ST211 Group

2024-05-09

Candidate numbers:

30468

34144

xxxxx

xxxxx

Report

Introduction

In this report we investigate the main predictors of a young person's level of debt and income. In addition, we also assess the impact that the mother has on their child's future debt and the effect of education on a person's income. The data set we used is the Next Steps (NS) which is a study of ~16,000 people in England born between 1989 and 1990. It began in 2004 (cohort at age 14), with the original sample size of 15,770 people. The cohort was surveyed annually until 2010, and once more in 2015-16. This data contains many potential predictors that cover the respondents' education, mental health and home environment as they grew up.

We found that the father's highest qualification, whether they went to an independent school, their attitude towards debt, their mental health now, if they've experienced ethnic discrimination and their current tenure were the main predictors of debt. On the other hand, we found that a variety of predictors were significant in predicting income such as the family's socio-economic background (ethnic group, single-parent household), the young person's school experience and their current qualifications and work experience.

As for our additional research questions, we found that a mother's situation and qualifications have an insignificant effect on their child's debt, whereas education had an effect on income but not a substantial one.

Exploratory analysis

Generally accepted economic theories have confirmed the positive impact of education on labour markets, and specifically on income (Card, 1999); due to knowledge and skills in higher education institutions. The educational predictors we used also include educational variables relating also to pre-university education (e.g. no. of GCSE's, A-Levels). Our approach involved the comparison of 3 models: a generalised model, one including educational predictors, and one without any educational predictors.

After removing predictors with 10% or more missing values or `NA` from the dataset, we ran an initial linear regression model including all variables. This indicated that the following variables could be significant predictors of income:

- Variables related to the highest qualification of their parents: `W1hiqualmum` & `W1hiqualdad`
- Variables which indicate the economic status of their family: `W1condur5MP`, `W1usevcHH`
- Socio-demographic factors (`W1ethrgrYP`, `W1famtyp2`, `W1disabYP`, `W2disc1YP`)
- Socio-economic factors: `W1nssecfam` (National statistics socio-economic classification), `W1hous12HH`

Out of the seemingly significant predictors listed above, we expect that variables relating to higher education attainment (incl. those of the parents), the type of school they attended, and the number of GCSE/A-Levels they undertook could be crucial, i.e. `W5EducYP`, `W6gcse`, `W6UnivYP`, `W5Apprent1YP`, `W1hiqualmum`. Noticeably, `IndSchool` (a categorical variable indicating attendance to an Independent School), did not appear significant and was disregarded from the following models. This hinted at educational predictors overall not being wholly important, and that interactions could be present.

The initial analysis of educational predictors indicated that `W6UnivYP` is correlated with `W5EducYP`, `W6gcse` and `W1heposs9YP.`

We also felt that certain levels of categorical predictors did not demonstrate enough variance between themselves to justify them being separate, and merged their levels accordingly. These were `W1wrk1aMP`, `W1empsmum` and `W8DACTIVITY` (see Appendix A).

Simple outlier analysis confirmed that there were not enough outliers amongst predictors in all 3 models to warrant any removal of the data points. By plotting each of the predictors against the outcome variable, we also did not see any need for transformations.

Plotting the predictors against the outcome variable and the linear regression models we tested indicated that the following predictors affected debt:

- Racism and discrimination
- The type of school respondent went to
- The GHQ-12 score
- Tenure of respondent

We merged the highest qualification and the employment status of the mum and dad by taking the lower value (i.e. the higher education qualification) from the two predictors and storing it in a new variable (i.e. `W1hiqualMP` and `W1empsMP`). This was because there was a lot of missing data in the variables to do with the father, so we created new variables to look at the main parent, which takes both single and double parent households into account.

The tenure of the respondent was found to be the most significant predictor of debt but doesn't offer much insight, because house ownership is directly related to debt through mortgages.

