Hutcheson, G.D., Black, L., Davis, P., Hernandez, P., Nicholson, S., Pampaka, M., Wake, G., Williams, J.S. (2008a). **ASTrad and UoM courses: factors influencing enrolment.** TLRP working paper.

Introduction

The aim of this paper is to describe the students who have enrolled on the ASTrad and UoM courses and the factors that might have influenced this enrolment. The modelling process is used to identify the underlying factors that might affect course selection and also to select a subset of variables that can be used for a predictive model of course selection.

It can be hypothesised that a number of "underlying" factors might be influencing course enrolment. For example, course enrolment might be associated with the socio-economic status of the student and previous examination grades. One can also hypothesise other "factors" that may also be associated with course such as the family history of University attendance, disposition to continue to higher education and other variables such as ethnicity and gender. These factors are represented in the data set by single and multiple variables and the factors also interact with each other (for example, the two distinct factors "ethnicity" and "socio-economic indicators" are associated with each other).

The following analysis attempts to identify the important influences on course choice and include these in a predictive model to assess their relative importance. The models are then compared to see which are the best-fitting and also the most useful for descriptive and practical purposes. The aim is to select a final model that is not only good-fitting, but also interpretable with regards to the underlying factors that the variables represent.

Methods

The following analyses considered all variables that possibly influenced the students' choice of course. The list of variables considered is described in Table 1. It is likely that these correlated variables represent a smaller group of "factors" that may indicate important relationships. It is important to realise that some of the variables clearly attempt to measure the same underlying construct, making it likely that only one variable from this cluster will enter into a regression model even though all may be significantly related at the binary level. The art of model-building in this case is to carefully select the variable that best represents the underlying factor, whilst realising that this might not be the most significant variable.

Table 1: Variables considered and their description

Variable Name	Description [categories]
Gender	[Male, Female]
Language	Language of First choice [English, Bilingual, Other]
EMA	Education Maintenance Allowance [YES, NO]
firstgenerationHE	***First generation at HE*** [YES, NO] The student will be in the first generation in his family to go at university
uniFAM	***Family at university*** [Firstgeneration, Parents, Siblings] This variable is a combination of the above including as a category the same generation (i.e. siblings) who go to university
TierGrade.cont	A numeric variable [1 to 6] with the numbers defined by Tier and grade at GCSE, as follows: 1_intC [Intermediate C] 2_HigherC [Higher C] 3_intB [Intermediate B] 4_HigherB [Higher B] 5_A [A] 6_A* [A*]
Ethnicity	[Asian, Black, Chinese, Other, White]
LPN	***Low Participation Neighbourhood*** [YES, NO] The student gave a home address located in an area for which the participation in HE is less than two thirds of the UK average (based on postcode analysi).
HEFCE_ social_group	***HEFCE Social Group*** [values 1-4] The values indicate affluence of a postcode and are simply quartiles based on the Claritas demographics data. (1 represents postcodes classified with the most affluent Claritas clusters; 4 represents postcodes classified with the least affluent Claritas clusters – from postcode analysis).
UNIoptionDP1	The ***University preferred option*** [outcome] variable for each DP. Categories are described below: -STEM (just STEM) -STEM-M (Mathematically demanding STEM) -M-nonSTEM (non STEM but mathematically demanding) -nonSTEM -undecided -notUNI (they will not go to university) -NA (no information available – missing data)
MSE1	***Mathematics Self- Efficacy*** at DP 1 [Min: -4.55, Max: 6.05]
MHEdisp1	***Disposition to Study (more) Mathematically demanding subjects*** at DP1 [Min: -5.46, Max: 5.44]
Hedisp1	***Disposition to go into HE*** at DP1 [Min: -6.84, Max: 4.32]
VocationalCourse	The students are enrolled in a vocational course [YES, NO]

We have used an optimal subsets regression technique (see Fox, 2002) to select the variables that are most associated with course enrolment. The technique used assesses

model-fit on the basis of the BIC statistic (Schwarz, 1978; Akaike, 1978) which aims to achieve a relatively parsimonious model by penalising those models with larger numbers of parameters. The optimal subsets regression computes all regression models from the available data (a very substantial number) and outputs these models to a graph. Figure 1 shows a restricted set of analyses where only the best 3 models are provided for solutions containing 4 to 7 parameters. It is important to note here that Figure 1 was selected after a more exhaustive search through greater numbers of models and a wider range of parameters. The "best-fitting" models were found to use between 4 and 6 parameters. It is also important to note that the models are compared on individual parameters, so categorical variables such as ethnicity, appear as a number of individual parameters rather than a single parameter illustrating the effect of the variable overall.

