Hutcheson, G.D. (2008b). Dispositions towards studying Mathematically-Demanding subjects in HE. TLRP working paper.

Aims and objectives:

The aim of this working paper is to model (and hence illustrate the modelling approach/heuristics in the project) student disposition to study mathematically-demanding subjects in HE (MHEdisp) over three time points (represented as MHEdisp1, MHEdisp2 and MHEdisp3, which broadly corresponds to the start of AS, the end of AS and during A2). This paper will describe dispositions at each time point and also investigate which variables are associated with these. Using ordinary least square (OLS) regression models we can obtain a prediction of MHEdisp at any point in time. For example, given information about the course a student is enrolled on, their language use and TierGrade, one can make a prediction of MHEdisp at time point 1 using the OLS regression model:

$$MHEdisp1 = Course + Language + TierGrade$$

These models will help us to answer questions such as: "Is the course a student enrols on associated with their MHEdisp score at time 1?", or "Is disposition associated with TierGrade, even after taking into account the effect of course and language use?". Such regression models will determine the existence of associations in the data, which may give indications of causal relationships. In addition to these analyses, it is also useful to investigate the changes in MHEdisp over time. For example, the change in MHEdisp between time points 1 and 2 may be modelled using an OLS regression

$$(MHEdisp2 - MHEdisp1) \sim Course + TierGrade + Pedagogy$$

These models will help to answer questions such as "does having a teacher with a high pedagogy score affect changes in MHEdisp over time?" and "are changes in disposition associated with tiergrade and course enrolled on?".

The data include a large number of interacting variables that may influence dispositions. It is also the case that disposition scores are highly variable over time and may also be subject to considerable random variation, or at least variation due to variables that have not been recorded. In these circumstances, there may be a large number of similarly-fitting models that can be applied. A major challenge for this analysis is therefore to select appropriate models to represent the relationships in the data.

Model selection

A number of "underlying factors" might be influencing the student's disposition and changes in dispositions over time. These individual factors may be represented by more than one variable in the data set which can result in multicollinearity and the associated problems this causes for model-selection. For example, if socio-economic status is related to someone's disposition to study a mathematically-demanding subject at HE (through a complex interaction of attitudes, opportunities and abilities), including a measure of socio-economic status in a model of disposition would be appropriate. Problems occur, however, when socio-economic status is measured using more than one indicator, as typically only a single or small subset of the socio-economic indicators can enter the model at one time and still provide a useful parameter estimate. This can make it difficult to select which indicator of socio-economic status should be retained in the model and it also makes it difficult to interpret the "true" effects of socio-economic status.

The following analysis attempts to identify the important influences on disposition to study mathematically-demanding subjects in HE and include variables that represent these in predictive models in order to assess relative importance. An optimal subsets regression technique (see Fox, 2002) is used to identify all models and from these select the group of "best-fitting" models from which a final model is chosen. The final model selection is based on a number of criteria including practical and theoretical considerations as well as statistical significance. The aim is to select a final model that is not only "good-fitting", but is also interpretable with regards to the underlying factors that the variables represent.

Variables to be considered as explanatory variables in the analyses:

There are many variables that might be of use when predicting dispositions and changes in dispositions over time. Table 1 below shows a number of variables that are to be considered for inclusion in a model of MHEdisp. These variables will be included in an optimal subsets regression model and a final model selected from one of the best-fitting, according to the BIC statistics. It is likely that a number of variables are highly related leading to relatively few variables from each "factor" entering into any regression model. Selecting the final model from the list of best-fitting involves considering a number of aspects of the model including the number of missing data points, the amount of variance explained and the effect size of the variable in question. Also, some non-statistical decisions are made about the likely importance of the variable theoretically, the ease with which it has been measured and can be interpreted, and its usefulness in relation to other models.

Table 1: Explanatory variables considered for models of MHEdisp

categories	**Variables**
Background and SES variables:	Ethnicity, Language, Gender
Family University history	FirstgenerationHE, uniFAM
Socio-economic indicators	EMA, LPN, HEFCE_social_group
"Achievement" variables:	TierGrade (in maths), GCSEEnglishJB.cont, GCSEtotalJB, MSE1, MSE2, MSE3
Teaching/course variables:	Course, VocationalCourse, OtherSubjectsMathsDP2, OtherSubjectsMathsDP3, AveragePed

Changes in MHEdisp over time:

MHEdisp scores at different time points are correlated as shown in Table 2 below. MHEdisp scores are more highly correlated the closer they are together temporally (i.e., MHEdisp1 is more highly correlated with MHEdisp2 than MHEdisp3). Although all correlations are highly significant, there is a degree of variation in the scores over time, with considerable variation between individuals at different time points (see Figure 1).

