How does previous education, household environment and family background affect the young person’s financial position?

Candidate number: 19811

Table of Contents

[Introduction 3](#_Toc39668081)

[Comments about the data 3](#_Toc39668082)

[Exploratory analysis to final model – W8DINCW. 3](#_Toc39668083)

[Results W8DINCW 5](#_Toc39668084)

[Comments about the analysis - W8DINCW 8](#_Toc39668085)

[Exploratory analysis to final model W8QDEB2 9](#_Toc39668086)

[Results - W8QDEB2 11](#_Toc39668087)

[Comments about the analysis W8QDEB2 12](#_Toc39668088)

[Exploratory analysis to final model – W8QMAFI 13](#_Toc39668089)

[Comments about the analysis – W8QMAFI 15](#_Toc39668090)

[Interpretation and Conclusions for a lay audience 15](#_Toc39668091)

[Appendix 18](#_Toc39668092)

## Introduction

This report analyses a subset of 5793 students from the Next Steps Study (NS) between 2004 and 2015; Waves 1-6 were collected yearly between 2004 when students were 14 and 2010 when students were 20. Wave 8 was collected in 2015 when they were 25. The survey taken with the main intent of “ Next Steps was set up to study young people’s experiences of **secondary school** and their transitions from compulsory schooling to further education, training or the labour market.” \*1. The NS is a large-scale survey that focuses on English students, asking about an array of subjects around their education, household environment and upbringing. This report is comprised of three regressions built around the variables W8DINCW: Continuous weekly income, W8QDEB2: Total amount owe, W8QMAFI: How managing finically these days. The data is collected by a questionnaire sent to the children and parent(s).

The results of the models indicate that a young person’s financial position is affected by many factors from their household environment and family background

For the baseline values of each model please refer to the appendix.

The equation for W8DINCW, continuous weekly income, shows that lots of factors have an effect, ethnicity has a big effect favouring white to all others, the log income of a white person in 0.21 bigger than all other ethnicities. Another big factor is gender favouring female’s income by log 0.08 to males.

The equation for W8QDEB2, total amount a person owes, shows that the factors that come into play the most is the occupation of the young person’s parent at aged 14, as well as whether or not the person has a child at age 20. This model appears to not be accurate as shown by its R2 values.

Our logistic regression model for W8QMAFI, how someone thinks they’re doing financially shows that occupation and last take home pay are two significant factors as to whether a person is likely to judge themselves as doing financially well or financially poorly.

## Comments about the data

We saw many missing values in our dataset, many of these seem systematic in the fact that the missing values tend to correlate with the factors people would be more likely to lie about. In our continuous weekly pay model, looking at W4AlcFreqYP, those with missing data go on to have the lowest income. This could be down to those who drink the most don’t want to fess to it in the survey so lie/don’t answer. This is seen with several of the predictors with W8DINCW as it’s the same data set. Looking at W1famtyp2, Missing leads to the most debt. This could be due to the fact people who are in care homes for example wouldn’t want to share that detail and they are less educated about finances due to not living with parents, leading to more debt.

The systematic missing data could affect our results however, some of the missing data is revealing in itself. If you were to introduce a ‘prefer not to answer category’, this might be more revealing as it could separate the response bias from the other data and allow for more accurate results.

We’re likely to experience some non-response bias to. While people are invited to join NS survey, they are not obliged. This means that the participants might not represent the population perfectly.

## Exploratory analysis to final model – W8DINCW.

To begin making the model some exploratory analysis was needed. We made all scatter plots for all the continuous variables and boxplots for all the categorical variables to see the relationship. Next, data was analysed, this showed that W4Childck1YP, W6Childliv, W6NEETAct, W8DAGEYCH and W8PUSA all contained less than 30% of the data. To avoid making the sample too small these variables were eliminated from the data set. NSID was also eliminated as was a randomly assigned number to each candidate. The next step was to eliminate the other predictors of income as these too closely relate to our outcome variable of W8DINCW; these predictors eliminated were W8GROW, W8NETW, W8NETA. As W8QDEB2 was the focus of the other multiple linear regression, that too was removed. The remaining data was reformatted: removing all the rows that contained a missing value for any of the continuous predictors as continuous predictors must have all their data to run a regression. This left the data with 2355 rows which is sufficiently large to continue using.

For the categorical values, the missing data was merged into a ‘missing’ category for each predictor. This made the data easier to analyse later as some of the predictors had many different types of missing data, some with small sample sizes so couldn’t be accurately compared to merge. In the data set the categorical predictors were coded as numbers. This meant R would process them as numeric values so it had to be specified that they were factors, meaning that the number represents something categorical like gender or employment status. To make the values more interpretable, the baseline value for each variable was changed to be the one that appeared most often.

The processed data was used to make the first linear regression with all the remaining variables. Using the ‘Anova’ function, the significance was checked for each variable. It showed that none of the continuous variables were significant. At no point for W1DINC did we see any dominating predictors so no interactions were included. As a large number of rows had been deleted from the continuous variables for having missing values for factors that are now know to not be significant, all these rows were added back in, and the continuous variables removed entirely; This leaves the data with lots of unprocessed rows. The categorical predictors were reprocessed, all the missing values were recoded as ‘missing’, and the values were changed to factors again.

The second regression was made with the data and ‘Anova’ was ran again. W1famtyp2 and W8DMARSTAT were both not significant so they were removed. The ‘vif’ function was also ran – VIF checks how easily a predictor is measured from a linear regression of the other predictors, the larger the values the worse it is. W1marstatmum, W1famtyp2 and W1wrk1aMP had a VIF much larger than all the other values, the vast majority were in the range 1-8, but these were 32, 18 and 18 respectively. For this reason they were removed from the data.

The third regression was ran and Anova showed that all the values were significant to 5%. The VIF was ran and again and variables were eliminated.