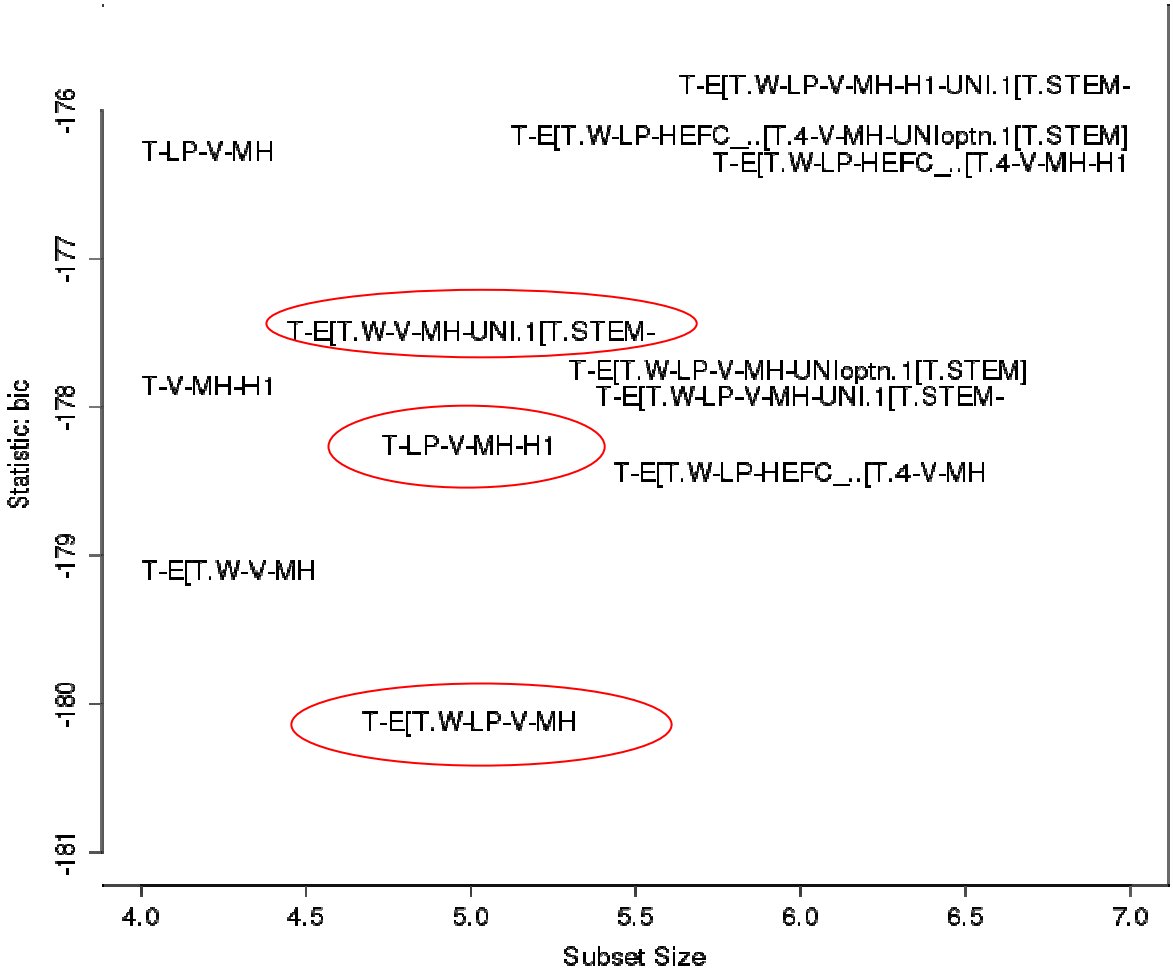

Figure 1. Optimal subsets regression model of course choice.

