Understanding the impact of institutional financial support on student success: Coding workbook

Analysing your bursary data

Part 1: readying the dataset

At this point, you will hopefully have two large datasets (for 2009 and 2012 starters) extracted from your student records system and stored in Excel (or similar). The next stage is to recode these data into a standardised format for analysis.

Because of the differences in the ways in which universities code and store their data, it is likely that this process will throw up some new questions along the way. You may also want to use different definitions/codings, but for the purposes of the pilot, the aim is to keep these as similar as possible. Clearly you can also undertake your own recodings in parallel if you wish.

It is your choice whether to recode the data within Excel or SPSS (which will be used for the analysis). There are pros and cons of each, based around the data manipulation tools which the two packages offer.

Stage 1 – checking the frame

Hopefully only data on relevant students will have been extracted, but this needs checking at this point. Your 2012 dataset should meet the following criteria for inclusion:

- English domiciled students (specifically excluding S/W/NI)
- HEFCE funded (excludes medics and some others)
- Full-time
- Pursuing a first degree – i.e. no sub-degree qualifications and no graduates
- In their first year of programme/award/course/degree (not necessarily of study)
- Did not withdraw prior to 1st December
- Their reason for leaving (if any) was not completion of the course (this can happen with top-up degrees and transfers), death or serious illness
- On the 2012 funding regime

Your 2009 dataset should meet the following criteria for inclusion:

- English domiciled students (specifically excluding S/W/NI)
- HEFCE funded (excludes medics and some others)

- Full-time
- Pursuing a first degree – i.e. no sub-degree qualifications and no graduates
- In their first year of programme/award/course/degree (not necessarily of study) in 2009
- Did not withdraw prior to 1st December in their first year
- Reason for leaving was not death or serious illness

From experience, it is useful at this point to critically 'eyeball' the datasets to see whether there are any obviously erroneous groups of students present. These have appeared in the initial institutional datasets due to unusual degrees, miscoding in the student records system and errors in the extraction process.

Stage 2 – recoding

It is important that the coding for the variables is consistent across institutions. In particular, it is vital that the 0/1 codings are correct as they will otherwise invalidate the analysis.

Control variables	Coding and notes
Sex	0 = Female, 1 = Male (in the institutions to date, there have been no transgender individuals – if any arise in the pilots, this will need consideration)
Nationality	0 = UK, 1 = Other
Age on entry	1 = 20 and under, 2 = 21 to 24, 3 = 25 to 29, 4 = 30 and over
First year accommodation	1 = institutional / private halls, 2 = parental home, 3 = own home, 4 = other rented, 5 = other (including not known and not in attendance)
Ethnicity	The extensive list of categories have been aggregated to ensure a reasonable coverage for analysis. As always, this is not optimal. 1 = White, 2 = Black Caribbean, 3 = Black African, 4 = Indian, 5 = Pakistani, 6 = Bangladeshi, 7 = Chinese, 8 = Mixed Heritage, 9 = Other, 10 = Unknown
Disability	1 = No known disability, 2 = Disability and receiving DSA, 3 = Disability and not receiving DSA

