Exploring factors influencing excellent reading skills among Finnish teenagers: a logistic regression analysis using PISA 2018 data

Jamin Kiukkonen

October 2022

1 Description

The research question for this paper was to find out, using a logistic regression model, which background variables explain excellent reading skills of Finnish teenagers. The dataset used for this paper was taken from the PISA 2018 study. The variables of the dataset included reading skills, location of school, gender, language spoken at home, Sosio-economic status, reading for own pleasure, and use of information and communication technologies. The variables chosen for the logistic regression were reading skills (lukut), gender (sukup), reading for own pleasure (lukem), and sosioeconomic status (SES). Key graphical descriptions of the chosen variables are shown on figures 1 and 2.

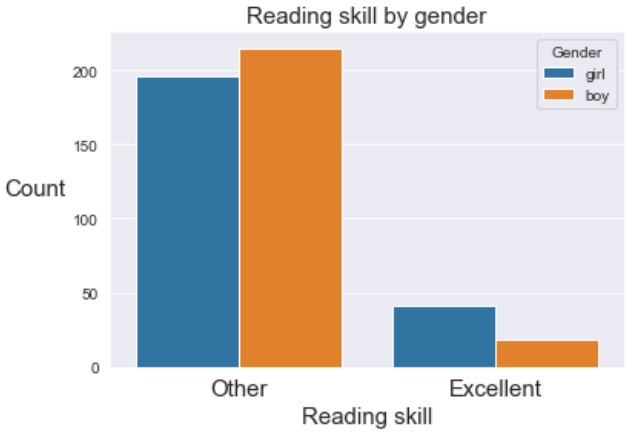

Figure 1: Reading skill by gender.

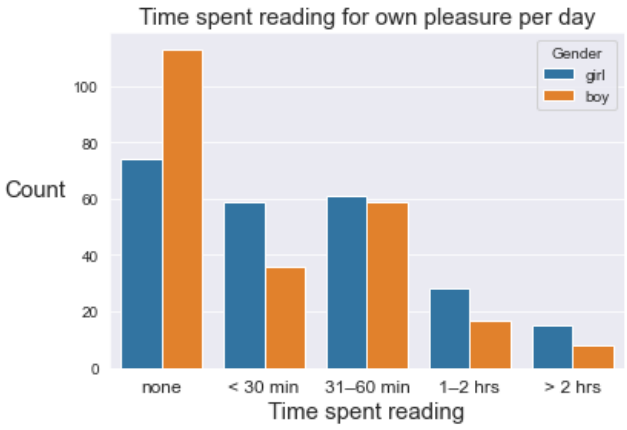

Figure 2: Gender-wise comparison of time spent reading in a day.

2 Methods

The model assumptions for the logistic regression were as follows; the response variable is binary, the errors are independent, the observations are independent, little or no multicollinearity between the explanatory variables, and linear relationship of explanatory variables and log odds. The variables chosen to answer the research question were chosen based on AIC value of different models. The lowest AIC was obtained when the location of school variable was also as a predictor in the model. However, when the location of school and gender variables were both as predictors, it yielded in a high variance inflation factor for both the location of school and the gender variables. The location of school variable was dropped due to it having lower correlation with the response variable. Now, when the explanatory variables were tested again for multicollinearity, the VIF value for all the variables were less than 3, which indicates that there is little multicollinearity between the explanatory variables. For the logistic regression, reading skill (lukut) was used as the response variable, and gender (sukup), reading for own pleasure (lukem), and sosio-economic status (SES), were used as the covariates. Interaction term for gender and sosio-economic status was included, since the goodness of fit for this model was better than for the model without the interaction term. The model equation was as follows:

$$logit(\pi_{lukut}) = \beta_0 + \beta_1\, sukup[girl] + \beta_2\, lukem[none] + \beta_3\, lukem[< 30\, min] + \beta_4\, lukem[31 - 60\, min] + \beta_5\, lukem[> 2\, hrs] + \beta_6\, SES + \beta_7\, sukup[girl]\, SES.$$

3 Results

The intercept term β_0 can be interpreted with odds, and the rest of the β coefficients can be interpreted with a ratio of odds. Odds, ratio of odds, and the corresponding confidence intervals are presented in table 1. The odds and ratio of odds can be acquired by exponentiating the regression coefficients.

	OR	Lower CI	Upper CI
Intercept	0.174	0.066	0.456
sukup[girl]	1.240	0.533	2.884
lukem[none]	0.117	0.037	0.375
lukem[< 30 min]	0.643	0.262	1.581
lukem[31-60 min]	0.625	0.250	0.156
lukem[> 2 hours]	1.642	0.499	5.403
SES	1.221	0.603	2.469
sukup[girl]:SES	2.776	2.285	3.266

Table 1: Odds and ratios of odds for the regression coefficients.

