London School of Economics and Political Science

ST211 Applied Regression

"What are the key predictors that influence income and debt at age 25?"

End of Term Group Project

ST211 Group:

22555

21379

28766

21588

WORD LIMIT: 3000

(excluding R code or the appendix)

WORD COUNT: 3289

(excluding table of contents, main headings, and appendix)

TABLE OF CONTENTS

I: INTRODUCTION

This project aimed to identify factors that influence weekly income and debt of individuals. To achieve this, we utilised two datasets extracted from the Next Step Study (NS). This longitudinal study consisted of a cohort of 15,770 individuals born in England between 1989 and 1990, who were initially interviewed in 2004, at age 14. They were interviewed yearly until 2010 with an additional interview in 2015-2016 aged 25 producing 8 waves of data. The two datasets, W8DINCW and W8QDEB2, were derived to facilitate our analysis and identify the driving factors for income and debt, respectively, at age 25.

The final models revealed that over half of the significant factors influencing income were from wave 1, with the most influential being ethnic group, socioeconomic class of family, and married status of the mother, alluding to the notion that early life circumstances play a vital role in financial earnings later in life. This was further confirmed by the debt model. Two of the three most significant variables were also from wave 1, namely whether the participant was in an independent school and their mother's highest qualification.

II: EXPLORATORY ANALYSIS

Plots

We initially plotted variables that we believed would significantly affect the outcome. One key insight extracted from this dataset was that categorical variables with many levels were ordinal. Therefore, the boxplots showing increasing or decreasing patterns for these variables suggested potential significance. Another indicator of significance was when a predictor with multiple levels showed subgroups clustered around similar output values. For example, the boxplot for *W1wrk1aMP* illustrated a notable difference between levels, with those below 5 showing higher *W8DINCW* values and those above showing lower ones. Significance was also evident when one level of a categorical variable had a different output value than the other levels, while still having a decent sample size. For instance, the "white" ethnic group in *W1ethgrpYP* showed higher income levels (*W8DINCW*) than the rest. Finally, if a binary predictor showed a difference in output value between its levels, with both levels having a similar sample size, it showed potential significance. For example, both *W5JobYP* and *W1condur5MP* displayed notable differences in output values between levels. However, in *W5JobYP*, the sample sizes for each level were similar, whereas *W1condur5MP* had a big gap between the sample sizes of its levels. Consequently, the first is potentially significant while the latter may not be.

Merging Levels

We prioritised merging levels of variables that had more than 5. We checked boxplots and analysed which levels appeared to be close. Summary() of variables was checked to see the sample sizes of each value. We mostly merged levels with smaller counts with ones with bigger counts, as well as smaller ones with smaller ones. The final decision to merge levels depended on whether it increased the variable's significance, assessed through the F-statistic value in the Anova() function. Check Appendix 2.1 for merged levels.

Dealing with Missing Values

Continuous variables with more than 60% values missing, and those with more than 30% missing but non-significant, were removed from the dataset. The significant ones with between 30% and 60% missing values were converted to categorical variables with a "missing" level and their significance was assessed. Examples include the *W1GrssyrMP* variable in W8DINCW and *W1GrssyrHH* in W8QDEB2 – Appendix 2.3.

We removed categorical variables from the dataset with more than 90% missing values. Rows with less than 10% missing values in the W8DINCW dataset were deleted, reducing it from 5792 to 4035, while in the W8QDEB2 dataset, those with less than 5% missing values were deleted, reducing it from 2878 to 2608. This is because removing the rows with less than 10% missing for the W8QDEB2 dataset would result in 39.36% of the data being lost while with 5%, it was only 9.38%. Next, we converted all NAs in the categorical variables to a “missing” level and proceeded with the analysis. In our final model for W8DINCW, the only continuous predictors used were *W2ghq12scr* and *W8DGHQSC*, resulting in an additional 231 missing rows to be ignored, leaving 3804 rows. For W8QDEB2, the continuous variable *W1GrssyrHH* was converted to a categorical variable and *W6DebtattYP* did not have any missing values leading to no further rows being ignored.

Finally, we replaced this “missing level” with NAs and ran a separate model. When comparing the models, we focused on the diagnostics and the number of significant predictors. For W8DINCW, we observed that R^2 decreased from 0.75 to 0.66, the residual standard deviation increased from 35.92 to 36.87, and seven previously significant variables became non-significant. For the W8QDEB2 dataset, the changes included the residual standard deviation increasing from 28258.31 to 30401.28 and the variable *W6DebtattYP* becoming less significant. This step also decreased rows by 66.9%. This was done by excluding the *W1GrssyrHH* variable as it had 46.56% missing values. Hence, we chose our initial model with a “missing level” for both datasets.

Dealing with Multicollinearity

To address multicollinearity, we used the vif() function. However, encountering an error due to some variables being aliased, we turned to the aliased() function for identification. For aliased pairs of variables, we included either the more significant one or the easier-to-interpret one. Check Appendix 3 for details.

III: FROM INITIAL TO FINAL MODEL

Chronological Steps (Appendix 1)

First, we loaded datasets “W8DINCW.csv”, “W8QDEB2.csv" and checked their summaries. Secondly, we replaced all the negative values with NA. We then removed variables with more than 90% missing values: *W4Childck1YP* for both datasets, as well as *W6NEETAct* for W8DINCW and *W6Childliv* for W8QDEB2. Then, we plotted relevant predictors against the outcome variable, using boxplots for categorical and scatterplots for continuous ones. We identified categorical columns with less than 10% missing values and removed those values – Appendix 4. Next, we converted categorical variables to factors and changed all NAs to a “missing” level. After running our first model, we dealt with multicollinearity. We added an extra code to solve the aliased coefficient problem for the W8DINCW data – Appendix 3. Next, we releveled the baseline of categorical variables to the most common level, to better interpret the coefficients. We merged the levels of some categorical variables (Appendix 2.1) and tried converting some continuous variables into categorical variables (Appendix 2.3). The categorical *W1GrssyrHH* was kept for W8QDEB2, whereas none were kept in the income model. We removed *W1GrssyrMP* and

W1GrssyrHH from the entire W8DINCW dataset and compared both approaches, choosing the removed version – Appendix 5. We performed backward elimination of non-significant variables. After that, we checked any significant interactions and performed outlier analysis. Then, we applied a logarithmic transformation to the outcome variable in our final model and compared it with the one without it. For the W8QDEB2 dataset, the square root transformation was also tried. In the final model, we changed the "missing" level back to NA and compared them, choosing the initial "missing level" model – see Exploratory Analysis. Finally, we ran cross-validation.