Subject	The need here is to collapse the three JACS codes for the degree into a single variable which reasonably represents the student's study. The principle used is that where a subject clearly dominates (100% or 67:33) then this is used, but if the components are equal (50:50 or 33:33:33) then a 'mixed' group is used. An Excel spreadsheet is provided to automate this coding process, although it may need some local adaptation. Due to its partial implementation, any 'I' JACS codes should be collapsed into 'G'. 1 = A, 2 = B, 3 = C, 4 = D, 5 = F, 6 = G, 7 = H, 8 = J, 9 = K, 10 = L, 11 = M, 12 = N, 13 = P, 14 = Q, 15 = R, 16 = T, 17 = V, 18 = W, 19 = X , 20 = Z (MIXED)
POLAR	As per the POLAR quintiles – i.e. 1 to 5.
Distance from home to university	As a continuous variable – i.e. 0 to ∞.
NSS score for degree	As a continuous variable – i.e. 0 to 100. There is likely to be a significant proportion of missing data which will challenge the analysis. Missing values should therefore be replaced with a mean – ideally for the department/faculty, but otherwise for the university as a whole.
Degree size (all students in year)	As a continuous variable – i.e. 1 to ∞.
Franchise course	0 = No, 1 = Yes
Clearing entrant	0 = No, 1 = Yes
Entry qualifications	This is a complex process and quite possibly not an optimal one. The first stage is to collapse the numerous different categories into five basic ones: 1. A Levels and similar (including IB) 2. BTEC diplomas and other vocational qualifications 3. Access to HE courses 4. Prior HE experience (below full degree) 5. Other / not known This is further complicated as the HESA codes for 2009 and 2012 cohorts differ. The following lists may not be exhaustive (as they are based on the initial institutions) and so more fitting may be necessary): **2009** 1 = P50, P62, P63 2 = P41, P42, P46, P80, P92, P93, P94 3 = X00, X01 4 = CXX, JXX 5 = Q80, X02, X04 (NB: DXX/HXX/MXX are previous degree level qualifications and so should not be in the sample) **2012** 1 = 37-40, 47 2 = 41, 43

	3 = 44-45 4 = 21-31 5 = 55-99 (NB: 1-16 are previous degree level qualifications and so should not be in the sample) The second stage is to subdivide the first category (i.e. A Levels and similar) into tariff groups. The tariff score recorded for other qualifications is not reliable and should be discarded. (NB: Attention is needed here not to conflate empty/missing tariff data with a zero tariff.) Clearly the tariff scores typical of different institutions are likely to vary widely with the entry profile of the institution. The approach taken for this stage to date has been to calculate quartiles for the tariffs (for A Level and similar qualifications) and to use these quartile thresholds to split this group into four sub-groups, with a fifth for students with A Levels, but no recorded tariff score (it is unclear why this occurs, but it seems relatively common to have some). The resulting coding is therefore: 1. A Levels and similar – Q1 tariff 2. A Levels and similar – Q2 tariff 3. A Levels and similar – Q3 tariff 4. A Levels and similar – Q4 tariff 5. A Levels and similar – unknown tariff 6. BTEC diplomas and other vocational qualifications 7. Access to HE courses 8. Prior HE experience (below full degree) 9. Other / not known
Degree result (2009 cohort only and only for employability analysis)	0 = Lower second, third class or pass degree, 1 = First or upper second class degree
Principal variable	**Coding and notes**
Bursary / household income combination	This is the key variable within the analysis and the hardest to specify and code, partly due to the variety of different bursary schemes in operation and the nature of the data available. It is important to pause at this point to consider the purpose of bursaries (in the terms of this study, at least) and the underpinning epistemology of what they are expected to achieve. Broadly speaking, their purpose is to provide a means of ameliorating the predicted educational disadvantage derived from coming from a low income background or being a member of a group felt likely to suffer disadvantage in higher education. In other words, bursaries are awarded as the university expects that the student will otherwise have poorer outcomes than other students in terms of retention, degree result and/or employability. This issue of comparison (to other students) is therefore important. The overriding principle in coding is therefore the creation of a 'comparison' group of students who are similar to bursary holders, but who did not receive a bursary, either through a prioritisation process or ineligibility. This is best

illustrated through examples:

1. A university makes bursaries available to all students with a household income of £25,000. The comparison group might therefore be students with household incomes between £25,001 and £42,600 (the upper threshold for the student maintenance grant for 2012 cohort).

2. A university makes a limited number of bursaries available to students with a household income of £25,000, but prioritises those from certain geographical areas. In this instance, the comparison group(s) might be those students with a household income of £25,000 outside of the target areas and/or those with household incomes between £25,001 and £42,600.

3. A university awards bursaries to students on the basis of criteria which are not means-tested – e.g. ethnicity, care leavers, disabled people, coming through an access route. The most comparison group might be comprised of students with low household incomes who might also be expected to have lower-than-average outcomes.

This is yet further complicated by many universities having multiple bursary schemes with very different criteria or priorities.