The variables in the graph are identified by the codes given in Figure 2. The graph shows the 3 best fitting models on the basis of the BIC criteria for models with 4,5,6 and 7 parameters. Those models which appear lower down in the graph are the ones that fit better. For example, the best fitting model with 4 parameters (excluding the constant) is T-E[T.W-V-MH, which indicates the 4 parameters TierGrade, Ethnicity(white), Vocational and

disposition to study a mathematically-demanding subject. These variables indicate quite distinct underlying factors. It is interesting to note that the best-fitting model according to our criteria is a 5-parameter solution that also includes the socio-economic variable LP (low-participation neighbourhood). It is interesting to note the strong representation of the variables T-V-MH in all models. A slightly poorer fitting 5-parameter model includes the disposition to study in HE rather than Ethnicity(white). Both of these variables are quite strongly related to course choice (although more weakly than T, V, MH and LP) and may enter the model as the fifth variable. The third best-fitting model displaces the socio-economic variable and replaces it with the choice of subject. As the difference in BIC scores between these three models is quite small, the choice of which one(s) to interpret will depend on the research questions (for example, if information about subject choice is particularly interesting, one might consider interpreting the third best-fitting model). For this analysis we will interpret the best-fitting 5-parameter model.

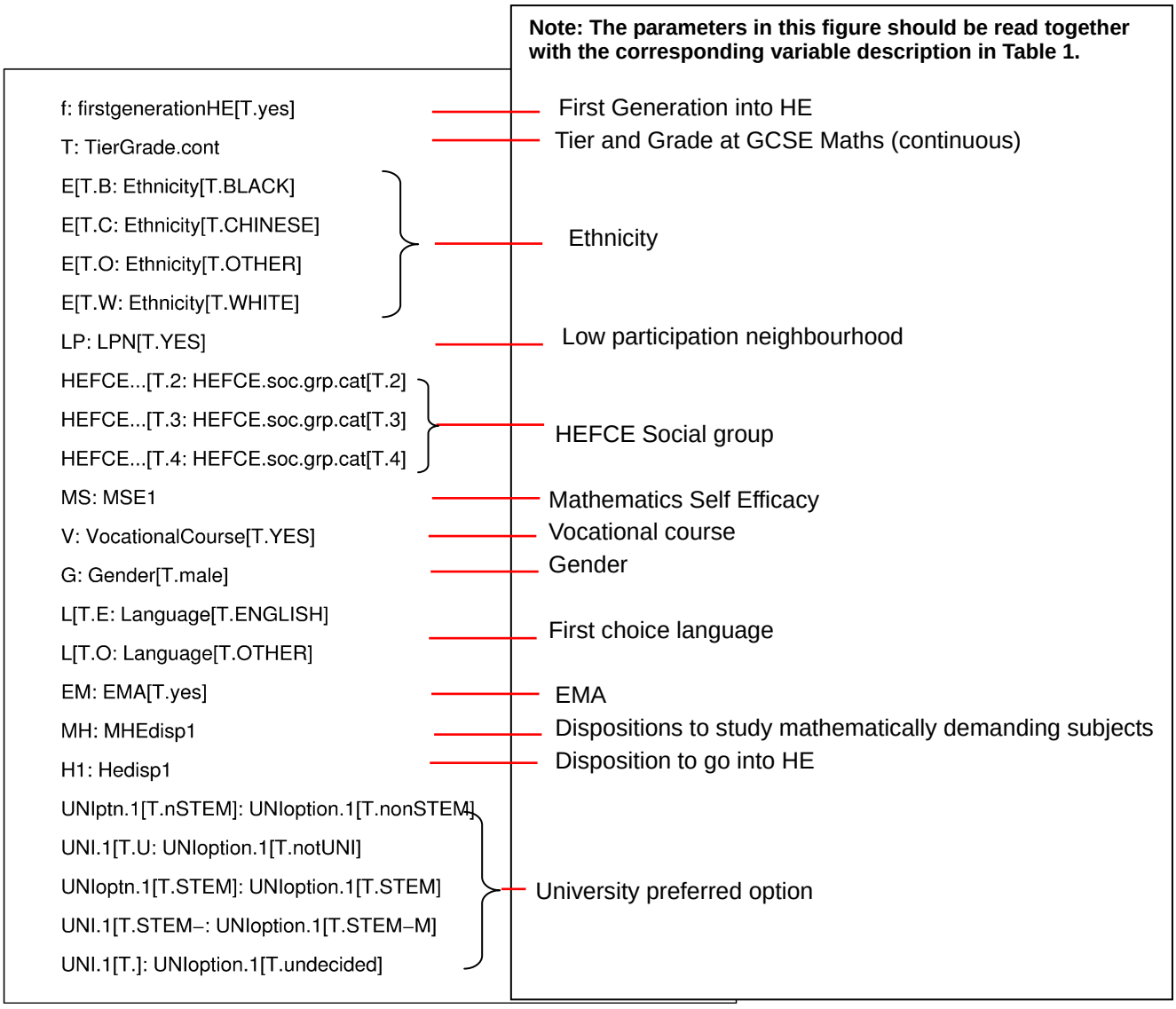

