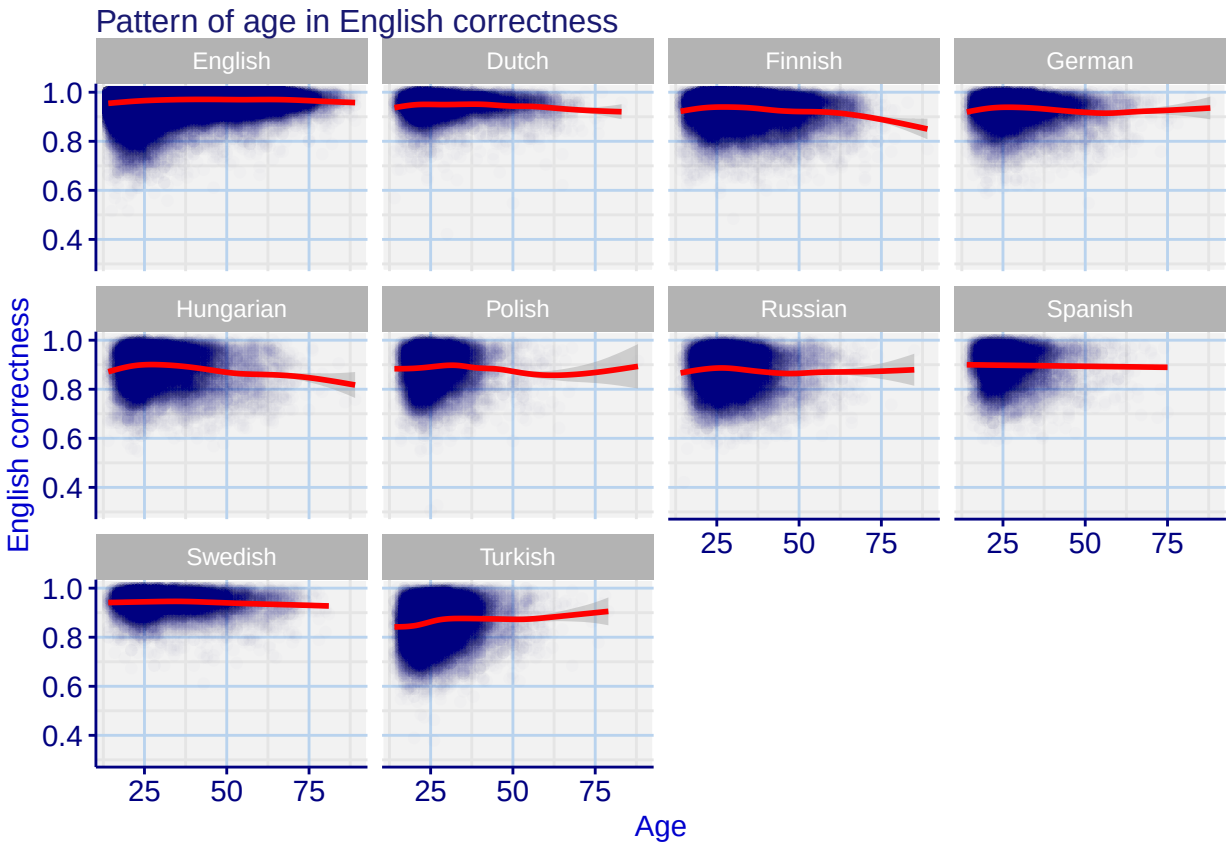

Figure 7: Smoothed patter