Interactions

We first identified the three most significant variables for W8DINCW: *W1hiqualmum*, *W1nssecfam*, and *W1ethgrpYP*. We ran separate regressions for each pair and observed that *W1hiqualmum&W1nssecfam* and *W1nssecfam&W1ethgrpYP* had significant interactions. However, when we added them to the final model, they were nonsignificant, so we did not include them. The same methodology was applied to *W1_GrssyrHH_category, IndSchool,* and *W8TENURE* for W8QDEB2 with *W1_GrssyrHH_category&IndSchool* and *W1_GrssyrHH_category&W8TENURE* being significant interactions. When added to the final model, the model improved since the interactions remained significant.

Outliers

For the outlier analysis, we first looked at standardised residuals over 3, cooks distances, DFFITS, and the leverage values. In both datasets, we found no outliers after doing these tests, so we did not further investigate any outlying values.

Residual Plots & Transformations

Firstly, we looked at the residual vs. fitted plot and observed a funnel shape, indicating heteroscedasticity, where the variance of residuals changes with fitted values. Since the outcome variables, *W8DINCW* and *W8QDEB2,* represent income and debt respectively, which are non-negative, this funnel shape suggests that the model's assumption of constant variance is violated.

For W8DINCW, the QQ plot is close to a straight line and the Histogram of Residuals shows a close-to-normal distribution, indicating normally distributed errors. For W8QDEB2, the QQ plot indicated that, compared to a normal distribution, the residuals have a heavier tail, and the histogram of residuals showed positive skewness. Before applying the logarithm transformation, the zero values were replaced with a small positive number since log(0) produced errors.

For W8DINCW, applying a logarithmic transformation reduced the funnel shape in the Residual vs. Fitted Plot, with no significant change in other plots. Overall, the model's fit improved with the decrease in heteroscedasticity and improved diagnostics – Appendix 8. Hence, we decided to keep this transformation.

For W8QDEB2*,* the diagnostics worsened after the transformation, four predictors became nonsignificant, and the residual plots did not improve. So, we decided not to keep the transformation.

Cross-Validation

For cross-validation, we separated each dataset into two subsets where the training set comprised 90% of the data and the remaining 10% was used for prediction named the test set. The process produced two essential plots outlined in Appendix 9. The first plot, “Predicted vs Original Plot”, assessed our model’s ability to predict the target variable by comparing them to actual observations. For the W8DINCW dataset, the points are scattered around the diagonal line but not perfectly on it. This could indicate that the model captures the underlying patterns in the data but with some degree of error. For the W8QDEB2 dataset, three iterations were taken as one split ran the chance of being unrepresentative. Contrastingly, the points were clustered at zero on the y-axis along the x-axis suggesting the model’s predictions were unbiased on average however spread along the x-axis shows it performs differently with varying subsets of data.

The second graph, “Predicted vs Error Plot”, assessed the model’s fit. The W8QDEB2 plot showed that the model's predictions are unbiased, with errors concentrated between -e+05 and 0 on the y-axis. This indicated consistent performance across different predicted values, suggesting the model effectively captures underlying patterns in the data. However, for the W8DINCW dataset we can observe a decreasing pattern which implies that the residuals are not randomly distributed but instead vary systematically with the predicted value.

When did you decide to remove a predictor?

We removed variables with more than 90% missing values and the ones combined with other variables. We also removed some of them to eliminate aliased coefficients and high VIF values. Finally, we removed the rest using backward elimination for non-significant predictors, starting with the least significant ones and continuing until all variables left were significant. See Appendix 6.

IV: RESULTS

Table showing the Output for the final model of *W8DINCW*

Variable	Sum Sq	Df	F value	Pr(>F)	Significance
W1ethgrpYP	24.207	7	270.3652	< 2.2e-16	***
W1wrk1aMP	10.030	11	71.2920	< 2.2e-16	***
W1marstatmum	8.063	6	105.0682	< 2.2e-16	***
W1nssecfam	7.776	7	86.8456	< 2.2e-16	***
W1hiqualmum	4.458	15	23.2377	< 2.2e-16	***
W4CannTryYP	1.083	1	84.6446	< 2.2e-16	***
W1disabYP	0.999	2	39.0521	< 2.2e-16	***
W6UnivYP	0.910	1	71.1133	< 2.2e-16	***
W1heposs9YP	0.701	2	27.4061	1.581e-12	***
W6JobYP	0.549	1	42.9212	6.620e-11	***
W8DDEGP	0.484	2	18.9281	6.731e-09	***
W5Apprent1YP	0.451	1	35.2811	3.162e-09	***
W2ghq12scr	0.270	1	21.0759	4.585e-06	***
W5EducYP	0.246	1	19.2470	1.186e-05	***
W1hwndayYP	0.341	5	5.3244	7.060e-05	***
W2disc1YP	0.151	1	11.8124	0.0005959	***
W1hous12HH	0.192	3	5.0151	0.0018063	**
W8DACTIVITY	0.358	10	2.7984	0.0018753	**
W1hea2MP	0.131	1	10.2317	0.0013940	**
W8DGHQSC	0.058	1	4.5419	0.0331509	*
Residuals	40.763	3187			

