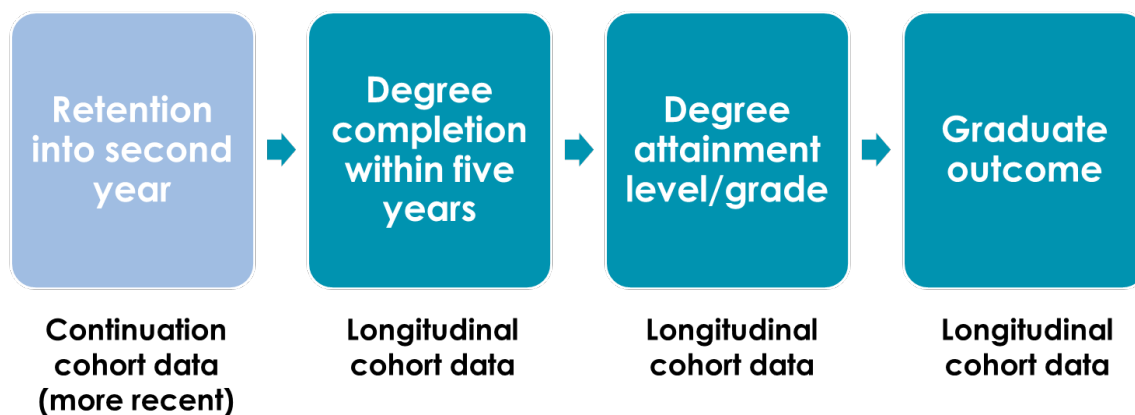


Financial support evaluation toolkit: technical coding workbook for statistical tool

Part 1: readying the data set

You will need data for all of the cohort years you wish to evaluate. The analysis uses the most recent available data to analyse the effects on each of the four relevant outcomes as illustrated below:



When you have compiled your data, you will have a minimum of two datasets, with:

- one or more 'continuation cohort' dataset(s) using more recent retention data and
- one or more 'longitudinal cohort' dataset(s) using long term data including graduate outcomes.

You are recommended to analyse at least two sequential years of data for each cohort to reduce the chance of drawing erroneous conclusions (particularly for your first analysis). To support this, since 2019 HESA has released longitudinal cohort data in respect of:

- Degree and graduate outcomes data for 2010-11 to 2016-17 student cohorts
- Continuation outcomes for 2012-13 to 2019-20 student cohorts.

Institutions with around 300 to 800 financial support recipients are unlikely to have enough data from one cohort year, so will need to compile their datasets with data from multiple years. However, this will only be possible where the financial support arrangements have remained the same or very similar for those years.

If combining data from multiple years, you may either include the year as part of the coding for your principal financial support variable, or include the year as an additional control variable to identify any differences between years.

The existing data you will need to collect to create your dataset(s)

HESA data release (specifically designed for this tool)

HESA data is only available to HESA subscribing institutions. If you subscribe to HESA, you will need to ask your institution's designated named HESA Student Record Contact to call 01242 211 482 to request the data extract to evaluate financial support impact. Your institution's contact will then receive a PIN code to download the data release through the secure HESA system (their normal process). If you are unsure who your designated HESA contact is, contact HESA on the same number.

Check the summary details of students included and excluded below, to ensure this data will be suitable for your institution's own context.

The HESA data release includes:

- full-time, first degree entrants
- sandwich students
- English domiciled students (under the same funding regime)
- HEFCE funded (includes HEFCE funded Initial Teacher Training students).

The HESA data release excludes:

- students whose reason for leaving was death/serious illness
- students not funded by HEFCE (FUNDCODE 1 and 7 are included 2, 3, 5 are excluded)
- students who qualified with a first degree in their year of entry (generally entrants from foundation degrees)
- students who left before 1 December
- part-time students
- students on degree courses over five years (e.g. some medical or architecture students)
- Scottish, Welsh and Northern Irish students.
- EU and international students.

If you cannot access the HESA data release or would prefer to collate the whole dataset manually, you can do this using the framework of variables and coding in this document. However, you will need to allow significant extra time for linking and coding this data.

Details of financial support provision (SLC HEBSS)

If your institution uses the Higher Education Bursary and Scholarship Scheme (HEBSS) to provide bursaries, the data tracking bursary amounts to recipients will be available through the HEBSS system. The Student Loans Company (SLC) has confirmed that institutions are covered contractually in using their HEBSS data internally to monitor and evaluate their own financial support packages. If your institution does not use HEBSS, you will need to find how and where this data is stored within your own institution.