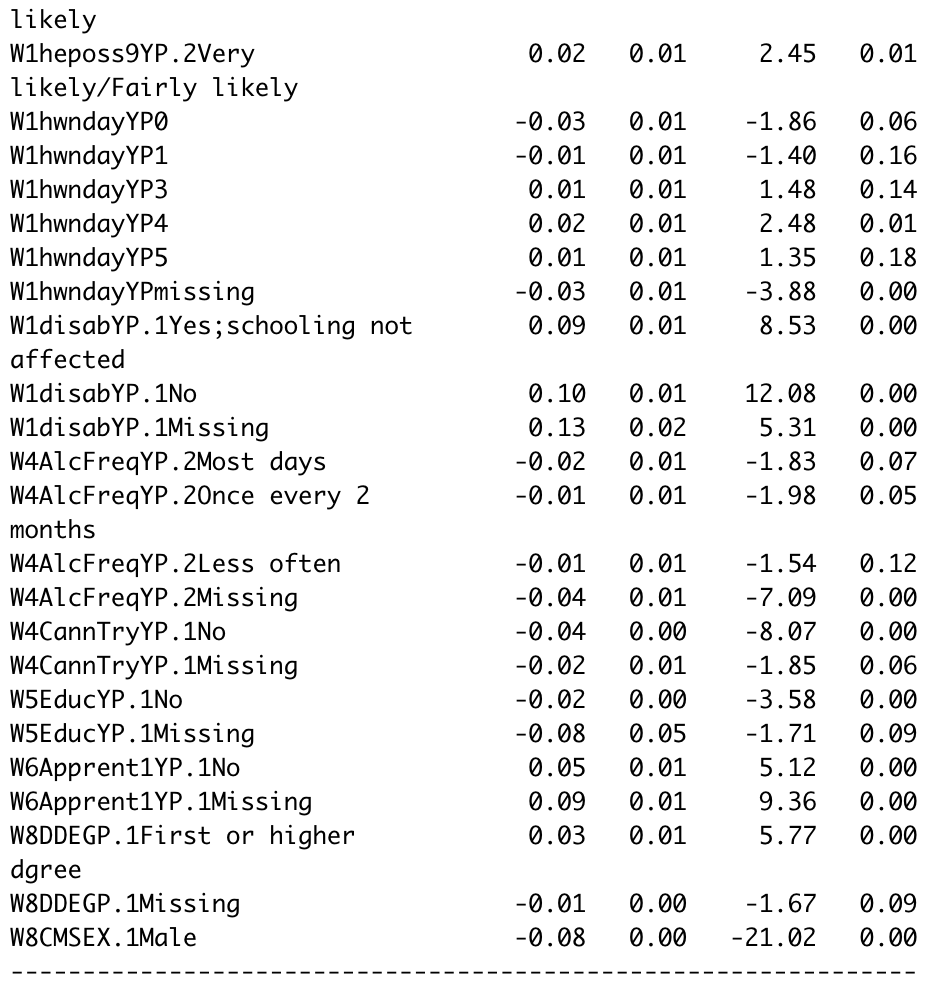
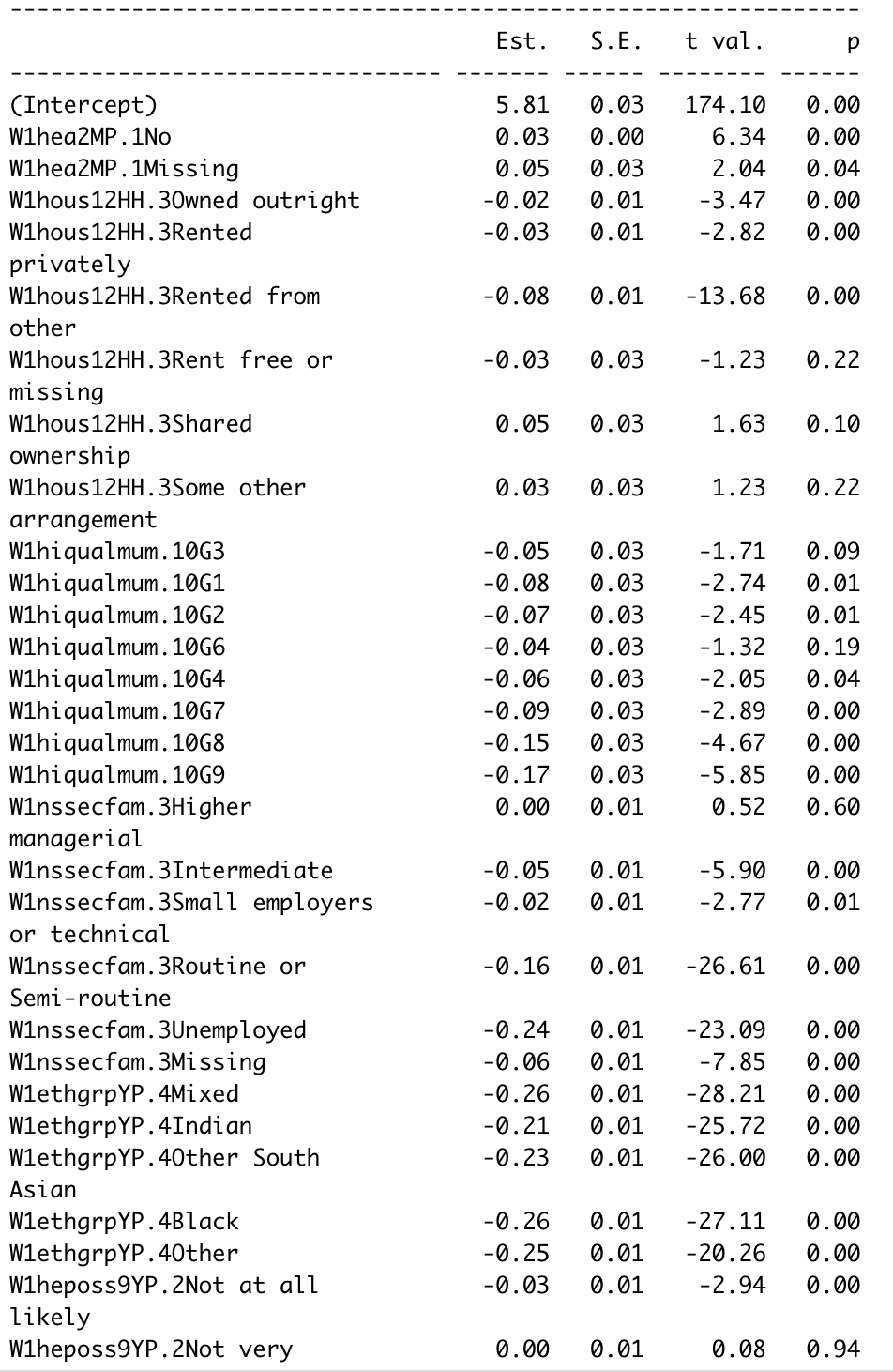
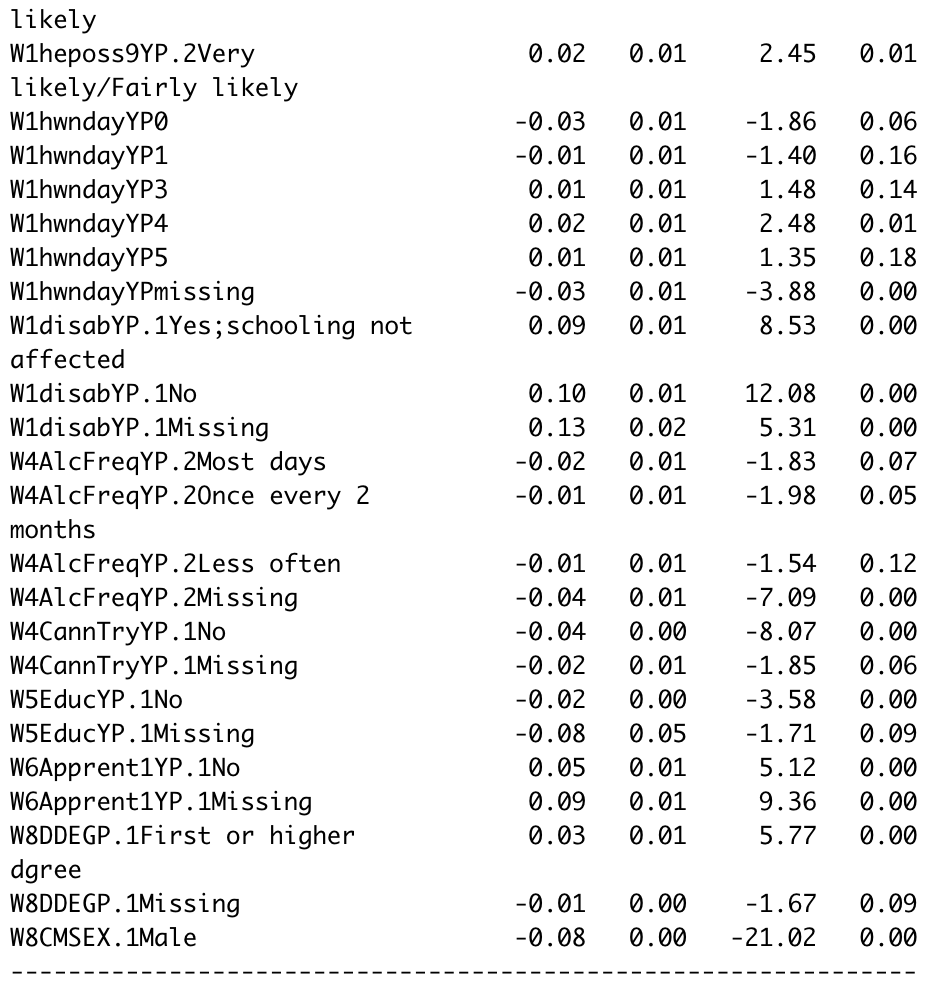
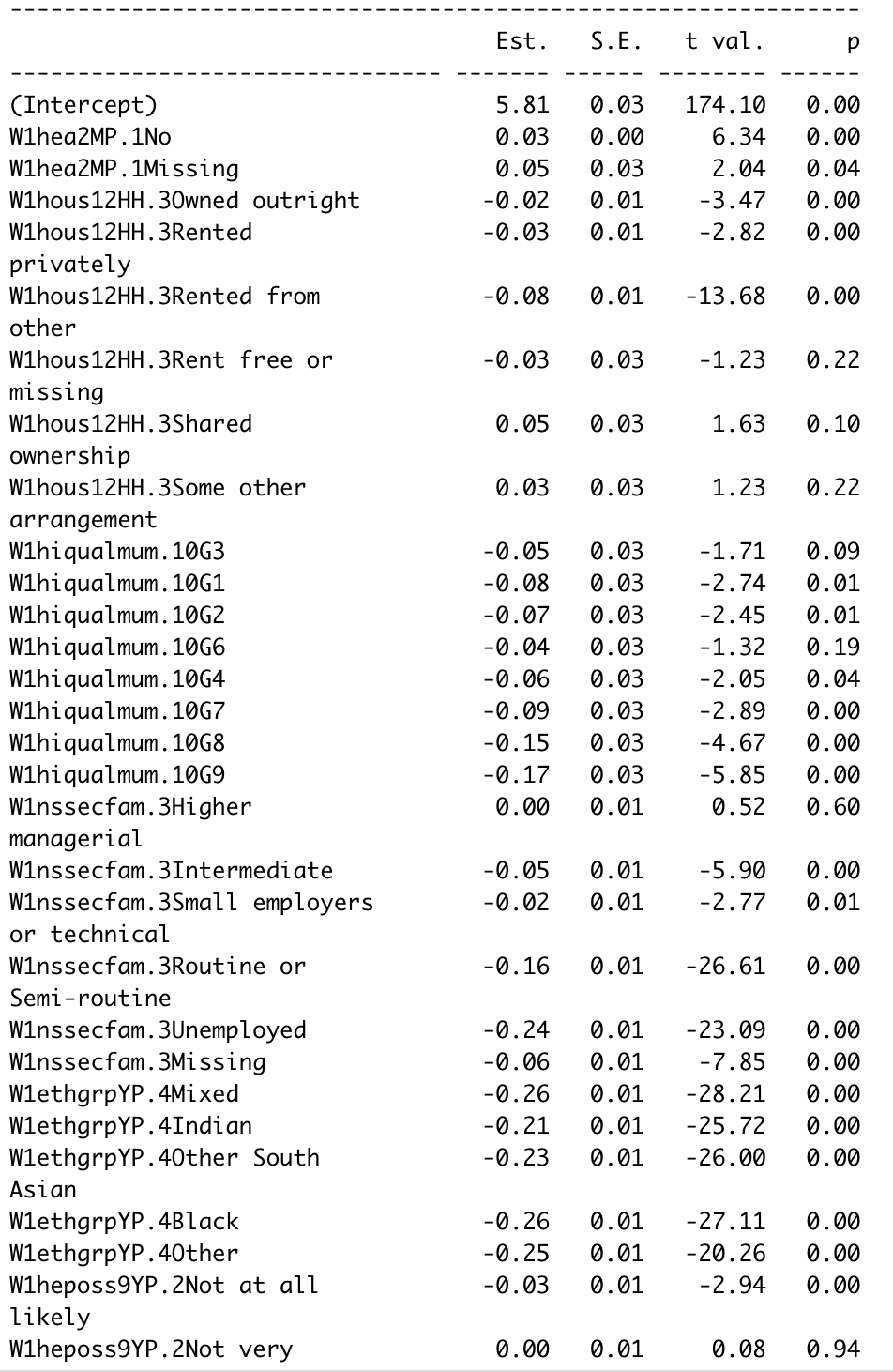
Now all the predictors are statistically significant, a boxplot was generated for each variable. to process the remaining data. When two factors of a variable had similar boxplots, we merged them when we deemed it logical for the factors to be merged (i.e. combining a having degree and GCSEs is not logical, combining having a level 3 BTEC and A-levels would be logical). This occurred for some the variables (See appendix for merged values in final regression). Throughout the process we saw numerous categories with outliers, none of these values seemed unfeasible or excessively extreme as to affect the data so they were left in.