<u>Table showing the Output for the final model of *W8QDEB2*</u>

Variable	Coefficient	P value	Significance	Overall Variable Significance Anova( )
(Intercept)	12920.66	0.006298	**	
W8TENURE	-1936.13	0.014246	*	***
IndSchool	4469.24	0.675360		***
W1GrssyrHH_category2	-7579.49	0.115331		***
W1GrssyrHH_category3	9847.87	0.082585		
W1GrssyrHH_category4	8895.51	0.237996		
W1GrssyrHH_category5	-41987.18	0.000369	***	
W1GrssyrHH_category6	29651.88	0.022337	*	
W1GrssyrHH_categoryMissing	460.35	0.912130		
father.qual2	-6249.37	0.000443	***	**
father.qualMissing	-6255.39	0.002028	**	
mother.qual2	4932.61	0.005048	**	**
W6DebtattYP	470.48	0.005163	**	**
new_depress3	3636.46	0.014035	*	*
new_depress4	-1629.35	0.381574		
W1GrssyrHH_category2:IndSchool	-2094.41	0.875839		***
W1GrssyrHH_category3:IndSchool	14583.95	0.296485		
W1GrssyrHH_category4:IndSchool	-11943.23	0.451040		
W1GrssyrHH_category5:IndSchool	86464.46	9.45e-09	***	
W1GrssyrHH_category6:IndSchool	9367.50	0.512961		
W1GrssyrHH_categoryMissing:IndSchool	1701.63	0.880980		
W1GrssyrHH_category2:W8TENURE	1174.31	0.267021		***
W1GrssyrHH_category3:W8TENURE	-1401.26	0.258810		
W1GrssyrHH_category4:W8TENURE	-1873.96	0.286445		

W1GrssyrHH_category5:W8TENURE	10564.85	1.83e-05	***	
W1GrssyrHH_category6:W8TENURE	-7744.91	0.013774	*	
W1GrssyrHH_categoryMissing:W8TENURE	-168.78	0.855640		

n = 2611, k = 27
residual sd = 27847.21, R-Squared = 0.08

The final model's predictors and their corresponding coefficients are shown in above tables in descending order of significance. Among the predictors influencing income, there were eight variables with equal significance.

Notably, a young person's ethnic group (*W1ethgrpYP)* was the most significant with a p-value of <2.2e-16. The associated coefficient indicated that with all other variables held constant; an individual income increased by £24.21. This implies that ethnic groups had an increasing effect at the baseline, defined as identifying as white. This finding aligns with the ethnicity pay gap where there is a disparity in the average pay between employees from a minority ethnic group compared to their white counterparts[1].

We can infer that the third to fifth most significant variables, which had an increasing effect on income prospects, focused on the mother's marital status and socioeconomic status in wave 1. An individual with a single mother who never married experienced a £8.06 increase in income. Further, the family NS-SEC Class, used to measure employment relations and conditions of occupation, had a coefficient of £7.78, with the baseline as Higher Managerial and professional occupation. This could be influenced by the highest qualification of the mother where a Higher Degree further increases income by £4.48.

The importance of wave 1 enforced the pivotal impact childhood circumstance has on adult poverty, defined as a household income less than 60% of the UK median[2] by the Office of National Statistics investigated and highlighted that individuals growing up in a workless household at age 14 were approximately 1.5 times more likely to experience poverty thereby influencing income prospects significantly.

This is mirrored in the debt analysis. Two of the three most significant variables were from wave 1: household gross income (*W1GrssyrHH)* and attendance to an independent school (*IndSchool).* However, the most significant was tenure (*W8TENURE)*, denoting the conditions in which the individual held their house. It is not surprising that this is the largest contributor to debt at the age of 25 given in 2014-15, the year before the interview, the majority of first-time buyers were aged 25-34[3]. Being a homeowner is a driver for debt accumulation.

V: COMMENTS ON THE DATA/ANALYSIS

Our method for dealing with missing values is detailed in the Exploratory Analysis section. When we tried to minimise the effects of missing values affecting the analysis, we observed that certain variables had more missing values than others. In the W8QDEB2 dataset, variables that indicated the young person's father's employment status, highest qualification, and full-time vs part-time employment had 20.60-26.41% missing values, while for mother's it was 2.85-4.79%. This could indicate a paternal non-response bias caused by societal pressure or cultural norms. The variable pertaining to gross annual salary in both datasets had over 40% of the data missing. This could also indicate a non-response bias due to privacy concerns or social desirability bias. Both of these non-response biases could indicate that the data is missing systematically, or Missing Not at Random (MNAR), which means certain groups were underrepresented like fathers with lower levels of education or people of a lower socio-economic status[4]. This could affect the validity of the analysis.

There were also counterintuitive results like in the W8DINCW dataset where the *W1hous12HH* variable showed that the income earned was less for people who owned their house outright compared to if it was bought on a mortgage or bank loan. This could indicate selection bias[5]. There could be other reasons why income earned was less for people who owned their house outright. An example is if they were semi-retired and working less compared to someone who has a bank loan and is working a more difficult job to pay off their loans. In this case, even though it is misleading, the fact that someone who owns a house has a lower income could be because of these other factors.

Overall, while we have attempted to handle the missing data to ensure our analysis is as representative as possible, there are still some limitations due to non-response and selection biases.

LAY REPORT

ST211 REPORT London School of Economics

"What are the key predictors that influence income and debt at age 25?"

28/04/2024 - In April 2024, the LSE Statistics Department published a report, led by a group of second-year students, analysing the factors that affected the weekly income and total debt of individuals at age 25. In this article, we will give a non-technical summary of this report.

The students used data from the Next Steps Study (NS), which focused on approximately 16,000 individuals in England born in the years 1989-90. At the start of the study, 2004, the participants were 14 years old. The cohort members were surveyed on an annual basis until 2010 and then one last time when they reached the age of 25 in 2015-16. The study was divided into "waves" from 1-8, each wave representing a year of the survey. Each survey contained questions about different aspects of their life, such as "number of siblings" in wave 1 and " whether the individual was currently at university" in wave 6.

Two separate datasets were derived from the larger NS dataset. Both datasets contained almost the same factors; however, each dataset had a specific outcome studied. The first study analysed the influence of different factors, such as the socio-economical class of their family, on their weekly income, whereas the second study looked at their total debt. Here, the focus was not on whether or not they were in debt, but it was on the amount of that debt. Both of these outcomes were measured in the final wave of the study when they were 25.

The students identified what factors had an effect on the outcomes and if they did, how strong it was compared to the other factors. Some factors had similar meanings, such as the data Number of A/A2/AS levels being studied and whether the individual was going to school or college during wave 6. Hence, only one of these was considered in the analysis.

The results showed that the young person's ethnic group was the most influential factor for the weekly income, followed by the socioeconomic class of the family (wave 1) and the married status of the mother (wave 1). On the other hand, the household income in the early stages of their lives had the biggest effect on their total debt at age 25. The model showed that the average weekly income was £319, the lowest was £129, and the highest was £491. The average debt was £10,610.80, the minimum was £0, and the maximum was £136,953.60.