The household income data are also problematic. Firstly, it is important to recognise that there is a difference between a zero figure (i.e. very low income and likely benefit-dependency) and a missing figure. Missing figures generally represent individuals who have not engaged with the student finance means-testing process. For the most part, this indicates individuals from very high income households (as they know they will not be eligible for means-tested grants/loans), but there are some individuals who do not engage for cultural reasons or who engage, but refuse permission for their income details to be passed to universities. It is therefore obvious why the zero and missing groups should not be conflated; in transferring data between software packages, missing figures can be automatically replaced by zeroes and this needs careful management.

The formulation of the coding categories thus have to be specified by individual institutions, potentially with multiple bursary groups. There may also be an interest in examining the outcomes for students with higher household incomes. An example that may serve as a useful template was:

1. Household income between £25,001 and £42,600 (comparison group)
2. Students receiving a bursary due to household income under £25,000
3. Students receiving a bursary due to entry through access route, but with a household income over £25,000
4. Household income over £42,601
5. Household income missing

For analysis purposes, the comparison group should be coded as (1) – other codings are not important. Needless to say, it is important to ensure that the coding groups are exhaustive and mutually exclusive. Different household income threshold are likely to be needed for different cohorts – the equivalent figure to £42,600 for the 2009 cohort was £50,020).

Outcome variables	Coding and notes
Continued (2012 cohort only)	0 = No, 1 = Yes (i.e. did not enter second year at university)
Completion (2009 cohort only)	0 = No, 1 = Yes (i.e. achieved a full degree [not interim award] within five years) This measure has been problematic in several of the institutions to date due to the ways in which data is coded within the student records system and due to withdrawing students being re-registered for lower awards. This will therefore need careful specification within the data extraction and preparation stage. Ordinarily, institutions might expected 75-85% of students to complete, so variations outside of this range need extra scrutiny.
Degree result (2009 cohort only)	Two options depending on institutional relevance: Either: 0 = Lower second class degree and below, 1 = First and upper second class degree Or: 0 = Upper second class degree and below, 1 = First class degree The latter option is more likely to be relevant to institutions with higher entry requirements where a higher proportion of students achieve first class degrees.
Employment outcome (2009 cohort only)	0 = No, 1 = Yes (i.e. achieved a graduate-level outcome, as per standard definition)

From experience, it is useful at this point to quality check the datasets, specifically looking for large numbers of missing values within the recodings. The analysis rejects students with an missing values, so work may be needed to ensure the maximum possible data coverage. Also, it is helpful to undertake quick crosstab analyses to ensure that the coding of the data behaves as it might be expected – for example, that students with low household income are concentrated in POLAR quintiles 1 and 2, or that students living in their parental or own home live nearer to the campus on average than those renting accommodation.

Part 2 – analysis

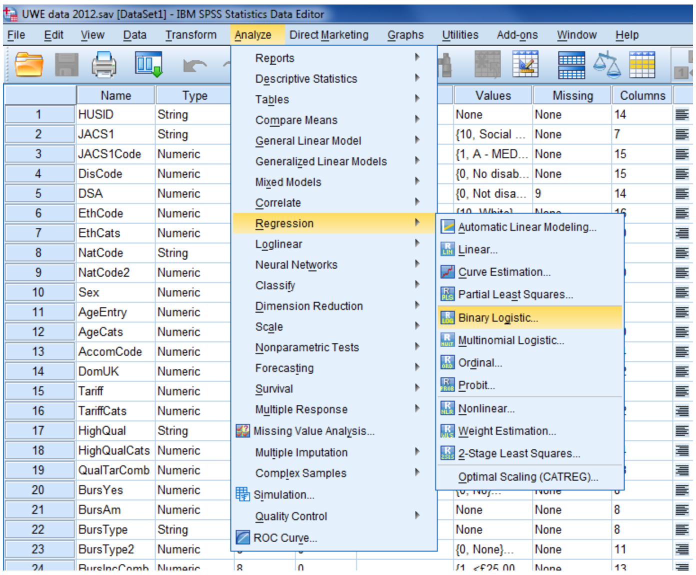
This project is predicated on *binary logistic regression analysis*. This was selected as it is a commonly-used and widely-known approach that is readily implemented using standard desktop computer software. It is not optimal (a multi-level approach or more formal discontinuity design would be stronger), but the underpinning principle of the project is to develop an analytical approach which can be readily replicated with fidelity across a wide range of institutions. Naturally, institutions are free to undertake alternative analyses outside of the project.

