Teaching Excellence and Student Outcomes Framework (TEF): Findings from the subject-level pilot 2018-19

Annex G: Logistic regression models

This report was completed in autumn 2019 following the conclusion of the pilot.

Contents

Introduction

1. Logistic regression modelling was undertaken to complement analysis of the outcomes arrived at in the second year of the Teaching Excellence and Student Outcomes Framework TEF subject-level pilots. This modelling investigates whether factors which were thought to influence subject ratings are indeed significant predictors of the final subject rating once other variables have been taken into account.

   - Key findings of the analysis are summarised in the report 'Teaching Excellence and Student Outcomes Framework: Findings from the subject-level pilot 2018-19'[1] to which this report is an annex.

   - This annex provides full details of the methodology used in each model (including the model assumptions that have been tested) and the findings relating to each variable considered.

2. OfS analysts fitted two types of logistic regression models in order to reflect upon two different response variables:

   a. Ordinal models to predict the final subject rating of subjects that were awarded Bronze, Silver or Gold

   b. Binary models to predict whether subjects were awarded no rating.

3. The models predicting final subject rating were constructed to identify:

   a. the metrics that had the largest effect on subject performance beyond their role in the initial hypothesis calculation

   b. any measurable factors that had a significant effect on subject performance other than the initial hypothesis and the metrics themselves

   c. whether or not subjects were treated consistently by the different joint panels.

4. The models predicting whether subjects received no rating were constructed to confirm our hypothesis that smaller subjects and subjects with fewer reportable metrics were more likely to not be rated, and test whether any other measurable factors had a significant effect on this outcome.

5. Initially, for both response variables, models were fitted with only the predictor variables thought mostly likely to have an effect. Other predictor variables were then tested one at a time. This approach is outlined in more detail below.

6. The predictor variables considered in this analysis are detailed in Table 1 in Annex G3.

7. We note that all of the modelling results reported in this annex are based on the relatively small sample of providers who participated in the 2018-19 pilot. Participating providers were selected for the pilot for their diversity, so it follows that the results given here may not be representative

---

[1] Available at www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

of those that might be observed among a final set of providers taking part in a future exercise. Moreover, readers will note that the variables used as predictors in the modelling include among them a number drawn from metrics data. As described elsewhere in these pilot findings, some significant limitations have been identified with regard to the metrics data at subject level. These limitations will also influence the reliability and generalisation of the modelling results reported here.

Annex G1: A model to predict final subject-level ratings

Methodology

8. OfS analysts fitted ordinal logistic regression models with data from the 587 subjects that received a final award in the subject pilot. The response variable in the models was the final award received by each subject (coded as 2 for Gold, 1 for Silver and 0 for Bronze).

9. The equations for these models take the form:

$$Logit\big(Pr(FinalAward \geq j)\big) = \alpha_j + \sum_i \beta_i x_i \qquad for\ j = 1,2\ and\ i = 1,2,\dots m,$$

where m is the number of predictor variables in the model, x_i is the ith predictor variable, α_j is the jth intercept and β_i is the coefficient for the ith predictor variable.

10. Initially, the following variables were chosen as predictors:

    - initial hypothesis

    - 9 core metric flags from the majority mode of study

    - 9 absolute value markers from the majority mode of study

    - the joint panel that assessed the subject

    - provider type.

11. There was no significant evidence from the initial model that any of the absolute value markers had an effect on the final ratings of subjects, beyond their role in the initial hypothesis calculation, so these predictors were removed.

12. 'Provider-level award' was later added as a predictor and found to have a significant effect at the 5 per cent level, with subjects at providers that went on to achieve a Bronze at provider level performing worse than those at Silver [2]) or Gold providers. Provider size, measured as total Full Person Equivalent (FPE) in both modes of study, also appeared to have a significant effect when added to the model, but this was diminished (and no longer significant) when it was included in conjunction with provider-level award. As a result, provider-level award was kept in the final model but provider FPE was removed.

---

[2] Reference levels for categorical variables are marked as 'ref'. These are not formally included in the model structure.

13. The following variables were also tested as predictors but not found to have significant[3] effects at the 5 per cent level:

- majority mode of the subject

- FPE in the majority mode of study of the subject

- batch of assessment

- proportion of students in the subject who are mature

- proportion of students in the subject who are local

- proportion of students in the subject who are black, Asian and minority ethnic (BAME)

- proportion of students in the subject who are disabled

- proportion of students in the subject who are female.

Model assumptions tested

14. For the final model, the proportional odds (parallel lines) assumption has been tested, finding insufficient evidence that the assumption had been violated (χ^2 = 41.1, DF=35, p=.22).

15. The final model has a c-statistic of 0.94, suggesting it has outstanding discriminative ability[4].

16. These findings generate no immediate concerns regarding the adequacy of the model. However, we acknowledge that testing goodness of fit within ordinal logistic regression models is not straightforward and that these tests of discriminatory ability do not definitively confirm that the model is well calibrated.

Results and conclusions

17. The parameter estimates from the final model are displayed below in Table 2 in Annex G3.

18. As might be expected, the initial hypothesis calculation (with absolute value markers considered) was a strong predictor of the final rating of the subject.