To better understand the influence of certain factors on the weekly income and total debt, the students created three fictional individuals. These individuals had different backgrounds and career paths, but all other personal circumstances, such as their health conditions and the marital status of their mother, were the same. For all fictional personas –James, Tania, Molly, and Sarah–, the students estimated weekly income and debt. Overall, James had the highest income and Molly had the highest debt. From this, we can see that different personal circumstances affect the outcomes. For example, Tania's ethnicity is associated with a decrease in her income of around £85. Interestingly, Tania's family having semi-routine occupations decreases Tania's income by £45, while this number is only £1.6 for other individuals with different socioeconomic classes.

Furthermore, over half of the influential factors from both datasets were from wave 1, illustrating the importance of the early stages of people's lives. These models suggest that the government has to support certain groups of people from an early age. This applies to children of single mothers, as well as minority ethnic groups. The report suggested that the government taking early action could have a significant impact on income and debt at age 25, providing a positive socio-economic change in the long run.

Although it was concluded that the models were good enough for decision-making, it is important to note some flaws. A limitation of the income model is that for certain factors, such as tenure in wave 8, there may be some biases due to respondent errors. The model suggested that owning a house outright had a negative impact on weekly income whilst owning one with the help of a mortgage had a positive one. This may result from individuals giving a socially acceptable response rather than the truth. In the future, it may be useful to research tenure further to see its impact on the weekly income. A surprising finding in the debt analysis suggested that if the individual grew up in a household with a higher income, they are more likely to have a higher debt. This could be explained by the fact that the individuals may have gotten used to living with a disposable income so they then used loans to fund their adult spending. They may have also gone to university so this could account for a large proportion of debt.

In conclusion, the LSE Statistics Department's report on weekly income and total debt offers valuable insights into the factors influencing these outcomes. The report highlights how different attributes and circumstances at different stages of life affect the income and debt of individuals.

PROFILES

Name	Ethnic Group	Family NS-SEC Class (wave 1)	Highest level of education (wave 5)	Weekly Income
 **Tania**	Bangladeshi	Semi-routine occupations	No degree	£325
 **James**	White	Higher Managerial and professional occupations	Has a Degree	£420

Name	Tenure	Independent School?	Household Income (wave 1)	Total Debt
 **Sarah**	Own - buying with help of mortgage or loan	Yes	Up to £40000	£6,954.271
 **Molly**	Rent free incl friends and family excl squatting	Yes	Up to £100000	£34,927.52

VII: APPENDIX

Appendix 1: Chronological Step of Building the Model

	Methodology
1	Loading and observing the data
2	Changing negative values with NA and removing more than 90% variables from the dataset
3	Plotting relevant variables
4	Removing all missing values from categorical predictors with less than 10% missing values
5	Converting all categorical variables to factor
6	Changing categorical variables' NAs to a "missing" level
7	Running initial model and solving multicollinearity problems
8	Re-levelling the baseline for categorical variables
9	Merging levels in categorical variables to make it more significant
10	Converting continuous W1GrssyrMP (W8DINCW dataset) and W1GrssyrHH (W8QDEB2 dataset) to categorical
11	Removing variables W1GrssyrMP and W1GrssyrHH from the entire dataset for W8DINCW dataset (Continuous with over 30% missing)
12	Backward elimination of non-significant variables
13	Checking interactions
14	Performing outlier analysis
15	Observing residual plots and trying out transformations
16	Replacing all "missing level" with NA and compare models
17	Performing cross-validation and analysing results

Appendix 2: Data Manipulation Details

Table 2.1 W8DINCW Merged Levels

Variable name	Merged Levels	F-statistic	Included in model?	Notes
W2depressYP	2&3	0.0868293	No	Increased significance but still not significant at 5% level

W4empsYP	1&2&5	0.314648	No	Increased significance but still not significant at 5% level
W6acqno	1&2, 3&4, 5&6, 7&8	0.038855 *	Yes	Non-significant after merging, became significant after backward elimination
W8DACTIVITY	4&6, 7&8, 9&10	0.005476 **	Yes	Merging made it significant initially
W8QMAFI	1&2, 4&5	0.073684	No	Merging made it significant initially
W1heposs9YP	1&2	4.030e-09 ***	Yes	Already significant, merging increased significance
W1hous12HH	1&2, 4&5&7, 3&8	0.001867 **	Yes	Merging decreased 0.22 to 0.07, backward elimination further increased
W1hiqualdad	1&2&3&4&5, 10&11, 14&16, 18&20	0.015	No	F-statistic increased during backward elimination
W1hiqualmum	2&3, 5&6&7, 9&10, 16&17&18	2.2e-16 ***	Yes	Already 2.2e^-16, so no noticeable change
W1wrk1aMP	6-11&10, 5&4, 2&3	2.2e-16 ***	Yes	Already 2.2e^-16, so no noticeable change

Table 2.2 W8QDEB2 Merged Levels

Variable name	**Merged Levels**	**F-statistic**	**Included in model?**	**Notes**
W1wrk1aMP	1-4, 5-12	0.8600925	No	Increased significance but still not significant at 5% level
W1disabYP	1-2	0.5011349	No	Made less significant
W1NoldBroHS	0-3, 4-7	0.3577073	No	Increased significance but still not significant at 5% level
W1depkids	1-5, 6-10	0.9729986	No	Made less significant

W1InCarHH	2-4	0.9644116	No	Increased significance but still not significant at 5% level
W1hous12HH	1-2, 4-8	0.8315373	No	Made less significant
W1hiqualmum	1-2, 3-20	0.0006183 ***	Yes	Increased significance from 0.3529595
W1hiqualdad	1-3, 4-20	0.0208460 *	Yes	Increased significance from 0.2642252
W1empsmum	1-2, 3-9	0.8546385	No	Increased significance but still not significant at 5% level
W1empsdad	1-3, 4-9	0.8350437	No	Increased significance but still not significant at 5% level
W1marstatmum	1&3-7	0.5764986	No	Increased significance but still not significant at 5% level
W1nssecfam	1-2, 3-4, 5-8	0.1822698	No	Increased significance but still not significant at 5% level
W1ethgrpYP	3-5, 6-7	0.2921471	No	Made less significant
W1heposs9YP	1-2, 3-4	0.1289891	No	Increased significance but still not significant at 5% level
W2ghq12scr	1-4, 5-8, 9-12	0.1627755	No	Increased significance but still not significant at 5% level
W2depressYP	1-2	0.2127400	Yes	Increased significance but still not significant at 5% level with merging. After backwards elimination it was significant
W4AlcFreqYP	4-6	0.6267828	No	Increased significance but still not significant at 5% level
W4empsYP	1-2, 3-5, 6-8	0.9127333	No	Made less significant
W6acqno	1-2, 3-4, 5-6, 7-8	0.7663742	No	Made less significant
W6gcse	1-2, 3-4	0.7533605	No	Made less significant
W6als	1-3	0.9360043	No	Made less significant

