Reading Week Project

GROUP 8

2022/3/3

Introduction

In this report, we will use data from a large-scale survey conducted for the Longitudinal Study of Young People in England (LSYPE), also known "Next Steps", which ran from 2004 until 2010. The survey data follows a sample of young people, which began when they were in year 9, collecting information about their academic experiences, socio-economic and demographic factors. For analysis, we use a subset of 13459 young people taken from LYSPE between 2005 and 2006 which data is linked to a National Pupil Database that tracks the examination results of all Students in the UK. The merging of the data allows us to focus on the study of various factors that might affect students' exam performance.

For the purpose of this study, a subset of 24 variables is taken from the originally merged data which contains about 114 variables. However, In this report we only focus on the important relationships between GCSE total score (ks4scores) and 9 explatonary variables fitted in a regression model, 5 A*-C GCSE grades (fiveac), mathematics proficiency level at the age of 14 (k3ma), the gender of the pupil (gender), their desire to further tertiary education (pupasp), their attitude towards school measured by school score band (atitude), their school exclusion status (exclude), their need for special education (sen), their mother highest qualification (hiquamum) and their parents' house ownership status (house). Exclude, pupasp and sen takes a binary value of yes and no.

We found that the regression model indicates the higher the math proficiency level of a pupil, the higher the GCSE total score. Also if a pupil is a female, interested in pursuing tertiary education, living in a non-rented house, achieved at least 5 A*-C grades in GCSE, has high attitude towards school, does not need special education and never being permananetlly or temporarily excluded from school tends to achieve higher GCSE total score on average relative to their counterparts. Pupils with mother holding a degree or equivalents are also on average score higher in GCSE relatives to those without.

Explanatory Analysis

All the explanatory variables are of a categorical type except for the deprivation index (idacn). Therefore, boxplots of all the categorical variables are produced to understand the distribution of the data. From the boxplots, we can see that most of the explanatory variables have at least a slight difference in the height of level boxplots suggesting a possible difference in ks4score on average. Moreover, 5 explanatory variables namely fiveem, fiveac, k3en, k3sc, k3ma shows a notable variability of boxplot height indicating a strong possible predictor to ks4score. However, they are also the ones with a serious number of outliers. Given that the median line is roughly in the middle across all levels for all categorical variables, we can interpret that the ks4score distribution for each level is symmetrical.

We also notice some identical trends from the boxplots across different explanatory variables suggesting a potential similar relationship to ks4score. For example, k3en, k3ma and k3sc showcase upward trends as the tier increases. Meanwhile, fiveem and fiveac boxplot patterns are almost identical to one another. The same case is also seen between absent, truancy and exclude as well as singlepar and house. This same trend might suggest that some explanatory variables are probably dependent on one another.

Merging leves of some categorical variables is also done here. We merged some levels of a few explanatory variables based on similar characteristics such as approximately equal median. For example, two levels, Inter-

mediate, Routine,_semi-routine_or_unemployed from secshort are combined as one level because they both have the same medians. The same concept also applies to the merging between GCE_A_Level_or_equivalent and HE_below_degree_level levels from hiquamum as one level. Meanwhile, homework, attitude and fsmband levels are all separated into two categories, low and high, respectively to reduce the dimension of the levels of the categorical variables for simpler model interpretation.

We decided to reduce the data to a complete set of data by removing pupils with missing values from statistically significant explanatory variables after backward elimination is run to avoid losing too much information. By doing this, we removed roughly 28%, which is a decent percentage, of the total given dataset.

Full model to Final Model

We ran the Full Model which included all 23 available variables. Goodness of fit statistics for this model are shown in Table 1 below. Coefficients with asterisks are significant at the 5% level. Variables except new_secshort, fsm, parasp and tuition were significant at the 5% level. The R2 and Adj R2 were just above 0.7. They were also within 1% of one another. The F-statistic was <2.2e-16 and the Standard Error of the model was 79.82. The residual plots of this regression indicated no kurtosis or heteroscedasticity. As new_secshort, fsm, parasp and tuition were non-significant, they were removed in the sequence of tuition, new_secshort, fsm and parasp based on backward elimination. This model was to be Model 1. The values for Model 1 were pretty close to those of Full Model.