Regression 3 was made with these updates to the data. Checking the residual analysis, we see funnel shaped data which indicates heteroskedasticity. Regression 4 takes the log of the variables and stops the data looking funnel shaped, while the QQ plot shows a slight bow shape which indicates normal distribution with small amount of skew but nothing to worry about.

Outlier analysis was performed using a function that found the values with the highest residuals, highest leverage values, highest diffits and highest Cook’s values. Two points had Cooks distances far bigger than the other points (4200 and 3981 respectively). Making a linear regression with the larger of these points removed and with both of them removed didn’t make much difference to the model so it was left in. Two points had a larger standard residual than the others. Again, a model was made with the larger and another with both of them removed at it made no significant difference so they were left in. Lots of the standard residuals were higher than 3, usually this means we should investigate them but due to how close all these values were to each other they can be left in.

## Results W8DINCW

Below is table 2, containing all the results for the final regression to two decimal places to make things clearer. Note, some of the variables are followed by ‘.’number, this was done in the processing of the data.



The majority of the results are highly significant in the final model. These results show how each variable compares to the baseline in each group which is the most commonly occurring value. For W1hea2MP those with missing data are the best off, followed by those with outcome No, then those with outcome yes. In terms of W1hous12HH those with shared ownership are best off and those renting from others are worst off. However, the p value for shared ownership is 0.1 so this result isn’t statistically significant. Surprisingly, those whose mother’s highest qualification (W1hiqualmum) is G5, which is AS levels or Scottish Higher Grade have the highest incomes while those who have no qualifications, G9 do the worst, closely followed by those with low level qualifications, G8, such as an NVQ1 or youth training. In terms of W1nssecfam, the highest earners are those with parents who are lower managerial or professional, whereas those who do the worst are those with parents who are unemployed. Ethnic group, W1ethgrpYP was one of the strongest factors, those who were white earner significantly more than those who were of any other ethnic group, Black and mixed earn slightly less than all the other groups. As expected, W1heposs9YP shows that those who think they’re very or fairly likely to go to university outperform everyone else, particularly those who are not at all likely who are the worst off. Looking at W1hwndayYP, those who have four pieces of homework a week are best off, even better off than those who have five, however, it should be noted that the result for five has a p value of 0.18 so is not significant to the 5% level. Those that are worst off are those that have data missing or have none. W1disabYP shows that those with a disability that affect their schooling go on to earn the least while those that have no data go on to do the best, followed by those with no longstanding illnesses. For W4AlcFreqYP, the more that a young person drinks, the less they go on to earn, those with no data go on to earn the lowest with the logged income being 0.04 less. Those who drink most days have a p value of 0.07, this isn’t significant to the five percent but perhaps would be with a larger sample size. Counterintuitively, W4CannTryYP shows that those who had tried cannabis, went on to earn more than those who had missing data, followed by those who hadn’t tried it. W5EducYP shows that those who were still in education earnt the most followed by those who didn’t. Those with no data had log income of 0.08 less than those who did. W6Apprent1YP shows that those who are not in apprentaships go on to have a higher income, those who aren’t have 0.05 higher log income than those who are. Those who have missing data do the best with 0.09 higher log earnings. W8DDEGP shows that those who achieve a first degree or better have the highest income. Those with missing data have the lowest but this result is not significant with a p value of 0.9. Lastly, W8CMSEX shows that females go on to have higher weekly income than those who are male. The log income difference is 0.08.