The analysis is therefore specified in SPSS as the most commonly-used statistics package in the social sciences. The instructions in this document are based around version 22, but should be relevant to versions dating back several years. Again, for consistency purpose, it is strongly recommended that SPSS is used and no support is available for other packages. It is assumed in these instructions that the reader has a reasonable level of proficiency with SPSS and statistical analysis, although they may not be specifically familiar with binary logistic regression.

In this context, binary logistic regression concerns itself with the likelihood that an individual has a particular dichotomous outcome – e.g. is retained (or not) into a second year or acquires a graduate job (or not). This likelihood is held to be predicted in part by the control variables and the main variable of interest (the combined bursary and household income variable). The analysis being undertaken here is a form of quasi-experiment, with a contrast between an experimental group (bursary holders) and a comparison group (other students from low/mid income households).

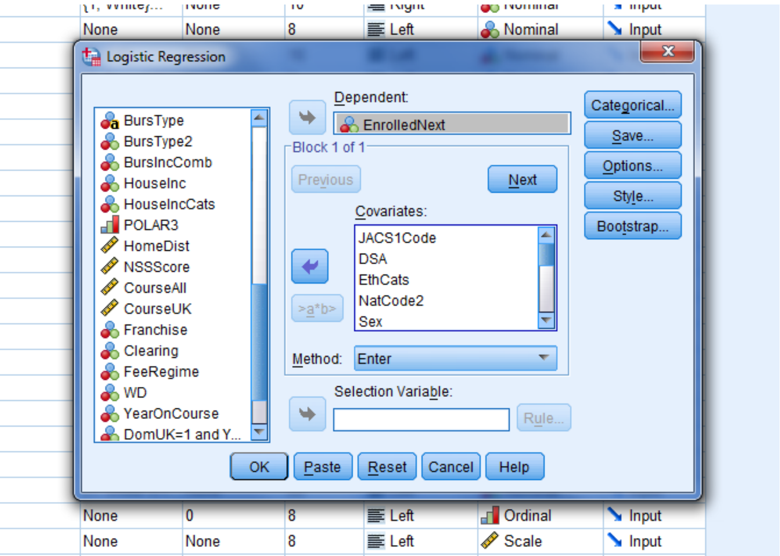
Needless to say, the first stage is to ready your dataset in SPSS if the recoding was done elsewhere. If you are importing data from Excel or another package, then a quality check will be required to ensure that none of the data have been altered in the process; as noted above, SPSS does sometimes render missing data as zeroes and this needs double-checking.

Binary logistic regression can be found under the menu options Analyze > Regression > Binary Logistic:



	Name	Type	Values	Missing	Columns
1	HUSID	String	None	None	14
2	JACS1	String	{10, Social ...	None	7
3	JACS1Code	Numeric	{1, A - MED...	None	15
4	DisCode	Numeric	{0, No disab...	None	15
5	DSA	Numeric	{0, Not disa...	9	14
6	EthCode	Numeric			
7	EthCats	Numeric			
8	NatCode	String			
9	NatCode2	Numeric			
10	Sex	Numeric			
11	AgeEntry	Numeric			
12	AgeCats	Numeric			
13	AccomCode	Numeric			
14	DomUK	Numeric			
15	Tariff	Numeric			
16	TariffCats	Numeric			
17	HighQual	String			
18	HighQualCats	Numeric			
19	QualTarComb	Numeric			
20	BursYes	Numeric			
21	BursAm	Numeric	None	None	8
22	BursType	String	None	None	8
23	BursType2	Numeric	{0, None}...	None	11