We also wanted to research the relationship between the parents' socio-economic status with the debt of the respondent, because of the intergenerational transmission of wealth and debt. We hypothesised that a respondent's parents would be very influential in their management and inheritance of wealth and debt, and that the respondent may also bear the burdens of their parents' debts, or inherit parents' approaches to debt management.

From initial to final model

To approach these questions, we made various models to assess the importance of the kinds of variables that we are interested in.

For example, to assess the importance of education-related predictors in determining the income outcome, we made these 3 models:

1. A general model that was made by firstly running a regression using all the predictors and then removing the predictors that were not significant at the 5% level. Thus, all the predictors that remain in this model are assumed to be important in determining the outcome variable which is income. The performance of this model will be used as a baseline for our assessments.
2. A model that has the same predictors as Model 1 but including the rest of the education-related predictors that we are interested in investigating.
3. A model that is similar to Model 1 but without any of the education-related predictors

The idea behind this approach is to compare the performances of these models by looking at the adjusted R-squared value to assess how important education-related predictors as a whole are in determining income. If the models do not differ much in their performance, then we can conclude at least that not all these predictors are important.

The education-related predictors that we have decided to investigate are:

Variable name	Meaning

W1hiqualmum	Mum's highest educational qualifications
IndSchool	Whether young person went to an independent school
W1heposs9YP	Likelihood of young person going to university
W1yschat1	Young person's attitude to school at wave 1
W4schatYP	Young person's attitude to school at wave 4
W5EducYP	Whether young person is going to school or college at wave 5
W6UnivYP	Whether young person is going to university at wave 6
W6acqno	Highest academic qualification at wave 6
W6gcse	Number of GCSEs studied at wave 6
W6als	Number of A Levels studied at wave 6

We approached the question on the debt outcome similarly with mother-related predictors. However, since mother-related predictors did not feature in the baseline model, we only used 2 models here. The mother-related predictors that we have decided to investigate are:

Variable Name	Meaning
W1wrkfullmum	Whether mother works full or part-time
W1empsmum	Employment status of mother
W1hiqualmum	Mum's highest educational qualifications
W1marstatmum	Marital status of young person's mother

Results

The baseline model for predicting income included these predictors:

Predictor	Estimate
(intercept)	320.93***
W1wrk1aMP	-3.84***
W1condur5MP	-9.46**
W1hea2MP	4.72**
W1hous12HH	1.92**
W1usevcHH	-7.83**
W1hiqualmum	-0.68***
W1wrkfullmum	13.54***
W1empsmum	-7.04***
W1marstatmum	-1.89*
W1famtyp2	-35.93***
W1nssecfam	-8.18***
W1ethgrpYP	-14.34***
W1heposs9YP	-1.96*
W1hwndayYP	1.60**
W1disabYP	12.23***
W2ghq12scr	1.76***
W2disc1YP	17.42***
W4CannTryYP	-14.06***
W4RacismYP	11.63***
W4schatYP	0.57**
W5JobYP	-13.25***
W5EducYP	-7.05***
W5Apprent1YP	20.34***
W6JobYP	-6.06***

W6UnivYP	-13.35***
W6gcse	6.10*

For the question on income, the residual plots for these 3 models do not differ significantly. Thus we use certain statistics as performance diagnostics for these models. The comparison table of these statistics are as follows:

Diagnostic	**Model 1 (baseline)**	**Model 2 (with all education-related predictors)**	**Model 3 (without any education-related predictors)**
Adjusted R-squared	0.6309	0.6313	0.6157
Residual standard error	43.35	43.29	44.24

Based on this table, model 2 is the best performing model but the gain in performance over model 1 is minimal. Model 2 has 4 predictors that were not included in model 1:

- `IndSchool`
- `W1yschat1`
- `W6acqno`
- `W6als`

None of these new predictors were significant at the 5% level in model 2. Thus we can conclude that these 4 education-related predictors are not important in determining income.