These results lead us to the equation

E(log(W8DINCW)) = 5.8064135 + W1hea2MP.1No \*0.0293666

W1hea2MP.1Missing \* 0.0530938 + W1hous12HH.3Owned outright \* (-0.0194249)+ W1hous12HH.3Rented privately \* (-0.0296964) + W1hous12HH.3Rented from other \* (-0.0773597) + W1hous12HH.3Rent free or missing \* (-0.0309656) + W1hous12HH.3Shared ownership \* 0.0465603 + W1hous12HH.3Some other arrangement \* (0.0324461) +W1hiqualmum.10G3 \* (-0.0513167) + W1hiqualmum.10G1 \* (-0.0812099) + W1hiqualmum.10G2 \* (-0.0723953 ) + W1hiqualmum.10G6 \* (-0.0387355) + W1hiqualmum.10G4 \* (-0.0613757 ) + W1hiqualmum.10G7 \* (-0.0867745) + W1hiqualmum.10G8 \* (-0.1472021 ) + W1hiqualmum.10G9 \* (-0.1728735) + W1nssecfam.3Higher managerial \* (0.0032250 ) + W1nssecfam.3Intermediate\*(-0.0472604) + W1nssecfam.3Small employers or technical \* (-0.0159619 ) + W1nssecfam.3Routine or Semi-routine \* (-0.1623843) + W1nssecfam.3Unemployed \* (-0.2425720 ) + W1nssecfam.3Missing \* (-0.0595784) + W1ethgrpYP.4Mixed \* (-0.2596910 ) + W1ethgrpYP.4Indian \* (-0.2115500 ) + W1ethgrpYP.4Other South Asian \* (-0.2277870 ) + W1ethgrpYP.4Black \* (-0.2560060) + W1ethgrpYP.4Other \* ( -0.2487989 ) + W1heposs9YP.2Not at all likely \* (-0.0333577) + W1heposs9YP.2Not very likely \* (0.0008266 ) + W1heposs9YP.2Very likely/Fairly likely \* (0.0229024) + W1hwndayYP0 \* (-0.0277488 ) + W1hwndayYP1 \* (-0.0101811) + W1hwndayYP3 \* (0.0082977 ) + W1hwndayYP4 \* (0.0159395) +W1hwndayYP5 \* (-0.0081656 ) + W1hwndayYPmissing \* (-0.0330295) + W1disabYP.1Yes;schooling not affected \* (0.0896102 ) + W1disabYP.1No \* (0.1026837) + W1disabYP.1Missing \* (0.1322620 ) + W4AlcFreqYP.2Most days \* (-0.0198883) + W4AlcFreqYP.2Once every 2 months \* (-0.0128439 ) + W4AlcFreqYP.2Less often \* (-0.0113032) + W4AlcFreqYP.2Missing \* (-0.0435453 ) + W4CannTryYP.1No \* (-0.0360158) + W4CannTryYP.1Missing \* (-0.0233641 ) + W5EducYP.1No \* (-0.0159142) + W5EducYP.1Missing \* (-0.0780184 ) + W6Apprent1YP.1No \* (0.0467067 ) + W6Apprent1YP.1Missing \* (0.0913392 ) + W8DDEGP.1First or higher dgree \* (0.0303067) + W8DDEGP.1Missing \* (-0.0080206 ) + W8CMSEX.1Male \* (-0.0790608 )

The other equations can be formed in a similar way. As the tables have been included this shall not be done in this report.

## Comments about the analysis - W8DINCW

W8CMSEX shows that females earn more than males, this goes against previous larger scale studies. This could be due to females doing better in schools meaning that the initially earn more as they’re in better jobs, then other factors like sexism come into play in later life and males go on to earn more. Another surprising result is those that have tried cannabis, W4CannTryYP, go on to have a higher income than those who haven’t and those with missing data. While this could be people who are more open to new things do better, I believe that this result needs investigating, which I cannot do due to lack of expertise.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 1 (full) | Model 2 | Model 3 | Model 4 | Model 5 (final) |
| R2 / Adj R2 | 0.737/0.7076 | 0.7873/0.7833 | 0.6752/0.6718 | 0.6982/0.695 | 0.6985/0.6957 |
| F stat significance | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 |
| Std. error | 34.52 | 33.96 | 41.31 | 0.1356 | 0.1378 |

This table contains a few of the key diagnostics for this model at various stages. The F stat remains very low for the duration implying at least one of the predictors presents must explain some of the outcome. This value might not be so reliable though as we see relatively high VIF numbers throughout which indicates multicollinearity issues, these were removed and implemented in model 3. One of the predictors removed from model 2 to form model 3 was W1wrk1aMP – interestingly if this remained in, model 3 the diagnostics would be Multiple R-squared:0.7435, Adjusted R-squared:0.7395.

While the R2 value does decrease from first to last model, what we do see is the R2 and Adjusted R2 values getting closer implying that a higher proportion of the variables explain variation of the dependent variables. They start off just over 4% different, by model 2 throughout to model 5 they remain well within 4% of each other.

An interesting observation is the huge change is std. error between model 3 and model 4. It should be noted that the values for Std.error are relative to the data. As we have taken the log of the dependant variable so you would expect this value to be much lower.

## Exploratory analysis to final model W8QDEB2

The first steps were the same as before, removing the variables with less than 30% of their data. Then, a boxplot was made for each categorical predictor and a scatter graph for each continuous to see how they correlate with W8QDEB2. A few outliers were spotted Predictors of income were kept in as they don’t capture the same idea as debt. Where continuous predictors had a missing value, the whole row was remove leaving us with 1827 rows which is large enough. The missing values for categorical values were merged. Where necessary, continuous predictors were cantered to make them more interpretable. The categorical variables were recoded as factors. The baseline was changed for the predictors to be the one that is most common, apart from W1hiqualdad and a few others where the most common was missing, so the second most common was used.

This is where the first regression was made with all the remaining variables. An Anova was ran and all the predictors that were not significant were removed. A second linear regression was then made and an Anova was ran. Few of the predictors were significant so an outlier analysis was ran. Two points were identified as outliers. Point 1541 had W8QDEB2 of £400,000 whereas the next highest was £275,000. While this is fusible, this is an extreme case which affects the diagnostics. As this is an exceptional case and not representative of the average as we are predicting for it is removed. Data point 3054 at 12 younger siblings. The next highest was 5, I don’t believe this value to be feasible so it is removed.

The third linear regression was made and this showed removing the outliers helped. Then by checking all three regressions with the alias functions we saw that W1empsmum and W1empsdad were colinear with two other significant predictors W1hiqualmum and W1hiqualdad respectively. W1NoOldBro was also colinear with W1disabYP so was removed. This led to the fourth regression being made. Few predictors were significant in the Anova test so vif was ran: W1hiqualmum, W1hiqualdad, W1depkids and W1famtpy had VIF scores far larger than the others with 11.7, 38.1, 21.3 and 58.7 respectively so they were all removed.

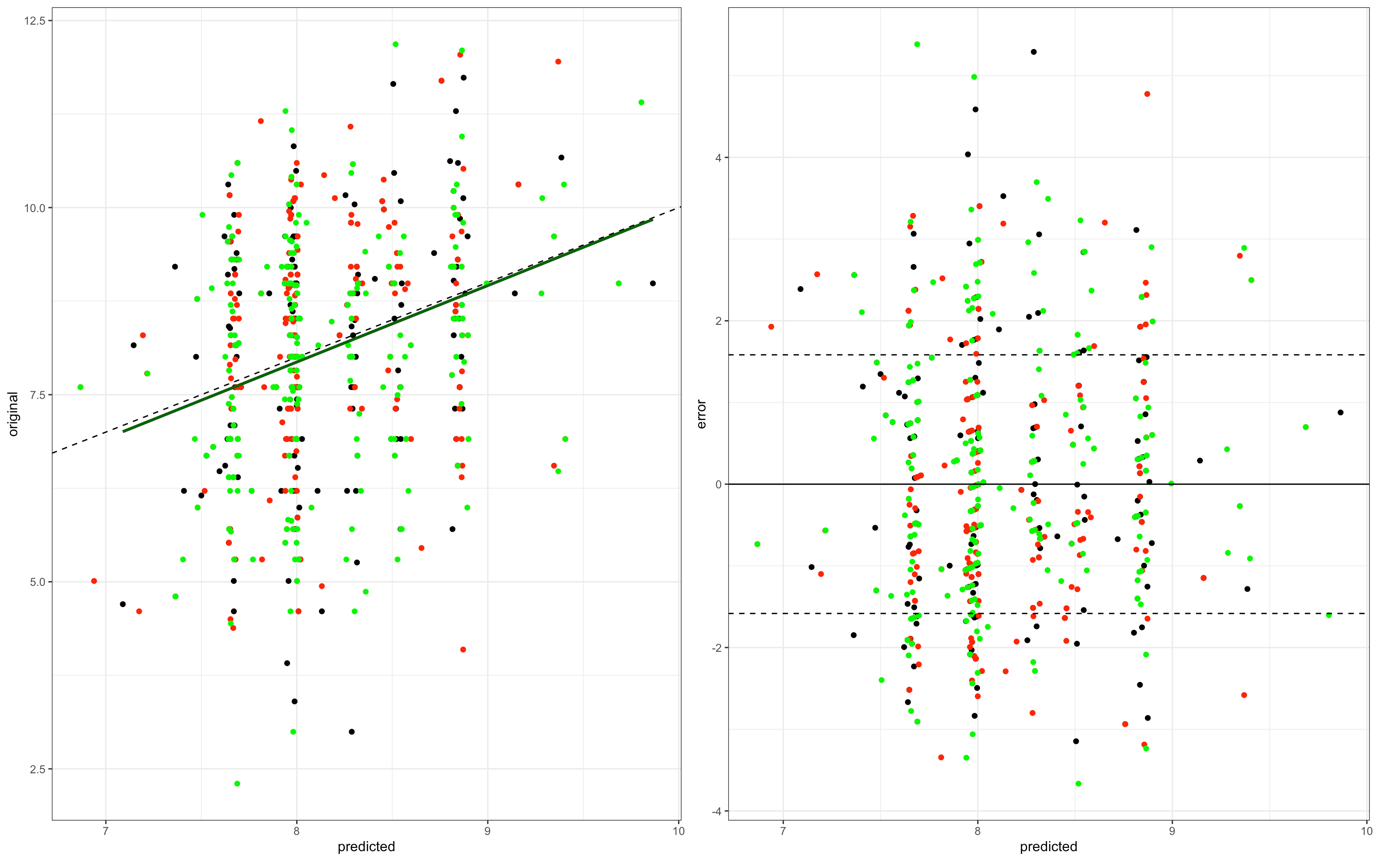
A fifth regression was made and W1nssecfam, W6gcse and W8qmafi were removed as they were non-significant and the signs other coefficients are the ‘wrong’. A sixth regression was ran and outlier analysis performed again, points 3111, 1541, 84 and 1678 had high DIFFTS and Standardised residuals so were eliminated to make a seventh linear model. The residual plots were checked and showed kurtosis so Regression 8 was made using a log transformation. It fixed the issue but the 0 values for W8QDEB2 had to be removed to enable this transformation. The new residual plots of this showed slight negative skew but nothing alarming.