Within the dialog box that appears, enter the relevant outcome variable into the box labelled Dependent and the control variables and bursary/income variable into the box labelled Covariates. Leave the Method box tagged as ‘Enter’:

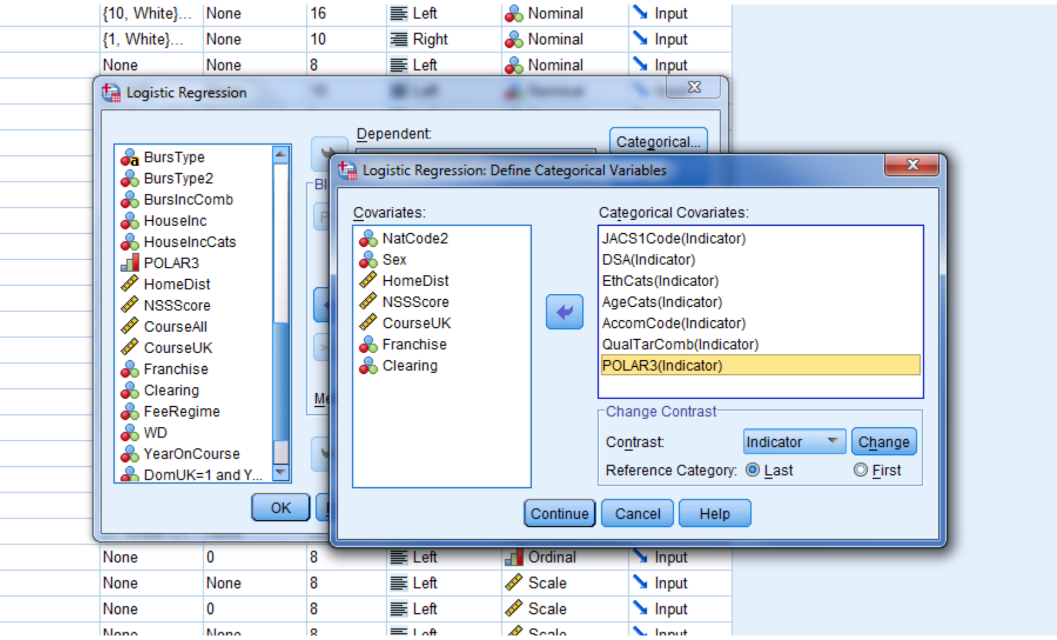


Dependent: EnrolledNext

Covariates:
JACS1Code
DSA
EthCats
NatCode2
Sex

Method: Enter

Now click on the Categorical link in the top right corner to define the many categorical variables within the model:



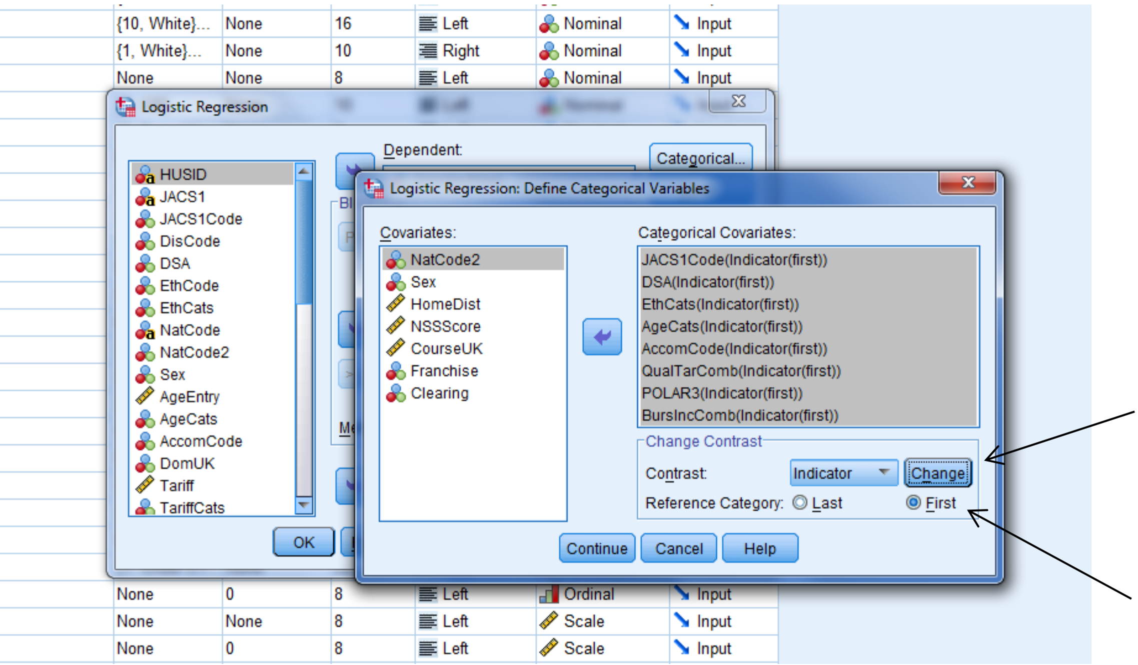
Covariates:
NatCode2
Sex
HomeDist
NSSScore
CourseUK
Franchise
Clearing