Exponentiating β_0 tells us that, for a male, who reads 1-2 hours per day for own pleasure, the odds for excellent reading skill is 0.174. Exponentiating β_1 tells us that the odds of excellent reading skill for a girl is 1.24-fold compared to a boy, when reading for own pleasure and sosio-economic status are adjusted. This means that the comparison is eligible only for boys and girls who are in the same category of reading for own pleasure and have the same sosio-economic status. Exponentiating β_2 tells us that, when comparing a student who doesn't read for own pleasure at all and a student who reads 1–2 hours per day, who are otherwise similar, then for the student who doesn't read for own pleasure the odds of having excellent reading skill is 0.117 times the odds of the student who reads 1–2 hours per day. Exponentiating β_3 tells us that when comparing a student who reads less than 30 minutes per day and a student who reads 1–2 hours per day, who are otherwise similar, then for the student who reads less than 30 minutes per day the odds of having excellent reading skill is 0.643 times the odds of the student who reads 1–2 hours per day. Exponentiating β_4 tells us that when comparing a student who reads 31–60 min per day and a student who reads 1–2 hours per day (reference category), who are otherwise similar, then for the student who reads 31–60 min per day the odds of having excellent reading skill is 0.625 times the odds of the student who reads 1–2 hours per day. Exponentiating β_5 tells us that when comparing a student who reads more than 2 hours per day and a student who reads 1–2 hours per day, who are otherwise similar, then for the student who reads more than 2 hours per day the odds of having excellent reading skill is 1.642 times the odds of the student who reads 1–2 hours per day. Exponentiating β_6 tells us that when comparing boys, who's SES value (sosio-economic status, standardized index) differ by 1, then for the group with the higher SES value, the odds for excellent reading skill is 1.221

times the odds of the group with the lower SES value. Note that the groups need to be otherwise similar for the comparison to be eligible. Lastly, exponentiating $\beta_6 + \beta_7$ tells us that when comparing girls, who's SES value (sosio-economic status, standardized index) differ by 1, then for the group with the higher SES value, the odds for excellent reading skill is 2.78 times the odds of the group with the lower SES value. As before, the compared groups need to be otherwise similar for the comparison to be eligible.

Residual analysis was conducted to asses the independency of the errors. Due to the binary nature of the response variable, the residuals will not be normally distributed and their distribution is unknow (Kutner, Nachtsheim, Neter, Li, 2004). The residuals assessed then are either the Pearson residuals, studentized Pearson residuals, and/or the deviance residuals. A plot that is helpful for diagnosing logistic regression model is to plot the studentized Pearson residuals, or the deviance residuals, against the estimated probability or linear predictor values with a Lowess smooth (Kutner, Nachtsheim, Neter, Li, 2004). Kutner, Nachtsheim, Neter, and Li (2004) show that under the assumption that the logistic regression model is correct, then the error (difference) between the observed value (Y_i) and the predicted value ($\hat{\pi}_i$) is equal to 0, i.e.

$$Y_i - \hat{\pi}_i = 0.$$

They conclude that this then suggests that a lowess smooth of one of the plots mentioned above would approximately be a horizontal line with zero intercept.

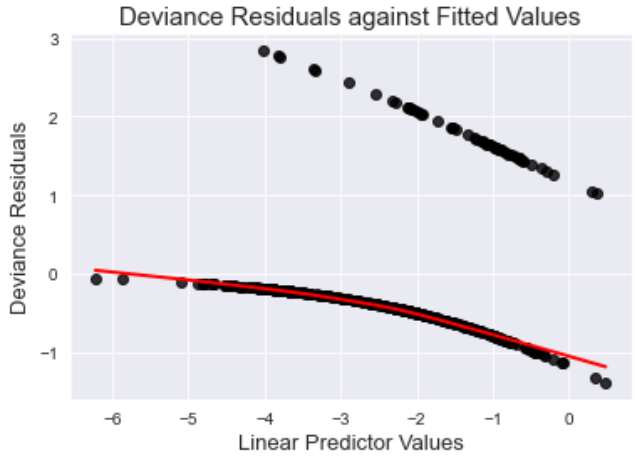

Figure 3: Deviance residuals against fitted values.

The lowess smooth is somewhat approximately on the zero intercept, but it does deviate from it when moving forward on the plot. Unfortunately, Kutner, Nachtsheim, Neter, and Li (2004) did not provide a suggestion of what "approximately" looks like.

4 Conclusions

Based on the ratio of odds, the model predicts that a girl will have a higher odds of having excellent reading skills compared to a boy with the same features otherwise. Also, the more one reads for own pleasure during leisure time, the higher is the odds of having excellent reading skills. Higher sosio-economic status resulted in an increase in the odds of having excellent reading skills.

In the future graphical description of the model's performance and the true observations needs to be evaluated more thoroughly. For this paper the theoretical knowledge for this task seemed to be insufficient.

5 References

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). Applied linear statistical models (5th ed.). New York, NY: McGraw-Hill Irwin.