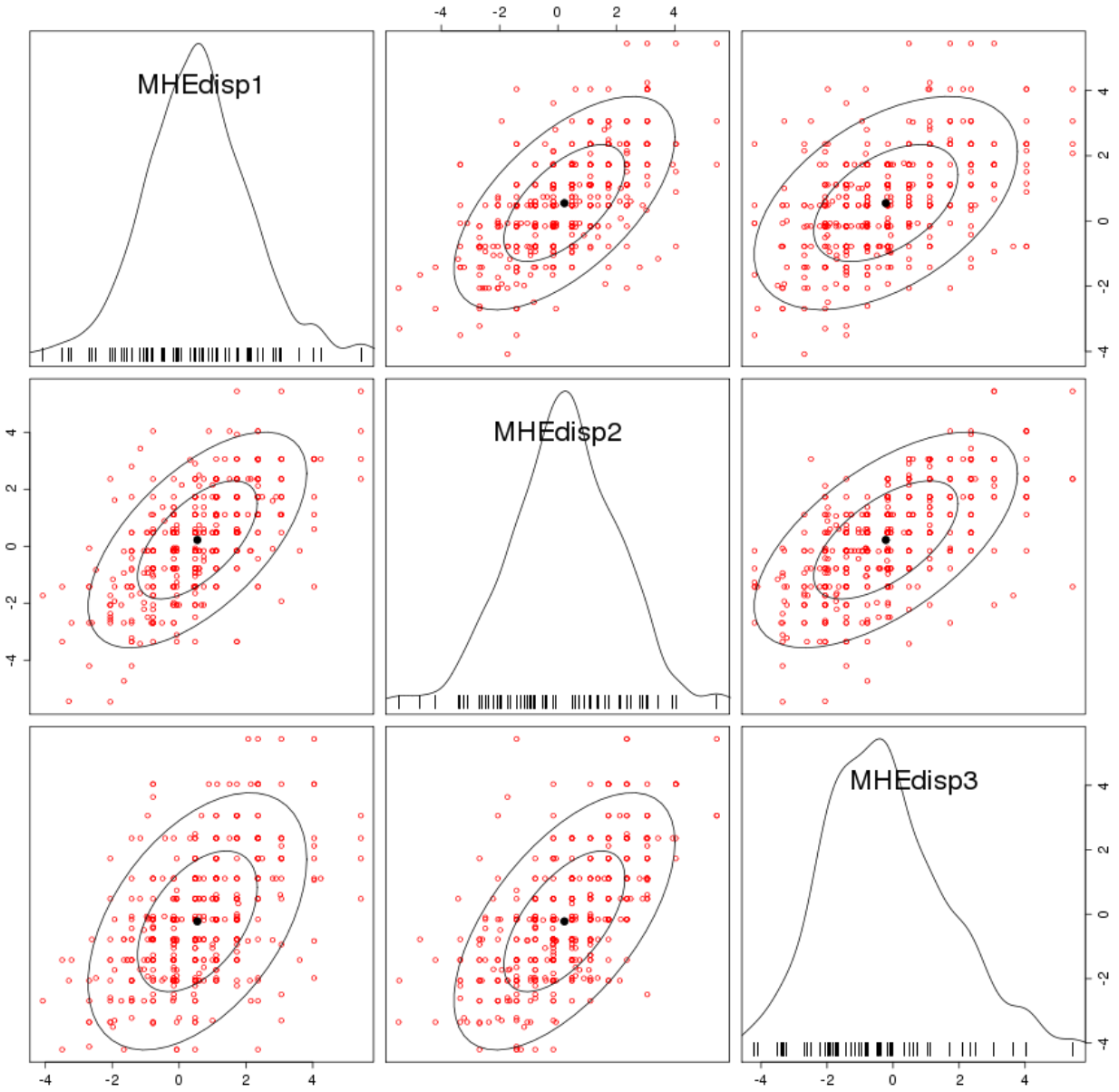
Table 2: correlations between MHEdisp scores at different points in time

	MHEdisp1	MHEdisp2	MHEdisp3
MHEdisp1	1.0000000	0.6174238	0.4814311
MHEdisp2	0.6174238	1.0000000	0.6160178
MHEdisp3	0.4814311	0.6160178	1.0000000

Modelling MHEdisp at time 1:

MHEdisp at time 1 was modelled using techniques that attempted to identify the important variables individually and then combined these (often highly related) variables into a single regression model to provide a prediction of the disposition score. In addition to this, all variables were analysed simultaneously in an optimal subsets model to check that the selection had provided one of the best-fitting models for the data. The model-building process was very detailed and can only be summarised very briefly in this paper. The authors have, however, tried to provide models that are useful for explanatory as well as predictive purposes.

Figure 1: correlation between disposition scores at three different points in time



