Update to association between characteristics of students (ABCS)

Annexes

Reference OfS 2020.54

Enquiries to official.statistics@officeforstudents.org.uk

Publication date 26 November 2020

Contents

Annex A: updating the modelling methodology

1. The introduction of the new characteristics to the ABCS FT continuation measure causes a substantial increase in the computational time taken to model the data. As a result, we have reconsidered the modelling approach we take in selecting and including interaction terms. The original ABCS methodology used a more complex method than is usual for including interactions, where some levels of interactions between two categorical variables could be included while others were excluded. The more standard approach is to include either all levels of interactions between two categorical variables, or none of them. To determine which modelling approach to take we have assessed three different techniques: the original ABCS model, the new ABCS model and a hybrid of the two.

2. The difference between the three approaches is in the way that we choose which two interactions to include. In all models, the main effects are included for all characteristics, and the choice of two-way interactions is made using stepwise regression. We use an entry and stay criterion of $\alpha=0.05$. This could be considered to be a very conservative criterion, but because of the size of the data we are using, we believe that is appropriate.

3. In the original method, interactions are not necessarily included for all possible values within a characteristic. For example, if we consider the characteristics of age and sex, it is possible that the interaction of being 18 and female is included, but the interaction between being 51+ and female might not be included. To determine which interactions to include, we create dummy variables for every possible combination of values from two characteristics. These dummy variables are then included in the model alongside the main effects and the stepwise selection method is used to determine which interactions to keep in the model. It is this element of the modelling that has become inefficient following the introduction of the four additional variables.

4. The new ABCS methodology does not allow for only some parts of interactions to be included (as in the example above) either all interactions between two characteristics are included, or none are. This results in the final model including some interactions which are not statistically significant.

5. The hybrid method we explored seeks to align with the original method more closely. We start with the new ABCS method described above, and having obtained the estimated parameters and statistical significance of each of the interaction between categories, we list all of the interactions between categories (such as aged 18 and female) that were shown to be statistically significant. We then re-run the model without the stepwise method (meaning that all variables put into the model will be part of the final model) with only the statistically significant interactions. This does not result in a model with no interactions which are not statistically significant, but it does remove a number of them.

6. The hybrid method is closer to the original method, but less efficient and more complex than the new ABCS method. To determine whether it would be plausible to replace the original methodologies with either of these new approaches, and if so, which one, we have compared the model fit and results from each of the approaches. We were looking for similar results for the three approaches, as this would allow us to pick the most efficient without concerns that we were impacting the results of the modelling, and in turn the ABCS FT continuation measure.

7. We found that there is little difference in statistical model fit across the three approaches, although the hybrid method did show a slightly better fit than the other two. Looking at the predicted continuation rates from the three models, the range is very similar across the three approaches – in particular between the original method and the new ABCS method. The new ABCS method has a slightly narrower range, with the lowest predicted continuation rate being around three percentage points higher than for the other two methods. The same comparison was made having removed all student groups with fewer than 10 students to ensure that small groups were not impacting these results. The ranges of predicted continuation rates across the three methods were even closer having removed small student groups.

8. We also wanted to see whether there were any student groups (that is, particular combinations of the 12 student characteristics) for which the predicted probabilities were very different across the three methods. To do this, we looked at the correlation between the predicted continuation rates from the three modelling approaches – finding that in all cases, there is a very strong correlation ($\rho>0.99$, $p<.0001$). We have also checked for any large student groups whose predicted continuation rate was very different between the three methods. Across the three modelling approaches, there is no group of more than 100 students whose predicted continuation rate differs by more than 3.2 percentage points.

9. These checks have led us to conclude that there is little difference in the results generated by the three different modelling approaches. This is why we have chosen to adopt the simplest and most efficient method (where we include interactions between whole characteristics only) as the new ABCS methodology.

The new ABCS methodology

10. The new ABCS modelling methodology is to use a binary logistic regression model with all characteristics forced in as main effects and every two-way interaction between categories tested using a stepwise regression method with entry and stay criteria of $\alpha=0.05$, and this decides which of the two way interactions are included in the final model.

