Homework 6

Ilya Schurov, Olga Lyashevskaya, George Moroz, Alla Tambovtseva

Deadline: 15 May, 23:59

In this home assignment you are suggested to work with the data set with the results of a psycholinguistic experiment dedicated to lexical decision. In studies of lexical decision paticipants (also called subjects) are asked to decide whether the word shown on the screen is a real word or not. In other words, whether a word exists in the language or it is just an artificical word created using grammatical rules.Then the reaction time is measured: how fast a person clicks on the button *word* or *non-word*.

This data set is taken from the library `languageR`, it contains lexical decision latencies elicited from 21 subjects for 79 English concrete nouns, with variables linked to subject or word. Data collected by Jen Hay, University of Canterbury, Christchurch, New Zealand, 2004.

Some variables of interest:

- `Subject`: participant's id;
- `RT`: logarithmically transformed reaction times;
- `NativeLanguage`: a factor with levels `English` and `Other`, distinguishing between native and nonnative speakers of English;
- `Correct`: a factor with levels `correct` and `incorrect` coding whether the word was correctly responded to as a word rather than a nonword;
- `Word`: word shown;
- `Frequency`: logarithmically transformed lemma frequencies as available in the CELEX lexical database;
- `FamilySize`: log-transformed count of a word's morphological family members;
- `SynsetCount`: log-transformed count of synonym sets in WordNet in which the word is listed;
- `Length`: word's length in letters.

The description of all the variables in this data set can be found here.

For brevity, below we will refer to variable `RT` as "reaction time" despite the fact that it is actually the logarithm of time measured in ms.

1. Lexical decision: correctness and native language

Imagine that you are suggested to conduct a small research on lexical decision. And before proceeding to more substantial analysis you want to check whether the correctness of decision (`Correct`) depends on the person's native language (`NativeLanguage`).

1.0. Load data (link) and look at the summary of the loaded data frame.

1.1. How many correct answers were provided by native English speakers? And by speakers of other languages? Answer the same questions, but for incorrect answers. Provide your R code used to answer these questions and answers as well.

1.2. Create a *mosaic plot* (via `vcd` library or `ggplot2`) that will show the same frequencies as above. Add graph title and correct group labels if necessary (see `?labelings` to find how to adjust, rotate labels, etc). Provide your R code.

1.3. State the method that is applicable to check whether the correctness of decision (`Correct`) depends on the person's native language (`NativeLanguage`). Explain your choice.

1.4. State the null hypothesis you are going to test. State the alternative hypothesis as well.

1.5. Perform the analysis using R. Provide your R code.

1.6. Based on the output obtained, can you conclude that the correctness of decision depends on the person's native language? Explain your answer.

2. Lexical decision: short words compared

Imagine that you have to check whether the average reaction time is different for different short words (less than 5 letters). The question is: is it true that it is harder to recognise some short words than others? If it is true, we can proceed to more sophisticated analysis and think of factors that can affect the reaction time.

2.1. Using `tidyverse` (`dplyr`) choose rows that correspond to correctly named words (column `Correct`) and words consisting of less than 5 letters (column `Length`). Save them to a new dataset, you should use this data set for this task.

2.2. Using `ggplot2` create boxplots of reaction time for different words (column `Word`). Provide your R code. Judging by the graph, report two words with the highest median value of reaction time.

2.3. Choose an appropriate statistical method to answer the question stated at the beginning of this task. Perform the analysis and provide your R code.

2.4. Interpret the output obtained.

2.4.1. State the null hypothesis you tested. State the alternative hypothesis as well.

2.4.2. Based on the output obtained, can you conclude that the reaction time differs for different words? Explain your answer.

3. Lexical decision: reaction time and word features

Now you are suggested to check how reaction time (`RT`) is related to several word features: word frequency (`Frequency`), word family size (`FamilySize`) and number of synonyms (`SynsetCount`). Here you should use the original data set, not the filtered one from the previous task.

3.1. Plot a scatterplot matrix (table of scatterplots) for the pairs of variables chosen at the previous step using `GGAlly` library. Which variables are positively associated? And negatively associated? The association between which variables is the strongest? Explain your answer.

3.2. Check whether the correlation between reaction time and word frequency is statistically significant.

3.2.1. State the null hypothesis you are going to test. State the alternative hypothesis.

3.2.2. Test the hypothesis stated above using R. Provide your R code.

3.2.3. Based on the output obtained, can you conclude that the reaction time is associated with word frequency? Explain your answer.

3.2.4. Report the correlation coefficient obtained. If it is statistically significant, interpret its value: state the direction of association (positive or negative) and the strength of association (approximately, strong or weak). There are no strict rules how to interpret the strength, for example, see here (p.9).

3.3. Create a bivariate linear model that will explain how reaction time is affected by word frequency.

3.3.1. State which variable is independent and which is dependent in our case.

3.3.2. Run this model in R and report your code.

3.3.3. How does (on average) reaction time change when word frequency increases by one?

3.3.4. Based on the output obtained, can you conclude that the word frequency affects the reaction time? Explain your answer. Can you conclude that higher word frequency leads to lower reaction time? Explain your answer.

4. Lexical decision: multiple regression

Now you are suggested to check how reaction time (`RT`) is affected by several word features: word frequency (`Frequency`), word family size (`FamilySize`), length of a word in letters (`Length`), word class (`Class`) and person's native language (`NativeLanguage`). Here you should use the original data set.

4.1. Using `tidyverse` (`dplyr`), choose rows that correspond to real words (see the column `PrevType`). Save them to a new dataset, you should use this data set for this task.