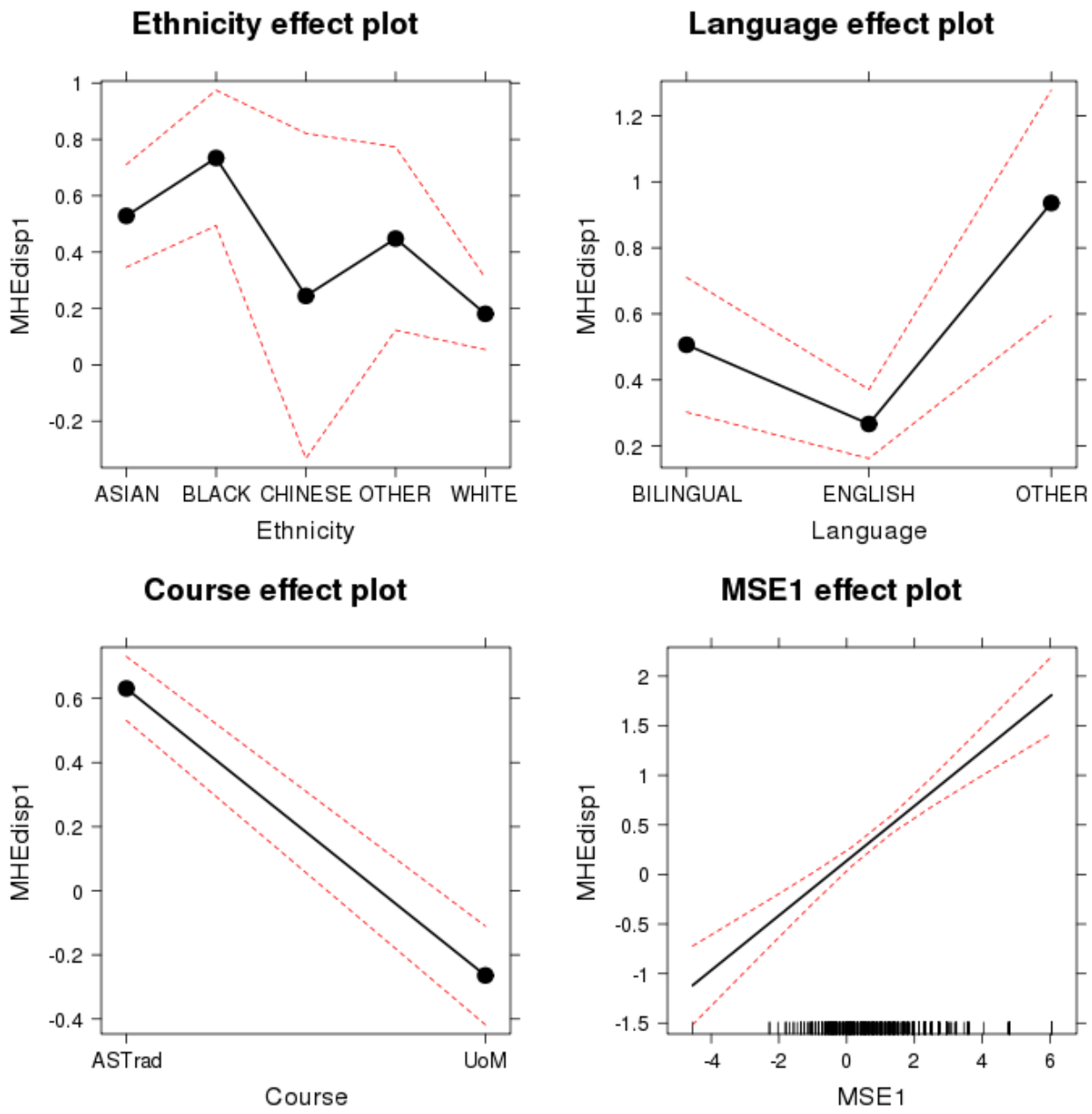
After comparing many models of MHEdisp it was found that a student's self efficacy score (which is highly related to previous examination results) and the course that they had enrolled on were the most significant predictors of MHEdisp. In addition to this, ethnicity and language also significantly predicted disposition scores. Models derived using the optimal subsets regression technique also indicate that a four-variable model is preferable, with MSE, course, ethnicity and previous examination results in English combining to form the best-fitting model according to the BIC statistics. These two model selection techniques provide very similar impressions, with similar underlying factors entering into the models. For this analysis we will choose the first model (which includes language rather than the GCSE examination result in English) as this model provides a slightly higher adjusted R-square and is an easy variable to interpret. The regression model is shown below along with the effects plots illustrating the relationships between the response variable and each explanatory.

```
Model:  MHEdisp1 = Ethnicity + Language + Course + MSE1

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.72228    0.10571   6.833 1.38e-11 ***
Ethnicity[T.BLACK]    0.20551    0.15119   1.359  0.17433
Ethnicity[T.CHINESE] -0.28416    0.30304  -0.938  0.34860
Ethnicity[T.OTHER]   -0.08057    0.18849  -0.427  0.66915
```

```
Ethnicity[T.WHITE]   -0.34762    0.12391  -2.805  0.00512 **
Language[T.ENGLISH]  -0.24001    0.12509  -1.919  0.05528 .
Language[T.OTHER]     0.42959    0.19180   2.240  0.02531 *
Course[T.UoM]        -0.89526    0.09476  -9.447  < 2e-16 ***
MSE1                  0.27628    0.03693   7.482 1.51e-13 ***

Multiple R-squared: 0.1856,      Adjusted R-squared: 0.1797
F-statistic: 31.15 on 8 and 1093 DF,  p-value: < 2.2e-16


Anova Table (Type II tests)

Response: MHEdisp1
           Sum Sq   Df F value    Pr(>F)
Ethnicity   35.58    4  4.5622  0.001172 **
Language    26.71    2  6.8497  0.001106 **
Course     174.00    1 89.2503 < 2.2e-16 ***
MSE1       109.12    1 55.9734 1.505e-13 ***
```

The regression diagnostics for the model above show little evidence of systematic variation and there are no suggested transformations that can be applied to the model that substantially improves the model fit (for example Box-Cox and Box-Tidwell transformations). For this model it is interesting to look at the effect plots in Figure 2 which give an indication of the direction of the effects for each of the explanatory variables on the response.

Figure 2: Effects plots for the MHEdisp1 model



From the effects plot in Figure 2 it can be seen that white, English speaking students have relatively low disposition scores and that those students who select the Uses of maths course and have relatively low self-efficacy scores also tend to have low disposition scores. Particularly striking is the differences between the ethnic groups even after the other factors have been taken into account. Table 3 shows the parameters for the different ethnic groups for the four-explanatory variable model shown above. We can see that the White students have significantly lower MHEdisp scores compared to other ethnic groups.