Annex B: Definitions

Characteristic definitions

11. The characteristics used are all based on variables available on the National Pupil Database (NPD), the Education and Skills Funding Agency's (ESFA's) individualised learner record (ILR) or the Higher Education Statistics Authority's (HESA's) student record or alternative provider (AP) student record. Income Deprivation Affecting Children Index (IDACI)[1] and the Index of Multiple Deprivation (IMD) quintiles are obtained from the Department for Communities and Local Government (DCLG) added to the NPD, ILR, HESA student and AP student records.

Age

12. Age, as used in the full-time and part-time continuation models, is defined as the age of the student on the 31 August in the year in which they started their course. This is calculated using their birth date as reported on the ILR or HESA student or AP student records.

Care experience

13. Care experience describes whether a student has ever spent time in the care of a local authority in England or Wales, or Health and Social Care Trust in Northern Ireland, or who has self-declared as in care for three months or more. The data is collected on the HESA record. The definition and reporting of care experience are complex. For details of the data used, see www.hesa.ac.uk/collection/c19051/a/careleaver.

Disability

14. Disability information is only used from the ILR and HESA student and AP student records. Disability is self-reported by students. Details of the disability types used can be found under the IPDISABLETYPE section in OfS' 2020 Core Algorithms document[2].

Ethnicity

15. Ethnicity information is taken from the NPD for the access model and from the ILR and HESA student and AP student records for the continuation model. For ABCS access, only the NPD is used as a source of ethnicity data, and so we use the groupings as they are collected, with the exception of Gypsy/Roma and Traveller of Irish heritage, which are combined to make a single group since the groups are very small when separated.

16. For continuation, we need to account for the fact that ethnicity is collected differently on the ILR and HESA records. To do this, a new grouping that covers the different levels of collection has been created. Table B1 below gives the details of the variables used and the codes included in each of the 16 ethnicity groups.

[1] See https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

[2] Available from: www.officeforstudents.org.uk/data-and-analysis/institutional-performance-measures/technical-documentation/

Table B1: ABCS continuation ethnicity definitions

ABCS continuation ethnic group	ILR code[3]	HESA code[4]
Asian or Asian British – Bangladeshi	41	33
Asian or Asian British – Chinese	42	34
Asian or Asian British – Indian	39	31
Asian or Asian British – Pakistani	40	32
Asian or Asian British – Other	43	39
Black or Black British – African	44	22
Black or Black British – Caribbean	45	21
Black or Black British – Other	46	29
Mixed – white and Asian	37	43
Mixed – white and black African	36	42
Mixed – white and black Caribbean	35	41
Mixed – Other	38	49
White	31, 32, 34	12, 13, 10, 11, 19
Gypsy, Roma or Traveller	33	14, 15
Other ethnic group	47, 98	50, 80
Unknown or refused	99	98, 99

FSM eligibility

17. FSM eligibility indicates whether the student was ever recorded as being eligible to receive free school meals in the six years prior to the March census date in their final year of key stage four (year 11).

18. The data on free school meal eligibility is produced by the Department for Education (DfE) as part of the National Pupil Database[5] and was linked onto data from the Education and Skills Funding Agency's (ESFA's) Individualised Learner Record (ILR) and the Higher Education Statistics Authority's (HESA's) student record and student alternative record.

IDACI

19. The Income Deprivation Affecting Children Index (IDACI) measures the proportion of children under the age of 16 in low income households for a particular area in England. It is calculated at lower-layer super output area (LSOA) level and is a supplementary measure to the Index of Multiple Deprivation (IMD), and therefore has been constructed for Department for Communities and Local Government (DCLG) by Oxford Consultants for Social Inclusion (OCSI). In this analysis, we use IDACI quintiles, where the most deprived areas are in quintile 1 and the least deprived are in quintile 5. For pupils or students domiciled outside of England, the value is set to not applicable, and data is not presented in the interactive tools.

