Analysis of time of solving the tasks from mathematics

A case study based on data from the PISA2015 test.

5 July 2017

Problem

We analyze the time of solving exercises from mathematics from the PISA2015 test. Our task was to choose a few independent variables from a wide range in given datasets. Datasets describe both exercises and a single student. Our task was also to build an optimal model based on selected variables.
Final model should as best describe the dependence between the time of solving exercises and our factors.

Data

In our work we use two datasets:

- `actionTimeScoreMath` consisting data about time of solving single sub-task.
- `Student questionnaire data file` consisting details about each student (e.g parents occupation, family status, gender, parents education level).

Data transformation

In the original dataset, we have the time of solving single sub-task, in our model we considered the time of solving all task. We also decided to select data from 5 countries (for faster calculations). We extract the information about a number of exercise and number of sub-exercise. For the information for each student, we divided variable specifying parents economical status into five intervals. We also grouped variables describing parents level of education. We also transform the dependent variable. We use `logtrans{MASS}` function to find an optimal shift and it was equal 45.

Chosen variables

In our model we consider following variables:

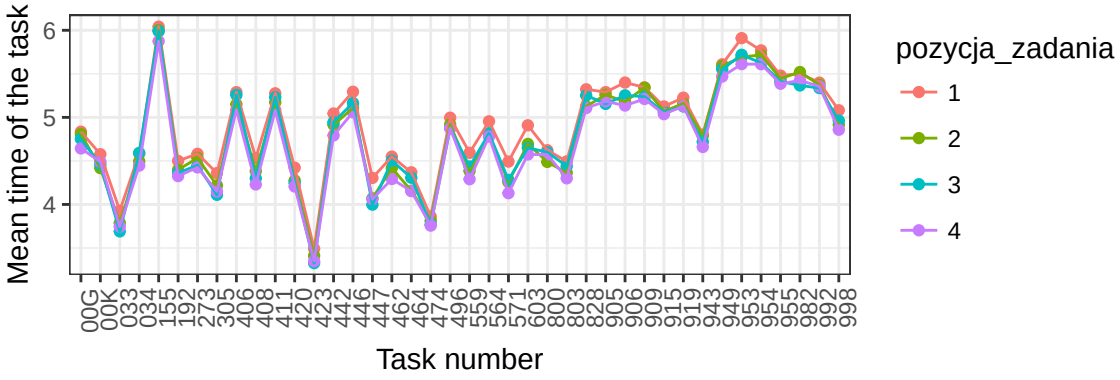
- task number (`zadanie`)
- the position of the task in a questionnaire (`pozycja_zadania`)
- number of a questionnaire (`id_kwestionariusza`)
- the month of birth of single student (`mies_ur`)
- identifier of the country (`id_kraju`)
- gender (`plec`)
- mother education level - grouped using post-hoc testing (`wyk_m_lsd`)
- father education level- grouped using post-hoc testing (`wyk_o_lsd`)
- mother occupation group (`gr_zawod_m`)
- father occupation group (`gr_zawod_o`)
- mother socioeconomic status (`stat_m`)
- father socioeconomic status (`stat_o`)
- school identifier (`id_szkoly`)
- student identifier (`id_ucznia`)

Interactions

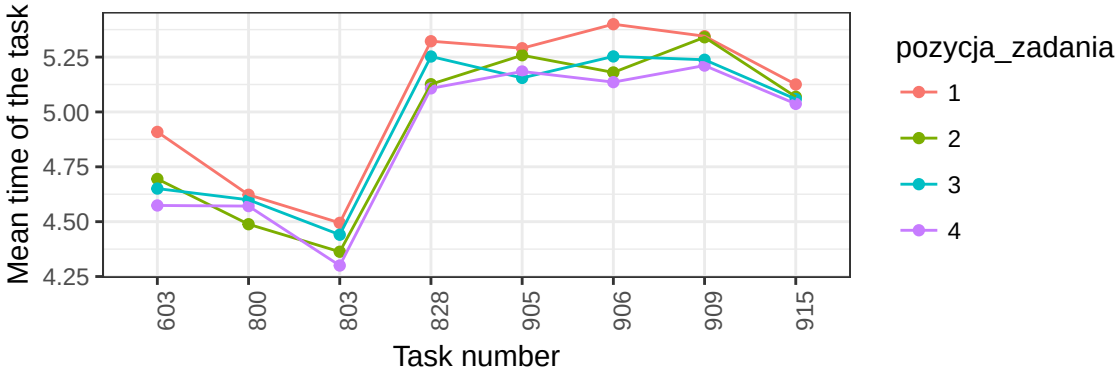
We want to find interactions between chosen variables. We consider following interactions.

Number of the task and its position in questionnaire

We think that the timing of a given task may depend on its position in the questionnaire. For example, a student devotes more time to a long task in the first position and skips over the last.



In the plot above we see that there are interactions for some tasks. Let's take a look at the few chosen tasks to better visualize them.