Table 3: Individual ethnicities compared to the average of all ethnicities

	Estimate	Std. Error	t value	Pr(>|t|)	
Ethnicity[S.ASIAN]	0.30620	0.10314	2.969	0.00305	**
Ethnicity[S.BLACK]	0.29373	0.12660	2.320	0.02051	*
Ethnicity[S.CHINESE]	-0.06800	0.25553	-0.266	0.79021	
Ethnicity[S.OTHER]	-0.07813	0.15687	-0.498	0.61853	
Ethnicity[S.WHITE]	-0.45380	0.09260	-4.901	1.09e-06	***

Modelling MHEdisp at time 2:

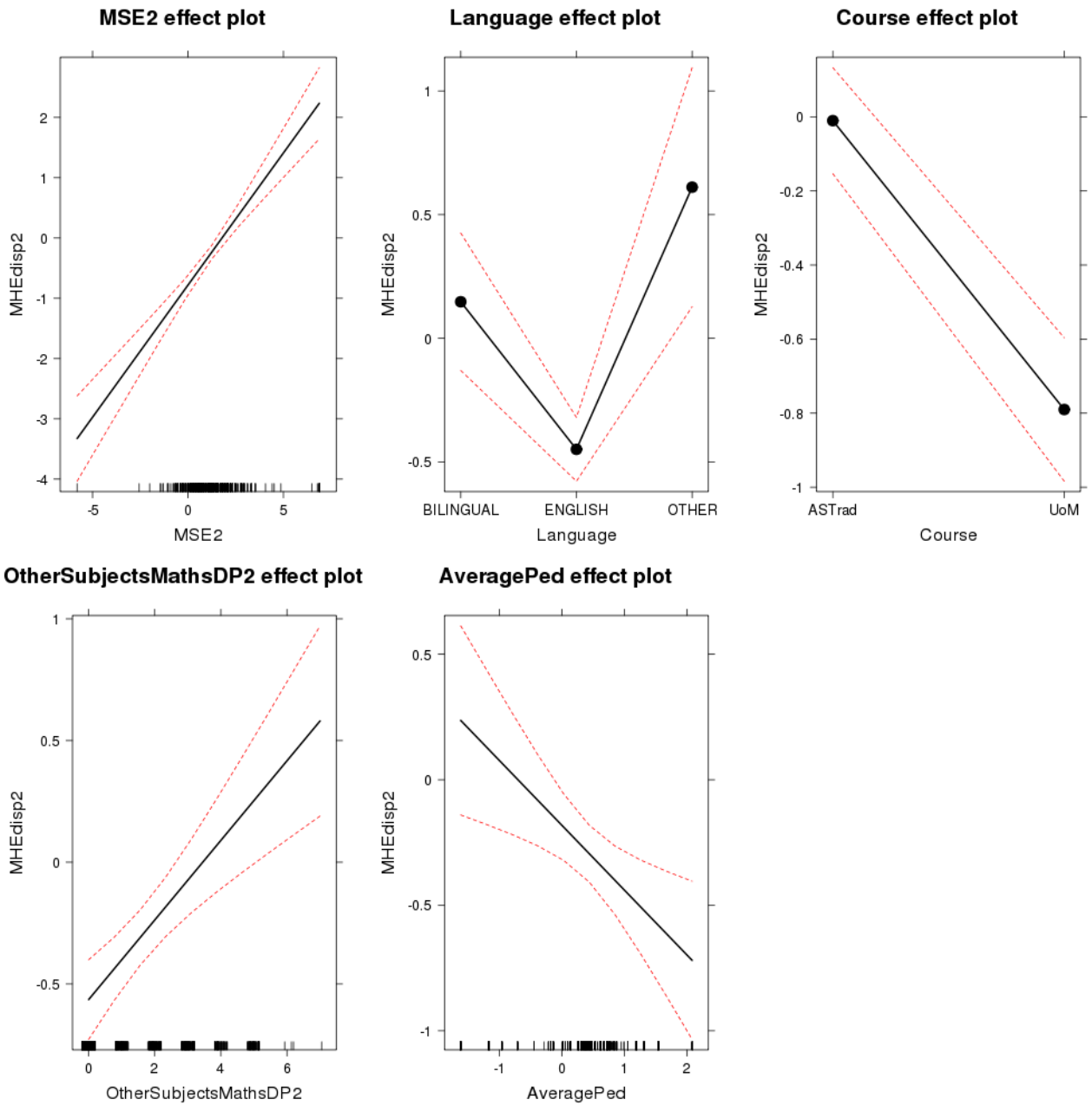
MHEdisp at time 2 can be modelled using similar information to that used for data point 1, but may also include the average pedagogy score of the class (although this variable is, perhaps, of most interest when used in a model looking at the change in MHEdisp scores) and OtherSubjectsMathsDP2 (how mathematically-demanding the other subjects are that the student is taking). Using a similar model to that used to model MHEdisp at time point 1 (apart from the addition of Average Ped and OtherSubjectsMathsDP2), we find that:

```
Model:  MHEdisp2 = Ethnicity + Language + Course + MSE2 +
                   AveragePed + OtherSubjectsMathsDP2

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.11448    0.19091  -0.600  0.54892
Ethnicity[T.BLACK]     0.13523    0.24637   0.549  0.58326
Ethnicity[T.CHINESE]  -0.09018    0.43301  -0.208  0.83508
Ethnicity[T.OTHER]    -0.26626    0.29450  -0.904  0.36626
Ethnicity[T.WHITE]    -0.36373    0.18212  -1.997  0.04620 *
Language[T.ENGLISH]   -0.42708    0.19118  -2.234  0.02581 *
Language[T.OTHER]      0.41815    0.28278   1.479  0.13968
Course[T.UoM]         -0.77903    0.12870  -6.053 2.35e-09 ***
MSE2                   0.46390    0.05344   8.681  < 2e-16 ***
AveragePed            -0.26541    0.09521  -2.788  0.00546 **
OtherSubjectsMathsDP2  0.14801    0.03604   4.107 4.50e-05 ***

Multiple R-squared: 0.268,Adjusted R-squared: 0.2572
F-statistic:  24.9 on 10 and 680 DF,  p-value: < 2.2e-16


Anova Table (Type II tests)

Response: MHEdisp2
                       Sum Sq  Df F value    Pr(>F)
Ethnicity               15.55   4  1.6863  0.151351
Language                27.11   2  5.8799  0.002939 **
Course                  84.47   1 36.6372 2.352e-09 ***
MSE2                   173.75   1 75.3611 < 2.2e-16 ***
AveragePed              17.92   1  7.7713  0.005457 **
OtherSubjectsMathsDP2   38.88   1 16.8640 4.504e-05 ***
```