	Full Model	Model 1	Model 2	Model 3	Model 4	Final Model
(Intercept)	*	*	*	*	*	*
factor(fiveac)yes	*	*	*	*	*	*
factor(fiveem)yes	*	*	*	*	NA	NA
k3en	*	*	*	*	NA	NA
k3ma	*	*	*	*	*	*
k3sc	*	*	*	*	NA	NA
factor(gender)Male	*	*	*	*	*	*
factor(new_secshort) professional		NA	NA	NA	NA	NA
factor(new_secshort) missing		NA	NA	NA	NA	NA
factor(new_hiquamum) Pre-university-equivalent	*	*	*	*	*	*
factor(new_hiquamum) GCSE_grades_A-C_or_equi	*	*	*	*	*	*
factor(new_hiquamum) missing			NA	NA	NA	NA
factor(new_hiquamum) No_qualification					*	*
factor(new_hiquamum) Other_qualifications	*	*			*	*
factor(singlepar)missing	*	*	NA	NA	NA	NA
factor(singlepar)no	*	*	*	*	NA	NA
factor(house)other/DK/Ref			NA	NA	NA	NA
factor(house)owned	*	*	*	*	*	*
factor(fsm)missing		NA	NA	NA	NA	NA
factor(fsm)no		NA	NA	NA	NA	NA
factor(parasp)missing		NA	NA	NA	NA	NA
factor(parasp)Yes		NA	NA	NA	NA	NA

	Full Model	Model 1	Model 2	Model 3	Model 4	Final Model
factor(computer)missing			NA	NA	NA	NA
factor(computer)Yes	*	*	*	*	NA	NA
factor(tuition)missing		NA	NA	NA	NA	NA
factor(tuition)Yes		NA	NA	NA	NA	NA
factor(pupasp)Yes	*	*	*	*	*	*
factor(new_homework)high	*	*	*	*	NA	NA
factor(new_homework)missing	*	*	NA	NA	NA	NA
factor(new_attitude)low	*	*	*	*	*	*
factor(new_attitude)missing	*	*	NA	NA	NA	NA
factor(sen)missing			NA	NA	NA	NA
factor(sen)No	*	*	*	*	*	*
factor(truancy)missing	*	*	NA	NA	NA	NA
factor(truancy)No	*	*	*	*	NA	NA
factor(absent)missing			NA	NA	NA	NA
factor(absent)No	*	*		NA	NA	NA
factor(exclude)missing	*	*	NA	NA	NA	NA
factor(exclude)No	*	*	*	*	*	*
IDACI_n	*	*		NA	NA	NA
factor(new_FSMband)high	*	*	*	*		NA
R2/Adj R2	0.7076	0.7075	0.6835	0.6832	0.6574	0.6572
	0.7067	0.7068	0.6827	0.6825	0.6569	0.6568
F-statistic significance	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16
Sd Error	79.82	79.81	78.69	78.71	81.82	81.83

Table 1

We removed the missing value of remaining predictors after backward elimination in order not to lose too much information. With the new set of data, we did the same regression and the result was shown as Model 2. A decline of 0.02 in R2 and Adj R2 led the value below 0.7. Also, the standard error dropped from 79.81 to 78.69. IDACI_n and absent became non-significant at the 5% level, so they were removed from the regression. Then the model was to be Model 3. The Adj R2 was 0.6832 and was close to the R2. Standard error increased to 78.71.

k3ma, k3en and k3sc had strong correlation. We used the regression of Model 3 three times, but only with one of the three k3s each time. With k3ma, highest R2 and lowest standard error were presented. Hence, K3en and K3sc was omitted. For the variance inflation factor of Model 3, multicollinearity was shown between fiveav and fiveem. We ran multiple linear regression (MLR) model combining k3ma with fiveav and fiveem respectively. As fiveac had higher R2, fiveem was eliminated. Truancy and exclude both represented the performance of a student at school by definition, so they were compared. Exclude was kept, as it got higher R2 when doing MLR model combining with k3ma. For homework and pupasp, which were both considered to reflect students' attitude towards the exam, comparison was made. Homework was dropped, as it had lower R2 in the MLR model combining with k3ma. Finally, as singlepar, house and computer all showed the financial circumstance of a student's family, we ran MLR model on them with k3ma respectively. Only house was taken, as it had the highest R2.

With the remaining predictors, we ran the regression and got Model 4. It showed that new_FSMband was non-significant at the 5% level, so it was eliminated. Then it turned out to be the Final Model. From Table 1, it can be seen that the R2 and Adj R2 were 0.6572 and 0.6568 so still satisfactory. The F-statistic was <2.2e-16 and standard error was 81.83, both indicating good fit of the model. The residual plots of this regression had no kurtosis or heteroscedasticity showing the model was fit.

