ST211 PROJECT

Candidate number 38772

Table of contents

How is the financial position of an individual at the age of 25 impacted by their ethnicity, education, socio-economic status of the family and health ?

Introduction

In this report, I analysed a subset of 5800 individuals taken from the Longitudinal Study of Young People in England (LSYPE). The Next Steps Study (NS), previously known as the LSYPE is a large scale survey following the lives of around 16,000 people born in 1989-90 in England. This study consists of 8 waves and the principal aim of the LSYPE was to advice educational policy making.

For this report I focus on how the financial position of an individual at the age of 25 is impacted by their ethnicity, education, socio-economic status of the family and health. The measures that I picked to reflect their financial position are their continuous weekly income (W8DINCW), total amount owed (W8QDEB2) and how they are managing financially (W8QMAFI). I conducted multiple linear regressions to predict W8DINCW, W8QDEB2 and a logistic regression to predict W8QMAFI.

After performing an exploratory analysis, I created the initial models, analysed the results obtained and developed my final models. I then drew my conclusions to the research question based on the same.

The results from the regression models indicate that education and socio-economic status of the family are expected to impact all three measures (W8DINCW, W8QDEB2 & W8QMAFI) picked to represent financial position. However, ethnicity of an individual is only expected to impact the continuous weekly income at 25 years (W8DINCW). Interestingly physical health was predicted to have some impact on the expected continuous weekly income at 25 years (W8DINCW) whereas mental health is predicted to impact how an individual is expected to manage financially (W8QMAFI). However, these results may not be reflective of the population of young people in England, as the sample considered was small and therefore unrepresentative.

Exploratory Analysis

I first familiarised myself with the data and observed that there were some variables that should be excluded from the models. I deleted NSID as this wasn't a predictor. I deleted W4Childck1YP, W6Childliv, W6NEETAct, W8DAGEYCH and W8PUSA from the dataset as these variables had less than 30% of the data and I wanted a larger, more representative sample.

Next, I examined the expected relationships between the predictors and the outcomes. I considered the expected sign and strength of the predictors as well as the expected correlation among predictors. Firstly I expected variables W1nssecfam **(**Family's NS-SEC class in wave 1), W1ethgrpYP (YP's ethnic group), W1disabYP (Whether YP has a disability/long term illness or health problem), W6EducYP (Whether YP was attending school or college in wave 6), W8DDEGP (Whether YP achieved a first degree or higher), W8DGHQSC **(**General Health

Questionnaire score), W8DWRK (Whether CM is employed in wave 8) and W8CMSEX (Gender) to have a significant impact on the continuous weekly income of an individual in wave 8 (W8DINCW). Secondly, I expected W1disabYP (Whether YP has a disability/long term illness or health problem), W6DebtattYP (YP's attitude to debt), W8DDEGP (Whether achieved a first degree or higher) and W8TENURE (Tenure) to have the largest impact on the total amount owed by a YP in wave 8 (W8QDEB2). Lastly, I expected W1disabYP (Whether YP has a disability/long term illness or health problem), W8DWRK (Whether CM is employed in wave 8), W8NETA (Last take-home pay (net)), W8DINCW and W8QDEB2 to directly impact how an individual is managing financially in wave 8 (W8QMAFI).

I then plotted all the variables against the outcomes. I did this by creating scatterplots for the continuous predictors and boxplots for the categorical variables. The aim of this was to observe the relationships between the outcomes and the predictors.

For the regression model predicting W8DINCW, I deleted W8GROW, W8NETW and W8NETA as these are measures of income and therefore should not be included as predictors. I also removed W8QDEB2 as I felt that it wouldn't predict W8DINCW. Further W8QDEB2 had over 2000 missing values. I then reformatted the data as follows. For continuous predictors I deleted all the rows that included missing values because it distorted the dataset. I always made it a point to add back such deleted rows if the continuous predictor wasn't significant at a 5% level. This resulted in a larger and more representative dataset. When dealing with categorical predictors, I merged the missing values to form a separate level. This made the missing values easier to analyze later. Afterwards, I centred the continuous predictors where necessary. I centred W1GrssyrMP (cent.W1GrssyrMP) and W1GrssyrHH (cent.W1GrssyrHH) as their lowest values were £250 and £260 respectively. Next, I coded the categorical predictors as factors. Then I changed the baseline for categorical predictors where necessary. The baseline for all categorical predictors was set to its modal level (most frequently occurring level). This was the case in all predictors but W1hiqualdad as its modal level was "missing". Therefore, the second most frequent level was set as a reference level for W1hiqualdad.

When creating my initial model predicting W8QDEB2, I first reformatted the data as mentioned above. I dealt with missing values, centred continuous predictors when necessary and set reference levels for categorical predictors using the same criteria as before.

W8QMAFI was a categorical predictor with 6 levels. I initially removed the missing data for W8QMAFI and then converted W8QMAFI into a binary variable. I grouped the "Living comfortably" and "Doing alright" categories together into a "Well" category (denoted by 1 in the model) and grouped the other levels together in a "Poorly" category (denoted by 0 in the model). I then deleted predictors W1GrssyrMP, W8DACTIVITY, W8GROW and W8NETW as they heavily overlap with other predictors. I then dealt with missing values, centred continuous predictors when necessary and set reference levels for categorical predictors using the same criteria as before.

The statistical significance of regression coefficients was the main criterion used to determine whether to remove a predictor or not. The initial regression models included all the predictors as I was hesitant to discard any predictors based on my instincts.

I didn't observe any dominant predictors in any of the three models. Therefore, I didn't add any interactions in any of the models.

I merged levels of predictors primarily based on the frequency of the levels and the boxplots corresponding to the levels. If two levels of a predictor had a low frequency and their boxplots were similar I merged them. I also considered the significance of estimated coefficients of the levels of predictors as a secondary factor. i.e. I only merged levels with estimated coefficients that were not significant at a 5% level.

I have also conducted an outlier and residual analysis for the final models predicting W8DINCW and W8QDEB2. This will be discussed at great length in the section explaining initial to final models.

From Initial to Final Models

W8DINCW (Continuous weekly income)

My initial model predicting the continuous weekly income of individuals in wave 8 had all 62 predictors whereas the fifth and final model (coloured in green in Table 1) included 15 predictors. Goodness of fit statistics for all models are shown in Table 1. Coefficients with asterisks are significant at the 5% level. The R^2 and Adjusted R^2 were above 0.65 for all the model and this signifies that all the models were satisfactory. The Adjusted R^2 was also within 1% of the R^2 for all models but the first. (For the first model it was within 4.03%). This tells us that there weren't many non-significant predictors in the models. The F-statistic was highly significant in all models. A highly significant F-statistic indicates an overall good fit of the model. The standard error of the models should be compared with the width of the range of the outcome variable.

