Annex B: Statistical modelling

Methodology overview

1. This annex outlines the methodology used for the statistical modelling of the attainment of first and upper second class degrees combined, and of first class degrees alone.

Graduate population

2. The graduate population in this analysis comprises UK-domiciled undergraduates (first degree and degrees including a postgraduate component) who were studying full-time at the start of their course, were registered at higher education providers in England and graduated in the academic years from 2010-11 to 2022-23 with a classified honours degree.

3. This population is summarised in Table 3 of Annex A and can be rebuilt using the following fields described in the OfS publication 'Technical algorithms for student outcome and experience measures':[1]

   - DFAPAPPEXCL = 0

   - IPDOQUALPOP = 1

   - IPCOUNTRY = 'E'

   - IPSTARTMODE = 'FT'

   - IPBASEYEAR in (2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022).

4. In our analysis this population was further limited to those who qualified from English providers awarding at least 10 classified honours degrees in each of the academic years considered.

Method to determine 'unexplained' attainment

5. Mixed-effects logistic regression modelling was employed to investigate the extent to which changes in degree attainment between 2010-11 and 2022-23 can be accounted for by changes in the makeup of the graduate population in terms of the explanatory variables included in the modelling.

6. The modelling used to investigate degree attainment changes with time at the sector level includes explanatory variables relating to the provider at which the graduate was registered,

---

[1] See 'Technical algorithms for student outcome and experience measures', available at www.officeforstudents.org.uk/publications/description-and-definition-of-student-outcome-and-experience-measures/.

graduation year and various key graduate characteristics. The effects of the following were included as explanatory variables in the model:

- the provider at which the graduate was registered

- year of graduation

- subject of study

- qualifications on entry into higher education

- age on entry to higher education.

7. We interact subject of study, entry qualifications and age with year of graduation to account for changes in awarding, especially since the start of the coronavirus pandemic. These interactions can be seen in Equation B1.

8. Our model allows us to statistically predict the proportion of graduates awarded a first or an upper second class degree, or a first class degree, accounting for the effects of the explanatory variables.

9. To investigate and isolate the effect of graduation year on degree attainment, the following methodology was applied:

    a. The models provide the probability of an individual with given characteristics attaining a first or upper second class degree, or a first class degree.

    b. The predicted probability for a given group of individuals (e.g. mature students graduating in 2011-12) may then be determined by taking the mean of the predicted probabilities of the individuals in that group.

    c. To investigate the effect of graduation year on degree classification attainment, the model is applied to the entire reported graduate population, but with the academic year of graduation for all graduates in the population changed to 2010-11.

    d. The observed value for the proportion of graduates attaining a first or upper second class degree, or a first class degree, in each academic year is then compared with the model's predicted value for the same graduates, had they graduated in 2010-11.

    e. Any differences between the predicted and observed values is said to be 'unexplained', and a result of unobserved effects between academic years that have not been accounted for and have not been included as explanatory variables in the model. It is not possible to determine from this analysis what these additional unobserved factors are.

10. In summary, we estimate the 'unexplained' difference in the proportion of graduates attaining a first or upper second class degree, or a first class degree, had they graduated in 2010-11, compared with the actual year of their graduation.

Model details

11. Here we detail the model used to describe the attainment of first and upper second class degrees combined, and of first class degrees alone.

12. Mixed-effects logistic regression has been used to model the probability of graduate i attaining a first or an upper second class degree, or a first class degree, from provider j, accounting for the effect of the explanatory variables outlined in paragraph 6 of this annex.

13. Following the changes in attainment of particular groups of students during the coronavirus pandemic we added 'year' interaction terms to better capture year-on-year sector changes in attainment for age on entry, qualifications on entry and subject of study.

14. The specification of the statistical model used in the analysis is displayed in Equation B1.

Equation B1: Mixed-effects logistic regression model for graduate degree attainment

$$first \ or \ upper \ second \ class \ OR \ first \ class \sim Binomial(n_{ij}, \pi_{ij})$$

$$logit(\pi_{ij}) = \beta_{0j} + u_{0j} + \sum_{Y=1}^{13}(\beta_Y + u_{Yj})X_{Yij} + \sum_{Sbj=1}^{21}\beta_{Sbj}X_{Sbjij} + \sum_{Q=1}^{23}\beta_Q X_{Qij} + \sum_{A=1}^{2}\beta_A X_{Aij}$$

$$+ \sum_{(Q*A)=1}^{46}\beta_{(Q*A)}X_{(Q*A)ij} + \sum_{(Y*Sbj)=1}^{273}\beta_{(Y*Sbj)}X_{(Y*Sbj)ij} + \sum_{(Y*Q)=1}^{299}\beta_{(Y*Q)}X_{(Y*Q)ij}$$

$$+ \sum_{(Y*A)=1}^{26}\beta_{(Y*A)}X_{(Y*A)ij}$$

Note: A number of the terms included in this equation are zero, as they refer to a reference category (see Table B1 for detail of reference categories).