W8DMARSTAT	3-4, 7-9	0.2765299	No	Increased significance but still not significant at 5% level
W8DACTIVITY	1-4&12, 5-8, 9-10	0.7006620	No	Increased significance but still not significant at 5% level

2.3 Converting Continuous Variables to Categorical

W1NoldBroHS: (W8DINCW)
This variable shows the number of younger siblings that the young person has. It was previously a continuous, numerical variable; however, we changed it to a categorical variable with levels 0, 1, 2, 3+. Considering that there is very little data of someone having more than 3 siblings, we merged all the values more than 2 under 3+. This change has not changed the significance level much, but it made it much easier to interpret.

W1depkids (to binary): *(W8DINCW)*
This variable shows the number of dependent children in household in Stage 1. We tried various manipulations such as merging levels over 2, but the highest increase in significance was seen when we turned this numerical variable into a binary variable with 1 indicating having more than one child in household and 0 indicating having just one child. Since we're looking at the child's household, it is guaranteed to have one child for this variable in this dataset.

W1GrssyrMP to *W1GrssyrMP_category*: *(W8DINCW)*
W1GrssyrMP was a significant continuous variable with about 44.82% missing values. To retain its importance while minimising the issue of high percentage of missing values, we tried to convert it to a categorical variable. New variable W1GrssyrMP_category has 6 levels including a "missing level", income less than 5K, 5-10K, 10K-15K, 15-20K, and +20K. Other level combinations have also been tried, but this appeared to produce the most significant result. Even though this categorical variable was significant in the initial model, including all variables, it became non-significant in the final model. As a result, it was removed from the model.

W1GrssyrHH to *W1GrssyrHH_category*: *(W8QDEB2)*
Similar to above, W1GrssyrHH was a significant continuous variable with 46.56% missing values. We converted it to a categorical variable which had the following levels, less than 20k, 20k-40k, 40k-60k, 60k-80k, 80k-100k, "Missing" level. Unlike the previous dataset, this variable remained significant throughout so it was kept in the final model.

2.4 Combining Predictors

W6als and *W6EducYP*:

These two variables W6als and W6EducYP have a very similar meaning. The first one shows the number of A Levels studied at Wave 6, whereas the second one is a binary variable showing if the person is going to school at Wave 6. Since at this age going to school means mostly indicates that the person is studying A-levels, the variables share a very similar meaning. Observing that both variables cause the aliased coefficient error, we inferred that these variables are highly collinear with each other. We further observed that W6EducYP is more significant, however it has more missing values than W6als, which has zero missing values. Therefore, we decided to combine these predictors instead of removing one from the model. To perform this, we first converted A Levels into a binary variable, named “new_W6als”, taking 0 if none A Levels are studied, and 1 if any number of A Levels are studied. We then created the new combined binary variable which takes 1 if the person either studied A levels or went to a school, and it will take 0 if the person did not do either. The new variable “*combinedvar*” solved the aliased coefficient problem and increased significance.

Appendix 3: Multicollinearity

We observed that *W6Apprent1YP* was highly aliased with *W6UnivYP* in the *W8DINCW* dataset. Both of their significance was checked within the model, and only the more significant one, *W6Apprent1YP*, was included in the model. For both datasets, *W1empsmum* was aliased with *W1wrkfullmum* and *W1empsdad* was aliased with *W1wrkfulldad*. Since *W1empsmum* and *W1empsdad* had 8 levels, whereas the other two had 3 levels; therefore, only *W1wrkfullmum* and *W1wrkfulldad* were included to make the model simpler and more interpretable. Following the same reasoning, W1depkids was kept while *W1ch0_2HH*, *W1ch3_11HH*, *W1ch12_15HH*, and *W1ch16_17HH* were removed, as the first variable represents the sum of the other four. Appendix 2.1 gives the explanation for the combined predictors W6als and W6EducYP under *combinedvar. W1famtyp2* was also removed since it had a VIF of 7, whereas all the other ones had around 1 in the *W8DINCW* dataset. For the *W8QDEB2* dataset, *W6EducYP* and *W6UnivYP* were aliased so *W6EducYP* was removed as it had over 30% missing rows. There were also multiple variables with high VIF over 10 that were removed as seen in the appendix 6.2. In some instances by removing one high VIF variable, another variable that was potentially higher correlated VIF was reduced such as *W6ApprentYP* and *W6UnivYP.*

Appendix 4: Step 4 Explanation

At this point in the process we only excluded continuous predictors while removing categorical variables with less than 10% missing value, but after we reached our final model, we went back to this step and excluded all the non-significant predictors which has not been used in our model to use as much data as possible. At first the data was reduced from 5792 to 3267 (about 44% decrease), whereas when we excluded all non-significant ones, it reduced to only 4035 (about 30% decrease). Continuous predictors were identified from the data dictionary and excluded manually. For the W8QDEB2 dataset, the data was first reduced from 2878 to 1745 but after excluding non-significant ones, it only reduced to 2608.

Appendix 5: Step 11 Explanation

Removing W1GrssyrHH and W1GrssyrMP from the W8DINCW dataset made the model better since it both changed some previously non-significant variables to significant (*W1hea2MP*, *W1hous12HH*, *W1usevcHH*, *W6Apprent1YP*, *W8TENURE*, *new_W6acqno*) and improved the model diagnostics, such as the R^2 increasing from 0.67 to 0.75 and decreasing the residual standard deviation from 37.02 to 35.58. For the W8QDEB2 dataset, we did not remove the W1GrssyrHH variable as it reduced R^2 from 0.05 to 0.04 and made the W1hiqualmum variable less significant.