Table 1: W8DINCW- Initial to final model

	Model 1 (W8DINCW.lm.1)	**Model 2 (W8DINCW.lm.2)**	**Model 3 (W8DINCW.lm.3)**	**Model 4 (W8DINCW.lm.4)**	**Model 5 (W8DINCW.lm.5)**
Coefficients					
W1hea2MP	*	*	***	***	***
W1hous12HH	*	*	***	***	***
W1hiqualmum	***	***	***	***	***
W1nssecfam	***	***	***	***	***
W1ethgrpYP	***	***	***	***	***
W1heposs9YP	**	***	***	***	***
W1hwndayYP	**	***	***	***	***
W1disabYP	***	***	***	***	***
W2disc1YP	*	**	*	*	*
W4AlcFreqYP	*	***	***	***	***
W4CannTryYP	***	***	***	***	***
W5EducYP	***	***	***	***	***
W6Apprent1YP	***	***	***	***	***
W8DDEGP	***	***	***	***	***
W8CMSEX	***	***	***	***	***
R2/Adj R2	0.7371 / 0.7074	0.7873 / 0.7833	0.6884 / 0.6842	0.6844 / 0.6814	0.7048 / 0.7019
F-statistic significance	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
Std Error	34.53	33.96	41	41.18	0.1373
Comments					
Range of outcome	156.2-480.0	115.51-491.71	115.51-491.71	115.51-491.71	4.749357 - 6.197889
Number of data points in the dataset used to create the model	2354	5791	5791	5791	5791
Number of predictors in the model	62	19	15	15	15

After running the initial model, I eliminated 43 predictors that weren't significant at the 5% level. It was important to consider that none of the continuous predictors were significant at the 5% level as observed in the exploratory analysis. Therefore, I added back the previously deleted rows to increase the dataset and ran the second regression.

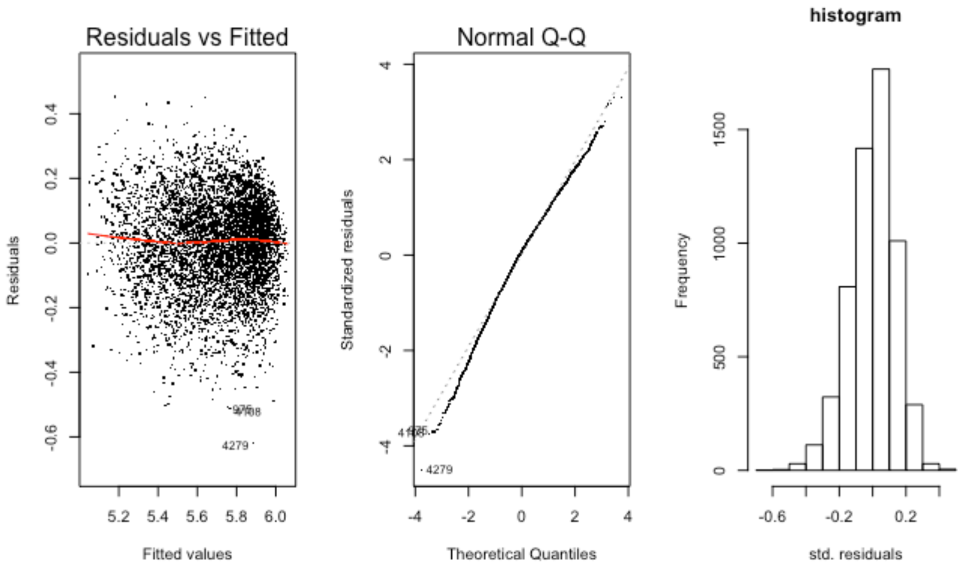
I removed W1famtyp2 and W8DMARSTAT as they weren't significant at the 5% level and I also removed W1wrk1aMP and W1marstatmum as they had very high GVIF values (17.7 and 32.7 respectively). This suggested that there was multicollinearity in the model. After removing the said predictors, I created the third model.

I then generated boxplots for the existing predictors and merged the levels of W1hous12HH, W1hiqualmum, W1nssecfam, W1ethgrpYP, W1heposs9YP and W4AlcFreqYP. The criterion used to merge levels was explained in the exploratory analysis and the aim of merging levels was to remove clutter in the model. (Explained in the appendix) I then created the fourth model.

I conducted a residual analysis on model 4 to see if the residual assumptions were violated. Based on the Q-Q plot and the histogram of standardized residuals I concluded that the assumption of normality had not been violated. However, I observed heteroscedasticity among the errors as the residuals vs fitted plot displayed a funnel shape. As the assumption of constant variance (homoscedasticity) was violated and as the lowest value of W8DINCW was £115.51 (W8DINCW was non-negative) I decided to use a log transform on the outcome of model 4 to create model 5.

I then conducted a residual analysis on model 5. As per Figure 1, the residual vs fitted plot looks like a random scatter of points. Therefore, I concluded that the residual assumption of homoscedasticity was not violated. The Q-Q plot and the histogram of standardised residuals reveals that the residuals are approximately normally distributed. However, I observed a slight kurtosis as the Q-Q plot deviates from the diagonal on the far left and right but this was not worrying. Therefore, I concluded that the residual assumptions of normality, independence and homoscedasticity were not violated for this model.

Figure 1: Residual plots for model 5



Then I investigated the possibility of outliers in the data as the exploratory analysis revealed possible outliers. The criteria followed to identify outliers were Residuals, Cook's distance, DFFITS and Leverage values. There were no points that violated Cook's distance but there were 2 points that violated residuals, DFFITS and leverage values (1597 and 5552). As the Cook's distance of the points were below 1, I decided not to exclude the points. I looked closely at these two points to see if there any similarities and observed that the two points were in the same level for seven out of the nine predictors in wave 6. I also noticed that 5552 had missing values in 10 predictors. Finally, I concluded that there were no outliers in the dataset. Based on the above I concluded model 5 as the final.

<u>W8QDEB2 (Total amount owed)</u>

My initial model predicting the total amount owed by an individual in wave 8 had all 66 predictors whereas the final model (coloured in green in Table 2) included 7 predictors.

After running the first model, I realized that it had only 4 predictors significant at the 5% level (W4CannTryYP, W4schatYP, W8TENURE and W8NETA). Based on the results of Anova and the scatterplots/boxplots of predictors I decided to remove 25 predictors. Thus, I created my second model with 41 predictors.

I observed that this model had aliased coefficients for some levels of the predictors. I observed that some levels of W1NoOldBro, W1empsmum, W1empsdad, W1famtyp2 and W1nssecfam exhibited perfect collinearity with some levels of W1disabYP, W1hiqualmum, W1hiqualdad, W1depkids and W6JobYP respectively. Therefore, I removed W1empsmum, W1empsdad and W1nssecfam as they were less significant than W1hiqualdad, W1depkids and W6JobYP. Thereafter I removed W1NoOldBro & W1famtyp2 as they had more missing values than W1disabYP & W1depkids. Thus model 3 was created.

I noticed that W1hous12HH, W1hiqualmum and W1hiqualdad had a very high GVIF (Greater than 10). This suggested that there was multicollinearity in the model. Thus, I eliminated these predictors and created the fourth model.

In order to increase the size of the dataset I decided to remove all the continuous predictors with a p-value greater than 0.5. This made the dataset larger and more representative. As a result of deleting W6DebtattYP, cent.W1GrssyrMP, cent.W1GrssyrHH and cent.W8DINCW the number of rows with complete information increased from 1118 to 2270 data points. I created the fifth model using this dataset.