Figure 2: Variables considered for the optimal subsets regression model

Results

Course enrolment is modelled using a logistic regression model. The 5-parameter model identified above as the best-fitting is shown in Tables 2 (Regression model) and 3 (ANOVA) below:

Table 2: A Logistic Regression Model for course choice

Coefficients	Estimate	Std. Error	Z value	Pr (>|z|)
(Constant)	1.4690	0.4472	3.285	0.00102 **
TierGrade.cont	-0.9699	0.1110	-8.738	< 2e-16 ***
VocationalCourse[T.YES]	1.5614	0.3640	4.289	1.79e-05 ***
MHEdisp1	-0.4091	0.0963	-4.249	2.15e-05 ***
LPN[T.YES]	-0.7741	0.3344	-2.315	0.02064 *
Ethnicity[T.BLACK]	0.7324	0.4661	1.571	0.11611
Ethnicity[T.CHINESE]	0.2032	1.2381	0.164	0.86962
Ethnicity[T.OTHER]	1.3386	0.5303	2.524	0.01160 *
Ethnicity[T.WHITE]	1.0860	0.3580	3.034	0.00242 **

Null deviance: 558.41 on 496 degrees of freedom
Residual deviance: 358.34 on 488 degrees of freedom

Table 3: ANOVA Table (Type II tests)

Response=Course	LR Chisq	Df	Pr(>Chisq)
TierGrade.cont	105.610	1	< 2.2e-16 ***
VocationalCourse	18.843	1	1.420e-05 ***
MHEdisp1	19.614	1	9.477e-06 ***
LPN	5.696	1	0.01701 *
Ethnicity	11.895	4	0.01815 *

The model above can be interpreted precisely by computing the odds ratios for each of the parameters (for example, those students who come from low-participation neighbourhoods (LPN) are less than half as likely (odds ratio = 0.46 ($e^{-0.7741}$)) to select an ASTrad course when compared to those not from a low-participation neighbourhood. The effects plots shown below illustrate the relationship between each explanatory variable and course choice to provide a comprehensive view of the how each variable affects the outcome.