[3] See www.gov.uk/government/publications/ilr-specification-validation-rules-and-appendices-2017-to-2018

[4] See www.hesa.ac.uk/collection/c17051/a/ethnic

[5] The DfE does not accept responsibility for any inferences or conclusions derived from the NPD data by third parties

IMD

20. The Index of Multiple Deprivation (IMD) is a measure of levels of deprivation for a small area within England. It is calculated at lower-layer super output area (LSOA) level and uses a number of different measures to determine levels of deprivation. They have been constructed for the Department for Communities and Local Government (DCLG) by Oxford Consultants for Social Inclusion (OCSI) [6]. In our analysis, we use IMD quintiles, where the most deprived areas are in quintile 1 and the least deprived are in quintile 5. For pupils or students domiciled outside of England, the value is set to not applicable, and data is not presented in the interactive tools.

Local or distance learner

21. Local learners are identified by comparing home travel to work area with study travel to work area, which are calculated from home postcode and study postcode respectively. Local students are those whose home address is in the same travel to work area as their provider. Distance learners are those who are not in attendance at the provider for the vast majority of their course. That is, they are studying at a distance from their provider. These definitions use a combination of the IPDL and IPLOCAL fields described in our 2020 Core Algorithms document[7]

NS-SEC

22. NS-SEC[8] classifies the socio-economic background of students. For those students who are under 21 at the start of their course NS-SEC is based on the occupation of their highest earning parent. For those students aged 21 or over at the start of their course NS-SEC is based on the occupation of the student themselves prior to higher education. Occupations are coded using the Standard Occupation Classification (SOC) and then grouped further into these categories[9]:

    a. Higher managerial, administrative and professional occupations

    b. Intermediate occupations

    c. Routine and manual occupations

    d. Never worked and long-term unemployed.

[6] See www.gov.uk/government/statistics/english-indices-of-deprivation-2015

[7] See footnote 2.

[8] Data or this characteristic comes from the HESA variable SEC. See www.hesa.ac.uk/collection/c19051/a/sec.

[9] See Section 7 of the ONS' webpage titled The National Statistics Socio-economic classification (NS-SEC) available from: www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010

Parental higher education

23. Parental higher education[10] is taken from the HESA record. It gives the student's response to the question 'Do any of your parents have any higher education qualifications, such as a degree, diploma or certificate of higher education?'. HESA defines parents as 'This includes natural parents, adoptive parents, step-parents or guardians who have brought you up.'

POLAR4

24. POLAR4 is a measure of the proportion of the young population that participates in higher education[11]. It is most applicable to students under the age of 21, and so POLAR4 is recorded as 'mature' for those aged 21 or over at the start of their course.

Sex

25. Sex is reported as either male or female. Very few records show a response of 'other' or are recorded as unknown. To avoid unnecessarily discarding data from the statistical model or having a group that is too small to use, responses other than male or female are grouped in with females (the larger of the two groups).

School types

26. School types are defined using the variable NFTYPE from the NPD. Our access model only includes schools that we define as 'main-stream state funded' schools. This includes the following school types:

- Academy 16-19 Converter
- Academy 16-19 Sponsor Led
- Academy – converter
- Academy – Sponsor-led
- City Technology College
- Community School
- Foundation School
- Free School – 16-19
- Free School – Mainstream
- Free School – Studio School
- Free School – UTC
- Further Education Sector Institution
- Voluntary Aided School
- Voluntary Controlled School.

[10] Data for this characteristic comes from the HESA variable pared. See www.hesa.ac.uk/collection/c19051/a/pared

[11] For more details of the POLAR measure, see www.officeforstudents.org.uk/data-and-analysis/polar-participation-of-local-areas/

Annex C: ABCS Access – statistical model and results

27. The access outcome measures the proportion of 18- or 19-year-olds entering higher education (sometimes referred to as young participation). Data regarding these students is taken from the DfE's National Pupil Database (NPD) from the summer in which they obtained their key stage four (KS4) qualifications – most commonly, GCSEs. We have then tracked these students through to the start of higher education, where we can determine whether they are in the higher education records two or three years later at the age of 18 or 19. This will capture any level or mode of undergraduate study.