Household income data (likely to be via SLC HEBSS)

As above, this will be available through SLC HEBSS data and/or within your own institution's processes. **Be very careful manipulating this data – blanks (income unknown, but probably very high) must not be converted to zeros (income very low).**

Student record data

This data is available within your own institution's student record data systems. This data should provide details about whether students entered through the clearing process. You may also need to refer to this data for details of financial support provision (e.g. criteria, amounts).

NSS data

This data is available within your own institution. Public NSS data by institution is available at: www.officeforstudents.org.uk/advice-and-guidance/student-information-and-data/national-student-survey-nss/nss-data-provider-level/. Ideally, you can use the NSS data for the relevant cohort year (i.e. the survey data those students provided in their final year for their degree). However, if this is not possible, choose a sensible alternative, considering any changes at your institution that may affect these results.

The pilot research team chose 'Overall satisfaction score' as a general indicator of the 'quality' of the course as they experience it, as this might be hypothesised to have a strong effect on student outcomes. If there is no course score, you may use the faculty/department average score. Bear in mind that records that do not have data will not be used in the analysis, so do find a relevant score to include.

Compiling your full dataset

You now need to combine the different datasets to create a new dataset for each cohort year, where each student has data for all of the required variables. The HESA data release includes the Person identifier (HUSID), higher education provider's own identifier (OWNSTU), and Student Support Number (SSN) to facilitate this data linking.

If you find duplicate student IDs in your data, we recommend using what looks like the largest or most important instance of study. Often, you may find students are only doing a single module of another course (e.g. languages).

Records that have any missing data will not be used in the analysis, so you should backfill this with a comparable acceptable data or create an unknown category.

Finally, when you have your dataset ready for coding, it is useful at this point to critically 'eyeball' the datasets to see whether there are any obviously erroneous groups of students present or absent. Issues can occur due to human error, unusual degrees, miscoding in the student records system, or errors in the extraction process. It is also a good idea to do some simple tallies and crosstabs to compare the dataset with what is generally known about your institution as a final check.

Coding/re-coding the dataset

At this point, you will hopefully have a large dataset for each cohort year stored in Excel (or similar). The next stage is to recode this data into a standardised format for analysis.

It is likely that this coding process will throw up new questions along the way. Coding decisions are important because they may affect the internal validity of the analysis and also shape what you can

learn from it. We would recommend that you do not make major changes to the model unless you have experts involved who can understand and assess the impact of changing the coding framework.

You will need to make judgements about what is a good group size when (re)coding the data due to the unique features of your institution – in general, aim for sizes above 50 as a minimum, although smaller groups can be used. As a general rule, having large groups will make it more likely that you will identify significant relationships with the outcome variables. However, larger groups will tend to lose granularity as you combine students together. For example, you could analyse age-on-entry by individual years (i.e. 18, 19, 20 etc), but you would be very unlikely to find that students of a given age-on-entry had radically different outcomes to those a year older or younger. We therefore group ages together (e.g. '18 to 20') to provide a broader picture in which meaningful relationships are likely to arise. The same principles apply to other variables in the model, like subject of study, entry qualifications or type of bursary.

One weakness in the statistical tool is that it necessarily precludes an understanding of very small groups (e.g. transgender students) who either have to be combined with other groups or omitted altogether. If you are specifically interested in very small groups, then you may want to considering pooling your datasets over multiple years or using qualitative methods.

You can recode the data within Excel or SPSS (which you will probably use for the analysis) as you prefer. **Beware that blanks must not be converted to zeros when manipulating your data.**

Variables and code framework

The following tables instruct how to compile and code your dataset. In this table, cell shading indicates the data is available pre-coded by HESA (for subscribing institutions).