Ethnicity is not significant for this model, but has been included to enable comparisons with the model at time point 1. As we have found, there is a complex relationship between ethnicity and course selection which affects the significance of ethnicity. An optimal subsets regression technique also identifies a 5-parameter model as giving the best model fit (according to the BIC statistics) and identifies the same variables as in the model above (MSE, Course, Language, other subjects and pedagogy).

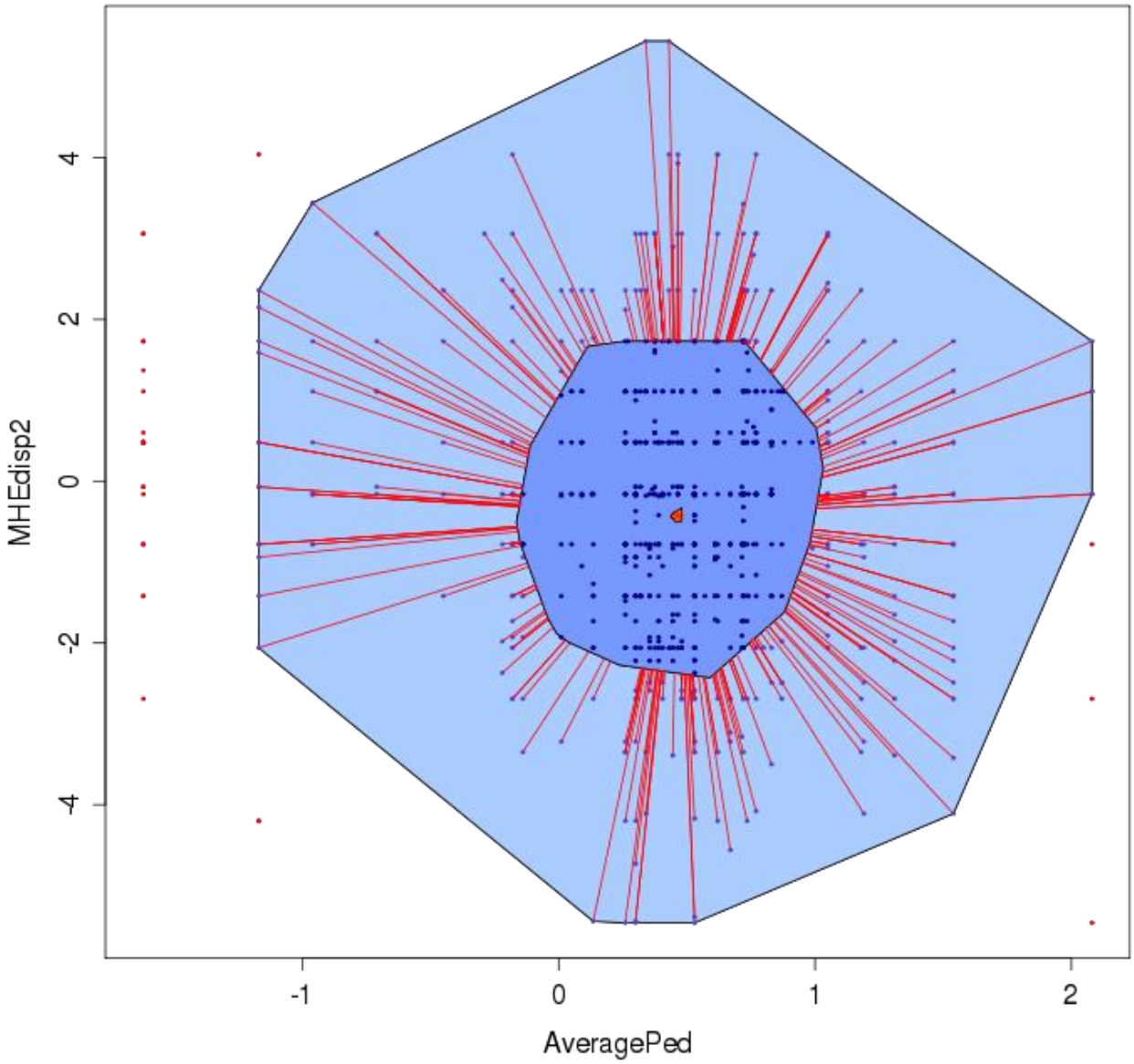
A very similar picture of MHEdisp at time 2 is obtained to that at time 1. From the effects plots shown in Figure 3 below it can be seen that those students who have a high self-efficacy score and take other mathematical subjects also tend to have a high disposition score. These students also show a preference for the ASTrad course. Those students with English as a first language show relatively low disposition scores as do those exposed to high pedagogy scores.