28. We have taken data for pupils who obtained their KS4 qualifications in the summers of 2011, 2012, 2013, 2014 and 2015 (that is, in the academic years 2010-11 to 2014-15) from the NPD. Using KS4 cohorts up to 2014-15 allows us to capture the most recent 18- and 19-year-old entrants into higher education in the academic year 2018-19. We use KS4 cohorts because they give almost complete coverage of all 16 year olds in England. In addition, combining data from five cohorts allows us to carry out robust analysis, ensuring that there are sufficient students in each of the characteristic groups to allow us to carry out analysis regarding their access behaviour.

29. The model includes data on 2,766,655 school pupils, 1,134,005 of which entered higher education aged 18 or 19.

The model

30. We have used a binary logistic regression model to predict the probability of entering higher education. The model includes the following characteristics: ethnicity, IDACI, IMD, FSM eligibility, POLAR4 and sex. Table C1 shows the categories within each of the characteristics and the proportion of the students in the model who are in each of these categories.

Table C1: Characteristics in the ABCS access model

Characteristic	Category	Number of individuals	Percent
Ethnicity	Bangladeshi	37,290	1.4%
	Chinese	10,420	0.4%
	Indian	65,135	2.4%
	Pakistani	87,440	3.2%
	Any Other Asian Background	36,790	1.3%
	Black - African	76,770	2.8%
	Black Caribbean	38,140	1.4%
	Any Other Black Background	14,645	0.5%
	White - British	2,103,205	76.0%
	White - Irish	9,330	0.3%
	Any Other White Background	95,940	3.5%
	Gypsy, Roma or Traveller	4,710	0.2%
	White and Asian	21,395	0.8%
	White and Black African	10,830	0.4%
	White and Black Caribbean	34,705	1.3%

	Any Other Mixed Background	35,710	1.3%
	Any Other Ethnic Group	35,090	1.3%
	Refused or unknown	49,120	1.8%
FSM eligibility	Not eligible for FSM	2,046,315	74.0%
	Eligible for FSM	688,650	24.9%
	N/A	31,695	1.2%
IDACI	Quintile 1 (most deprived)	654,745	23.7%
	Quintile 2	563,190	20.4%
	Quintile 3	529,040	19.1%
	Quintile 4	515,820	18.6%
	Quintile 5 (least deprived)	503,860	18.2%
IMD	Quintile 1 (most deprived)	660,880	23.9%
	Quintile 2	547,910	19.8%
	Quintile 3	518,580	18.7%
	Quintile 4	507,740	18.4%
	Quintile 5 (least deprived)	531,545	19.2%
POLAR4	Quintile 1 (least represented)	589,280	21.3%
	Quintile 2	550,300	19.9%
	Quintile 3	545,410	19.7%
	Quintile 4	552,440	20.0%
	Quintile 5 (most represented)	529,230	19.1%
Sex	Female	1,367,640	49.4%
	Male	1,399,020	50.6%

31. Stepwise selection has been used to determine which two-way interactions to include in the model with entry and stay criteria of α=0.05. This has resulted in the inclusion of the following interactions: sex*ethnicity, sex*POLAR4, ethnicity*POLAR4, ethnicity*IMD, POLAR4*IMD, ethnicity*IDACI, POLAR4*IDACI, IMD*IDACI, sex*FSM eligibility, POLAR4*FSM eligibility, IDACI*FSM*eligibility.

32. The model is:

$$logit(\pi_i) = \beta_0 + \tilde{\beta}_1 ethnicity_i + \tilde{\beta}_2 IDACI_i + \tilde{\beta}_3 IMD_i + \tilde{\beta}_4 FSM_i + \tilde{\beta}_5 POLAR4_i + \beta_6 sex_i + interactions$$

Where i is an individual, π_i is a binary response variable which takes the value of 1 if the individual accessed higher education aged 18 or 19 and 0 otherwise, $\tilde{\beta}$ represents vectors of different sizes and the interactions are as listed above.