Results

	Full Model	P-values	Model1	P-values	Model2	P-values	Final Model3	P-values
(Intercept)	-94.3041	< 2e-16	-93.6356	< 2e-16	-88.105	< 2e-16	23.5744	0.001112
factor(fiveac)yes	116.9768	< 2e-16	117.0093	< 2e-16	110.063	< 2e-16	133.6160	< 2e-16
factor(fiveem)yes	6.5121	0.006054	6.6621	0.004925	9.249	0.000949	NA	NA
k3en	21.1644	< 2e-16	21.2625	< 2e-16	21.144	< 2e-16	NA	NA
k3ma	19.1377	< 2e-16	19.1611	< 2e-16	19.698	< 2e-16	39.1814	< 2e-16
k3sc	15.8917	< 2e-16	15.9203	< 2e-16	16.767	< 2e-16	NA	NA
factor(gender)Male	-10.0738	9.87e-12	-10.2354	4.04e-12	-8.370	1.54e-06	-14.8412	< 2e-16
factor(new_secshort) professional	1.3169	0.461540	NA	NA	NA	NA	NA	NA
factor(new_secshort) missing	0.6952	0.731275	NA	NA	NA	NA	NA	NA
factor(new_hiquamum) Pre-university-equivalent	-7.1689	0.009582	-7.4621	0.006496	-6.172	0.035373	-11.8472	9.72e-05
factor(new_hiquamum) GCSE_grades_A-C_or_equiv	-9.4991	0.000673	-10.0325	0.000224	-8.386	0.004081	-16.4612	4.73e-08
factor(new_hiquamum) No_qualification	-4.7713	0.123278	-5.1859	0.082554	-2.680	0.418132	-12.9941	0.000106
factor(new_hiquamum) Other_qualifications	-7.3806	0.034262	-7.9738	0.019211	-6.760	0.072537	-17.0753	1.09e-05
factor(new_hiquamum) missing	-3.3313	0.410543	-3.6142	0.365659	NA	NA	NA	NA
factor(singlepar)no	10.4401	8.90e-09	10.2241	9.78e-09	7.692	0.000525	NA	NA
factor(singlepar) missing	19.4714	0.045440	19.5670	0.041758	NA	NA	NA	NA
factor(house)owned	7.1409	7.1409	6.8612	0.000157	7.106	0.001067	13.0732	3.42e-10
factor(house) other/DK/Re	-4.2160	0.504808	-4.4726	0.478021	NA	NA	NA	NA
factor(fsm)no	-1.3717	0.539262	NA	NA	NA	NA	NA	NA
factor(fsm)missing	8.1955	0.676863	NA	NA	NA	NA	NA	NA
factor(parasp)Yes	2.8522	0.175790	NA	NA	NA	NA	NA	NA
factor(parasp)missing	5.7149	0.776158	NA	NA	NA	NA	NA	NA
factor(computer)Yes	12.5529	2.72e-07	12.4869	2.67e-07	17.603	5.20e-08	NA	NA
factor(computer) missing	-15.8189	0.535421	-7.4865	0.266889	NA	NA	NA	NA
factor(tuition)Yes	-1.1029	0.606366	NA	NA	NA	NA	NA	NA
factor(tuition)missing	5.5237	0.816178	NA	NA	NA	NA	NA	NA
factor(pupasp)Yes	11.1128	1.08e-07	11.9715	2.24e-09	13.242	8.84e-08	18.6440	3.14e-13
factor(new_homework)	5.7316	0.000286	5.7381	0.000277	6.918	0.000144	NA	NA

	Full Model	P-values	Model1	P-values	Model2	P-values	Final Model3	P-values
high factor(new_homework)	-10.5633	0.000354	-10.6354	0.000321	NA	NA	NA	NA
missing factor(new_attitude)	-12.1553	1.17e-15	-12.2411	6.77e-16	-12.074	1.92e-12	-16.0001	< 2e-16
low factor(new_attitude)	-15.8385	3.18e-08	-15.8639	2.96e-08	NA	NA	NA	NA
missing factor(sen)No	26.8364	< 2e-16	26.7913	< 2e-16	24.601	2.65e-08	33.1023	4.67e-13
factor(sen)missing	NA	NA	9.3386	0.634069	NA	NA	NA	NA
factor(truancy)No	20.7963	< 2e-16	20.8030	< 2e-16	16.274	1.72e-09	NA	NA
factor(truancy)missing	12.8617	0.000254	12.9411	0.000229	NA	NA	NA	NA
factor(absent)No	11.5325	0.006056	11.6923	0.005353	NA	NA	NA	NA
factor(absent)missing	8.0021	0.328106	8.0767	0.323371	NA	NA	NA	NA
factor(exclude)No	34.9954	< 2e-16	35.0832	< 2e-16	34.585	< 2e-16	41.6812	< 2e-16
factor(exclude)missing	41.4191	1.72e-07	42.0350	1.04e-07	NA	NA	NA	NA
IDACI_n	2.3151	0.012490	2.3651	0.009953	NA	NA	NA	NA
factor(new_FSMband) high	7.5618	1.13e-05	7.7391	6.22e-06	9.677	8.92e-08	NA	NA