Categorical Covariates:
JACS1Code(Indicator)
DSA(Indicator)
EthCats(Indicator)
AgeCats(Indicator)
AccomCode(Indicator)
QualTarComb(Indicator)
POLAR3(Indicator)

Contrast: Indicator

Reference Category: ● Last ○ First

The variables that need including are subject, disability, ethnicity, age, accommodation, POLAR, entry qualifications and the bursary/income variable. The dichotomous categorical variables (nationality, sex, franchise and clearing) do not need to be included. Next highlight all of the categorical variables, click the radio button marked First and then click Change immediately above. This specifies that the first category is the reference category for these variables. This is largely arbitrary in terms of the analysis, but provides consistency and a shared basis for comparison. The list should now look something like this:



Logistic Regression

Logistic Regression: Define Categorical Variables

Covariates:
- NatCode2
- Sex
- HomeDist
- NSSScore
- CourseUK
- Franchise
- Clearing

Categorical Covariates:
- JACS1Code(Indicator(first))
- DSA(Indicator(first))
- EthCats(Indicator(first))
- AgeCats(Indicator(first))
- AccomCode(Indicator(first))
- QualTarComb(Indicator(first))
- POLAR3(Indicator(first))
- BursIncComb(Indicator(first))

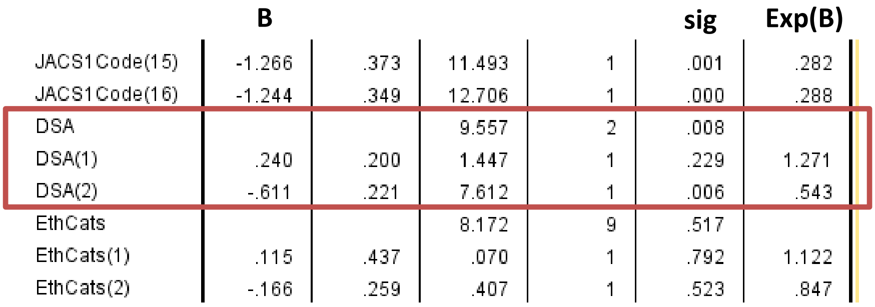
Change Contrast

Contrast: Indicator

Reference Category: ○ Last ● First

Click Continue to return to the main dialog box and click OK to run the analysis. The output from SPSS is rather opaque and hard-to-interpret. The table marked Model Summary is of interest as it provides the R-squared estimates that provide an indication of the proportion of the variation in the outcome variable explained by the other variables in the model. However, the main results are in the large table marked Variables In The Equation and the key columns of interest are those labelled ‘B’, ‘sig’ and ‘Exp(B)’. (Please note the examples in this document are fictional.)