Model results

33. The coefficient estimates for each of the factors and for all the two-way interactions included in the final model can be found in the Excel/CSV files at www.officeforstudents.org.uk/publications/update-to-associations-between-characteristics-of-students/

Annex D: ABCS FT continuation – statistical model and results.

34. Full-time continuation is measured one year and 14 days after a student starts their studies. Students are defined as continuing if they are continuing with or have completed their studies or have transferred to another higher education provider to continue their studies[12] .

35. We have used continuation data for UK-domiciled undergraduate students who started their courses between the academic years 2013-14 and 2017-18 and were studying full-time at an English provider. We have used data from the ESFA individualised learner record (ILR) and the HESA student record and alternative provider student record. Data regarding FSM eligibility has been taken from the NPD. Combining data from five cohorts allows us to carry out robust analysis, ensuring that there are enough students in each of the characteristic groups to allow us to carry out analysis regarding their continuation behaviour.

36. The model includes data on 2,024,560 students, 1,829,740 of which are recorded as continuing.

The model

37. We have used a binary logistic regression model to predict the probability of entering higher education. The model includes the following characteristics: age, care experience, disability, ethnicity, FSM eligibility, IDACI, IMD, local or distance learner, NS-SEC, parental education, POLAR4 and sex. Table D1 shows the categories within each of the characteristics and the proportion of the students in the model who are in each of these categories.

Table D1: Characteristics in the ABCS FT continuation model

	Category	Number of individuals	Percent
Age	18 and under	907,945	44.9%
	19	430,445	21.3%
	20	152,760	7.6%
	21-25	249,860	12.3%
	26-50	263,825	13.0%
	51+	19,725	1.0%
Care experience	Care experienced	10,765	0.5%
	Not care experienced	1,185,640	58.6%
	Unknown or N/A	828,155	40.9%
Disability	Cognitive or learning difficulties	118,480	5.8%
	Mental health condition	48,995	2.4%
	Multiple impairments	40,580	2.0%
	Sensory, medical or physical impairments	41,900	2.1%
	Social or communication impairments	12,960	0.6%

[12] For details of how we calculate continuation measures from the data, see the OfS access and participation data methodology and rebuild instructions: www.officeforstudents.org.uk/data-and-analysis/access-andparticipation-data-dashboard/guide-to-the-access-and-participation-data-resources/

	No known disability	1,761,655	87.0%
Ethnicity	Asian or Asian British - Chinese	15,405	0.8%
	Asian or Asian British - Indian	73,805	3.7%
	Asian or Asian British - Pakistani	79,180	3.9%
	Asian or Asian British - any other background	47,820	2.4%
	Asian or Asian British – Bangladeshi	40,945	2.0%
	Black or black British - African	142,240	7.0%
	Black or black British - Caribbean	43,005	2.1%
	Black or black British - any other background	11,710	0.6%
	Gypsy or Traveller	340	0.0%
	Mixed or multiple - any other background	25,505	1.3%
	Mixed or multiple - white and Asian	26,115	1.3%
	Mixed or multiple - white and black African	11,980	0.6%
	Mixed or multiple - white and black Caribbean	26,365	1.3%
	Other ethnic background	36,385	1.8%
	White	1,412,410	69.8%
	Refused or unknown	31,355	1.6%
FSM eligibility	Eligible for FSM	226,305	11.2%
	Not eligible for FSM	1,185,790	58.6%
	N/A	612,470	30.3%
IDACI	Quintile 1	415,345	20.5%
	Quintile 2	380,515	18.8%
	Quintile 3	362,900	17.9%
	Quintile 4	371,750	18.4%
	Quintile 5	428,090	21.1%
	N/A	65,965	3.3%
IMD	Quintile 1	409,860	20.2%
	Quintile 2	381,645	18.9%
	Quintile 3	361,660	17.9%
	Quintile 4	377,765	18.7%
	Quintile 5	427,670	21.1%
	N/A	65,965	3.3%
Local or distance learner	Distance	4,770	0.2%
	Local	584,375	28.9%
	Neither	1,435,415	70.9%
NS-SEC	Higher managerial, administrative and professional occupations	430,735	21.3%
	Intermediate occupations	205,485	10.2%
	Routine and manual occupations	242,150	12.0%
	Never worked and long-term unemployed	4,905	0.2%
	Unknown or N/A	1,141,290	56.4%
Parental higher education	Higher education qualification held by parent(s)	761,555	37.6%
	No higher education qualification held by parent(s)	767,325	37.9%
	Unknown or N/A	495,680	24.5%
POLAR4	Quintile 1	172,680	8.5%