Outcome variables	Coding	Notes
Continued into second year of study (continuation cohort only)	0 = No, 1 = Yes: continued on first degree or higher course, 2 = Continued on other undergraduate course	Source: HESA, pre-coded. Students who continued studying at another institution are coded as not continuing due to data protection. Institutions with very small numbers of students not continuing may struggle to understand this outcome in the statistical model.
Degree completion by #YEAR# (longitudinal cohort only)	0 = No, 1 = Yes, 2 = Gained other undergraduate qualification, 3 = Still studying	Source: HESA, pre-coded. 1/Yes counts only full degree completion, not interim award, within five years (linked using HUSID & INSTID).
Degree result (longitudinal cohort only)	Either: 0 = Lower second class degree and below, 1 = First and upper second class degree, 2 = Did not obtain degree with applicable classification	Source: HESA, pre-coded. You can choose to use either the 'Degree result' variable or 'Degree result (higher)' depending your own institution's context.
Degree result (higher) (longitudinal cohort only)	0 = Upper second class degree and below, 1 = First class degree, 2 = Did not obtain degree with applicable classification	Source: HESA, pre-coded. The 'Degree result (higher)' outcome is more likely to be relevant to institutions with higher entry requirements where a higher proportion of students achieve first class degrees.
Positive / negative destination marker (Graduate outcome) (longitudinal cohort only)	PRO_EMP = Professional employment FURTHER_STUDY = Primarily studying OTHER_POSITIVE = Other activity considered positively NON_PRO_EMP = Non professional employment EMP_SOC_MISSING = Employment with missing SOC code	Source: HESA, pre-coded. Methodology is consistent with that used for the graduate progression measure within OfS publication of Projected completion and employment from entrant data (Proceed) in 2021 ¹ , and consulted on as the definition of progression measures within the OfS consultation on the construction of student outcome and experience indicators for use in OfS regulation ² .

¹ See [Projected completion and employment from entrant data \(Proceed\): Updated methodology and results - Office for Students](#)

² See [Outcome and experience data - Office for Students](#), including the 'Core algorithms' supporting technical document for data definitions related to progression indicators.

Outcome variables	Coding	Notes
	UNEMPLOYED = Unemployed or due to start work OTHER_NEGATIVE = Other activity considered negatively UNKNOWN = Unknown activity Not applicable = Not in the GO population or non response to survey	

Control variables	Coding	Notes
Gender/Sex	1 = Male, 2=Female, 3 = Other	Source: HESA, pre-coded.
Nationality	0 = UK, 1 = Other (incl. unknown)	Source: HESA, pre-coded. As data is restricted to only England domicile students, 'Other' can identify students with dual nationality, or who have acquired UK citizenship.
Age on entry	1 = 20 and under, 2 = 21 to 24, 3 = 25 to 29, 4 = 30 and over, 5 = Unknown	Source: HESA, pre-coded.
Ethnicity	1 = White, 2 = Black Caribbean, 3 = Black African, 4 = Indian, 5 = Pakistani, 6 = Bangladeshi, 7 = Chinese, 8 = Mixed, 9 = Other, 10 = Unknown	Source: HESA, pre-coded. The extensive list of categories have been aggregated, this is not optimal but ensures a reasonable coverage for analysis.
Disability and Disabled Students Allowance receipt group	1 = No known disability, 2 = Disability and receiving DSA, 3 = Disability and not receiving DSA, 4 = Disability/DSA unknown	Source: HESA, pre-coded.
Main subject of Degree (Principal subject JACS codes) (longitudinal cohort only)	Data is the XJACSA01 code with the additional code M for multiple subjects	Source: HESA, pre-coded. Where a subject clearly dominates (greater than 50%) then this is used, but if the components are equal or all less than 50% then a 'mixed' group is used. More details on JACS: https://www.hesa.ac.uk/support/documentation/jacs Collapse these groups down so that you have less groups. Make an informed choice based on the disciplines, departments, and number of students at your institution.