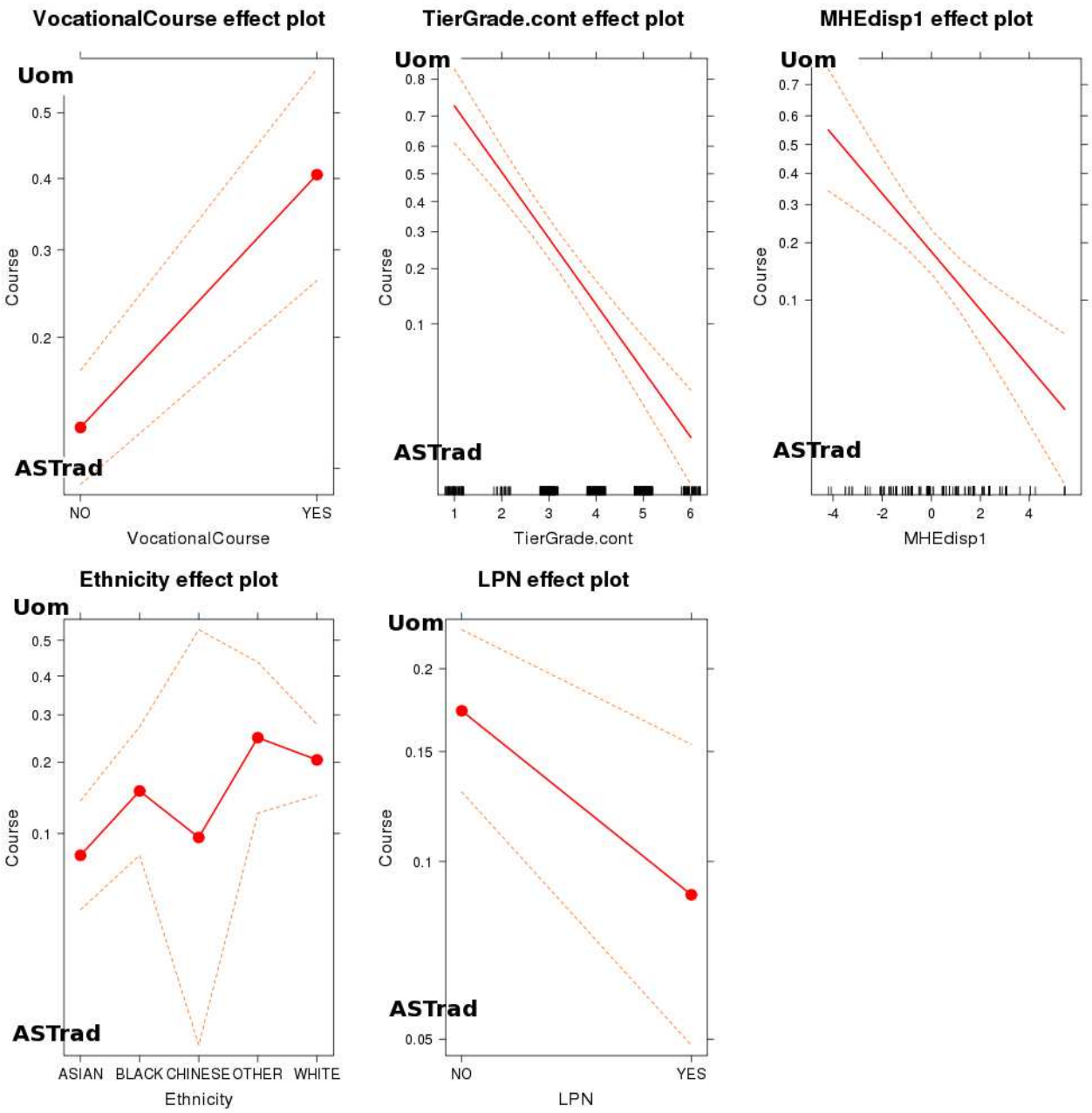

Figure 3: Effects plots for the best-fitting 5-parameter model

In general, course enrolment is associated with a number of factors. Particularly, those with higher grade in GCSE maths (tiergrade) show a strong preference for ASTrad courses as do those with stronger dispositions to study mathematically-demanding subjects in HE. White and other ethnicity groups show a preference for UoM courses compared to Asian students. Those choosing vocational courses tend to select UoM courses and those from low-participation neighbourhoods tend to select ASTrad.

Comparing those students with lower TierGrades

It should be noted that there are far more A and A* students who have selected ASTrad courses. The following table shows the large discrepancy in the numbers.

Table 4: Course enrolment by GCSE Tier and Grade

Course	GCSE Tier/Grade					
	$C^{INTERMEDIATE}$	C^{HIGHER}	$B^{INTERMEDIATE}$	B^{HIGHER}	A	A*
ASTrad	108	53	285	331	322	105
UoM	166	44	143	75	22	2

When the optimal-subsets regression is repeated after the A* and A pupils have been removed from the analysis, the results are very similar, with Tiergrade, vocational courses and disposition to study mathematically-demanding subjects all represented in most good-fitting models.