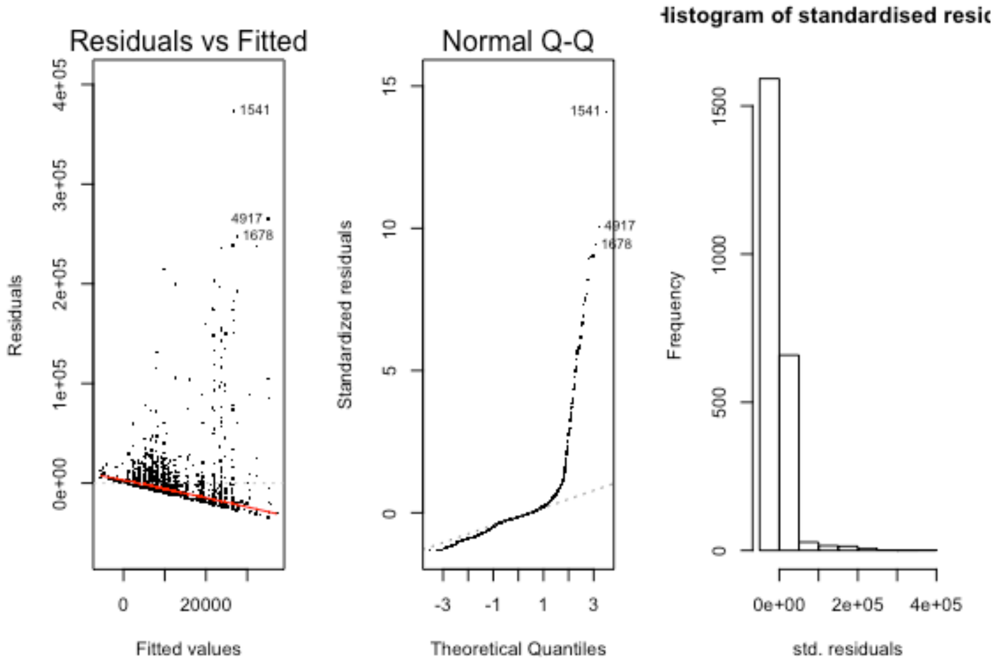
The fifth model included 29 predictors. I then got rid of all the categorical and binary predictors that weren't significant at a 50% level. After removing 15 predictors I obtained my sixth model.

This model contained 14 predictors, 4 of which were significant at a 5% level (IndSchool, W1heposs9YP, W8CMSEX and W8TENURE). I then proceeded to remove all the predictors that weren't significant at a 30% level. As a result, I removed W1hea2MP, W1hwndayYP, W1bulrc, W2depressYP from model 6 and created model 7.

IndSchool, W1heposs9YP, W8CMSEX and W8TENURE continued to be significant at the 5% level in model 7. I removed W1yschat1 from the model as it was the least significant predictor and created model 8.

I then removed W4yschat to create the ninth model and W4CannTryYP to create the tenth model as they were the least significant predictors in each model. Next, I conducted a residual analysis on model 10. A funnel shape was not observed in the fitted vs residual plot. However, the Q-Q plot and the histogram of standard residuals revealed that the model was heavily positively skewed. (Refer Figure 2)

Figure 2: Residual plots for model 10



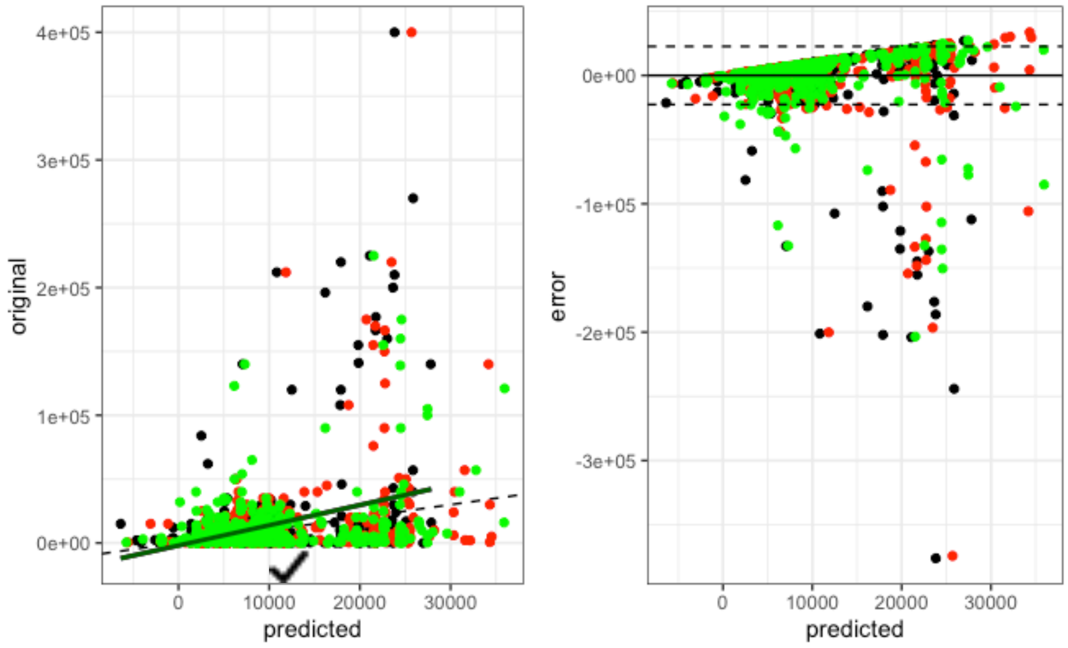
I decided to remove data points containing zero debt to create the eleventh model although I was aware that this would skew the dataset. This reduced the dataset from 2317 to 2249. I then applied a log transform to W8QDEB2 (the outcome) of model 10. Normality of residuals wasn't violated in model 11 and the errors had a constant variance (homoscedasticity). However, Table 2 reveals that coefficients estimates of most predictors were not significant in this model. The goodness of fit statistics were also better for model 10: model 10 had a marginally better R^2 and Adjusted R^2 compared to model 11 and the standard error of model 10 relative to the size of the range was better than that of model 11. The residual standard deviation of model 10 was 26506 and this was compared with the width of the range of W8QDEB2 (400,000 – 0 = 400,000). The standard error of model 11 was 1.634 and this was compared with the width of the range of log(W8QDEB2) (12.8992 - 1.6094 = 11.2898). The F-statistic p-value was low for both models. Therefore, from the diagnostics and the residual plots I concluded that model 10 was marginally better. However, I was aware that there was a possibility for predictions to become unreliable if the normality of residuals was violated.

Table 2: W8QDEB2 – Initial to final model

	Model 1 W8QDEB 2.lm.1	Model 2 W8QDEB 2.lm.2	Model 3 W8QDEB 2.lm.3	Model 4 W8QDEB 2.lm.4	Model 5 W8QDEB 2.lm.5	Model 6 W8QDEB 2.lm.6	Model 7 W8QDEB 2.lm.7	Model 8 W8QDEB 2.lm.8	Model 9 W8QDEB 2.lm.9	Model 10 W8QDEB 2.lm.10	Model 11 W8QDEB 2.lm.11
Coefficients											
W1condur5MP											
IndSchool					*	**	**	**	**	**	
W1heposs9YP					*	*	*	*	*	**	
W6OwnchiDV											**
W8CMSEX						*	*	*	*	*	
W8TENURE	***	***	***	***	***	***	***	***	***	***	***
W8NETA	**	**	**	**							
R2/Adj R2	0.2016 / -0.00083	0.1677 / 0.02046	0.1546 / 0.03058	0.1148 / 0.0317	0.09852 / 0.05217	0.08608 / 0.07177	0.08092 / 0.07234	0.08004 / 0.07197	0.07832 / 0.0707	0.07779 / 0.07097	0.05221 / 0.04498
F-statistic significance	0.5067	0.1265	0.03448	0.01138	2.90E-10	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
Standard Error	28678.8	27812.2	27668.12	27652.1	26657.2	26380.2	26372.1	26571.4	26510.09	26506.12	1.634
Comments											
Range of outcome	0 - 400000	0 - 400000	0 - 400000	0 - 400000	0 - 400000	0 - 400000	0 - 400000	0 - 400000	0 - 400000	0 - 400000	1.6094 - 12.8992
Number of datapoints in the dataset used to create the model	1068	1118	1118	1118	2270	2270	2270	2300	2316	2316	2248
Number of predictors in the model	66	41	36	33	29	14	10	9	8	7	7

I then performed some cross validation on model 10 and the results obtained are provided in Figure 3. Most points in the original – predicted plot were close to the regression line (dark green line). The regression line was close to the x-y diagonal (black dotted line). The predicted vs error plot shows that a majority of points were within one standard deviation of the 0 line. The data seems to be broadly similar across each split. Thus, I concluded that model 10 was acceptable but not great.