Control variables	Coding	Notes
Main subject of degree (CAH level 1 subject codes) (continuation cohort only)	Data is supplied as the XCAH01 code with the additional code CAH99 for multiple subjects	Source: HESA, pre-coded Where a subject clearly dominates (greater than 50%) then this is used, but if the components are equal or all less than 50% then a 'mixed' group is used. For more details on CAH: https://www.hesa.ac.uk/support/documentation/hecos/cah
Entry qualifications	1 = 01-AAAA, 02-AAA, 03-AAB, 04-AAC, 05-ABB, 2 = 06-ABC/BBB, 07-ACC/BBC, 3 = 08-BCC/CCC, 4 = 09- >290, 10 >260, 5 = 11>230, 12>200, 6 = 13>160, 14>100, 15>0, 7 = 16-COMB3, 17 – LEV3, 8 = 20 – FOUND (HE & FE level), 21 -ACCESS, 9 = 18 – GNVQ/NVQ, 22 – BTEC/ONS, 10 = 23 – HEPG, 24- HEFD, 25-HEOUG, 11 = 26- NONE, 27- OTHERS, 28 – Unknown, 12 = 19 – BACC	Source: HESA, pre-coded. You can choose to use either the either the 'Entry qualifications' or 'Tariff score' variable depending on your own institution's context. Collapse these groups down so that you have around 5-6 groups. Make an informed choice based on the entry qualification student profile at your institution. 'Entry qualifications' is suitable for institutions whose students enter with a range of different qualification types.
Tariff score	As a continuous variable representing tariff score – i.e. 0 to ∞.	Source: HESA, pre-coded. This variable is restricted to those with Level 3 tariffable qualifications and thus 'Tariff score' is only appropriate for institutions where nearly all first degree students enter with traditional tariffable qualifications.
Franchised degree marker	0 = Other, 1 = Fully franchised	Source: HESA, pre-coded.
POLAR quintile (QYPR)	As per the POLAR quintiles, i.e. 1 = Lowest participation, to 5 = Highest participation	Source: HESA, pre-coded.
Accommodation type	1 = Institutional/private halls, 2 = Parental home, 3 = Own home, 4 = Other rented, 5 = Other (including not known and not in attendance)	Source: HESA, pre-coded. First year accommodation
Distance travelled from home address to	As a continuous variable – i.e. 0 to ∞.	Source: HESA, pre-coded. In kilometres.

Control variables	Coding	Notes
higher education Provider		
Degree size	As a continuous variable – i.e. 1 to ∞.	Source: HESA, pre-coded. Number of full-time students on each course, based on CourseID returned to HESA.
Industrial placement location marker (continuation cohort only)	0 = Other, 1 = Placement year	Source: HESA, pre-coded. 1/Placement year includes HESA codes for full placement year or year working abroad (LOCSDY = D, LOCSDY = T, MOBTYP is only 02).
NSS score for degree (can be used as a quality marker for both cohorts)	As a continuous variable – i.e. 0 to 100.	To avoid losing a large proportion of students from the analysis due to missing data, you must replace any missing values with a mean score. Ideally use a department/faculty score, but otherwise for the university as a whole.
Clearing entrant	0 = No, 1 = Yes	Source: Institution's own student record data. You can remove this variable if you don't have enough to make a reasonable group size or it is not relevant to your institution.

Principal variable	Coding and notes	
Financial support/household income combination	<p>**EXAMPLE CODING ONLY**</p> <ol style="list-style-type: none"> 1. COMPARISON GROUP: Household income between £25,001 and £42,600, students are not eligible for a bursary. 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 (assuming high income) 	<p>See notes below to inform this coding, as this critical variable where you define your comparison group(s) must be relevant to your institution's own context.</p> <p>You must not confuse or replace a zero with missing/blank data in your household income data.</p> <p>The comparison group should be coded as '1'.</p>

Choosing your comparison (or 'reference') group

This is the most important variable to specify and code within the analysis. You will now need to carefully think through how you define your comparison group and other bursary groups.

We can pause at this point to consider the purpose of bursaries and the principles that underpin their goal(s). Generally, financial support provision aims to 'level the playing field', so bursaries are awarded as the university expects that the bursary student will otherwise have poorer higher education outcomes than other students. This point of comparison with other students is therefore crucial.

The main aim in this analysis for the principle variable is to create a 'comparison' or 'reference' group of students who are as similar to bursary holders as possible, but who did not receive a bursary, due to prioritising particular characteristics or chosen eligibility criteria at your institution. The comparison group will vary depending on the nature of your financial support provision. Your bursary comparison group should be a reasonable size, ideally between 500-1000, although fewer can be used if necessary. Your coded groups must have:

- one comparison group who did not receive financial support but are considered similar to recipients
- at least one group of financial support recipients.

You may want to include multiple groups for recipients of different types of financial support packages and associated support, to examine the outcomes for students with higher household incomes.

For analysis purposes, the comparison group should be coded as '1', but the other group's assigned codes are not important. Needless to say, it is important to ensure that the coding groups are exhaustive and mutually exclusive. Different household income thresholds may be necessary for different cohorts depending on the year (e.g. the equivalent threshold to £42,600 in 2012 was £50,020 in 2009).

This is illustrated in the examples below.