4.2. Create a multiple linear model that will explain how reaction time is affected by word features stated above. Run this model in R and report your code.

4.3. Interpret the R output.

4.3.1. Write the equation of the model using the R output obtained.

4.3.2. All else equal, how does the reaction time change (on average) when the length of a word increases by one?

4.3.3. Interpret the coefficient of `Class` (explain what does it show).

4.3.4. Interpret the coefficient of `NativeLanguage` (explain what does it show).

4.4. Perform some model diagnostics.

4.4.1. Report the R^2 of this model.

4.4.2. Plot any graphs that can show whether there are some patterns in the residuals distribution. Can we conclude that residuals of this model are scattered randomly (with no patterns)?

5. Lexical decision: mixed-effects model

5.1. Take the filtered data set from task 4 (only words, without non-words) and choose rows that correspond to the following subjects:

```
subj <- c("A1", "B", "C", "D", "I", "J", "K", "M1", "P", "R1", "S", "T1", "V", "W1", "Z")
```

In this task you are supposed to work with this filtered data set.

Hint: consider the following example, it might be helpful:

```
6 %in% c(1, 3, 6, 8)
```

```
## [1] TRUE
```

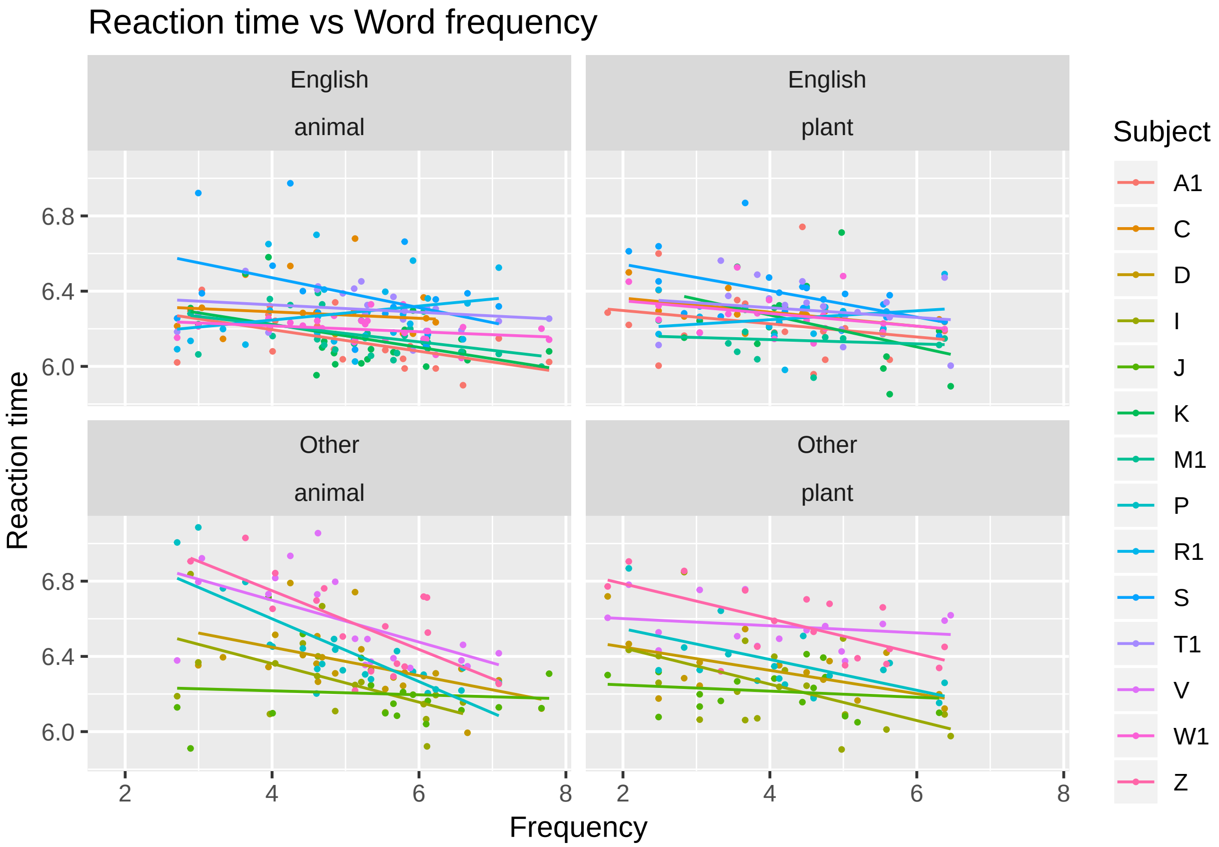
5.2. Using `tidyverse` (`dplyr`), get the following descriptive statistics for every subject: number of trials corresponding to each subject, mean reaction time, standard deviation of reaction time and median reaction time. Can we say that the distribution of reaction time is different for every subject?

5.3. Run the model from the previous task, but now include the random effect on the intercept supposing that the intercept is different for each individual participating in the experiment (`Subject`).

Hint: use the library `lme4`.

5.4. Write the equation of this model using the estimates of fixed effects.

5.5. Replicate the following graph:



Hints: facet grouping is done by the native language of a participant (`NativeLanguage`) and the word type (`Class`), point size is set to 0.5 and line width is also set to 0.5.

Based on these graphs, can you conclude that the intercept is different for every subject?

5.6. Report the intraclass correlation (ICC), i.e. the share of variance that is explained by grouping based on subjects. It can be computed directly based on the R output with random effects or via the function `icc()` from the `sjstats` (see here).