Figure 3: Results of cross validation for model 10



W8QMAFI (How individuals are managing financially)

My first logistic regression model predicting how individuals are managing financially in wave 8 (W8QMAFI) had 62 predictors. The Anova function revealed that this model only had 7 predictors significant at the 5% level. Therefore, I removed the predictors that were not significant and ran the second model.

As W1empsmum wasn't significant at the 5% level, I eliminated this predictor and ran the third model.

Model 3 is nested in model 2 as the predictors in model 3 are a subset of the predictors in model 2. Similarly models 2 and 3 are nested in model 1. As models 2 and 3 are created from a dataset with the same rows they can be compared. The residual deviance of model 3 and 2 are 3873.1 and 3862.1 respectively (refer Table 3). The difference between the deviances is 11 (δ_d). δ_d is distributed according to the χ^2 distribution on 1 degree of freedom (model 2 has 7 predictors and model 1 has 6 predictors). As $\delta_d > 3.84$ (Critical value of χ^2 on 1 degree of freedom) model 2 is better.

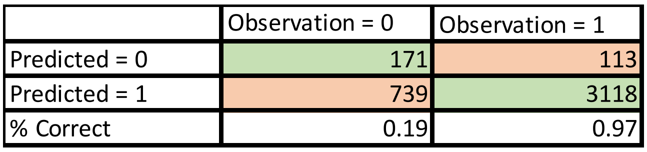
Table 3: W8QMAFI – Initial to final model

	Model 1 W8QMAFI.glm.1	**Model 2 W8QMAFI.glm.2**	**Model 3 W8QMAFI.glm.3**
Coefficients			
W1wrk1aMP	***	**	***
W4NamesYP	*	*	*
W4empsYP	**	***	***
W8DGHQSC	***	***	***
W8TENURE	***	***	***
W8NETA	*	***	***
Null deviance (dev0)	1368.83	4361.3	4361.3
Residual deviance (devm)	989.91	3862.1	3873.1
Comments			
Number of datapoints in the dataset used to create the model	1221	4140	4140
Number of predictors in the model	62	7	6

I also used classification tables to plot the predicted vs observed outcomes as a measure of how well the models fits the data. I assumed that a predicting probability exceeding 0.5 corresponded to a predicted outcome of 1. Model 2 predicted correctly 79.35% of the time whereas model 3 predicted correctly 79.43% of the time (refer Table 4).

Therefore, I concluded that model 3 was marginally better as model 2 contains a predictor that was not significant at the 5% level and model 3 was marginally more accurate.

Table 4: Classification table of model

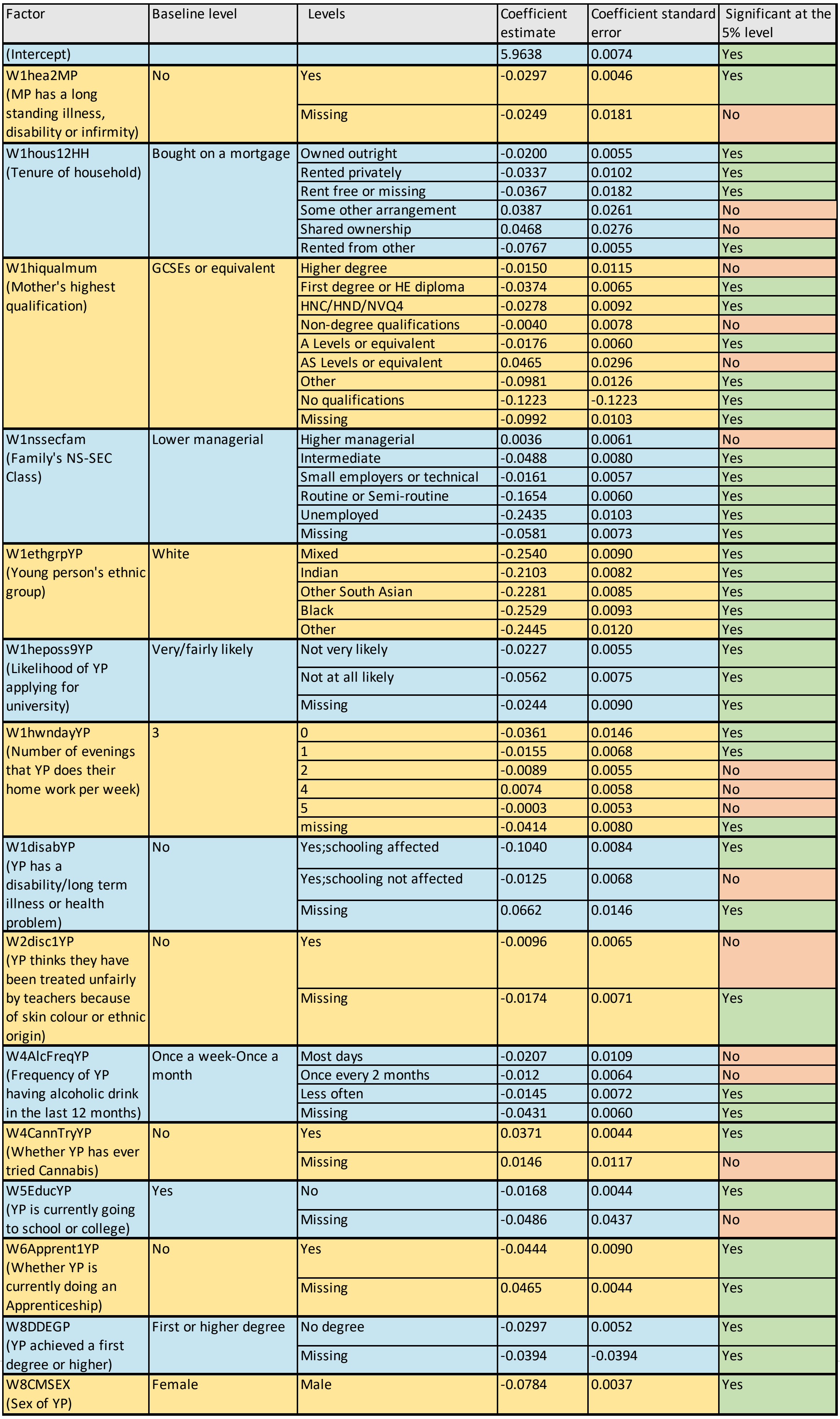


	Observation = 0	Observation = 1
Predicted = 0	171	113
Predicted = 1	739	3118
% Correct	0.19	0.97

Results

Results: W8DINCW (Continuous weekly income)