Continuous predictors were standardised and a scatter plot was made for each one. W8NETW still had a noticeable outlier so it was removed as it doesn’t represent the average which this regression is made to predict for.

A boxplot was made for each categorical predictor. W1hous12HH and W8DDEGP were both removed as each factor’s median was very similar, not aiding prediction, the latter also had a high VIF. Numerous levels were merged for other predictors when medians and distributions were similar and it was logical (as before).

Regression 11 saw anova and vif eliminated predictors again. Regression 12 was made and was checked for interactions were removed. The predictors that were not significant were removed which took us to model 14, the predictors that weren’t significant were removed again.

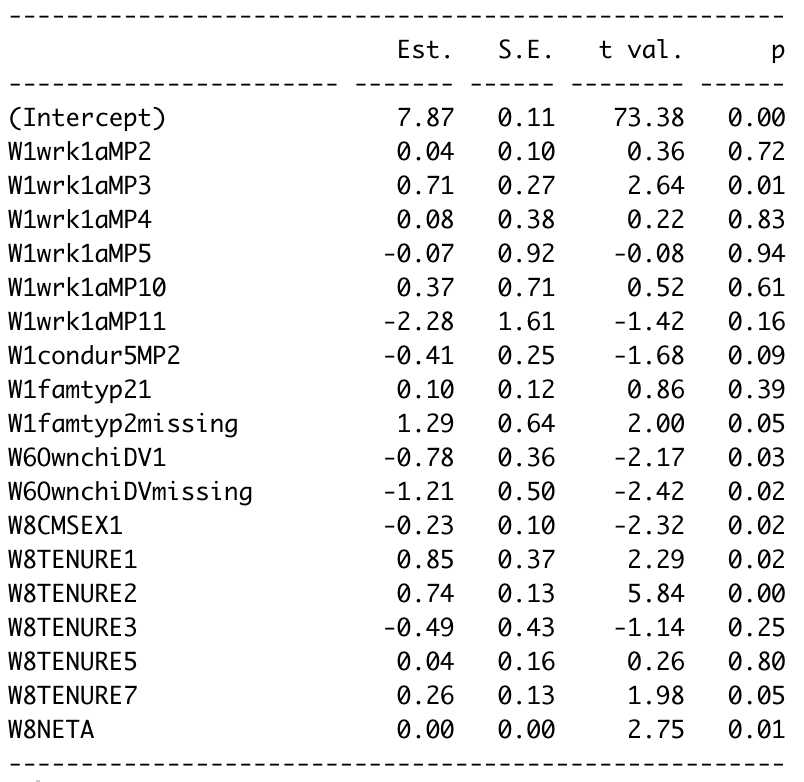
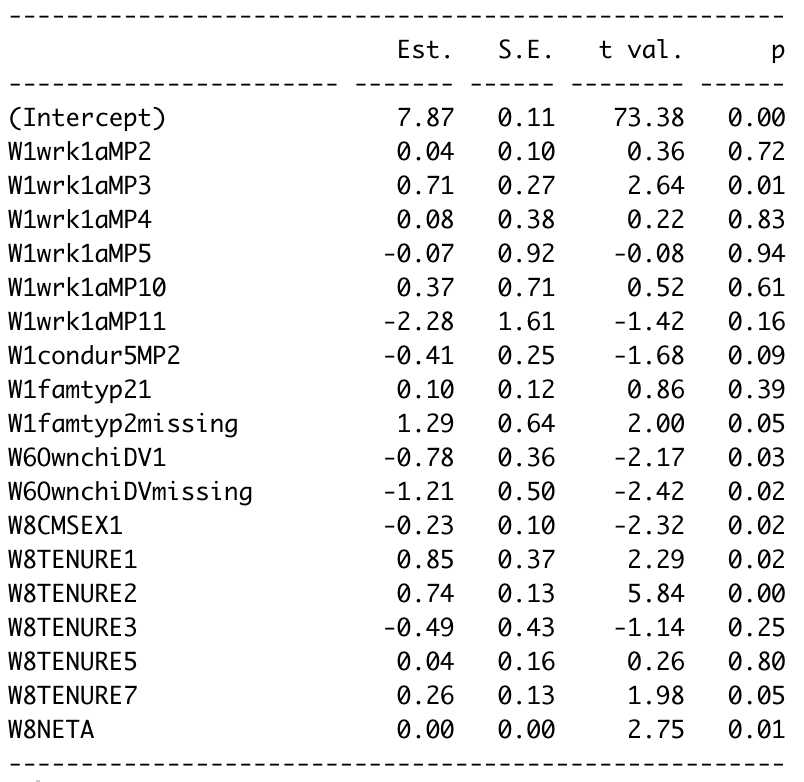
Model 15 remains which only contains mostly significant predictors. At this point cross-validation was performed with three 80/20 splits.



This predicted against original plot shows lots of variation in the y-direction suggesting that there is more variation in the W8QDEB2 than the model is accounting for. The two lines are very close on the same graph suggesting that there is no non-linear trend (and hence the log transformation was effective). The predicted against error plots show decent amount of the data are beyond the one standard deviation lines, however, these points are evenly distributed across the range of predictions. The data seems to be broadly similar across each split.