19. At the 5 per cent level, the following core metrics appeared to have a significant effect on the final rating of a subject, beyond their role in the initial hypothesis calculation:

- Academic support
- Assessment and feedback
- The teaching on my course
- Student voice
- Continuation

---

[3] To counteract the multiple comparisons problem when testing significance of these variables in turn, the Šidák correction was applied. Unfortunately, this considerably reduces the power of these tests. Šidák, Z. K. (1967). 'Rectangular Confidence Regions for the Means of Multivariate Normal Distributions'.
[4] Hosmer, D. W., and Lemeshow, S. (2000). Applied Logistic Regression.

- Above median earnings threshold or higher study.

20. In general, the model shows that:

    a. positive flags on these metrics have a positive effect on the final subject award, beyond their role in the initial hypothesis calculation

    b. negative flags on these metrics have a negative effect on the final subject award, beyond their role in the initial hypothesis calculation.

21. The model suggests that 'teaching on my course' has the largest effect of these six metrics.

22. Provider-level award is a significant predictor of subject performance, as discussed in paragraph 12 above.

23. The model also shows provider type to be a significant predictor of subject performance, with subjects at further education colleges (FECs), or non-FECs with fewer than 10 subjects, performing worse than those at non-FECs with 10 or more subjects.

24. There is evidence that subjects assessed by the Medical Sciences, and Nursing and Allied Health subject panel were likely to achieve worse ratings than those assessed by the Natural Sciences and Engineering & Technology panel, when accounting for the other predictors. This effect remains when UKPRN is added[5] as a predictor in the model, suggesting that subjects were not treated consistently within each provider.

Annex G2: A model to predict subjects receiving no rating

Methodology

25. OfS analysts fitted binary logistic regression models for the 629[6] subjects that were assessed in the subject-level pilot. The response variable in the models was coded as 1 if the subject received no rating and 0 if the subject received a rating of Bronze, Silver or Gold.

26. The equations for these models take the form:

$$Logit\big(Pr(Subject\ received\ no\ rating)\big) = \beta_0 + \sum_i \beta_i x_i \qquad for\ i = 1,2,\dots m,$$

where m is the number of predictor variables in the model, x_i is the ith predictor variable, β_0 is the intercept and β_i is the coefficient for the ith predictor variable.

27. To mitigate the effects of quasi-complete separation in the model, Firth's penalised likelihood method[7] was used to calculate the parameter estimates.

---

[5] The final model was subsequently adjusted to explore the issue of within-provider consistency. To do this, provider type and provider award were removed from the model as predictors, and UKPRN was added. While we acknowledge that this adjustment may increase the risk of multi-collinearity (on account of the associations between provider and the other factors), this is primarily an issue of statistical power and should not alter the conclusions drawn from the model where significant effects have been identified.

[6] One subject was removed as an outlier.

[7] This method was originally proposed by Firth (1993 Biometrika 80:27-38).

28. Initially the following variables were chosen as predictors:

- the number of non-reportable metrics in the majority mode of study of the subject

- subject size

- the majority mode of study of the subject

- the joint panel that assessed the subject.

29. Later, the following variables were tested as predictors but not found to have significant effects on the model:

- provider type

- proportion of students in the subject who are mature

- proportion of students in the subject who are local

- proportion of students in the subject who are BAME

- proportion of students in the subject who are disabled

- proportion of students in the subject who are female.

Model assumptions tested

30. The final model passed a Hosmer and Lemeshow goodness of fit test (χ^2 = 5.9905, DF=7, p=0.541) providing no significant evidence to reject the model fit.

31. The final model also has a c-statistic of 0.96, signifying outstanding discriminative ability.

32. These findings generate no immediate concerns regarding the adequacy of the model.

Results and conclusions

33. The parameter estimates from the final model are shown below in Table 3 in Annex G3.

34. As might be expected, the number of non-reportable core metrics and subject size were significant predictors. Subjects with more non-reportable metrics were more likely to receive no rating. Larger subjects were less likely to receive no rating.

35. The final model shows that the joint panel is a significant predictor. The model suggests that the Arts and Humanities, and Business and Law and Education and Social Care panels were less likely to award no rating than the Natural Sciences, and Engineering and Technology panel and the Medical Sciences, and Nursing and Allied Health panel.

36. The majority mode of study in the subject area did not have a significant effect on the likelihood of a subject being awarded no rating.

Annex G3: Predictor variables and parameter estimates

Table 1: Predictor variables tested throughout the modelling

Variable type	Variable(s)	Description
Categorical[8]	Initial hypothesis	This is the Step 1a calculation of the initial hypothesis but with absolute value markers considered in place of neutral flags, as panel members would have done as part of Step 1b. • Silver (ref) • Bronze • Bronze/Silver • Silver/Gold • Gold
	Nine core metric flags	The core flags for each of the nine metrics for the majority mode of study in the subject, each grouped as: • Neutral (ref) • Positively flagged • Negatively flagged • Not reportable
	Nine absolute value markers	The absolute value markers for each of the nine metrics for the majority mode of study in the subject, each grouped as: • Not flagged (ref) • Positively flagged • Negatively flagged
	Joint panel	The joint panel that assessed the subject: • Natural Sciences, and Engineering and Technology (ref) • Arts and Humanities • Business and Law, and Education and Social Care • Medical Sciences, and Nursing and Allied Health • Natural and Built Environment, and Social Sciences
	Provider type	The type of provider that taught the subject: • Non-FEC with 10 or more subjects (ref) • FEC • Non-FEC with fewer than 10 subjects
	Majority mode	The majority mode of study in the subject: • Part-time (ref) • Full-time

---

[8] Reference levels for categorical variables are marked as 'ref'. These are not formally included in the model structure.