Bursary group	Potential comparison group
All students with a household income of £25,000	Students with household incomes between £25,001 and £42,600 (the upper threshold for the student maintenance grant for 2012 cohort).
Students with a household income of £25,000, but priority given to those from certain geographical areas, but a limited number of bursaries are available.	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.
Students meeting specific non-means tested criteria (e.g. ethnicity, care leavers, disabled people, access entry route).	Students with low household incomes who might also be expected to have lower-than-average outcomes.

You must not confuse or replace a zero with missing or blank data in your household income data. Zero figures generally represent very low or no income, while missing data indicates those who have not engaged with the student finance means-testing process who are likely to be from high income backgrounds. Be careful that your software program (e.g. SPSS) does not change blank fields to zeros when you transfer or copy data.

We recommend you assume that people who refuse to share their data (blanks or missing), fall in the high income bracket as these students have chosen to be excluded from automatic consideration for financial support. In elite institutions, the social make-up of the student body may mean that there are few or no students with household income data who have not received a bursary. In this instance, using the group with unknown household income as a comparator group may be a good approach.

If you have multiple assessments for bursary applicants/holders, we recommend using the first assessment for simplicity.

If you have a bursary eligibility criteria at your institution that is also one of the control variables in the model (e.g. all clearing students get bursaries), this may compromise the analysis and undermine your results. This can be dealt with by a statistician; however, this is not simple so you need to just remove the variable unfortunately.

You will need to use your informed judgement to construct appropriate financial support and comparison groups as they need to be tailored to your own institution's context. Defining the groups depends to a large degree on elements like the social mix of students, the number of bursaries given and the criteria used. This may require some modelling before the final analysis to ensure that you are able to draw conclusions at the level of detail required. For example, if you have more (but smaller) financial support groups, you may be able to explore the impact of different types of bursary, but you are less likely to be able to make firm claims about effectiveness.

Finally, before moving on to analysis...

From experience, it is useful at this point to quality check the datasets, specifically looking for large numbers of missing values within the (re)coding. Student records with missing data will be excluded from the analysis so it is worth exploring and correcting any significant omissions (for example, by tracking down missing data or through creating an additional 'unknown' code).

You may wish to undertake quick crosstab analyses to ensure that the coding of the data behaves as you might expect. For example, check 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.

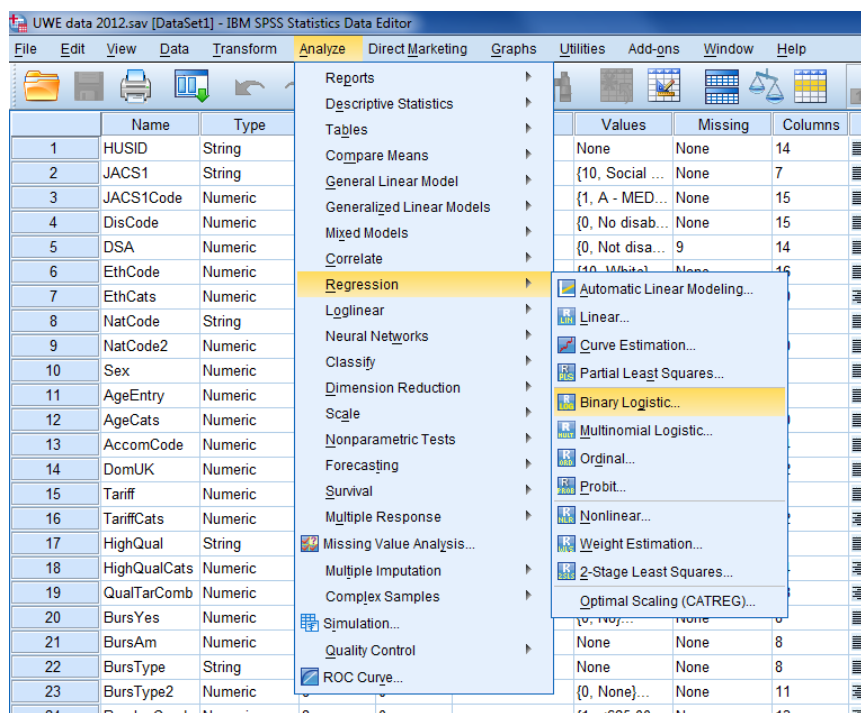
Part 2: SPSS analysis guidance

A binary logistic regression analysis was chosen in the development of the statistical tool to support an analytical approach which can be readily replicated with fidelity across a wide range of institutions. 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).