	Quintile 2	236,890	11.7%
	Quintile 3	283,415	14.0%
	Quintile 4	343,550	17.0%
	Quintile 5	454,615	22.5%
	N/A	533,410	26.4%
Sex	Female	1,129,130	55.8%
	Male	895,430	44.2%

38. Stepwise selection has been used to determine which two-way interactions to include in the model with entry and stay criteria of α=0.05. This has resulted in the inclusion of the following interactions: FSM eligibility*NS-SEC, FSM eligibility*care experience, FSM eligibility*parental education, age*FSM eligibility, age*NS-SEC, age*care experience, age*parental education, care experience*parental education, ethnicity*NS-SEC, ethnicity*age, ethnicity*care experience, ethnicity*disability, ethnicity*local or distance learner, ethnicity*parental education, ethnicity*IMD, ethnicity*POLAR4, disability*FSM eligibility, disability*care experience, disability*local or distance learner, disability*parental education, local or distance learner*FSM eligibility, local or distance learner*NS-SEC, local or distance learner*age, local or distance learner*care experience, local or distance leaner*parental education, parental education*NS-SEC, IDACI*age, IDACI*disability, IMD*FSM eligibility, IMD*NS-SEC, IMD*local or distance learner, IMD*parental education, IMD*IDACI, POLAR4*NS-SEC, POLAR4*age, POLAR4*disability, POLAR4*local or distance leaner, POLAR4*parental education, POLAR4*IMD, sex*NS-SEC, sex*age, sex*care experience, sex*ethnicity, sex*disability, sex*local or distance learner and sex*POLAR4.

39. The model is:

$$logit(\pi_i) = \beta_0 + \tilde{\beta}_1 age_i + \tilde{\beta}_2 \text{care experience}_i + \tilde{\beta}_3 disability_i + \tilde{\beta}_4 ethnicity_i + \tilde{\beta}_5 FSM\ eligibility_i + \tilde{\beta}_6 IDACI_i + \tilde{\beta}_7 IMD_i + \tilde{\beta}_8 local\ or\ distance_i + \tilde{\beta}_9 NS-SEC_i + \tilde{\beta}_{10} parental\ highr\ education_i + \tilde{\beta}_{11} POLAR4_i + \beta_{12} sex + interactions$$

where i is an individual, π_i is a binary response variable which takes the value of 1 if the individual continued into the second year of their course and 0 otherwise, $\tilde{\beta}$ represents vectors of different sizes and the interactions are as listed above.

Model results

40. The coefficient estimates for each of the factors and for all the two-way interactions included in the final model can be found in the Excel/CSV files at www.officeforstudents.org.uk/publications/update-to-associations-between-characteristics-of-students/

Annex E: ABCS PT continuation – statistical model and results

41. Part-time continuation is measured two years and 14 days after a student starts their studies. Students are defined as continuing if they are continuing with or have completed their studies or have transferred to another higher education provider to continue their studies[13] .

42. We have used continuation data for UK-domiciled undergraduate students who started their courses between the academic years 2012-13 and 2016-17 and were studying part-time at an English provider. We have used data from the ESFA individualised learner record (ILR) and the HESA student record and alternative provider student record. Combining data from five cohorts allows us to carry out robust analysis, ensuring that there are enough students in each of the characteristic groups to allow us to carry out analysis regarding their continuation behaviour.