The diagnostics also tell us that model 3 performs slightly worse than model 1. The predictors that were removed from model 1 to make model 3 were:

- `W1hiqualmum`
- `W1heposs9YP`
- `W4schatYP`
- `W5EducYP`
- `W6UnivYP`
- `W6gcse`

These predictors are significant at the 5% level which indicate that they are useful in determining income. However, since the model still performs relatively well without these predictors, we would not say that these education-related predictors are important in determining income. In other words, even without taking education into consideration, we can still make a model that performs decently in predicting the young person's income. This can be attributed to the fact that the baseline model includes many other significant predictors.

The baseline model for predicting debt included these predictors:

Predictor	**Estimate**

(Intercept)	27,570.6***
`W1hiqualdad`	-309.3**
`IndSchool`	11,768.9***
`W6DebtattYP`	513.9*
`W8DGHQSC`	-608.8**
`W8TENURE`	-1,912.9***
`W2disc1YP`	-5,495.7*.

For the question on debt, the residual plots for these 2 models do not differ significantly. Thus we use certain statistics as performance diagnostics for these models. The comparison table of these statistics are as follows:

Diagnostic	**Model 1 (baseline)**	**Model 2 (including mother-related predictors)**
Adjusted R-squared	0.03183	0.03045
Residual standard error	31,130 (1972 dof)	31,070 (1991 dof)

Based on this table, Model 1 is the best performing model but by an insubstantial margin with it having a 0.00138 greater adjusted R-squared and a 60 lower residual standard error.

The only somewhat significant factor when the mother-related predictors were included was `W1wrkfullmum` with a p-value of 0.06694 making it significant at the 10% level.

The predictors for the debt model seem unrelated to the mother and more dependent on other factors such as mental health (which may be made worse due to high levels of debt), attitude towards debt and their household's situation. As the model works better without the mother-related predictors, the mother appears to have very little significant impact on the level of debt that her child will accrue debt.

Comments about the data/analysis

The models for the debt dataset show most predictors have high standard errors, suggesting imprecise and unstable estimates. Moreover, very few predictors are statistically significant and the p-values and adjusted R-squared values are also very small, indicating the model explains only a small proportion of the variance in the dependent variable. This suggests our model is not well-fitted to the data, and wouldn't be used to generalise on the population.

There are categories of missing values which allows us to investigate why some variables have a lot of missing values. For example, in the predictors that are to do with the father, a lot of the missing data stems from the father not being present. This suggests that values in key predictors are not randomly missing and are reasonably explained. We attempted to overcome this by combining mother and father variables to look at the main parent, as aforementioned in the exploratory analysis.

We believe that our analysis may have suffered from large amounts of missing data from important predictors, such as household gross annual salary. Those predictors could've been significant, and would have allowed us to build better fitted models.

In the debt analysis, some predictors we didn't expect to be very influential were found to be more significant than predictors we expected to be. For example, the racism and discrimination predictors were more significant than the predictors related to education and the parents. This does not suggest the results are counter-intuitive, because it may highlight structural inequalities and discrimination in the UK and how it impacts debt, even if we may not see the relationship directly.

We found that GHQ-12 score had a negative correlation with debt, which seems counter-intuitive. We expected those with more mental health issues (higher GHQ-12 score) would have more debt. However, the GHQ-12 score is a significant and relatively stable predictor. It may suggest young people with mental health issues don’t request or secure debt.

Respondents may have lied or outright refused, in some of the questions they were asked, such as the drugs, alcohol and mental health predictors, because those questions are of a sensitive and personal topic. This results in both response and non-response bias in the data.

We liked to have data on the geography of the respondents. In the UK, aspects of inequality distribution are geographic, where a North-South divide exists. Thus, we would’ve liked to infer the impact this had on income and debt, as this would’ve made the data more holistic, which potentially could give us a better fitted model.