Table 5: Results of the final model for W8DINCW



Factor	Baseline level	Levels	Coefficient estimate	Coefficient standard error	Significant at the 5% level
(Intercept)			5.9638	0.0074	Yes
W1hea2MP (MP has a long standing illness, disability or infirmity)	No	Yes	-0.0297	0.0046	Yes
		Missing	-0.0249	0.0181	No
W1hous12HH (Tenure of household)	Bought on a mortgage	Owned outright	-0.0200	0.0055	Yes
		Rented privately	-0.0337	0.0102	Yes
		Rent free or missing	-0.0367	0.0182	Yes
		Some other arrangement	0.0387	0.0261	No
		Shared ownership	0.0468	0.0276	No
		Rented from other	-0.0767	0.0055	Yes
W1hiqualmum (Mother's highest qualification)	GCSEs or equivalent	Higher degree	-0.0150	0.0115	No
		First degree or HE diploma	-0.0374	0.0065	Yes
		HNC/HND/NVQ4	-0.0278	0.0092	Yes
		Non-degree qualifications	-0.0040	0.0078	No
		A Levels or equivalent	-0.0176	0.0060	Yes
		AS Levels or equivalent	0.0465	0.0296	No
		Other	-0.0981	0.0126	Yes
		No qualifications	-0.1223	-0.1223	Yes
		Missing	-0.0992	0.0103	Yes
W1nssecfam (Family's NS-SEC Class)	Lower managerial	Higher managerial	0.0036	0.0061	No
		Intermediate	-0.0488	0.0080	Yes
		Small employers or technical	-0.0161	0.0057	Yes
		Routine or Semi-routine	-0.1654	0.0060	Yes
		Unemployed	-0.2435	0.0103	Yes
		Missing	-0.0581	0.0073	Yes
W1ethgrpYP (Young person's ethnic group)	White	Mixed	-0.2540	0.0090	Yes
		Indian	-0.2103	0.0082	Yes
		Other South Asian	-0.2281	0.0085	Yes
		Black	-0.2529	0.0093	Yes
		Other	-0.2445	0.0120	Yes
W1heposs9YP (Likelihood of YP applying for university)	Very/fairly likely	Not very likely	-0.0227	0.0055	Yes
		Not at all likely	-0.0562	0.0075	Yes
		Missing	-0.0244	0.0090	Yes
W1hwndayYP (Number of evenings that YP does their home work per week)	3	0	-0.0361	0.0146	Yes
		1	-0.0155	0.0068	Yes
		2	-0.0089	0.0055	No
		4	0.0074	0.0058	No
		5	-0.0003	0.0053	No
		missing	-0.0414	0.0080	Yes
W1disabYP (YP has a disability/long term illness or health problem)	No	Yes;schooling affected	-0.1040	0.0084	Yes
		Yes;schooling not affected	-0.0125	0.0068	No
		Missing	0.0662	0.0146	Yes
W2disc1YP (YP thinks they have been treated unfairly by teachers because of skin colour or ethnic origin)	No	Yes	-0.0096	0.0065	No
		Missing	-0.0174	0.0071	Yes
W4AlcFreqYP (Frequency of YP having alcoholic drink in the last 12 months)	Once a week-Once a month	Most days	-0.0207	0.0109	No
		Once every 2 months	-0.012	0.0064	No
		Less often	-0.0145	0.0072	Yes
		Missing	-0.0431	0.0060	Yes
W4CannTryYP (Whether YP has ever tried Cannabis)	No	Yes	0.0371	0.0044	Yes
		Missing	0.0146	0.0117	No
W5EducYP (YP is currently going to school or college)	Yes	No	-0.0168	0.0044	Yes
		Missing	-0.0486	0.0437	No
W6Apprent1YP (Whether YP is currently doing an Apprenticeship)	No	Yes	-0.0444	0.0090	Yes
		Missing	0.0465	0.0044	Yes
W8DDEGP (YP achieved a first degree or higher)	First or higher degree	No degree	-0.0297	0.0052	Yes
		Missing	-0.0394	-0.0394	Yes
W8CMSEX (Sex of YP)	Female	Male	-0.0784	0.0037	Yes

All 15 predictors included in the model were significant at the 5% level. Although all predictors are significant at the 5% level, I did not observe any dominant predictors.

The coefficient estimate of the intercept tells us that the predicted log(W8DINCW) would be 5.9638 when all predictors are at the reference level. That is expected W8DINCW would be £389 ($e^{5.9638}$) when all predictors equal their baseline levels. (Refer Table 5 for the baseline levels of the predictors)

The coefficient estimate of a level of a predictor in this model can be interpreted as the percentage increase/decrease in W8DINCW caused by the said predictor changing from the reference level (given that all else is equal). Let me interpret the coefficients of the levels "0" and "4" of the predictor W1hwndayYP (Number of evenings a week spent on homework) as an example. As per Table 5, the baseline level of this predictor is "3" and the coefficient estimates of the levels "0" and "4" are -0.0361 and 0.0074 respectively. This means that there is a 3.54% ((1-$e^{-0.0361}$) * 100%) expected decrease in W8DINCW for someone who does their homework 0 times a week and a 0.74% (($e^{0.0074-}$1) * 100%) expected increase in W8DINCW for someone who does their homework 4 times a week compared to someone who does their homework thrice a week.

Unsurprisingly white individuals are predicted to have the highest log(W8DINCW) whereas mixed individuals are predicted the lowest. As expected academic achievement positively correlated with log(W8DINCW). Students who said they were very/fairly likely to apply to university in wave 1 (W1heposs9YP), individuals who attended school or college in wave 5 (W5EducYP) and respondents who obtained a first or higher degree (W8DDEGP) are expected to earn a higher log(W8DINCW). As also expected healthier respondents (respondents with no disability/ long term illness or health problem) are predicted a higher log(W8DINCW). Interestingly females are predicted a higher W8DINCW than their male counterparts at 25. This is probably impacted by their relatively better academic performance at university level. i.e. 1207 females obtained first or higher degrees compared to the 925 males. Students whose parents were from a higher managerial NS-SEC class are predicted the highest log(W8DINCW) whilst students who lived in rented houses had a lower predicted log(W8DINCW) compared those living in a house bought on a mortgage.

Diagnostic statistics are used to determine how well a model fits the data. The R^2 value for this regression is 0.7048. This value is relatively high and this means that the model is adequate but can still be improved. The Adjusted R^2 value for this model is 0.7019 and this was within 0.42% of the R^2 value. This tells us that there weren't many non-significant predictors in this model. The residual standard deviation is 0.1373 and this is relatively small when compared to the width of the range of the data (6.197889 - 4.749357 = 1.448532). This increases the credibility of the model. The F-statistic was highly significant in this model as well. A highly significant F-statistic indicates an overall good fit of the model.

Results: W8QDEB2 (Total amount owed)

Table 6: Results of the final model for W8QDEB2

Factor	Baseline level	Levels	Coefficient estimate	Coefficient standard error	Significant at the 5% level
Intercept			6,745.5412	1238.3222	Yes
W1condur5MP (Have a computer in household)	Yes	No	-3,977.4636	2458.8794	No
		Missing	-6,374.2655	6483.9543	No
IndSchool (Was YP at an independent or maintained school at sampling stage)	Maintained	Independent	8,598.1572	2799.2892	Yes
W1heposs9YP (Likelihood of YP applying for university)	Very likely	Fairly likely	-4,515.9594	1276.4935	Yes
		Not very likely	-1,808.6658	1704.6524	No
		Not at all likely	-2,967.3281	2397.0742	No
		Missing	1,097.2821	3087.8281	No
W6OwnchiDV (Do respondents have own Child/ Children)	No	Yes	-3,889.9629	3941.0239	No
		Missing	-6,031.778	4994.5621	No
W8CMSEX (Sex of YP)	Female	Male	2,765.474	1131.096	Yes
W8TENURE (Tenure of house)	Rent including housing benefits	Owned outright	5,786.0803	3275.8111	No
		Bought on a mortgage	16,734.0912	1470.5423	Yes
		Part rent/mortgage	4,896.5052	4393.9955	No
		Rent free	2,760.4235	1698.7351	No
		Other	1,338.3054	1554.3052	No
		Missing	-890.5943	26642.2722	No
W8NETA (Last take-home pay (net))			0.1504	0.1321	No

Out of the 7 predictors included in the final model, only IndSchool, W1heposs9YP, W8CMSEX and W8TENURE were significant at a 5% level.