Variable type	Variable(s)	Description
	Batch	Batch of assessment: • Batch two (ref) • Batch one
	Provider award	Final award received by the provider: • Silver (ref) • Bronze • Gold
	UKPRN	The UKPRN of the provider.
Ordinal	Number of non-reportable metrics	The number of non-reportable metrics in the majority mode of study for the subject, weighted by their Step 1a weightings.
Continuous	Provider size	FPE at the provider.
	Subject size	FPE in the majority mode of the subject.
	Proportion mature	The proportion of students in the subject who were mature.
	Proportion local	The proportion of students in the subject who were local.
	Proportion BAME	The proportion of students in the subject who were BAME.
	Proportion with reported disability	The proportion of students in the subject area who had a reported disability.
	Proportion female	The proportion of students in the subject area who were female.

Table 2: Parameter estimates from the final model to predict final subject ratings

Parameter		Estimate	Standard error	Wald Statistic	P-value
Intercept	2	-2.271	0.396	32.960	<.0001
	1	3.426	0.436	61.652	<.0001
Initial hypothesis	Bronze	-2.792	0.718	15.112	0.0001
	Gold	3.474	0.648	28.772	<.0001
	Gold/Silver	1.481	0.418	12.555	0.0004
	Silver/Bronze	-1.303	0.423	9.508	0.002
Academic support	Positively flagged	0.933	0.387	5.806	0.016
	Negatively flagged	-0.222	0.458	0.234	0.6288
	Not reportable	-1.257	1.091	1.327	0.2493
Assessment and feedback	Positively flagged	1.024	0.366	7.843	0.0051
	Negatively flagged	-0.602	0.396	2.311	0.1284
Continuation	Positively flagged	0.891	0.380	5.507	0.0189
	Negatively flagged	-0.738	0.538	1.883	0.17
Highly skilled employment or higher study	Positively flagged	0.356	0.330	1.161	0.2813
	Negatively flagged	-0.580	0.377	2.369	0.1238
	Not reportable	0.161	0.532	0.091	0.7629
Sustained employment or further study	Positively flagged	0.742	0.460	2.601	0.1068
	Negatively flagged	-0.407	0.461	0.779	0.3774
	Not reportable	2.610	1.365	3.656	0.0559
Above median earnings threshold or higher study	Positively flagged	0.490	0.342	2.049	0.1523
	Negatively flagged	-0.456	0.330	1.906	0.1674
	Not reportable	-3.047	1.333	5.223	0.0223
Learning resources	Positively flagged	0.365	0.378	0.930	0.3349
	Negatively flagged	0.057	0.385	0.022	0.8833
	Not reportable	-0.916	0.894	1.052	0.3051
Student voice	Positively flagged	-0.172	0.376	0.210	0.647
	Negatively flagged	-1.176	0.395	8.857	0.0029
The teaching on my course	Positively flagged	1.420	0.393	13.024	0.0003
	Negatively flagged	-1.161	0.432	7.220	0.0072
Joint panel	Arts + Humanities	-0.302	0.345	0.768	0.381
	Business and Law + Education and Social Care	-0.657	0.375	3.068	0.0798

Parameter		Estimate	Standard error	Wald Statistic	P-value
	Medical Sciences + Nursing and Allied Health subjects	-0.972	0.360	7.294	0.0069
	Natural and Built Environment + Social Sciences	-0.253	0.383	0.437	0.5085
Provider type	Further education college (FEC)	-1.733	0.414	17.517	<.0001
	Non-FEC (under 10 subjects)	-2.560	0.572	20.059	<.0001
Provider award	Bronze	-0.755	0.296	6.494	0.0108
	Gold	0.308	0.320	0.923	0.3368

Table 3: Parameter estimates from the final model to predict whether subjects receive no rating

Parameter		Estimate	Standard error	Wald Statistic	P-value
Intercept		-0.630	0.790	0.637	0.4249
Number of non-reportable metrics		0.935	0.183	26.079	<.0001
Joint panel	Arts + Humanities	-1.534	0.720	4.536	0.0332
	Business and Law + Education and Social Care	-1.833	0.867	4.476	0.0344
	Medical Sciences + Nursing and Allied Health subjects	0.774	0.559	1.916	0.1663
	Natural and Built Environment + Social Sciences	-0.180	0.600	0.090	0.7637
Subject FPE		-0.018	0.005	14.687	0.0001
Majority mode of study	Full-time	-0.512	0.542	0.894	0.3445

OGL