## Results - W8QDEB2

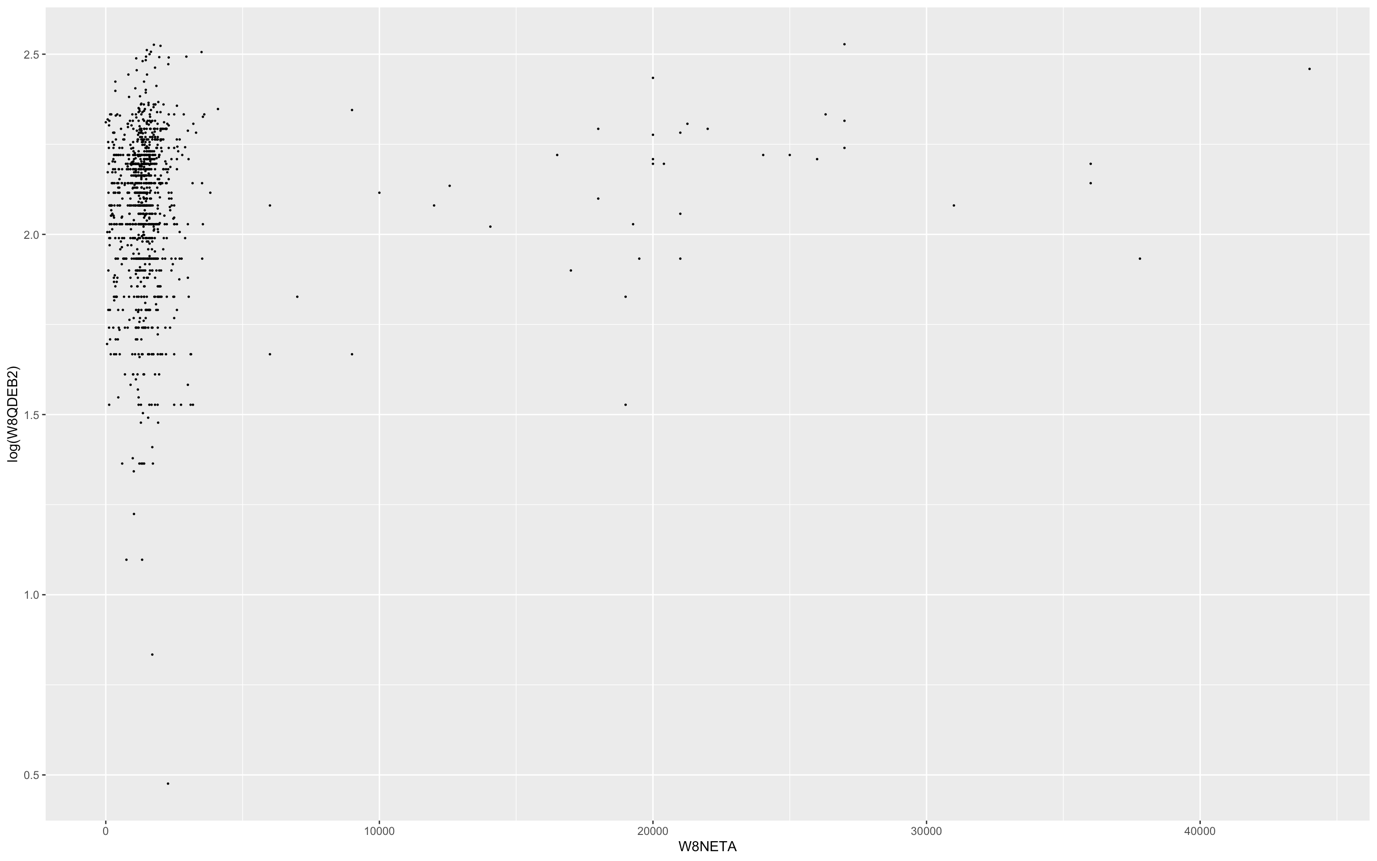
Below shows table Four with the results from our final regression to two decimal places.



The results show that around half of the factors are significant. A couple of predictors such as W1condur5MP2 with a p value of 0.09 might benefit from a larger sample size. The results show that in terms of W1wrk1aMP, the current working status of their main parent has a large impact on their total debt. If their parent is self-employed (W1wrkaMP3), their log debt is 0.71 more than the baseline of a full-time employee. This result is highly significant. Those with the least debt are those whose parents are retired by 2.28 in log debt compared to the baseline. Surprisingly, W1condur5MP2 shows that those who do not have a computer at their house go on to have 0.41 less log debt. W1famtyp2 shows that those who live in a single parent household go on to have 0.1 more log debt than those who live with both of their parents and interestingly, those who do the best are those with missing data, who have 1.29 more log debt. W6OwnchiDV shows that those who have their own children have 0.78 less log debt than those who don’t (who are the baseline). Those who have not data for this category do the best by having 1.21 less log debt than the baseline. Looking at W8CMSEX1, it shows us that the baseline females have on average 0.23 log debt more. W8TENURE shows us that the people who have the most debt are those that own their house outright followed by those who own their house thanks to a loan. The best off are those that pay part of rent or mortgage by 0.49 log debt compared to the baseline which is renting (including housing benefits). The table is somewhat misleading for W8NETA as it only presents to two decimal places to make the categorical values easier to read. Per change in W8NETA you expect to see 3.315\* difference in the log debt.

## Comments about the analysis W8QDEB2

A lack of expertise on my behalf led me to struggle with this data analysis, I struggled with eliminating predictors as they were almost all not significant. If I were more knowledgeable, I feel like I would have made a better model here.



This graph shows last take home pay against log(debt). While the model indicated that they have a positive correlation, this is possibly not the case. The vast majority of the data points form seemingly a vertical line, then a few dozen points form a positive correlation with a wide enough range to have an impact on the regression. I feel like a large sample would give more clarity weather there is a relation or not here. I’d hypothesise that there is only a relation around the 15000 mark and upwards, but this rules out most of the data so wouldn’t help us make a model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model 1 (full) | Model 7 | Model 8 | Model 15 (Final) |
| R2 / Adj R2 | 0.1193/-0.006486 | 0.1137/0.04189 | 0.1339/0.06228 | 0.07903/0.06291 |
| F stat significance | 0.9484 | 1.583 | 1.869 | 4.901 |
| Std. error | 22640 | 25530 | 1.569 | 1.567 |

Here we have the diagnostics for four of the models. The R2 values decrease throughout the process, however we do see that R2 and adjusted R2 get closer together. These R2 values are worry and raise a red flag.

Model 7 and 8 were included to highlight the large drop in standard error when the log of a model is taken.

## Exploratory analysis to final model – W8QMAFI

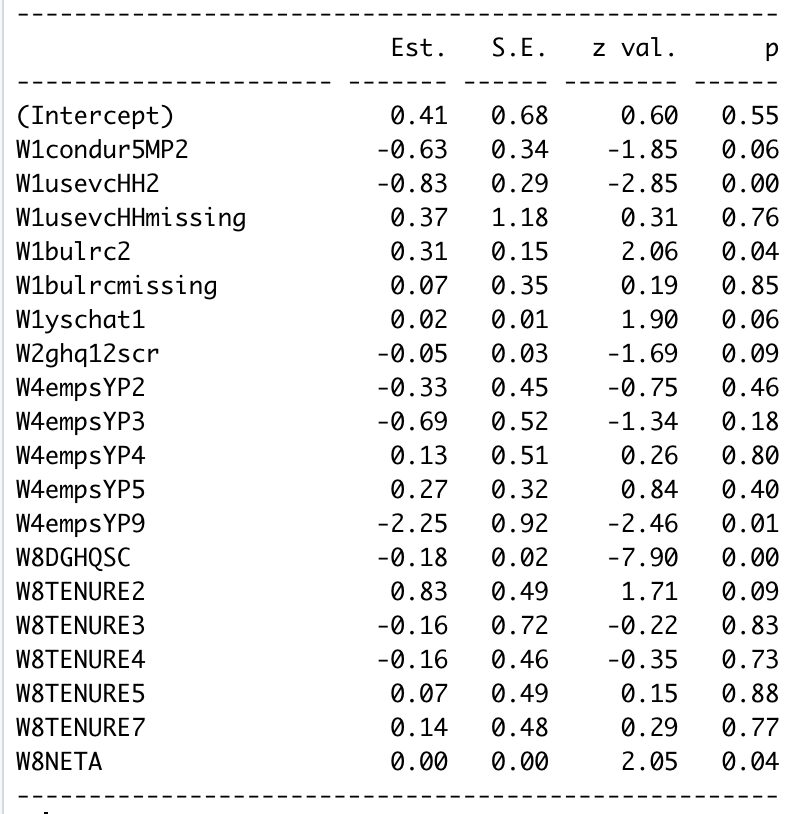
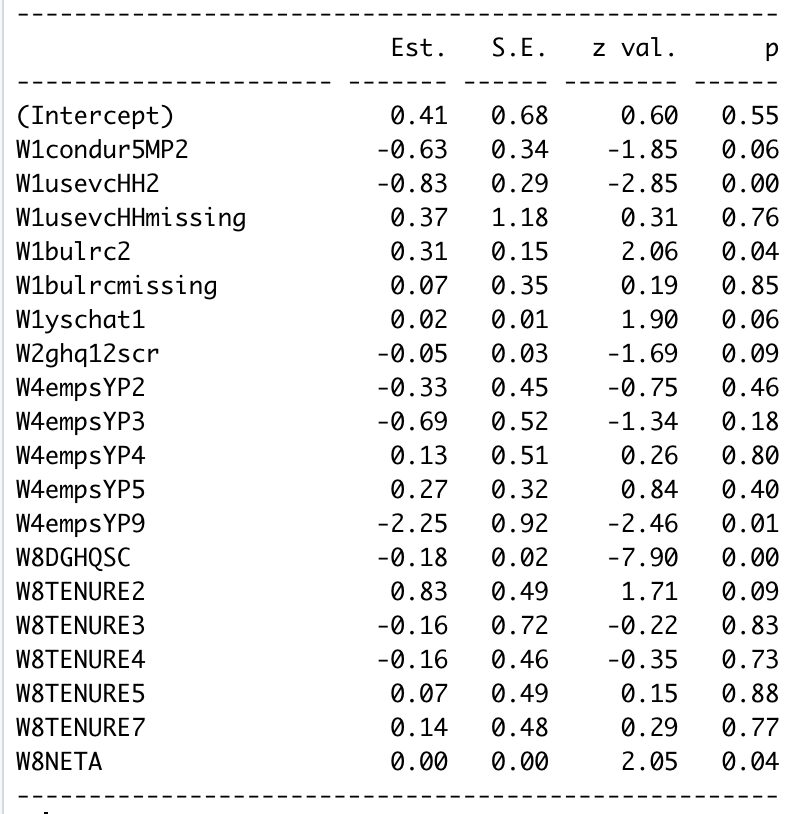
To start, a box plot was formed of W8QMAFI against a few of the predictors to see how it stacks up and the size of each factor was found. Due to there being such few missing values, these are not representative of the standard person and therefore they should be excluded from the model all together. As a logistic model is being made, W8QMAFI needed to be made into a binary predictor so all the positive outcomes were grouped, and all the negative outcomes were group into people doing financially well, and people not doing financially well respectively.