Appendix 6: Removed Variables and Reasonings

Table 6.1 Removed variables for *W8DINCW*

Removed Variable Name	Reasoning for Removal
W4Childck1YP	90%+ missing values
W6NEETAct	90%+ missing values
W6Childliv	90%+ missing values
W1ch0_2HH	Aliased with W1depkids & for easier interpretability
W1ch3_11HH	Aliased with W1depkids & for easier interpretability
W1ch12_15HH	Aliased with W1depkids & for easier interpretability
W1ch16_17HH	Aliased with W1depkids & for easier interpretability
W1empsmum	Aliased with W1wrkfullmum
W1empsdad	Aliased with W1wrkfulldad
W6UnivYP	Aliased with W6Apprent1YP

W6als	Combining predictors (combinedvar)
new_W6als	Combining predictors (combinedvar)
W6EducYP	Combining predictors (combinedvar)
W1famtyp2	High VIF (7)
W8DMARSTAT	Backward elimination (non-significant)
W1GrssyrMP_category	Backward elimination (non-significant)
W8QMAFI	Backward elimination (non-significant)
W1alceverYP	Backward elimination (non-significant)
W1hiqualdad	Backward elimination (non-significant)
W6OwnchiDV	Backward elimination (non-significant)
W4schatYP	Backward elimination (non-significant)
combinedvar	Backward elimination (non-significant)
W1depkids	Backward elimination (non-significant)
W1bulrc	Backward elimination (non-significant)
W4NamesYP	Backward elimination (non-significant)
W1truantYP	Backward elimination (non-significant)
IndSchool	Backward elimination (non-significant)
W4empsYP	Backward elimination (non-significant)
W6gcse	Backward elimination (non-significant)
W2depressYP	Backward elimination (non-significant)
W1yschat1	Backward elimination (non-significant)
W4RacismYP	Backward elimination (non-significant)
W1condur5MP	Backward elimination (non-significant)
W1wrkfullmum	Backward elimination (non-significant)
W1InCarHH	Backward elimination (non-significant)
W6DebtattYP	Backward elimination (non-significant)
W1NoldBroHS	Backward elimination (non-significant)
W4AlcFreqYP	Backward elimination (non-significant)
W1wrkfulldad	Backward elimination (non-significant)

Table 6.2 showing the removed variables for*W8QDEB2*

Removed Variable Name	**Reasoning for Removal**
W4Childck1YP	90%+ missing values
W6Childliv	90%+ missing values
W1empsmum	Aliased with W1wrkfullmum
W1empsdad	Aliased with W1wrkfulldad
W6EducYP	Aliased with W6UnivYP
W1NoldBroHS	High VIF
W1wrkfullmom	High VIF
W1famtyp2	High VIF

W1yschat1	High VIF
W6Apprent1YP	High VIF
W6acqno	High VIF
W6als	High VIF
W4schatYP	Backward elimination (non-significant)
W6UnivYP	Backward elimination (non-significant)
W1wrk1aMP	Backward elimination (non-significant)
W1condur5MP	Backward elimination (non-significant)
W1hea2MP	Backward elimination (non-significant)
W1disabYP	Backward elimination (non-significant)
W1depkids	Backward elimination (non-significant)
W1InCarHH	Backward elimination (non-significant)
W1hous12HH	Backward elimination (non-significant)
W1usevcHH	Backward elimination (non-significant)
W1wrkfulldad	Backward elimination (non-significant)
W1marstatmum	Backward elimination (non-significant)
W1nssecfam	Backward elimination (non-significant)
W1ethgrpYP	Backward elimination (non-significant)
W1heposs9YP	Backward elimination (non-significant)
W1hwndayYP	Backward elimination (non-significant)
W1truantYP	Backward elimination (non-significant)
W1alceverYP	Backward elimination (non-significant)
W1bulrc	Backward elimination (non-significant)
W2ghq12scr	Backward elimination (non-significant)
W2disc1YP	Backward elimination (non-significant)
W4AlcFreqYP	Backward elimination (non-significant)
W4CannTryYP	Backward elimination (non-significant)
W4NamesYP	Backward elimination (non-significant)
W4RacismYP	Backward elimination (non-significant)
W4empsYP	Backward elimination (non-significant)
W5JobYP	Backward elimination (non-significant)
W5EducYP	Backward elimination (non-significant)
W5Apprent1YP	Backward elimination (non-significant)
W6JobYP	Backward elimination (non-significant)
W6gcse	Backward elimination (non-significant)
W6OwnchiDV	Backward elimination (non-significant)
W8DGHQSC	Backward elimination (non-significant)
W8DMARSTAT	Backward elimination (non-significant)
W8DACTIVITY	Backward elimination (non-significant)

Appendix 7: Final Model Results

Table 7.1 Output for the final model of *W8DINCW*

Variable	Sum Sq	Df	F value	Pr(>F)	Significance
W1ethgrpYP	24.207	7	270.3652	< 2.2e-16	***
W1wrk1aMP	10.030	11	71.2920	< 2.2e-16	***
W1marstatmum	8.063	6	105.0682	< 2.2e-16	***
W1nssecfam	7.776	7	86.8456	< 2.2e-16	***
W1hiqualmum	4.458	15	23.2377	< 2.2e-16	***
W4CannTryYP	1.083	1	84.6446	< 2.2e-16	***
W1disabYP	0.999	2	39.0521	< 2.2e-16	***
W6UnivYP	0.910	1	71.1133	< 2.2e-16	***
W1heposs9YP	0.701	2	27.4061	1.581e-12	***
W6JobYP	0.549	1	42.9212	6.620e-11	***
W8DDEGP	0.484	2	18.9281	6.731e-09	***
W5Apprent1YP	0.451	1	35.2811	3.162e-09	***
W2ghq12scr	0.270	1	21.0759	4.585e-06	***
W5EducYP	0.246	1	19.2470	1.186e-05	***
W1hwndayYP	0.341	5	5.3244	7.060e-05	***
W2disc1YP	0.151	1	11.8124	0.0005959	***
W1hous12HH	0.192	3	5.0151	0.0018063	**
W8DACTIVITY	0.358	10	2.7984	0.0018753	**
W1hea2MP	0.131	1	10.2317	0.0013940	**
W8DGHQSC	0.058	1	4.5419	0.0331509	*