Figure 3: Effects plots for the MHEdisp2 model



It is worth noting that AveragePed is negatively associated with MHEdisp, although this relationship is quite weak as illustrated by a simple bagplot shown in Figure 4 (see Rousseeuw, Ruts and Tukey, 1999). Figure 4 suggests that even though there is a significant association, this association may be due to the scarcity of low disposition and low pedagogy scores. In deed, removing the low pedagogy scores (below -.5) renders the relationship insignificant.

Figure 4: A bagplot of MHEdisp2 and AveragePed


MHEdisp2
4
2
0
-2
-4
-1
0
1
2
AveragePed

Modelling MHEdisp at time 3:

Using a similar model to that used to model MHEdisp at time point 2 we find that disposition at time point 3 can be predicted using self efficacy scores, other subject choice, English marks, Language (or Ethnicity) and Tiergrade. This 5-parameter model is also one of the best fitting models selected using the optimal subsets procedure. The major difference at time 3 is that course and pedagogy are not significant predictors when other variables are taken into account. For this model, we have decided to include Ethnicity rather than language (the two models are very similar with respect to model-fit) as Ethnicity is, perhaps, easier to interpret.

```
Model:  MHEdisp3 = MSE3 + GCSEEnglish.cont + TierGrade.cont +
                   OtherSubjectsMathsDP3 + Ethnicity

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.11518    0.41013  -0.281 0.778993
MSE3                   0.40793    0.06763   6.032 3.81e-09 ***
GCSEEnglish.cont      -0.31358    0.09160  -3.424 0.000684 ***
TierGrade.cont         0.20912    0.06693   3.124 0.001917 **
OtherSubjectsMathsDP3  0.18981    0.05492   3.456 0.000609 ***
Ethnicity[T.BLACK]    -0.40165    0.32175  -1.248 0.212674
Ethnicity[T.CHINESE]   0.08327    0.83007   0.100 0.920148
Ethnicity[T.OTHER]    -0.80139    0.35290  -2.271 0.023709 *
Ethnicity[T.WHITE]    -0.73543    0.19103  -3.850 0.000138 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared: 0.2315,      Adjusted R-squared: 0.2155
F-statistic:  14.5 on 8 and 385 DF,  p-value: < 2.2e-16


Anova Table (Type II tests)

Response: MHEdisp3
                       Sum Sq  Df F value    Pr(>F)
MSE3                    94.81   1 36.3797 3.812e-09 ***
GCSEEnglish.cont        30.55   1 11.7209 0.0006844 ***
TierGrade.cont          25.44   1  9.7611 0.0019175 **
OtherSubjectsMathsDP3   31.13   1 11.9461 0.0006087 ***
Ethnicity               42.67   4  4.0934 0.0029183 **
Residuals             1003.33 385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

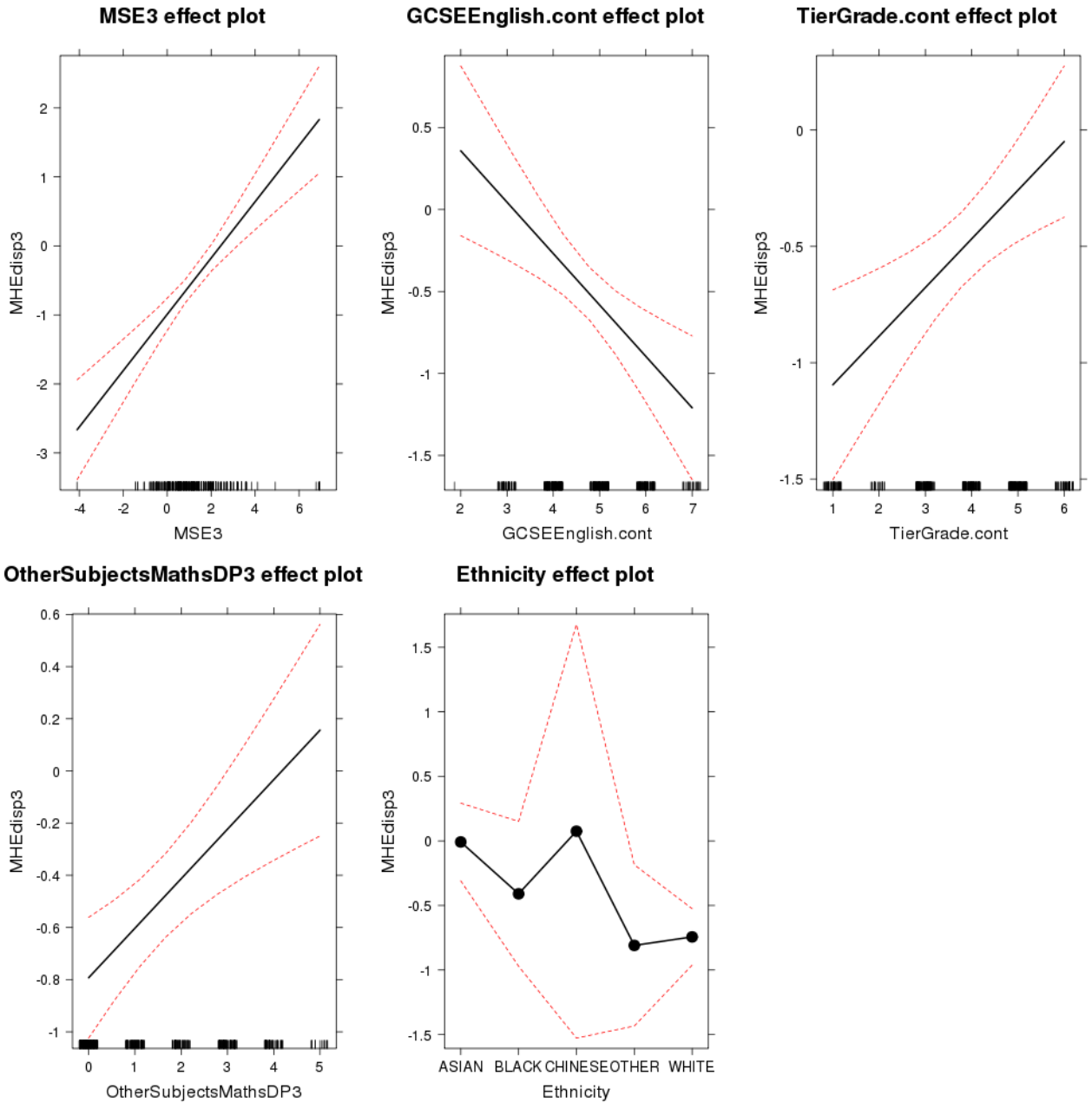