The coefficient estimate of the intercept is 6,746 and it represents the expected debt of a female, who at the age of 13-14 owned a computer, was at a maintained school and was very likely to apply to university. Further they did not have children when they were teenagers and also lived in a rented house at the age of 25.

The coefficient intercept of W8NETA was 0.1504 and this can be interpreted as expected W8QDEB2 increases by £ 0.1505 for a £1 increase in W8NETA.

Mortgages are expected to have the largest impact on W8QDEB2. This was as expected because mortgages are one of the main causes of debt. Interestingly individuals who owned a computer and attended an independent school are predicted to have a higher debt than their counterparts. This may be because the lurking variable in W1condur5MP and IndSchool was wealth. Individuals that come from wealthier backgrounds may be entitled to borrow more from their parents (As debt is defined as total amount it could also include borrowing from parents) as well as leverage on their family wealth to borrow more. Respondents who reported that they were very likely to attend university also had a higher expected debt than respondents that did not. 1238 respondents who said they were very likely to go to university went on to attend university in wave 6 (They were in the yes level for the W6UnivYP predictor). It is most likely that individuals who attended university had obtained student loans which results in a higher debt. 25 year old males were also expected to have approximately £2765 more debt than females of a similar age.

The diagnostics for this model are given in Table 2. The R^2 and Adjusted R^2 values were very poor for this model. This is possibly due to the limited number of observations provided for the W8QDEB2 outcome. However, this model had a highly significant F-statistic and this indicates an overall good fit of the model.

Results: W8QMAFI (How individuals are managing financially)

Table 7: Results of the final model for W8QMAFI

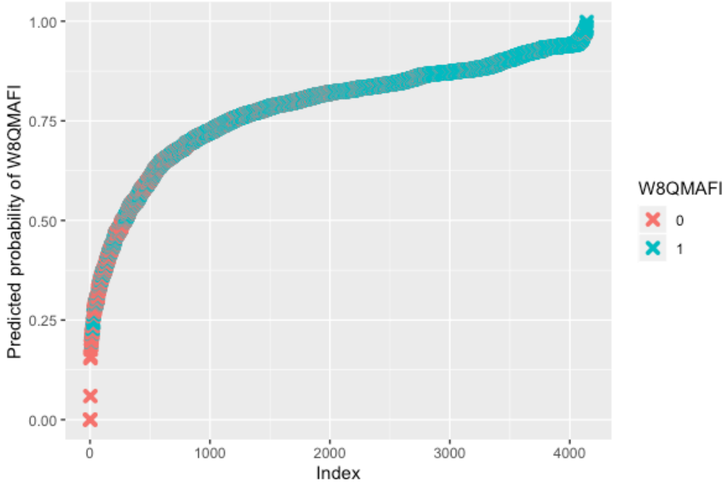
Factor	Baseline level	Levels	Coefficient estimate	Coefficient standard error	Significant at the 5% level
(Intercept)			1.48	0.09	Yes
W1wrk1aMP (Current working status of MP)	FT Paid	PT Paid	0.07	0.10	No
		FT Self employed	0.33	0.24	No
		PT Self employed	0.88	0.39	Yes
		Unemployed	-0.64	0.28	Yes
		FT Education	0.23	0.40	No
		Government scheme - employment training	-12.08	324.70	No
		Temporarily sick/ disabled	0.01	0.55	No
		Permanently sick/ disabled	-0.34	0.29	No
		Looking after family	0.13	0.12	No
		Retired	-0.86	0.36	Yes
		Other	-1.79	0.58	Yes
		Missing	-0.68	0.37	No
W4NamesYP (Has YP has been bullied, sworn at or insulted in the last 12 months)	No	Yes	-0.23	0.10	Yes
		Missing	0.38	0.31	No
W4empsYP (Employment status of YP)	FT Education	Paid work >30	-0.54	0.19	Yes
		Paid work <30	-0.31	0.21	No
		Unemployed	-0.73	0.24	Yes
		Training course	-0.44	0.21	Yes
		Looking after family	-1.40	0.82	No
		Other	-1.28	0.40	Yes
		Missing	-14.80	324.70	No
W8DGHQSC (General Health Questionnaire (GHQ12) score)			-0.20	0.01	Yes
W8TENURE (Tenure of household)	Rent including housing benefits	Owned outright	0.55	0.22	Yes
		Bought on a mortgage	1.12	0.14	Yes
		Part rent/mortgage	0.43	0.39	No
		Rent free	0.36	0.11	Yes
		Other	0.34	0.11	Yes
		Missing	-0.93	1.03	No
W8NETA (Last take-home pay (net))			0.00	0.00	Yes

All 6 predictors included in this model are significant at a 5% level.

W8DGHQSC (GHQ12 score) and W8NETA (Last take home pay -net) were highly significant in this model. The GHQ12 score is a measure of the mental health of an individual and the lower the score, the better the mental health of the individual is. Respondents engaged in full time education in wave 4 are predicted to be in a better financial position than their counterparts. The socio-economic background of respondents had an effect on the expected W8QMAFI. Individuals whose main parent was self-employed were predicted to do better financially than those whose main parent was full time paid. Students whose main parent was unemployed were predicted to do financially worse compared to those whose main parent was full time paid.

The logistic model is non-linear and this is clearly observed in Figure 4. As a result, the effect on the probability of W8QMAFI associated with an increase in one predictor depends on the starting value of the predictor and it is not the same across the range of the predictor.

Figure 4: Logistic model predicting W8QMAFI



1= Well
0 = Poorly

From table 3 the residual deviance of the model was 3873.1 whilst the null deviance of the model was 4361.3. The difference between these is 488.2 . This is much higher than 12.6, the critical value of χ^2 on 6 degrees of freedom. Therefore, this model is much better than the null model.

Comments About the Data/Analysis

Although approximately 16 000 people were initially part of the LSYPE only 5800 individuals completed wave 8. Even out of the individuals who completed wave 8, some refrained from answering all the questions provided. Therefore, my final model predicting W8QDEB2 used only 40% of the observations and the logistic regression used 71% of the observations. My final model predicting W8DINCW used all the observations provided. Even if all 5 800 respondents answered all the questions, the sample would still be unrepresentative of the entire population of young people in England. This is because the dataset with 5800 observations was a subset of the dataset corresponding to the NS survey which is in turn a subset of the dataset of the population of young people in England. Therefore, I may not be able to generalize my results to all young people in England.

The data for the survey was collected through multiple means. i.e. face to face interviews, online interviews and over the phone. As the data is self-reported the data may be biased due to many reasons. Firstly, respondents may have felt uncomfortable providing answers that weren't socially desirable. For example, respondents may have answered dishonestly for predictors such as W1alceverYP (Whether ever had proper alcoholic drink) and W4CannTryYP (Whether YP has ever tried Cannabis). As the respondents were 13-14 years of age during wave 1 and they were below the legal drinking age in England (18). Smoking cannabis is also illegal in England. Secondly, respondents may also have refused to reveal sensitive information such as W1GrssyrMP (Gross annual salary of MP) and W8QDEB2 (Total amount owed).