Residuals	40.763	3187			

Table 7.2 Output for the final model of *W8QDEB2*

Variable	Coefficient	P value	Significance	Overall Variable Significance (Anova ( ))
(Intercept)	12920.66	0.006298	**	
W8TENURE	-1936.13	0.014246	*	***
IndSchool	4469.24	0.675360		***
W1GrssyrHH_category2	-7579.49	0.115331		***
W1GrssyrHH_category3	9847.87	0.082585		
W1GrssyrHH_category4	8895.51	0.237996		
W1GrssyrHH_category5	-41987.18	0.000369	***	
W1GrssyrHH_category6	29651.88	0.022337	*	
W1GrssyrHH_categoryMissing	460.35	0.912130		
father.qual2	-6249.37	0.000443	***	**
father.qualMissing	-6255.39	0.002028	**	
mother.qual2	4932.61	0.005048	**	**
W6DebtattYP	470.48	0.005163	**	**
new_depress3	3636.46	0.014035	*	*
new_depress4	-1629.35	0.381574		
W1GrssyrHH_category2:IndSchool	-2094.41	0.875839		***
W1GrssyrHH_category3:IndSchool	14583.95	0.296485		
W1GrssyrHH_category4:IndSchool	-11943.23	0.451040		
W1GrssyrHH_category5:IndSchool	86464.46	9.45e-09	***	
W1GrssyrHH_category6:IndSchool	9367.50	0.512961		
W1GrssyrHH_categoryMissing:IndSchool	1701.63	0.880980		
W1GrssyrHH_category2:W8TENURE	1174.31	0.267021		***

W1GrssyrHH_category3:W8TENURE	**-1401.26**	**0.258810**		
W1GrssyrHH_category4:W8TENURE	**-1873.96**	**0.286445**		
W1GrssyrHH_category5:W8TENURE	**10564.85**	**1.83e-05**	***	
W1GrssyrHH_category6:W8TENURE	**-7744.91**	**0.013774**	*	
W1GrssyrHH_categoryMissing:W8TENURE	**-168.78**	**0.855640**		

n = 2611, k = 27
residual sd = 27847.21, R-Squared = 0.08

Appendix 8: Diagnostics

Table 8.1 Diagnostics for W8DINCW before and after log transformation

	A) Before log transformation	**B) After log transformation**
Residual standard error:	35.99 on 3719 degrees of freedom	0.1131 on 3187 degrees of freedom
Multiple R-squared:	0.7481	0.777
Adjusted R-squared:	0.7424	0.7715
F-statistic:	131.5 on 84 and 3719 DF	140.6 on 79 and 3187 DF,
p-value	< 2.2e-16	< 2.2e-16

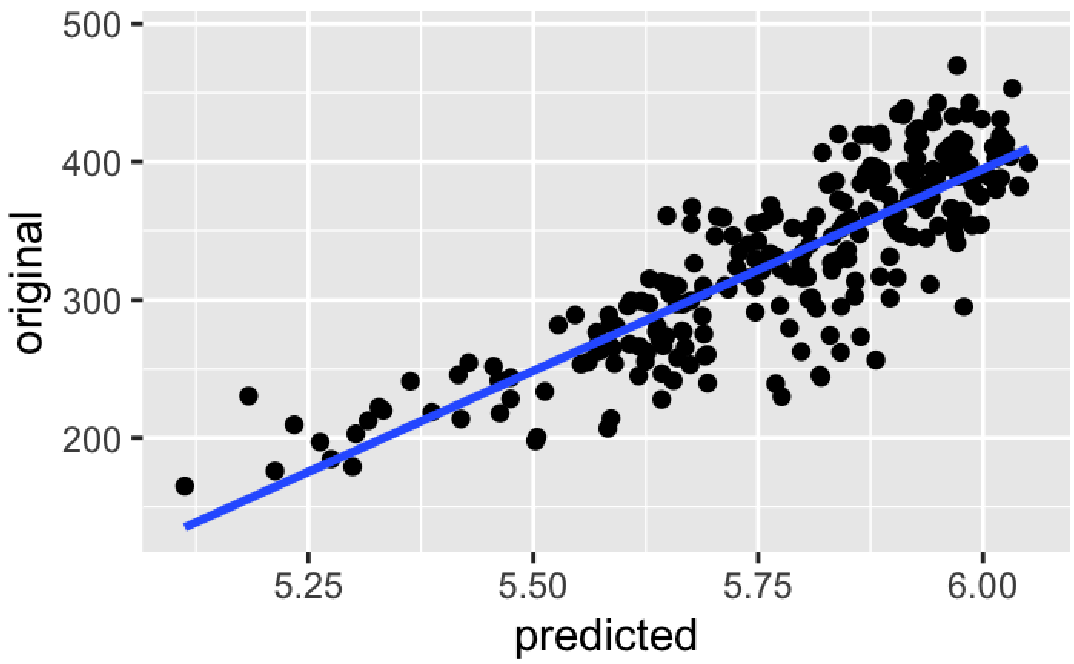
Table 8.2 Diagnostics for W8QDEB2 before and after log transformation

	A) Before log transformation	**B) After log transformation**
Residual standard error:	27850 on 2584 degrees of freedom	2.196 on 2584 degrees of freedom
Multiple R-squared:	0.08298	0.02056
Adjusted R-squared:	0.07375	0.0107
F-statistic:	8.993 on 26 and 2584 DF	2.086 on 26 and 2584 DF,
p-value	< 2.2e-16	0.00104