From the effects plots shown in Figure 5 below it can be seen that those students who have a high self-efficacy score, high Tiergrade and take other mathematical subjects also tend to have a high disposition score. Compared to Asians, White students show relatively low dispositions as do those who obtain relatively high scores in the English GCSE examinations.

Conclusion

A student's disposition to study a mathematically-demanding subject in HE is mainly related to their self-efficacy scores, choice of course and subjects and also their previous performance in mathematics and English. It is particularly striking that white English first language students have a significantly lower disposition to study mathematics in HE. There is some evidence that pedagogy is associated with disposition scores at time 2, but this relationship is weak and may be due to the effect of outliers.

It should be noted that the R-square statistics for all models are in the range 0.18 to 0.27, which indicates that there is a lot of unaccounted for variation in the models.

Figure 5: Effects plots for the MHEdisp3 model



Modelling changes in disposition over time.

It is useful to also look at which variables may be associated with changes in disposition over time. The following analysis models the change in disposition scores using similar techniques to those employed above. Regression models are to be carefully selected from a set of best-fitting models as determined by the optimal subsets procedure.

Three sets of models were considered to show the development of disposition scores over time; time 1 to time 2, time 2 to time 3 and time 1 to time 3. Optimal subsets regressions were computed on all models including all data available at the time of the observations (for example, the time 1 to time 2 model did not include any explanatory variables recorded at time 3). The best fitting models (at least according to the BIC statistic) all included 2 to 3 parameters and were constructed using the variables shown in Table 4 below.

Table 4: variables contributing to best-fitting models

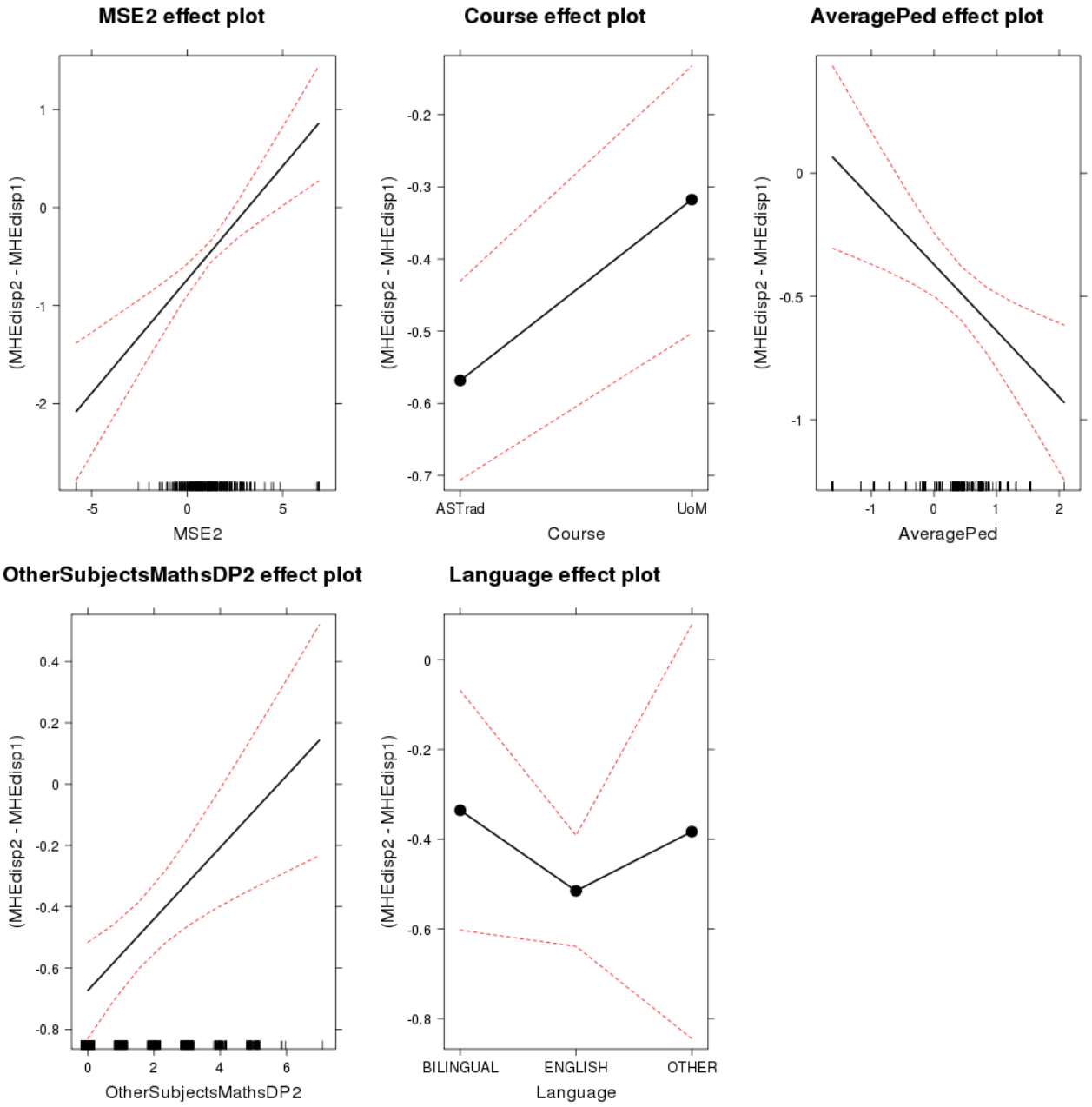
Time 1	Time 2	Time 3
MSE, Course, Average Pedagogy, Other courses, Language		
	MSE, Course, Average Pedagogy, Other courses, Language	
MSE, Course, Average Pedagogy, Other courses, Low Participation Neighbourhood		

We can see from Table 2 that similar variables are related to the changes in disposition over time and these are self efficacy scores, Course, Average Pedagogy, Other courses taken and Language. An example analysis containing these variables is shown below (the model here is change in MHE disposition scores between time 1 and 2).

