explaining excellent reading skill of Finnish teenagers using select background variables

Sasu Ilmo

30. October 2022

1 Introduction

The research question of this paper is to find out what background variables can be used to explain excellent reading skill in Finnish teenagers. Data used for this paper is the PISA study from 2018 [PIS18]. The study doesn't include all variables in the PISA study, but a select few. These are: Excellent reading skill as a binary value of 1 or 0, size of the town the school is as a factor (small town 3000 - 15000 inhabitants, city 15000 - 100000 inhabitants and metropolis more than 100000 inhabitants), gender, language used at home, socio-economic status as a standardised index (SES), whether the student reads for their own pleasure as a factor(lukem), and use of ICT in free time as a standardised index. There are 470 students in the data, out of which 375 speak Finnish at home and 95 speak something else. I decided to turn the language into a binary since some of the languages had small sample sizes, in some cases one and were unusable. The new variable was 1 if the student spoke Finnish and 0 if something else.

2 Methods

I used a logistic regression model to find out which variables were significant. Location of the school and usage of ICT turned out to be non-significant, so they were dropped out of the model. The assumptions for the model are that the variables are independent, variables have a linear relationship with response variable, there are no extreme outliers and that the sample size is big enough. Mathematical formula for the model is:

$$logit(\pi_i) = \beta_0 + \beta_1 Finnish + \beta_2 genderMale + \beta_3 SES +$$

$$\beta_4 lukem2 + \beta_5 lukem3 + \beta_6 lukem4 + \beta_7 lukem5$$

The estimates, standard errors and P-values of the model

Coefficients	Estimate	Std. Error	P-value
Intercept	-4.9256	0.7781	2.45e-10
Finnish	1.6132	0.6256	0.009918
genderMale	-0.7895	0.3320	0.017404
SES	0.7226	0.2388	0.002475
lukem2	1.7885	0.5297	0.000735
lukem3	1.6855	0.5411	0.001839
lukem4	2.4023	0.5901	4.69e-05
lukem5	2.3446	0.6881	0.000656

3 Results

The odds, odds ratios and their confidence intervals

Coefficients	Estimate(OR)	conf. interval 2.5%	conf. interval 97.5%
Intercept	0.007258021	0.001293075	0.02867279
Finnish	5.018894	1.709025323	21.56166167
genderMale	0.4540681	0.231572370	0.85715160
SES	2.059703	1.316841433	3.36959263
lukem2	5.980468	2.260677866	18.81398639
lukem3	5.395321	1.984039622	17.26501352
lukem4	11.0489	3.625024616	38.21846089
lukem5	10.42875	2.700190853	42.14205249

The intercept in this case represents the odds of the reference group, that is a girl from a small town who speaks Finnish, doesn't read for fun and has an SES score of zero having excellent reading skills versus not having excellent reading skills is 0.007258021. If the teenager speaks Finnish and everything else is equal than his odds of having excellent reading skills is 5.018894 times

the odds of those who speak other languages. When comparing men and women if everything is same then the odds of a man having excellent reading skills is 0.4540681 times the odds of women. When we compare individuals whose SES score differ by one but are equal in other ways, for those with higher score odds of having excellent reading skills is 2.059703 times the odds of the lower one. When we compare those who don't read at all for fun to those who read less than 30 min per day but are otherwise similar, for those who read for less than 30 minutes the odds of having excellent reading skills is 5.980468 times the odds of those who don't read. For those who read 31 - 60 min per day the odds are 5.395321 times the odds, for those who read 1 - 2 hours per day the odds are 11.0489 times and for those who read more than 2 hours per day the odds are 10.42875 times higher.

The residuals in figure 1 seem good, there appears to be no patterns and no outliers.