43. The model includes data on 460,065 students, 293,915 of which are recorded as continuing.

The model

44. We have used a binary logistic regression model to predict the probability of entering higher education. The model includes the following characteristics: age, disability, ethnicity, IDACI, IMD, local or distance learner and sex. Table E1 shows the categories within each of the characteristics and the proportion of the students in the model who are in each of these categories.

Table E1: Characteristics in the ABCS PT continuation model

Characteristic	Category	Number of individuals	Per cent
Age group	18 and under	17,465	3.8%
	19	17,345	3.8%
	20	19,655	4.3%
	21-25	102,985	22.4%
	26-50	270,840	58.9%
	51+	31,775	6.9%
Disability	Cognitive or learning difficulties	15,705	3.4%
	Mental health condition	8,120	1.8%
	Multiple impairments	16,425	3.6%
	Sensory, medical or physical impairments	10,065	2.2%
	Social or communication impairments	1,095	0.2%
	No known disability	408,655	88.8%
Ethnicity	Asian or Asian British - Chinese	1,410	0.3%
	Asian or Asian British - Indian	9,205	2.0%
	Asian or Asian British - Pakistani	6,405	1.4%
	Asian or Asian British - any other background	7,430	1.6%
	Asian or Asian British – Bangladeshi	2,250	0.5%

[13] See footnote 12

	Black or black British - African	18,105	3.9%
	Black or black British - Caribbean	9,015	2.0%
	Black or black British - any other background	1,985	0.4%
	Mixed or multiple - any other background	3,865	0.8%
	Mixed or multiple - white and Asian	2,130	0.5%
	Mixed or multiple - white and black African	1,495	0.3%
	Mixed or multiple - white and black Caribbean	3,810	0.8%
	White	377,775	82.1%
	Gypsy or Traveller	95	0.0%
	Other ethnic background	4,875	1.1%
	Refused or unknown	10,225	2.2%
IDACI	Quintile 1	81,435	17.7%
	Quintile 2	87,220	19.0%
	Quintile 3	86,660	18.8%
	Quintile 4	84,380	18.3%
	Quintile 5	77,190	16.8%
	N/A	43,175	9.4%
IMD	Quintile 1	78,490	17.1%
	Quintile 2	88,305	19.2%
	Quintile 3	86,155	18.7%
	Quintile 4	86,645	18.8%
	Quintile 5	77,295	16.8%
	N/A	43,175	9.4%
Local or distance learner	Distance	204,060	44.4%
	Local	121,925	26.5%
	Neither	134,080	29.1%
Sex	Female	273,870	59.5%
	Male	186,195	40.5%

45. Stepwise selection has been used to determine which two-way interactions to include in the model with entry and stay criteria of α=0.05. This has resulted in the inclusion of the following interactions: sex*ethnicity, sex*IMD, ethnicity*IMD, IMD*IDACI, sex*disability, ethnicity*disability, IDACI*disability, sex*local or distance learner, ethnicity*local or distance learner, IMD*local or distance learner, IDACI*local or distance learner, disability*local or distance learner, sex*age, ethnicity*age, IMD*age, disability*age and local or distance learner*age.

46. The model is:

$$logit(\pi_i) = \beta_0 + \tilde{\beta}_1 age_i + \tilde{\beta}_2 disability_i + \tilde{\beta}_3 ethnicity_i + \tilde{\beta}_4 IDACI_i + \tilde{\beta}_5 IMD_i + \tilde{\beta}_6 local\ or\ distance_i + \beta_7 sex + interactions$$

where i is an individual, π_i is a binary response variable which takes the value of 1 if the individual continued into the second year of their course and 0 otherwise, $\tilde{\beta}$ represents vectors of different sizes and the interactions are as listed above.

Model results

47. The coefficient estimates for each of the factors and for all the two-way interactions included in the final model can be found in the Excel/CSV files at www.officeforstudents.org.uk/publications/update-to-associations-between-characteristics-of-students/