```
Model: (MHEdisp2 - MHEdisp1) = MSE2 + Course + AveragePed +
                               OtherSubjectsMathsDP2 + Language

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.77137    0.16939  -4.554 6.23e-06 ***
MSE2                  0.23147    0.05080   4.556 6.16e-06 ***
Course[T.UoM]         0.25065    0.11964   2.095 0.036537 *
AveragePed           -0.26877    0.08909  -3.017 0.002648 **
OtherSubjectsMathsDP2  0.11660    0.03451   3.379 0.000769 ***
Language[T.ENGLISH]  -0.17960    0.15064  -1.192 0.233589
Language[T.OTHER]    -0.04758    0.27033  -0.176 0.860353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared: 0.06852,     Adjusted R-squared: 0.06042
F-statistic:  8.46 on 6 and 690 DF,  p-value: 7.059e-09


Anova Table (Type II tests)

Response: (MHEdisp2 - MHEdisp1)
                       Sum Sq  Df F value    Pr(>F)
MSE2                    44.21   1 20.7581 6.162e-06 ***
Course                   9.35   1  4.3889 0.0365374 *
AveragePed              19.39   1  9.1017 0.0026476 **
OtherSubjectsMathsDP2   24.32   1 11.4171 0.0007686 ***
Language                 3.36   2  0.7896 0.4544585
Residuals             1469.58 690
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The effects plot for this model is shown in Figure 6 below.

Figure 6: Effects plots for the change in dispositions between time points 1 and 2



For all models, there are a number of "common effects" that are summarised below:

- Although dispositions generally go down over time (Figure 5 shows much of the disposition change to be negative), the greatest reductions are from those with low self-efficacy scores. Mathematical dispositions are decreasing over time and mathematics seem to gain few converts.
- The decrease in disposition is more noticeable for those on the ASTrad courses (all other things being equal).
- High pedagogy (i.e. more transmissionist, teacher-centred) scores are associated with decreasing dispositions.

- Those students taking fewer math-related courses are associated with lower and decreasing dispositions.

- Generally, white, English students are associated with decreasing dispositions over time (all other things held constant).

It should be noted that the R-square statistics for all models predicting change in disposition are below 0.1, which indicates that there is a lot of unaccounted for variation in the models. In almost all these models, furthermore, the effect of mathematics self efficacy (MSE) serves as a black-box for sometimes significant effects of other variables that are not reported here.

The main purpose of the working paper has been to illustrate the modelling of the quantitative data – the completion of this analysis and the definitive, substantive interpretation of all these models is to come.

References

Fox, J. 2002. *An R and S-Plus companion to Applied Regression.* London: Sage Publications.

Rousseeuw, P. J., Ruts, I. and Tukey, J. W. (1999): The bagplot: a bivariate boxplot. *The American Statistician*, **53, 4:** 382–387.