The data could also be missing in a systematic way. It was observed that the missing levels of some predictors perfectly correlate. For example the "missing" levels of the predictors W1condur5MP and W1wrk1aMP were perfectly correlated for model 5 predicting W8QDEB2. Furthermore, those with missing data in the W4AlcFreqYP predictor have the lowest log continuous weekly income. Therefore, if the missing values were systematic and correlated the analysis would be biased.

There are several flaws in my analysis. One of the main issues that I faced was missing data and ability to process it. Lack of expertise on the subject matter was another limiting factor when conducting the analysis.

Interpretation and conclusion for a lay audience

I developed 3 models to determine how the financial position of a 25 year old in England is impacted by their ethnicity, education, socio-economic status of the family and health. The measures that I picked to reflect their financial position are their continuous weekly income, total amount owed and how they are managing financially. A separate model was developed for each of the 3 measures. These models were based on a sample of 5800 individuals who had responded to the Next Steps Survey between the years 2004 and 2016.

Model 1 - predicting the continuous weekly income of a 25 year old

To illustrate the impact of an individual's ethnicity, education, socio-economic status of the family and health on their continuous weekly income I will use the examples given in the table below (Table 8).

Most of the predictors can be categorized in to the 4 categories listed below and have been coloured in table 8 based on the colour coding given below:

Ethnicity
Education
Socio - economic status of family
Health

Table 8: Expected continuous weekly income of five individuals

	Student A	Student B	Student C	Student D	Student E
Does the main parent have a long standing illness, disability or infirmity? (aged 13-14)	No	No	No	No	No
Tenure of household (aged 13-14)	Shared ownership	Shared ownership	Shared ownership	Shared ownership	Rented from other
Mother's highest qualification (aged 13-14)	AS Levels or equivalent	AS Levels or equivalent	AS Levels or equivalent	AS Levels or equivalent	AS Levels or equivalent
Family's NS-SEC Class (aged 13-14)	Higher managerial	Higher managerial	Higher managerial	Higher managerial	Unemployed
Young person's ethnic group (aged 13-14)	White	Mixed	White	White	White
Likelihood of young person applying to university (aged 13-14)	Very/fairly likely	Very/fairly likely	Not at all likely	Very/fairly likely	Very/fairly likely
Number of evenings a week spent on home work (aged 13-14)	4.00	4.00	0.00	4.00	4.00
Does the young person have a disability/long term illness or health problem? (aged 13-14)	No	No	No	Yes	No
Does the individual think they have been treated unfairly by teachers because of their skin colour or ethnic origin? (aged 14-15)	No	Yes	No	No	No
Frequency of individual having an alcoholic drink in the last 12 months (aged 16-17)	Once a week - Once a month	Once a week - Once a month	Once a week - Once a month	Once a week - Once a month	Once a week - Once a month
Has the individual tried Cannabis? (aged 16-17)	Yes	Yes	Yes	Yes	Yes
Was the young person attending school or college? (aged 17-18)	Yes	Yes	No	Yes	Yes
Was the young person doing an Apprenticeship? (aged 18-19)	No	No	No	No	No
Has the young person achieved a first degree or higher? (aged 25)	First or higher degree	First or higher degree	No degree	First or higher degree	First or higher degree
Sex of respondent	Female	Female	Female	Female	Female
Predicted continuous weekly income in GBP (aged 25)	**448.2**	**344.4**	**387.3**	**404.0**	**309.4**

Individual A earned the highest possible continuous weekly income that the model predicted i.e. GBP 448. I will use her as the benchmark to compare the others. Individual A is a white female from a higher managerial NS-SEC class, doesn't have a disability/long term illness or health problem and has achieved a first degree or higher.

Student B has the same characteristics as A except those related to ethnicity. i.e she is of a mixed ethnicity. This difference in ethnicity has resulted in a significant reduction of £104 (£448-344) in her expected continuous weekly income. This is probably because of the possible unconscious bias employers have towards white individuals.

Student C had the same characteristics as A except those related to education. She had no aspirations of higher studies when in grade 9 and consequently had dropped out of school by grade 13. Her lack of interest in education was further evident as she had not done her weekly homework in grade 9. C's expected income was £61 (£448-387) lower than that of A. Surprisingly education had a relatively lower impact on an individual's expected continuous weekly income. Interestingly a white female with no university education is expected to earn £ 87 (£435-348) more than a mixed female with a first degree or higher (all else being equal to Student A)

Student D has the same characteristics as A except her health. i.e when surveyed in grade 9 she had stated that she had a disability/long term illness or health problem and it affected her schooling. D is expected to earn a continuous weekly income of £404 which is only £44 (£448-404) lower than A. Therefore, the health of an individual seems to have the least impact on the expected continuous weekly income. This is probably because the overall level of corporate social responsibility is high in the UK and employers make an effort to integrate those with disabilities/impaired health into the workforce.

Student E has the same characteristics as A except her socio-economic status. i.e at the age of 13-14, her main parent was unemployed and she lived in a rented accommodation. Both indications of a relatively difficult economic background. E's expected continuous weekly income was only £309 and the lowest out of the 5 students. Surprisingly socio-economic background had a greater impact than ethnicity and education.

My model suggests that there may be a possible unconscious bias in the workplace towards white individuals and those from a relatively stronger socio-economic background. These had a favourable impact on their expected continuous weekly income. Education and good health also have a favourable impact on the same.

Model 2 – predicting the total amount owed by a 25 year old

The model created to predict the above consisted of 7 criteria. These criteria were related to socio-economic status & education but not ethnicity & health. Interestingly it included gender, last take home pay at 25 and whether the responded had children when they were 18-19.

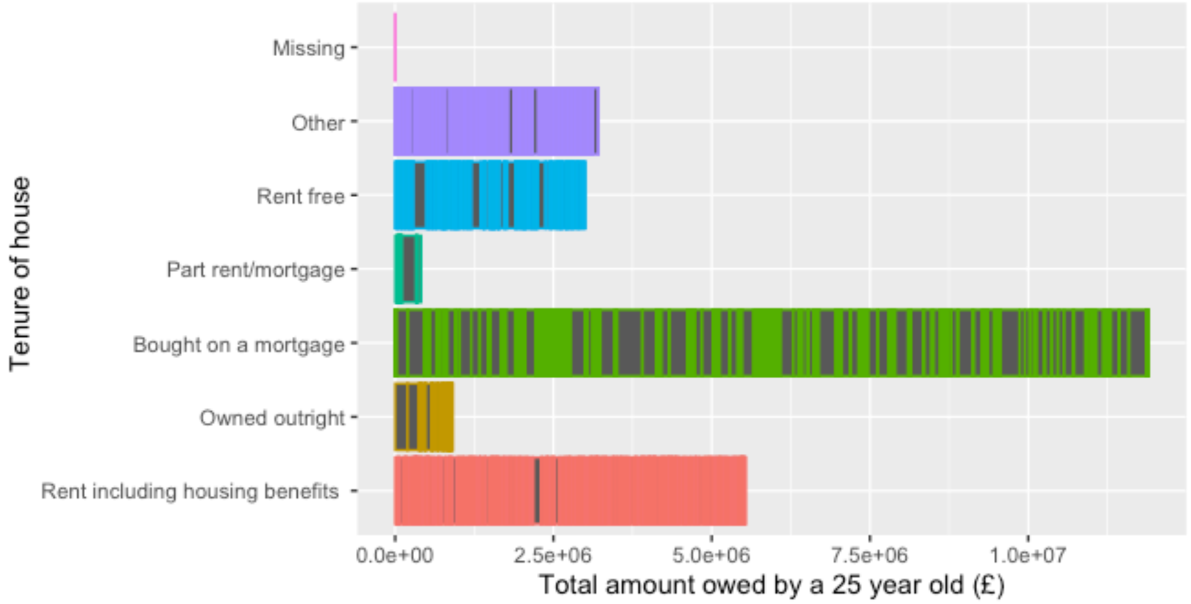
There were 2 criteria that were indications of family wealth i.e whether the individual attended an independent school and whether they owned a computer in grade 9. Interestingly, individuals who said yes to both were likely to have a higher debt than their counterparts. This is most probably influenced by their ability to borrow from their parents as well as the ability to leverage on their family wealth to borrow more.