The rest of the data had to be process as before removing the continuous predictors with less than 30% of their data, removing the rows corresponding to missing values in the other continuous predictors and then deleting the predictors that were heavily overlapping. The categorical predictors had their missing values combined into a missing category and were coded as factors.

The first logistic regression was then ran using all the remaining predictors. Then using the anova function a chi squared test was ran. This compares each predictor, one by one to W8QMAFI to dictate weather or not they are significant. We deleted all the predictors that were not significant and ran another chi squared test using anova on the new regression of just significant predictors. This was repeated one more time and left us with a model of highly significant predictors.

Results:

Looking at the table, the first thing noticed is the moderately large p values



The table shows with W1condur5MP2 that those who do have a computer in their house outperform those who don’t by 0.63. The p value of this result is 0.06 so not significant to the 5% level but perhaps would be with a larger sample size. Simarly, W1usevcHH2 shows that those who live in a household with a car (the baseline) outperform those who don’t by 0.83. This result is highly significant. Those who have missing values do better than the baseline by 0.37 however it this has a high p value. W1bulrc2 shows that those who had not been bullied in the past six months outperform those that had by 0.31, those who missing data for that category performed better than those who had by 0.07 but not as well as those who hadn’t. W1yschat1 shows that the better the young person’s attitude to school, the better they do by 0.02 per score in the attitude test. This result is significant to the 10% but not to the 5% so might benefit from a larger sample. W2ghq12scr shows that the higher a young person’s gcq12 score the lower their chance of financially doing well. For each change in gh12 score leads to a change of 0.05. W4empsYP shows that type of occupation of the young person goes on to impact their finatial standings in later life, however, few of these values are significant. Those that are most likely to judge themselves to be doing financially well are those that are still in full time education during wave 4. Those that are most likely to judge themselves to be doing finically poorly are those that put their employment status as ‘other’. W8DGHQSC shows that those who score higher on the general health questionnaire are less likely to be financially doing well. W8TENURE shows that a person’s tenure too has an effect with those that own a house thanks to a mortgage or loan are the ones most likely to deem themselves doing financially well whereas those that rent, share rent, or share a mortgage with someone else are most likely to deem themselves not financially stable. W8NETA has a misleading value in the table once again as values are rounded to two decimal place – a higher degree of accuracy would give the value as 6.856\*. This shows that those that have a higher last net pay are more likely to deem themselves as doing financially well.

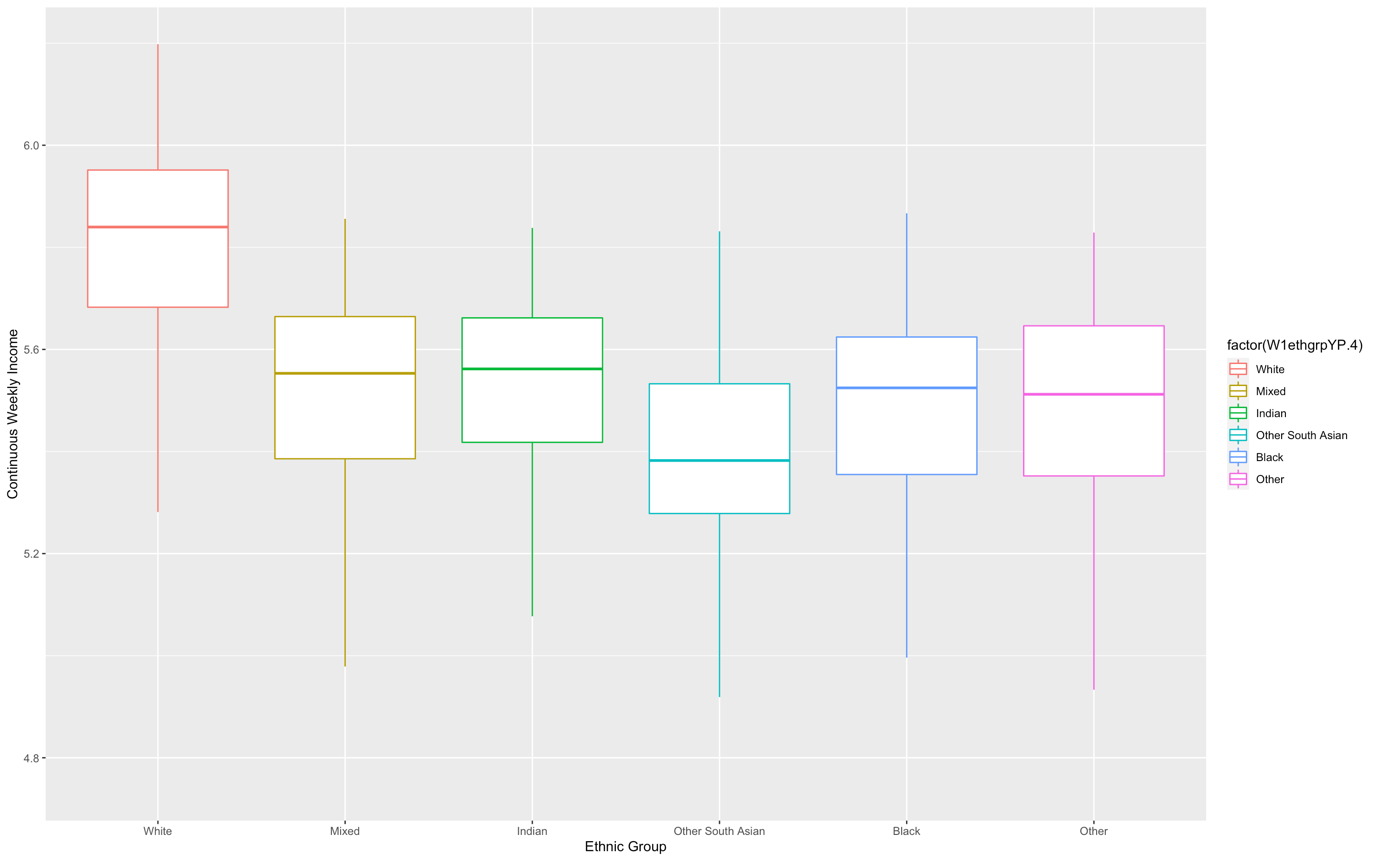
## Comments about the analysis – W8QMAFI

The Chi Squared value at 5% significance level is 12.59, the difference between our residual deviance and null deviance was 82 making our model highly significant. If you look at the table of