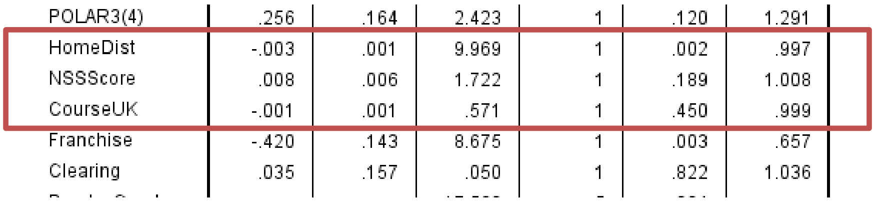
For example, focusing in briefly on the disability variable with three categories (no known disability, disabled with DSA and disabled without DSA):



	B				sig	Exp(B)
JACS1Code(15)	-1.266	.373	11.493	1	.001	.282
JACS1Code(16)	-1.244	.349	12.706	1	.000	.288
DSA			9.557	2	.008	
DSA(1)	.240	.200	1.447	1	.229	1.271
DSA(2)	-.611	.221	7.612	1	.006	.543
EthCats			8.172	9	.517	
EthCats(1)	.115	.437	.070	1	.792	1.122
EthCats(2)	-.166	.259	.407	1	.523	.847

The first of these is the reference category and therefore does not appear in the results – the other groups are compared to this one. In this example, disabled students with a DSA, labelled as ‘DSA(1)’, has a ‘sig’ of .229. This is the p-value and as it is over .050 (i.e. the conventional significant level), there is no significant difference in retention rates between this group and the reference group. However, disabled students without a DSA, labelled as ‘DSA(2)’, have a p-value below .050 and so there is a significant difference here. As the ‘B’ (estimated coefficient) is negative, this group can be inferred to be less likely to be retained into a second year, all else being equal. The ‘Exp(B)’ column provides the ‘odds ratio’ and a measure of the effect size for this relationship, with the odds of disabled students without a DSA being retained being just over half (.543) of the odds of students with no known disability[1].

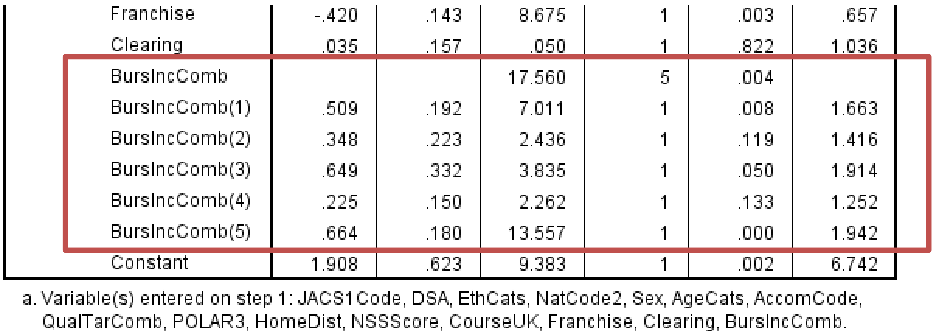
Similarly, if we take the continuous variables within the model:



POLAR3(4)	.256	.164	2.423	1	.120	1.291
HomeDist	-.003	.001	9.969	1	.002	.997
NSSScore	.008	.006	1.722	1	.189	1.008
CourseUK	-.001	.001	.571	1	.450	.999
Franchise	-.420	.143	8.675	1	.003	.657
Clearing	.035	.157	.050	1	.822	1.036

Here we can see that NSS score and course size have p-values over .050 and so are not significant predictors for retention. However, home distance has a p-value of .002 and a negative B coefficient, meaning that students whose permanent home is further from the university are more likely not to be retained, all else being equal. This translates (see footnote 1) into about a 3% lower likelihood of being retained for every 100km distance.

These examples are drawn from the control variables in the model. While these may be of wider interest, the principal focus is on the combined bursary/income variable as this is where the bursary and comparator groups can be contrasted, with all else being held equal:



Franchise	-.420	.143	8.675	1	.003	.657
Clearing	.035	.157	.050	1	.822	1.036
BursIncComb			17.560	5	.004	
BursIncComb(1)	.509	.192	7.011	1	.008	1.663
BursIncComb(2)	.348	.223	2.436	1	.119	1.416
BursIncComb(3)	.649	.332	3.835	1	.050	1.914
BursIncComb(4)	.225	.150	2.262	1	.133	1.252
BursIncComb(5)	.664	.180	13.557	1	.000	1.942
Constant	1.908	.623	9.383	1	.002	6.742

a. Variable(s) entered on step 1: JACS1Code, DSA, EthCats, NatCode2, Sex, AgeCats, AccomCode, QualTarComb, POLAR3, HomeDist, NSSScore, CourseUK, Franchise, Clearing, BursIncComb.