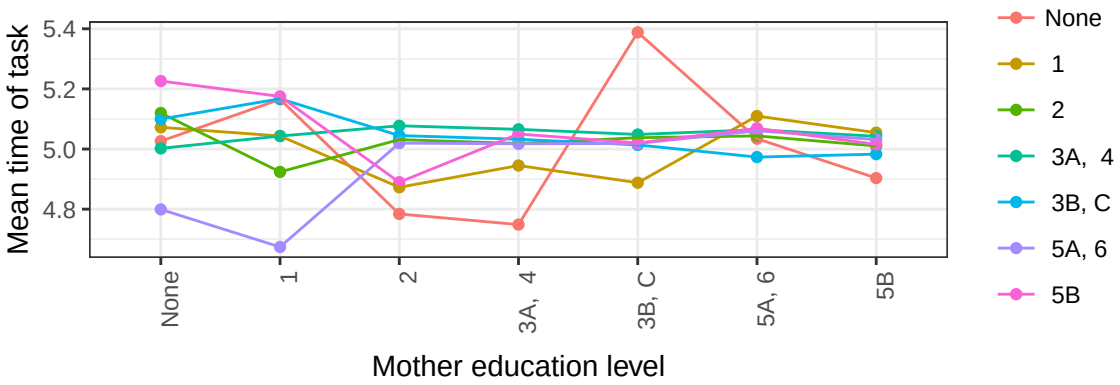
Actually, task 906 on positions 1 and 3, thus resolved in the first part of the test, is resolved longer than on positions 2 and 4. It is not for tasks 603, 905, 909 and 915.

The lack of parallelism between the curves for individual tasks indicates the presence of interaction.

Mother and father education level

We want to check if the high education of both mother and father has an effect on their child's behavior.Does it change when one parent has lower education, or both have a low education?

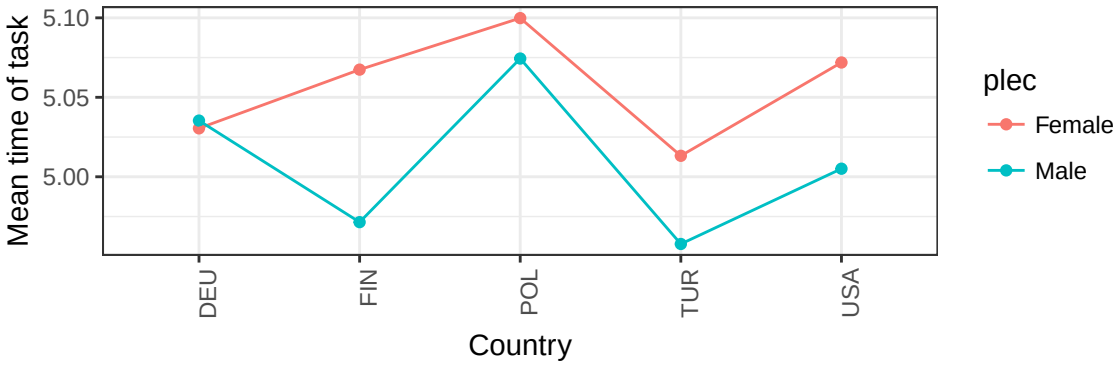
The level of education is from the lowest to the highest order and corresponds to ISCED levels.



The graph shows the existence of interaction. We notice the greatest interaction when at least one of the parents has a low education. We do not see, however, that the high education of one parent significantly influences the resolution time of the task (for higher education the lines have similar inclinations).

Country and gender

We also verified if exists a difference in the education of girls and boys depending on the country. Is it possible that there is a country where girls or boys have worse educational opportunities?



The plot above shows that the biggest differences in solving the problem between boys and girls are in Finland (and we thought it could happen in Turkey, which was partially confirmed). Based on the graph, we find the existence of interaction.

Nested variables

Also, we wanted to base our analyses on nested variables. In our opinion, a sensible choice was to use variables describing school number and the student identifier as a nested variable. Indeed, we know that a single student wrote test in one school, this dependence creates a hierarchical structure between school and student.

Random effects

Except for fixed effect, we add to our model some random effect. In our case random effects are:

- number of the questionnaire (`id_kwestionariusza`)
- identifier of the country (`id_kraju`)
- school identifier (`id_szkoly`)
- student identifier (`id_ucznia`)

It's obvious that the las two variables should be used as a random effect - we don't have data for every school or student. Similarly, for the country, we chose only five countries. In selected countries, also we don't have information for all questionnaires.

Final model

During our analysis, we created a dozen models. We chose Akaike and BIC for the optimal set of independent variables. The final model has the lowest value of the information criteria (BIC = 27158.1):

```
model <- lmer(log(czas_zadania +45)~zadanie*pozycja_zadania+(1|id_kwestionariusza)+mies_ur+
                (1|id_kraju)*plec+wyk_m_lsd*wyk_o_lsd+
                gr_zawod_m+gr_zawod_o+stat_m+stat_o+
                (1|id_szkoly/id_ucznia), data = dane_nowe, REML=F)
```