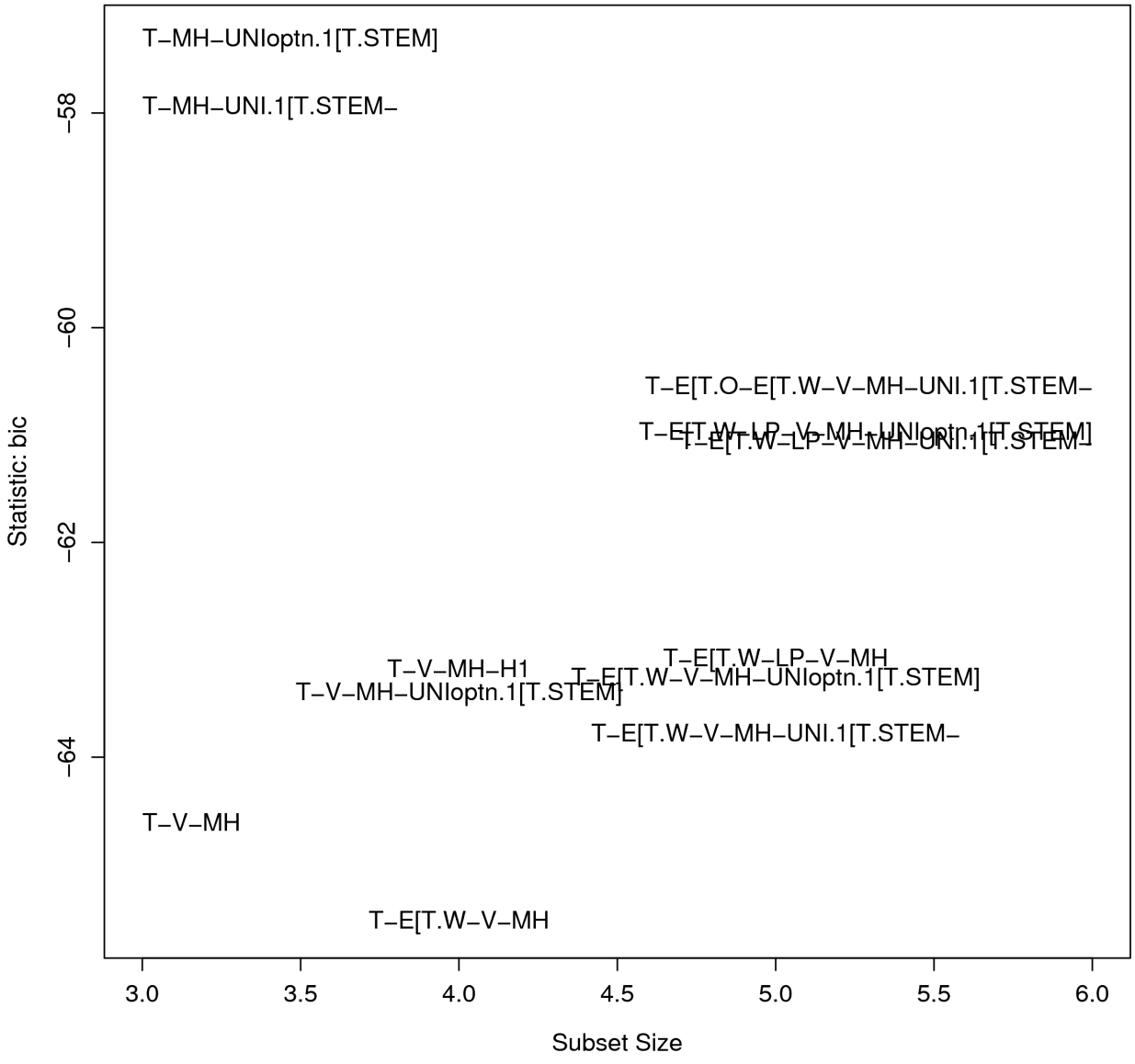

Figure 4: Optimal subsets regression model of course choice when excluding A and A*

Conclusion and discussion

Overall, we find that course enrolment is associated with factors like prior grades, dispositions to study mathematically demanding subjects in HE, ethnicity and vocational courses. Particularly students with higher grade in GCSE maths show a strong preference for ASTrad courses, similarly to those with stronger dispositions to study mathematically-demanding subjects in HE. White and other ethnicity groups are more probably to be enrolled at UoM courses compared to Asian students. Students enrolled in vocational courses tend to attend UoM courses and those from low-participation neighbourhoods tend to select ASTrad.

These patterns are what we might have expected from the case studies: We knew that those intending to progress with mathematics, and with stronger GCSE qualifications, would tend to be entered to traditional "AS mathematics" as this is the 'normal' progression to A2 etc.

The trend with regard to Asian students might have been guessed: we believe this trend is associated with professional aspirations.

Finally, the BTEC students have a need for mathematics that is traditionally satisfied within the BTEC course itself. The fact that our sample is biased towards the "Use of mathematics" may simply be due to our having small numbers of BTEC students, and having sought out a few classes of BTEC engineering students where teachers were doing interesting work with Uses of mathematics.

References

Akaike, H. (1978b). A Bayesian analysis of the minimum AIC procedure*. Ann. Inst. Statist. Math*., 30, 9-14.

Fox, J. (2002). An R and S-Plus Companion to Applied Regression. London: Sage Publications.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.