Interpretations and conclusions for a lay audience

Interpretation for a lay audience:

We have found that education is not the sole factor in determining a young person's future income. The most important education-related aspects of a young person's upbringing that could determine their future income generally relate to their attitude to school and their qualifications at higher education. However, there are many other aspects of a young person's upbringing that should not be overlooked, such as their mental health and other systemic factors like their socio-economic status and ethnicity.

We also found that very few of our relatively holistic factors can accurately predict the amount of debt a respondent takes on. We expected parental, socio-economic and educational factors to have a greater impact on the debt a respondent owes. However, we did find a strong relationship regarding the type of school a respondent went to, where it suggests that a respondent who went to an independent school took on far more debt than one who went to a maintained school.

Respondents who reported discrimination had higher levels of debt, which may suggest structural racial inequalities in the financial market.

We have some evidence to suggest that mental health has a relationship with a respondents' debt, through the GHQ-12 score. Our model suggests that respondents with more mental health issues take on less debt.

Is the model good enough for decision-making:

Our most accurate model is able to account for 63% of the variation in income: and we do not believe this is accurate enough for predicting someone's income. Despite this, the threshold at which we deem the model good enough for decision-making depends largely on the application of the model. We would not recommend using this model for critical decision-making, which could affect someone's livelihood.

The model for debt is relatively unstable and imprecise with the predictors we have; they are not effective in predicting the debt someone takes on. However, the model does show strong relationships from certain predictors, which still makes it of some use in decision making.

Recommendations to a government official:

Our results demonstrate the importance of socio-economic and socio-demographic factors in determining people's income, as opposed to solely educational factors. The government concerns itself with maximising the welfare of its citizens, and should seek to improve socio-economic and socio-demographic conditions of citizens with similar background to the participants of the study, to improve their income.

To put it simply, when we most accurately predicted a person's income, two of the five factors that had the largest effect on income were 'whether they thought that they had been treated unfairly by teachers due to their skin colour', and their ethnic group. To address the negative correlation between these two factors and income, the government should aim to invest in areas within the UK with larger ethnic-minority populations, and particularly the schools within these areas. 'Unfair treatment by teachers due to skin-colour', although unverified, would worsen the overall quality of education received by ethnic minorities and lead to a lower likelihood of attending university, and lower chances of obtaining a high-income skillset. Within the survey population, those belonging to ethnic groups other than 'White' were more likely to believe that they were discriminated against by teachers.

We can also reasonably deduce that higher levels of government intervention to support single-parent households could lead to higher income for the young person later on in life, as participants from a single-parent household had lower levels of income.

Based on our model for debt, we recommend the government investigate racial disparities in debt and finances; firms may be charging more interest to ethnic minorities, for example. Thus, the government should look at whether ethnic minorities are being discriminated against when taking out debt, and take steps to mitigate it.

Bibliography

Appendix

Appendix A: Merging levels of categorical variables

Predictor variable	**Post-merging results**	**Justification**
`W1wrk1aMP` (current working status of main parent) | ‘1,2,3,4’ merged to ‘1’ (full and part-time employment) | | | | | ‘5’ converted to ‘2’ (unemployed and seeking work) | | | | ‘6,7,8,9,10,11,12’ merged to ‘3’ (unemployed and not seeking work) |		
`W1empsmum` (Employment status of mother) | ‘1,2’ merged to ‘1’ (full- and part-time employment) | | | | | ‘3’ converted to ‘2’ (unemployment and seeking work) | | | | ‘4,5,6,7,8,9’ converted to ‘3’ (unemployed and not seeking work) |		
`W8DACTIVITY` (Current activity of CM) | ‘1,2,3,4,11,12’ merged to ‘1’ | | | | | (full- and part-time employment) | | | | ‘5’ converted to ‘2’ | | | | (unemployment and seeking work) | | | | ‘6,7,8,9,10,13,14’ merged to ‘3’ | | | | (unemployed and not seeking work) | | | | |		