In this example, six categories were used within the model:

1. Household income under £25,001 and no bursary (reference group)
2. Household income between £25,001 and £42,600
3. Household income between £42,601 and £62,125
4. Household income over £62,125

[1] Note Zhang and Yu's work that cautions against conflating 'odds ratios' with relative likelihood – the example result does not mean that disabled students without a DSA are half as likely to be retained, especially when the outcome (retention is common). In this instance, they are around 8% less likely. In other words, odds are not the same as probabilities.

Zhang, J. & Yu, K. (1998) What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes, *Journal of the American Medical Association*, 280(19), 1690-1691.

5. Household income missing
6. Household income under £25,001 and bursary

Three of the income bands (£42,601 to £62,125, over £62,125 and missing) were not significantly different from the reference group – although the middle one was right on the threshold. However, two groups did have significantly better retention than the reference/comparator group, which were those from mid income households (4% increased likelihood of retention) and those with bursaries (5% increased likelihood).

This example also usefully illustrates the relationship between subsample size, significance and effect size. The second largest positive effect size ('Exp(B)') is for the household income over £62,125 group. However, this is a small group as most students from high income household do not participate in the means-testing process. Therefore, despite the high estimated effect size, there is insufficient evidence to conclude that they have significantly higher retention rates than the reference group. Care is needed, therefore, in the interpretation and reporting of the results. This process of analysis can then be repeated for each of the cohorts and outcome variables, remembering to add degree result as a control variable for analysis of employment outcomes.

For simplicity, and cognisant of the relatively small subsample sizes, no interaction terms are used in the model. In other words, it is not possible to infer whether bursaries are associated with different effects for different groups (e.g. women or mature students). During the earlier phase of the project, interaction terms were explored, but no convincing relationships were identified – this was likely due, in part, to the increasingly small subsamples. It might be possible to perform this analysis more robustly by pooling cohorts over multiple years, should an institution wish.

Part 3 – interpretation of results

Once the analysis has been completed, there are effectively three main results that can be extracted about the impact of bursaries. These, and the inference to be drawn, are summarised in the table below:

Result 1: bursary holders have significantly weaker outcomes than the comparison group	Bursaries are either ineffective or insufficient in scale to overcome the underlying effects of financial disadvantage
Result 2: bursary holders have the same outcomes as the comparison group (i.e. no significant differences)	Bursaries are effective (or there is actually no impact of financial disadvantage on educational outcomes)
Result 3: bursary holders have significantly stronger outcomes than the comparison group	Bursaries are very effective (and possibly unfairly so with respect to the comparator group)

As can be seen, each result has a degree of ambiguity and complexity attached to it. Results 2 and 3 should lead to the inference that bursaries have helped to ‘level the playing field’ with respect to the comparator group, although the latter may suggest that the amounts have been so high as to cause a degree of unfairness. Result 1 is problematic. It is evidence that bursaries are not ‘levelling’, but it is necessarily silent on whether (a) this is because financial support is ineffective in overcoming educational disadvantage, or (b) the bursaries being offered are just too small. This result almost certainly requires further investigation.

It is important to appreciate that there are limits to the analysis described within this document. The interplay between financial and educational disadvantage is complex and the data available to institutions are limited in their reliability and/or validity. ‘Proof’ is not going to be forthcoming and the results will require a nuanced weighing of evidence. Furthermore, there may be important measures of impact which sit outside of the data available – e.g. around students’ mental health or wellbeing. The analytical framework is therefore intended to provide a balance between good evidence, the strength of inference it will bear and the ‘costs’ of generating it.