Those who responded that they were very likely to attend university, when surveyed in grade 9, were expected to have a higher debt than their counterparts. This is possibly related to student loans. i.e. 1238 respondents out of the 2227 individuals who responded that they were very likely attend university went on to do so later.

Individuals who had a higher take home pay at 25 were also expected to have a higher debt. This is most likely because of their increased ability to borrow. For every increase in £1 of take home pay an individual's debt increased by £0.1504 .

Unsurprisingly housing mortgages had the largest impact on the total debt of a 25 year old. The debt of an individual who lived in a house bought on a mortgage was £16 734 higher than that of a person who lived in a rented house. This is clearly illustrated in Figure 5

Figure 5: Total amount owed by a 25-year-old vs Tenure of house



The reasons for not having a discernible relationship between either the ethnicity or health and total amount owed, could be because the sample considered was small (and therefore unrepresentative of the population) and also because the model was poor.

<u>Model 3 – predicting how an individual is managing financially at 25</u>

The model created to predict the above consisted of 6 criteria. These criteria were related to education, socio-economic status and health but not ethnicity. Interestingly it included last take-home pay at 25 and tenure of house at 25.

Those with a better General Health Questionnaire (GHQ12) score (i.e. lower score) were expected to manage better financially. The GHQ12 is a measure of mental health of an individual and this indeed will impact the ability to manage one’s financials better.

As expected those with a higher last take home pay at 25 were expected to manage better financially.

An individual’s socio – economic background had an impact on how they were expected to manage financially at 25. Characteristic that impacted this outcome was the working status of the main parent when the individual was aged 13-14. Individuals whose main parent was self-employed were expected to do better than those whose main parent was full time paid. Also, students whose main parent was unemployed are expected to be financially worse compared to those whose main parent was full time paid.

Respondents engaged in full time education in grade 12 are predicted to be in a better financial position at 25 than their counterparts. This was most probably because they could continue their education beyond secondary level. 2089 respondents out of the 4855 individuals who were in full time education in grade 12 went on to attend university.

Overall Summary/Conclusion

Based on the results obtained through the 3 models that were discussed above, let me now conclude how the financial position of an individual at the age of 25 is impacted by their ethnicity, education, socio-economic status of the family and health.

Table 9: Comparison of results

	Ethnicity	Education	Socio Economic	Health
Continuous weekly income of a 25 year old	Yes	Yes	Yes	Yes
Total amount owed by a 25 year old	No	Yes	Yes	No
How an individual is managing financially	No	Yes	Yes	Yes

Ethnicity of an individual is only expected to impact the continuous weekly income but not their overall debt (amount owed) nor how they are expected to manage financially.

Education and more specifically those characteristics directly or indirectly impacting a first degree or higher are expected to impact all three measures picked to represent financial position.

Similarly, socio economic back ground of the individuals and more specifically characteristics that indicate wealth of the parents are expected to impact all three measures picked to represent financial position.

Physical health was predicted to have some impact on the expected continuous weekly income and mental health is predicted to impact how an individual is expected to manage financially.

However, the above results may not be reflective of the population of young people in England, as the sample considered was small and therefore unrepresentative.

Policy recommendations

Whilst I may not be able to generalize my results, if I am forced to provide recommendations to support policy decisions I would leverage on the model developed to predict the continuous weekly income as it was the most satisfactory out of the 3 models.

Recommendations made are set out below;

- There seems to be indications of discrimination based on ethnicity. The government needs to implement policies that encourage equality among diverse ethnic groups. For example, introduce a system to track and monitor income anomalies caused by ethnic bias. This can be complemented by a periodic survey. Government could also regulate the recruitment process so that practices that discourages discrimination can be promoted. (Recommend interview do's and don'ts)

- Students need to be encouraged to pursue higher education. Career talks, career counselling and higher education fairs etc. need to be conducted. Government itself can undertake these initiatives as well as encourage the private sector also to do the same.

- Take measures to reinforce the positive sentiments that the employers seem to be having over employment of people with disabilities/impaired health. Government could introduce tax incentives for those employing people with disabilities/ impaired health.

Bibliography

Retrieved from http://nesstar.ukdataservice.ac.uk/webview/index.jsp?v=2&mode=documentation&submode=abstract&study=http://nesstar.ukdataservice.ac.uk:80/obj/fStudy/5545&top=yes

Appendix: Merging of Variables

During modelling I merged the levels of some variables. I merged levels of predictors primarily based on the frequency of the levels and the boxplots corresponding to the levels. If two levels of a predictor had a low frequency and their boxplots were similar I merged them. I also considered the significance of estimated coefficients of the levels of predictors as a secondary factor. i.e. I only merged levels with estimated coefficients that were not significant at a 5% level.

<u>Merging levels of predictors for the model predicting W8DINCW</u>

1.W1house12HH

- Rent free , Missing = Rent free or missing
- Rented from Council , Rented from Housing Association = Rented from other

2.W1hiqualmum:

- 2,3 = First degree or HE diploma
- 5,6= Non-degree qualifications
- 7,8,9,10 = A Levels or equivalent
- 11,12 = AS Levels or equivalent
- 13,14,15,16 = GCSEs or equivalent
- 17,18,19,20 = Other

3.W1nssecfam

- Small employers , Technical =Small employers or technical
- Routine , Semi-routine = Routine or Semi-routine

4.W1ethgrpYP

- Pakistani , Bangladeshi = Other South Asian
- Black Caribbean , Black African = Black
- Other , Missing = Other or Missing

5.W1heposs9YP

- Very likely , Fairly likely = Very likely/Fairly likely

6.W4AlcFreqYP:

- 1-2 times a week , 2-3 times a month , Once a month = Once a week - Once a month

Merging levels of the outcome for the model predicting W8QMAFI

These levels were merged as instructed in the EOT Project instructions

W8QMAFI

- 1 , 2 = 1 (Well)
- 3 , 4 , 5 = 0 (Poorly)

Working circumstances

When I went back home on the 16th of March I was sent into a COVID-19 quarantine centre, directly from the airport. I spent 2 weeks there. This was in line with the strict regulations introduced by the government to control the spread of COVID-19. All passengers coming in to the country from "high risk" countries were mandated to be sent in to quarantine at quarantine centres run by the government. As a result, I was unable to carry out any academic activity during this period. I was only able to resume academic activity on the 1st of April as this was when I was released from the quarantine centre. Whilst I can't complain about the working conditions back home I have a bit of a challenge of a space restriction. Both my parents are also working from home and are occupying the area designated as the "study" where I used to do my studies. As a result, I now work from my bedroom and am compelled to work from my bed. At times this is a bit uncomfortable but now I've got used to it. Due to the limited bandwidth of the WIFI at home at times I have connectivity problems.

I worked in a group for the purpose of developing the models. The model predicting W8DINCW and the logistic regression was developed by me. The other model was developed as a group. However, I did some final tweaking on my own.

Despite the challenges above I thoroughly enjoyed this project and am happy with the outcome.