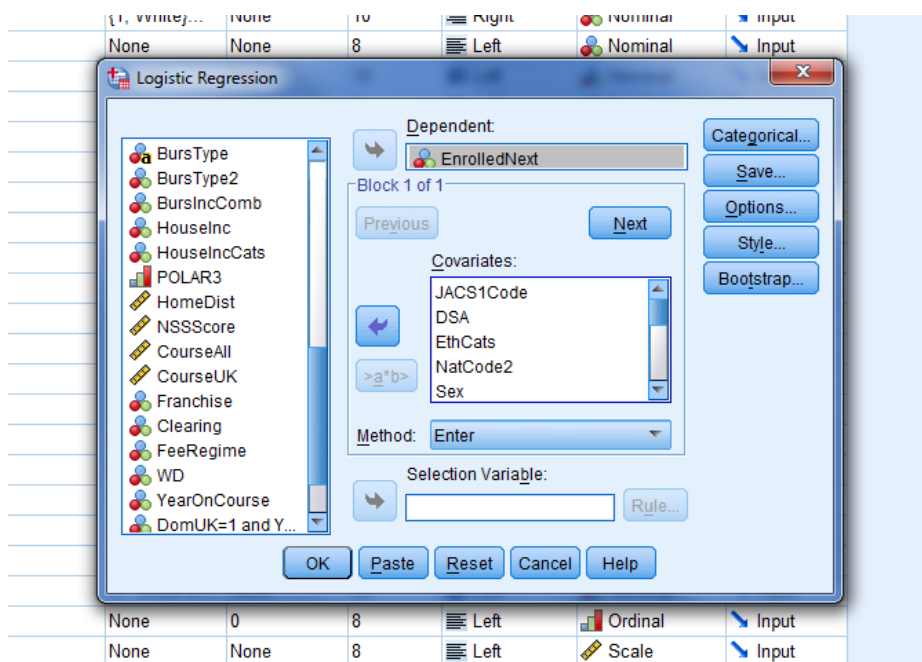
This guidance is for using SPSS, as the most commonly-used statistics package in the social sciences. The instructions in this document are based around version 22, but should be relevant to versions dating back several years. It is possible to use other statistics packages, but the advice provided in this document is based on using SPSS. 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.

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 has been altered in the process; as noted above, SPSS does sometimes render missing data as zeros and this needs double-checking.

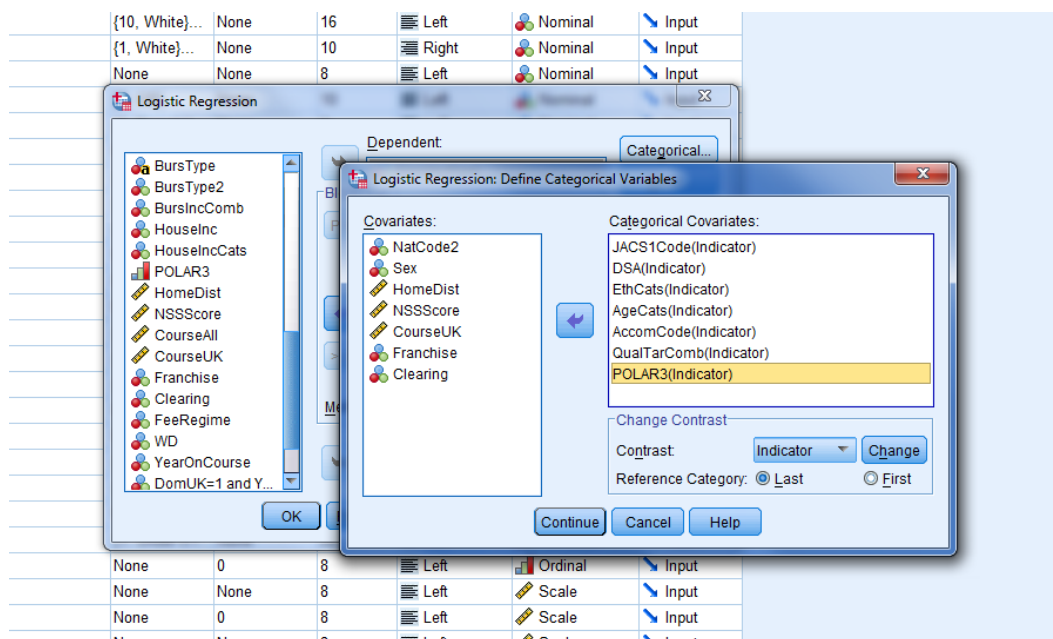
Binary logistic regression can be found under the menu options **Analyze > Regression > Binary Logistic** as shown in this screenshot:



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', as shown in this screenshot:



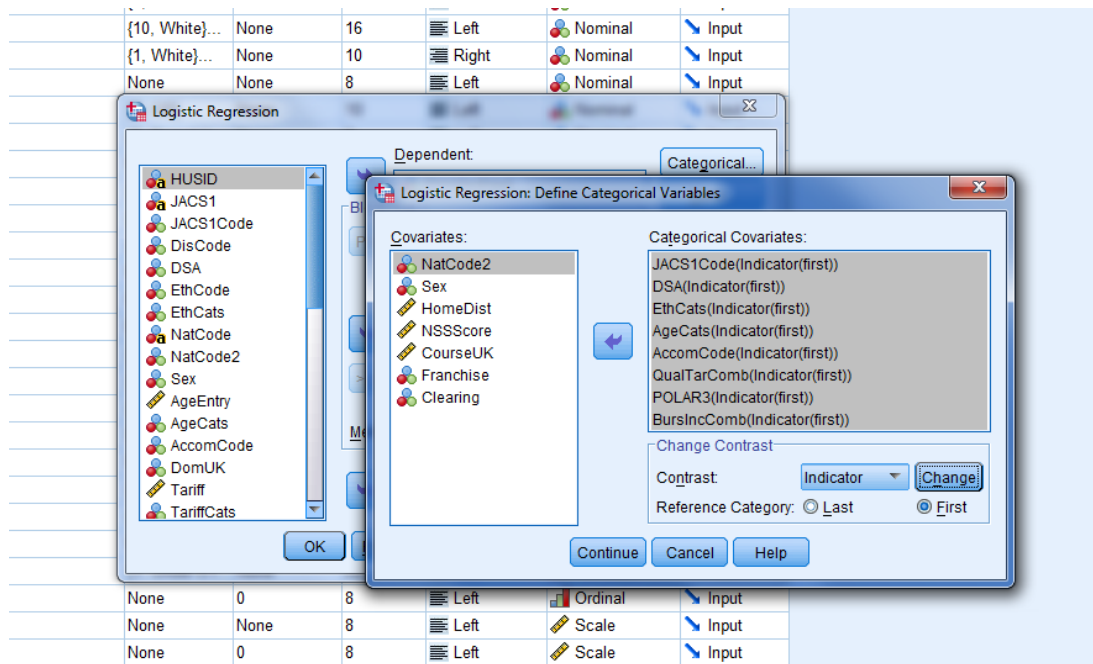
Now click on the **Categorical** link in the top right corner to define the many categorical variables within the model:



The variables that need including are those that are categorical with more than one category: i.e. 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:



Click **Continue** to return to the main dialog box and click **OK** to run the analysis.

The SPSS output 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				si	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 significance 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.

However, odds are not the same as probabilities, so be cautious about conflating ‘odds ratios’ with relative likelihood. For further details of a method to translate odds ratios into relative likelihoods, see Zhang, J and Yu, K ‘What’s the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes’.³ The formula proposed in this article is: “Relative likelihood = $OR / ((1-P) + (P \times OR))$ ”, where OR is the odds ratio (from the SPSS output) and P is proportion of non-bursary students achieving the outcome being analysed (e.g. being retained into second year or getting a first). A spreadsheet to perform the calculation for this formula is here: <https://www.officeforstudents.org.uk/media/a82f9ee8-8f0a-4713-941b-e38b94635897/odds-ratio-to-relative-likelihood-converter-1.xlsx>.

The example result shown here 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 per cent less likely.

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 into about a 3 per cent 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 comparison 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: JACS1 Code, 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

³ See Zhang, J and Yu, K ‘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), 1690-1 (1998), at: <https://www.ncbi.nlm.nih.gov/pubmed/9832001>.

4. Household income over £62,125
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/comparison group, which were those from mid income households (4 per cent increased likelihood of retention) and those with bursaries (5 per cent 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 households not participating 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. Take care in interpreting your findings and report your results clearly to avoid misinterpretation.

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, or by analysing subsets of the data (e.g. just women).