## Interpretation and Conclusions for a lay audience

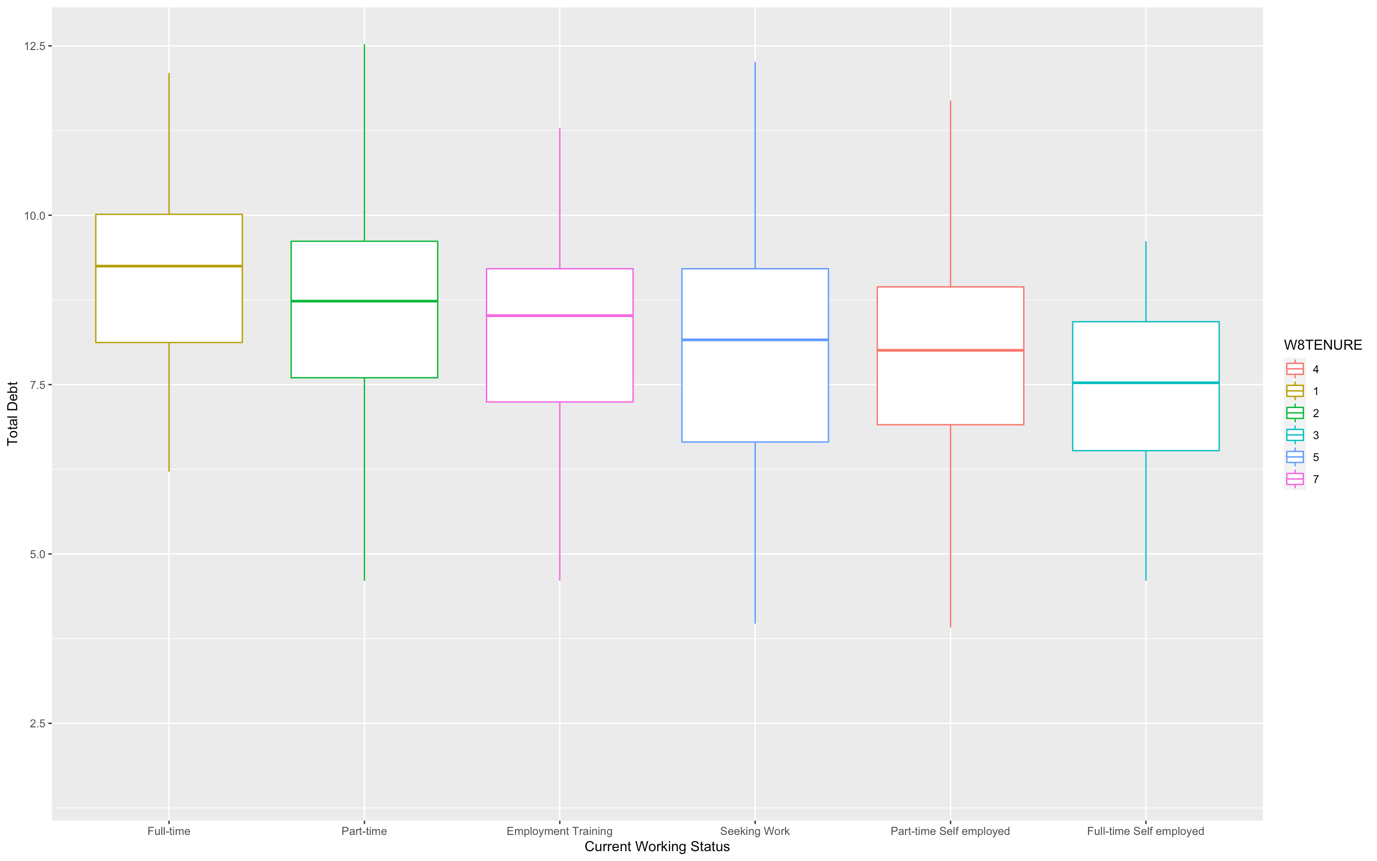
This report attempts to analyse how a person’s financial position in England is related/explained by numerous variables based on their education, household environment and family background go on to affect their financial position. Their financial position has been summarised by looking at their continuous weekly income, the total amount they owe, and how they think they’re coping financially. We had data on a sample of around 6000 students who had participated in the Next Step S between the years xxxx and xxxxx. We ran three regression analyses to try and determine the relationships to these three variables.

Looking at their continuous weekly income at age xxxxx, we saw a large number of factors making an impact. The most important predictor was the ethnicity of the young person. Someone who was white had expected earning much higher than someone of any other ethnicity. Another predictor that had a big impact was the highest qualification of the young person’s mother. Those whose mother highest qualifications were A-levels or Scottish Highers did the best, while those that do the worst are those that mothers have no qualifications. Other predictors that play a key role are whether a young person has a disability, or weather a young person is male or female.



The graph above shows how ethnicity of the individual goes on to affect continuous weekly income. It’s clear from this that White people have a much higher Continuous weekly income than any other ethnic group in England, while it appears that ‘Other South Asian’ go on to have the lowest. All the other categories are quite close together indicating that with a large enough sample they could be the same. It’s likely that this trend is the result of racism.

A young person’s debt at age 25, sees a smaller number of predictors than continuous weekly income. One of the most impactful predictors is the current occupation of a young person’s main parent at age 14. Those whose parents are retired often have the least debt whereas those who have the most debt is those that have parents who are self-employed. Another key predictor is weather a young person has their own child at age 20. Those that have missing data for this category go on to have on average the most debt, followed by those that have their own kids.



This graph shows a clear relation between total debt and their main parents current working status. This result is perhaps not what you would have expected with those whose parents are in full time working having a larger amount of debt when they’re older than anyone else.

It should be noted that the model looking at debt appears to not be very accurate.

Lastly, looking at how someone aged xxxx views their own financial position. Their occupation makes the largest difference to weather they believe they’re financially stable or not. Those that are Another large factor is the last net take home pay of the person. The more their last take home pay, the more likely they would be to judge themselves as doing well.

Our models suggest that a number of predictors surrounding a young person relating to their education, household environment and family background have an impact on their future financial status. While it’s obvious from our first model that certain factors have an impact on continuous income, due to low significance of our other models we have insufficient knowledge to answer the research question as continuous income

## Appendix

1. \*1 <https://nextstepsstudy.org.uk/home/about/purpose/>

W8DINCW

|  |  |  |
| --- | --- | --- |
|  | Variable Type | Baseline |
| W1hea2MP | Categorical | Yes |
| W1hous12HH | Categorical | Bought on a mortgage |
| W1hiqualmum | Categorical | G5 |
| W1nssecfam | Categorical | Lower managerial |
| W1ethgrpYP | Categorical | White |
| W1heposs9YP | Categorical | Missing |
| W1hwndayYP | Categorical | P2 |
| W1disabYP | Categorical | Yes; schooling affected |
| W4AlcFreqYP | Categorical | Once a week – Once a month |
| W4CannTryYP | Categorical | Yes |
| W5EducYP | Categorical | Yes |
| W6Apprent1YP | Categorical | Yes |
| W8DDEGP | Categorical | No degree |
| W8CMSEX | Categorical | Female |

Merging:

* W1hiqualmum:
  + 1,2 = G1
  + 3,4,7 = G2
  + 5,6= G3
  + 8,9,10 = G4
  + 11,12 = G5
  + 13,14,15 = G6
  + 16 = G7
  + 17,18,19 = G8
  + 20 = G9.
* W4AlcFreqYP:
  + ('1-2 times a week','2-3 times a month','Once a month')='Once a week-Once a month'"
* W1house12HH
  + ('Rent free','Missing')='Rent free or missing'"
  + ('Rented from Council','Rented from Housing Association')='Rented from other'"
* W1nssecfam
  + ('Small employers','Technical')='Small employers or technical'"
  + ('Routine','Semi-routine')='Routine or Semi-routine'"
* W1ethgrpYP
  + ('Pakistani','Bangladeshi')='Other South Asian'"
  + ('Black Caribbean','Black African')='Black'"
  + ('Other','Missing')='Other'"
* W1heposs9YP
  + ('Very likely','Fairly likely') ='Very likely/Fairly likely '"

W8QDEB2

|  |  |  |
| --- | --- | --- |
|  | Variable Type | Baseline |
| W1wrk1aMP | Categorical | 1 |
| W1condur5MP | Categorical | 1 |
| W1famtyp2 | Categorical | 0 |
| W6OwnchiDV | Categorical | 2 |
| W8CMSEX | Categorical | 2 |
| W8TENURE | Categorical | 4 |
| W8NETA | Continuous |  |

W8QMAFI

|  |  |  |
| --- | --- | --- |
|  | Variable Type | Baseline |
| W1condur5MP2 | Categorical | 1 |
| W1usevcHH | Categorical | 1 |
| W1bulrc | Categorical | 1 |
| W1yschat1 | Continuous |  |
| W2ghq12scr | Continuous |  |
| W4empsYP5 | Categorical | 1 |
| W8DGHQSC | Continuous |  |
| W8TENURE | Categorical | 1 |
| W8NETA | Continuous |  |

Merged:

* W8QMAFI
  + ('1','2')='1')
  + ('5','4','3')='0 ')

General

For the R code please refer to the other files.