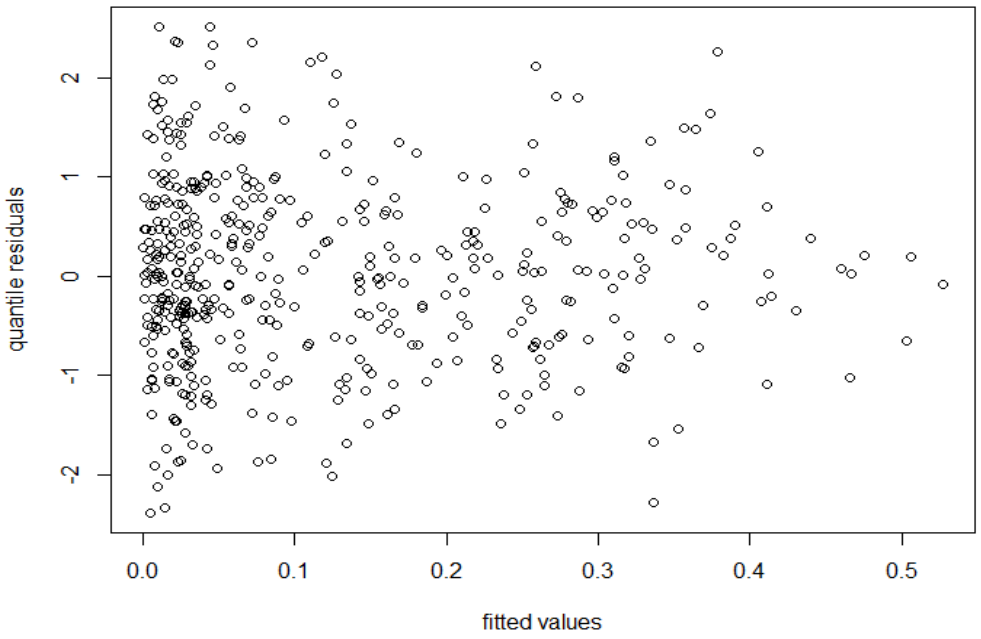

Figure 1: randomized quantile residuals of the model

The plot in figure 2 appears linear so I would say that the model is good.

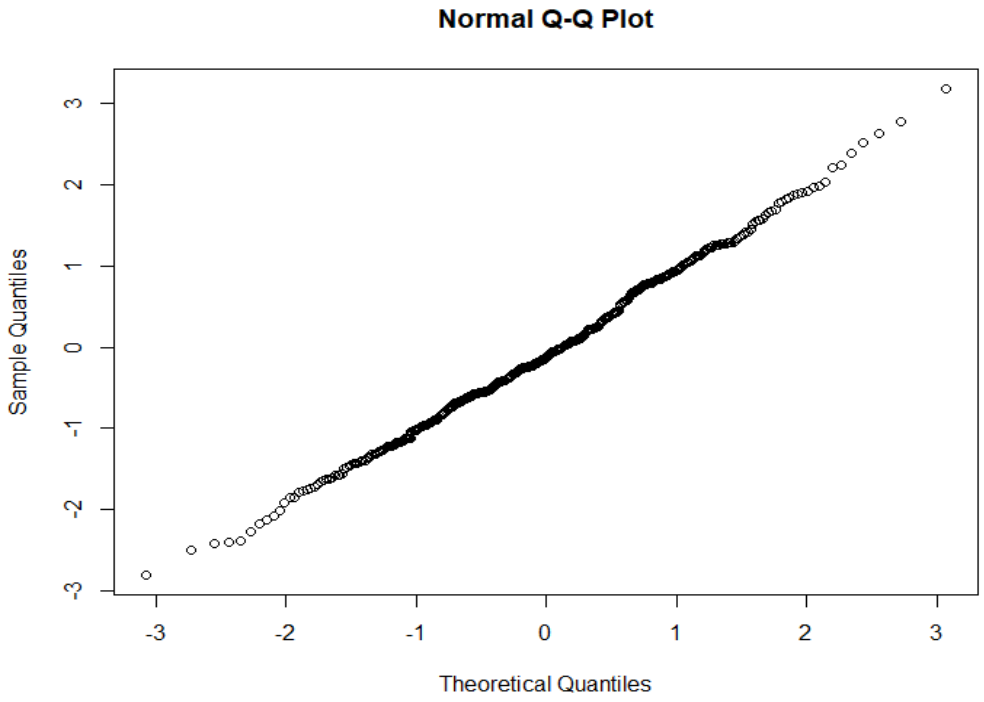

Figure 2: Q-Q plot of the residuals

4 Conclusions

I would say that the model is that I constructed in this paper is good. Acording to the model excellent reading skill can be explained with SES, gender, howmuch person read for fun and if the person speaks Finnish at home. The sample size of those who read more than 2 hours was small so maybe it would have been smart to combine the last 2 groups together since they seemed very similar. Some of the findings at least to me seem a bit obvious. If you read a lot your reading is better than those who do not and those who speak Finnish can read Finnish better than those who speak something else. The gap between gender seems very large though. Perhaps there might have been some kind of interactions that I missed and that could be explored in the future.

References

[PIS18] PISA. *PISA 2018 Database*. 2018. URL: https://www.oecd.org/pisa/data/2018database/.