15. In Equation B1 the βs represent the fixed effect coefficients that are common to individuals across all providers (the sector) and years,[2] and Xs (0 or 1) represent whether or not an individual has the characteristics:

- Y = academic year of graduation

- Sbj = subject of study

- Q = entry qualifications

- A = age on entry

- $Q * A$ = interaction between entry qualifications and age

- $Y * Sbj$ = interaction between academic year of graduation and subject of study

- $Y * Q$ = interaction between academic year of graduation and entry qualifications

---

[2] The summation term for academic year of graduation includes the reference year of 2010-11, as each provider has a random coefficient for all years but the fixed effect 2010-11 coefficient $\beta_1 = 0$ (reference categories for other explanatory variables are omitted from the model structure).

- $Y * A$ = interaction between academic year of graduation

- u_{0j} is the random intercept for provider j

- u_{Yj} represents the random coefficient for provider j in academic year Y)

with:

$$u_{0j} \sim N\left(0, \sigma_{u_0}^2\right)$$

$$u_{Yj} \sim N\left(0, \sigma_{u_Y}^2\right).$$

16. A full summary of the variables used in the model, and the categories within those variables, is given in Table B1. Table B1 does not include the various interaction terms featured in the model (539 non-zero estimates). Each year is interacted with the student characteristics independently, resulting in the large number of interaction terms.

17. Fixed effect coefficient estimates, standard errors and p-values for models can be found in Table 6 of Annex A. Estimates for the interaction terms are not included in this table. They are available on request.

Table B1: Variables used in the graduate degree attainment modelling (all categorical)

Model variable name	Description
Academic year (Y)	Academic year of graduation: 2010-11 (ref) 2011-12 2012-13 2013-14 2014-15 2015-16 2016-17 2017-18 2018-19 2019-20 2020-21 2021-22 2022-23
Subject of study (Sbj)	Subject studied: Agriculture, food and related studies Architecture, building and planning Biological and sport sciences Business and management (ref) Combined and general studies Computing Design, and creative and performing arts Education and teaching Engineering and technology

Model variable name	Description
	Geography, earth and environmental studies
	Historical, philosophical and religious studies
	Language and area studies
	Law
	Mathematical sciences
	Media, journalism and communications
	Medicine and dentistry
	Physical sciences
	Psychology
	Social sciences
	Subjects allied to medicine
	Veterinary sciences
Entry qualifications (Q)	Entry qualifications of the graduate:
	A-level: AAA and above (ref)
	A-level: AAB
	A-level: AAC
	A-level: ABB
	A-level: ABC
	A-level: ACC
	A-level: BBB
	A-level: BBC
	A-level: BCC
	A-level: CCC
	A-level: CCD
	A-level: CDD
	A-level: DDD
	A-level: Below DDD
	BTEC: DDD and above
	BTEC: DDM
	BTEC: DMM
	BTEC: MMM and below
	2 A-levels and 1 BTEC
	1 A-levels and 2 BTEC
	International Baccalaureate
	Other Level 3
	No Level 3 Equivalent
Age (A)	Age on entry
	Under 21 (Young) (ref)
	21 and over (Mature)

Note: Those categories marked with '(ref)' are the reference categories for each categorical or dummy variable and are not formally included in the model structure (they are equal to 0).

18. Estimates of the variance components and their standard errors for the random intercepts and random year coefficients are shown for first and upper second class and first class degrees in Table B2.

Table B2: Variance component estimates for the models for first or upper second class and first class only degree attainment

	Random effect	First or upper second class model estimate	First or upper second class model standard error	First class model estimate	First class model standard error
Intercept	$\sigma^2_{u_0}$	0.103	0.014	0.147	0.018
Year	$\sigma^2_{u_Y}$	0.038	0.002	0.040	0.002

19. Model fit statistics for first and upper second class and first class degrees can be found in Table B3.

Table B3: Model fit statistics for the models for first or upper second class and first class only degree attainment

Statistic	First or upper second class model	First class model
-2logLiklihood	3,137,473	3,432,615
Akaike information criterion	3,138,669	3,433,811