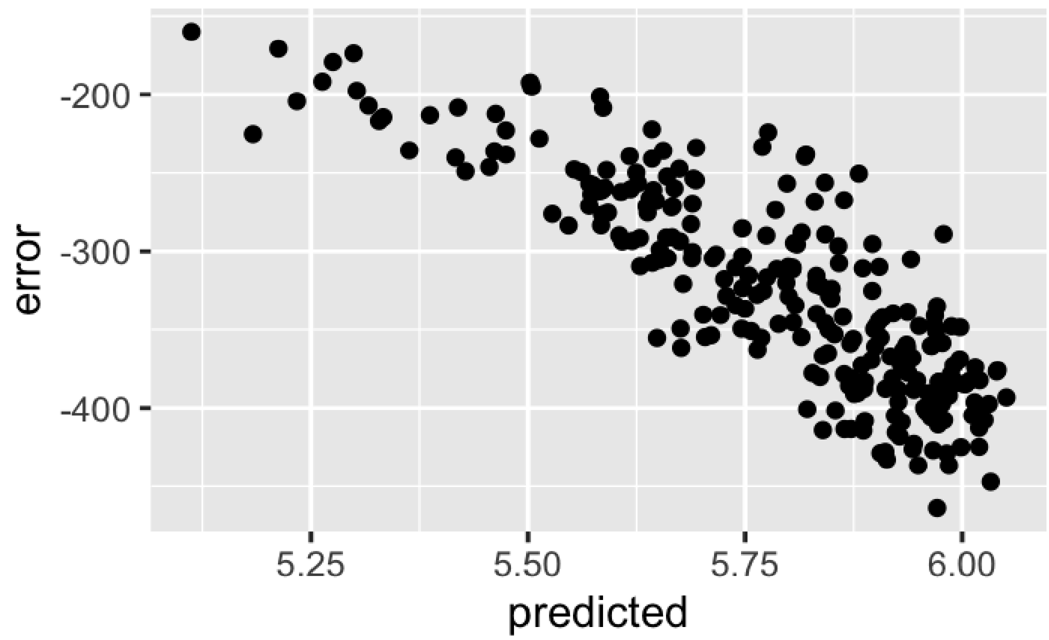
Appendix 9: Cross Validation Plots

9.1 Cross Validation Plots for W8DINCW

Graph 9.1.1 Predicted vs Original Points



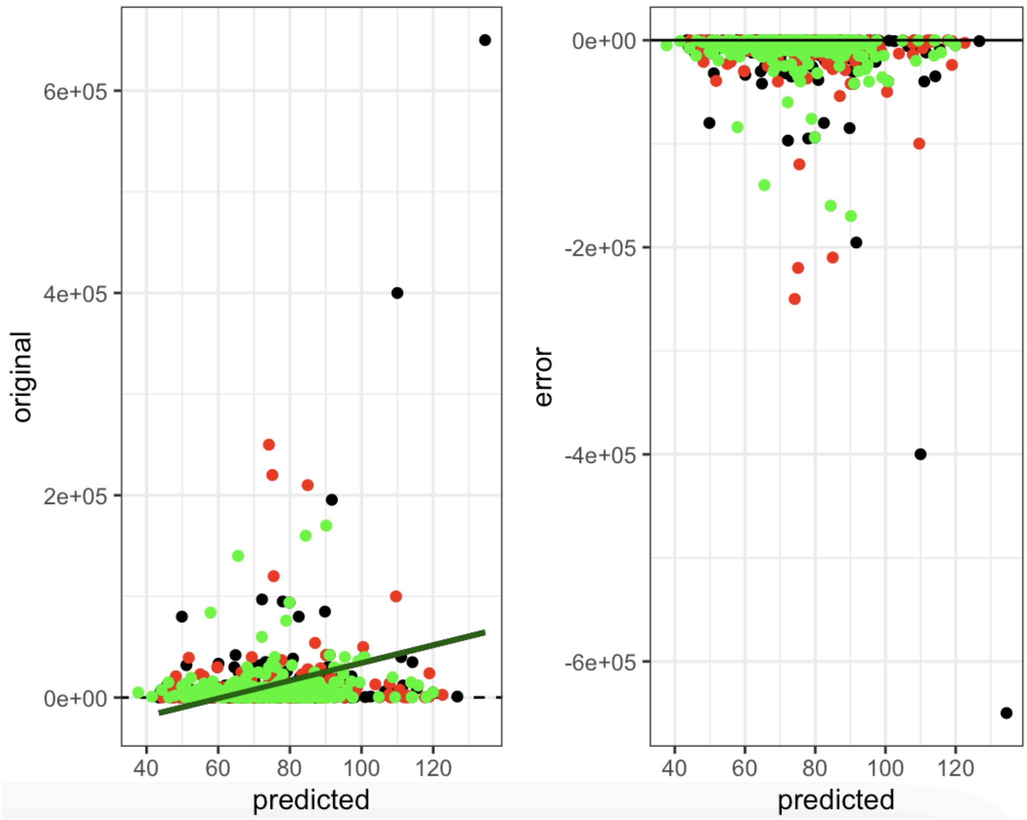
Graph 9.2.2 Predicted Points vs. the Prediction Error Associated with These Points



9.2 Cross Validation Plots for W8QDEB2

Graph 9.2.1 Predicted vs Original Points (Left)

Graph 9.2.2 Predicted Points vs. the Prediction Error Associated with These Points (Right)



Appendix 10: References

1. CMA. Ethnicity pay gap report: 1 April 2022 to 31 March 2023 [Internet]. 2023 [cited 2024 Apr 24]. Available from: https://www.gov.uk/government/publications/ethnicity-pay-gap-report-2022-to-2023/ethnicity-pay-gap-report-1-april-2022-to-31-march-2023#:~:text=The%20ethnicity%20pay%20gap%20shows,from%20an%20ethnic%20minority%20group.

2. ONS. How do childhood circumstances affect your chances of poverty as an adult? [Internet]. Office for National Statistics; 2016 [cited 2024 Apr 27]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/articles/howdochildhoodcircumstancesaffectyourchancesofpovertyasanadult/2016-05-16#:~:text=Childhood%20predictors%20of%20future%20poverty&text=Growing%20up%20in%20a%20workless%20household%20also%20appears%20to%20have,where%20one%20adult%20was%20working

3. Department for Communities and Local Government. English housing survey [Internet]. 2015 [cited 2024 Apr 27]. Available from: https://assets.publishing.service.gov.uk/media/5a82461bed915d74e6236b5d/First_Time_Buyers_report.pdf

4. Mack C, Su Z, Westreich D. Types of missing data [Internet]. U.S. National Library of Medicine; 2018 [cited 2024 Apr 26]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK493614/#:~:text=Missing%20not%20at%20random%20(MNAR,not%20measured%20by%20the%20researcher

5. Selection bias [Internet]. 2014 [cited 2024 Apr 27]. Available from: https://www.iwh.on.ca/what-researchers-mean-by/selection-bias