From the table, both the factor(fiveac) and k3ma are highly significant in all four models. We could expect that the future academic achievement is highly correlated with students' previous academic achievement in mathematics and achievement of 5 or more A* - C grades at GCSE. In the final model, it shows that students who achieve 5 or more A* - C grades at GCSE perform better than those without this achievement at around 134 points more. Also an increase in one in the ks3 score of mathematics leads to an expected increase of approximately 33 points in the ks4score. When other factors remain at the same level, male can expect to have on average 15 points less than their female counterparts, which means male pupils normally perform less well than female pupils. The higher mothers educational qualification, the higher ks4score their children got. This is especially true for those pupils' mothers with no qualification as they scored almost 13 points less than their mothers with degree or equivalent and 2 points less than those with pre-university-equivalent. If pupils live at a owned house, they could achieve 13 points higher than pupils at a rented house. Furthermore, Students want to continue in FTE after age 16 push them to achieve higher grades commonly at around 19 points more. The attitude for students to school score band also affect the ks4score. Obviously, the lower the attitude, the lower the score. As shown above, students with a low attitude are expected to have less than 16 points compared with students with a high attitude. Most of the time, students requiring school to seek external advice from appropriate support services always perform worse as School Action has not been able to help the child make adequate progress. Based that, relative to students required additional help, those not needed sen scored 33 points more.In comparison to pupils with one or more exclusions from school, pupils whose age from 11 to 14 without any exclusions are expected to gain a better grade, of which the ks4score could be 42 points more.

Comments about the data/analysis

Conclusions

In this analysis, we attempted to understand how the standardized score of pupils aged 16 (KS4 score) in England are explained by the following three variables: the GCSE grades, the gender of the students, and

whether or not a student has been excluded from school for more than once.

We had curated data on a sample of approximately 13,500 students who had taken part in the Next Steps Study and the National Pupil database (NPD) between the years 2005 and 2006. We ran a regression analysis to determine what these relationships were.

Based on the results of the regression analysis and diagnostics statistics, it is demonstrated that the GCSE performance of the students is the most important predictor of the KS4 score. On average, students who have obtained 5 or more GCSE grades (ranging from A* to C) score about 134 more points than the students who haven't achieved that. The grades on similarly designed standardized assessments (GCSE) is the most direct indicator of the assessment score being benchmarked, which is the KS4 score. Moreover, GCSE grades is an all-encompassing measure which includes various direct and indirect factors that ultimately contribute to standardized test results. Therefore, it makes sense that it is the most effective predictor of KS4 score.

The other factors that have also been influential factors in the final KS4 scores are gender of a student and school exclusion track record of the student. Both factors are categorized as binary outcomes: Male/Female for gender and Yes/No for frequent school exclusion. It can be observed that a male student on average scores 15 points lower than their female counterparts in their KS4 score. This is reflected in the box plot below, where the average KS4 score of female students is higher than that of male students.

Due to the large sample size (N = 13,500) of the dataset, it is safe to assume that reliable observations and conclusions can be drawn. We think several recommended solutions could be considered by government representatives. First of all, the deep-rooted perception of male students enthusiastic and passionate about learning/studying as 'uncool nerds' must be eliminated. This gender stereotype is an obstacle which prevents male students from fully engaging with the course materials, which naturally spills over to academic performance.

Besides, we would recommend the government to pay more attention to students who got excluded from school multiple times to reduce the numbers. In our sample at least 1000 students have been excluded from school more than once. While this population is just 7.5% of the student sample, it is clear that a concerning number of pupils do not value going to school. In general, exclusion from school can stem from multiple factors such as lackluster teaching process, family issues, etc. The root cause of these school exclusions should be taken care of in order to boost student interest in academics and thus boost test performance.

Last but not least, additional subsidized schemes should be offered to students so that they can familiarize themselves with the questions and formats of standardized tests. These schemes can include intensive class near exam periods, optional extra practices, and past paper practices with model answers, and many more.