Q-STEP SPSS 'HOW TO' GUIDES:

RECODING CATEGORICAL VARIABLES INTO DUMMY VARIABLES IN SPSS

Creator: Joshua Townsley

Say we wanted to run a regression model in SPSS with a categorical independent variable:

Normally, we would interpret each IV's coefficient in a regression as the change in the DV for a one unit increase in that particular IV. But, if we have a categorical variable such as Ethnicity or Age Group, then what would a one unit increase in that variable actually mean? It's either nonsensical or tricky to interpret.

So instead, we have to break that categorical variable up into 'dummy variables' for each particular value/answer. This is a nuisance, but it's really important for any running regression in SPSS that uses a categorical Independent Variable.

E.g. Let's say we have a variable – Age Group – coded as follows:

1 = Under 18
2 = 18-25
3 = 26-35
4 = 36-45
5 = 46-55
6 = 56-65
7 = 66+
:
We'll have to break it up into 'Dummy Variables' (coded 1 for 'yes'/0 for 'no') for each particular age group. E.g. AgeGroup1 - Under 18 (yes or no), AgeGroup2 - 18-25 (yes or no), AgeGroup3
- 26-35 (yes or no), and so on. So, we will end up with as many new variables as there are categories in our categorical variables (in this case, 7).

There are basically two ways to do this – an easier way and a trickier way. This guide runs through each one below, but the easy one is of course better (but may not be available for older versions of SPSS).
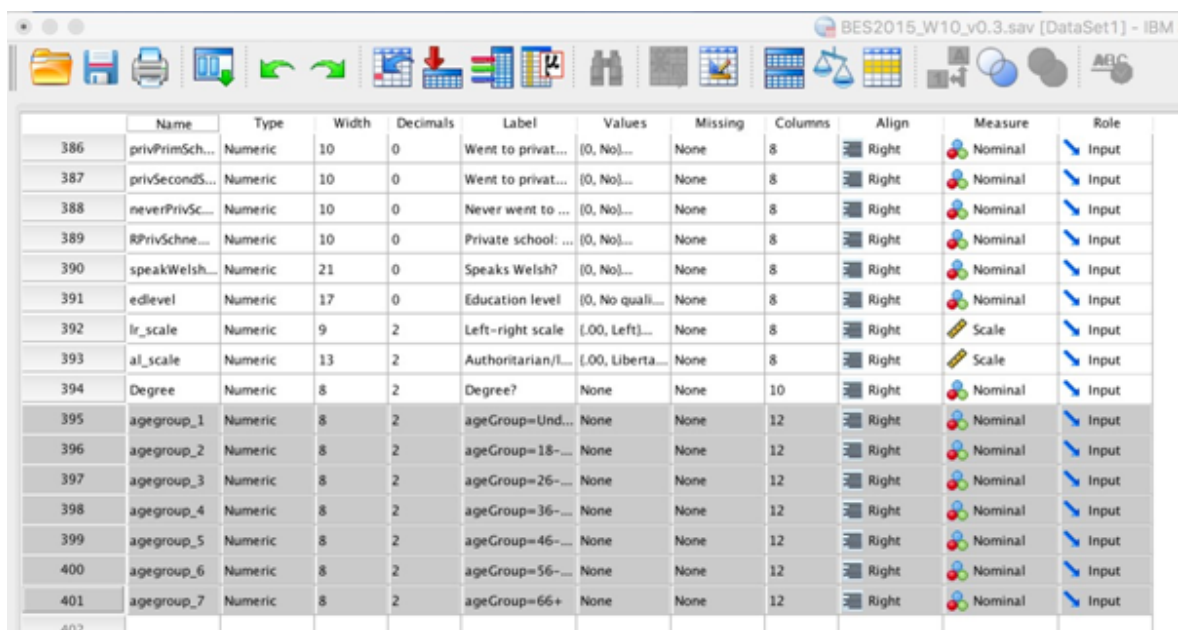
1. Method 1

Go through the following menus on SPSS:

Transform > Create Dummy Variables

1. Put your categorical independent variable into the top right box, and enter an intuitive name for the dummy variables into the 'Root Names' box
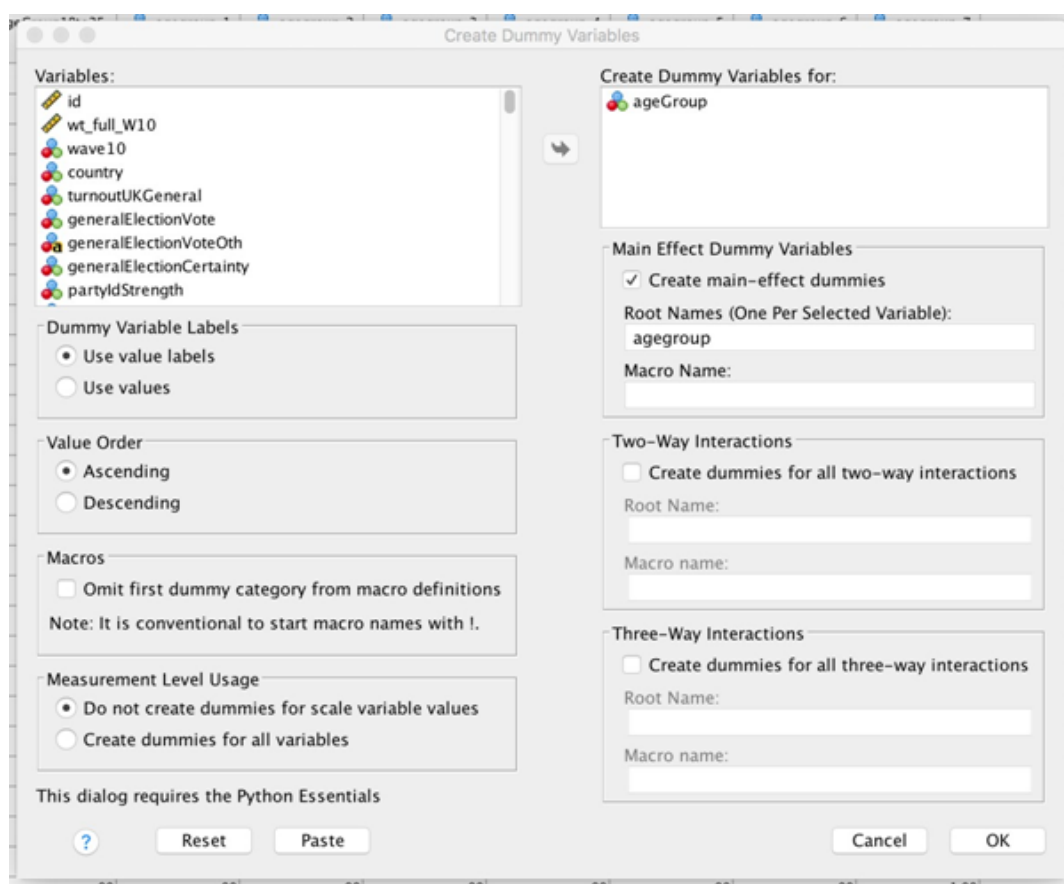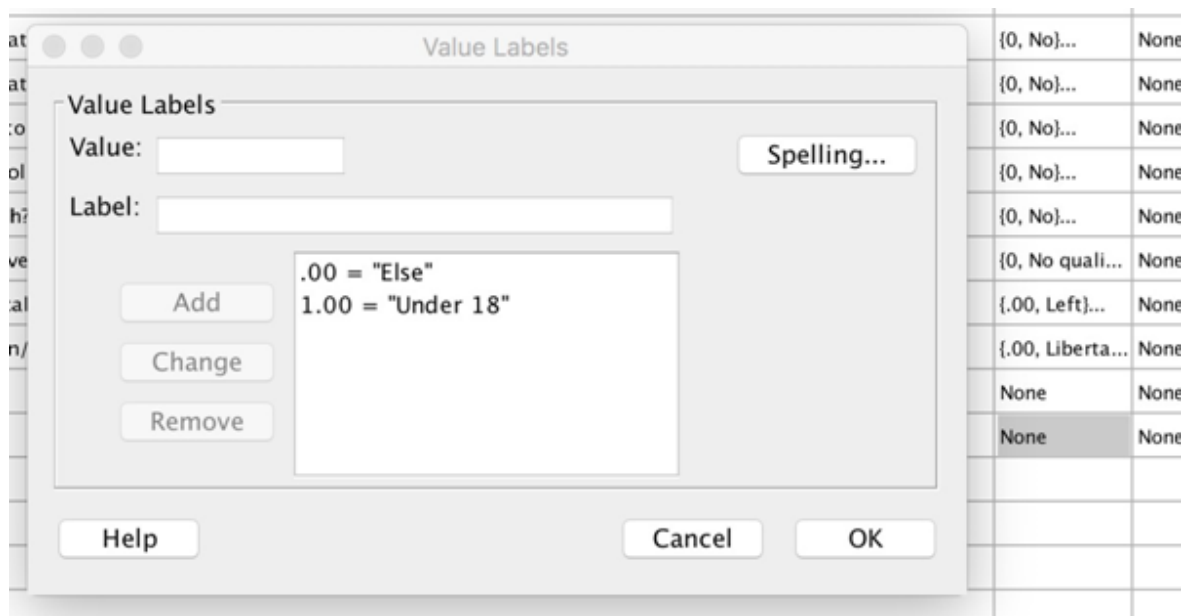

2. Select 'OK' to create the dummy variables. They will then appear at the bottom of the variable list:

Sheffield
Methods
Institute.

3. For each new dummy variable, click on the Values box and create the value labels as appropriate for its values. E.g.:

(And so on for each dummy variable)

Sheffield
Methods
Institute.

4. Once all of the dummy variables have been assigned Value Labels, you can then enter these into the Regression.

Note: when you do this, you enter all of the dummy variables EXCEPT ONE. This is important, because the one you leave out is the category that you compare the other dummy variables' coefficients to. E.g. if you enter the dummy variables for all the age group categories except the Under 18s, then the coefficients that they produce are telling you the difference between that age group compared to Under 18s.

(This YouTube video walks through this interpretation: https://www.youtube.com/watch?v=AJ88dheUieY).

2. Method 2

Go through the following menus on SPSS:

Transform > Recode into different variables

1. Put Age Group into the main box, then select 'Old and New Values'. To create the first dummy variable (for Under 18s), enter 1 into the value box on the left hand side, and 1 into the New Value box on the right hand side. Then, select 'All other values' on the left side, and enter '0' on the right side. It will end up looking like this:
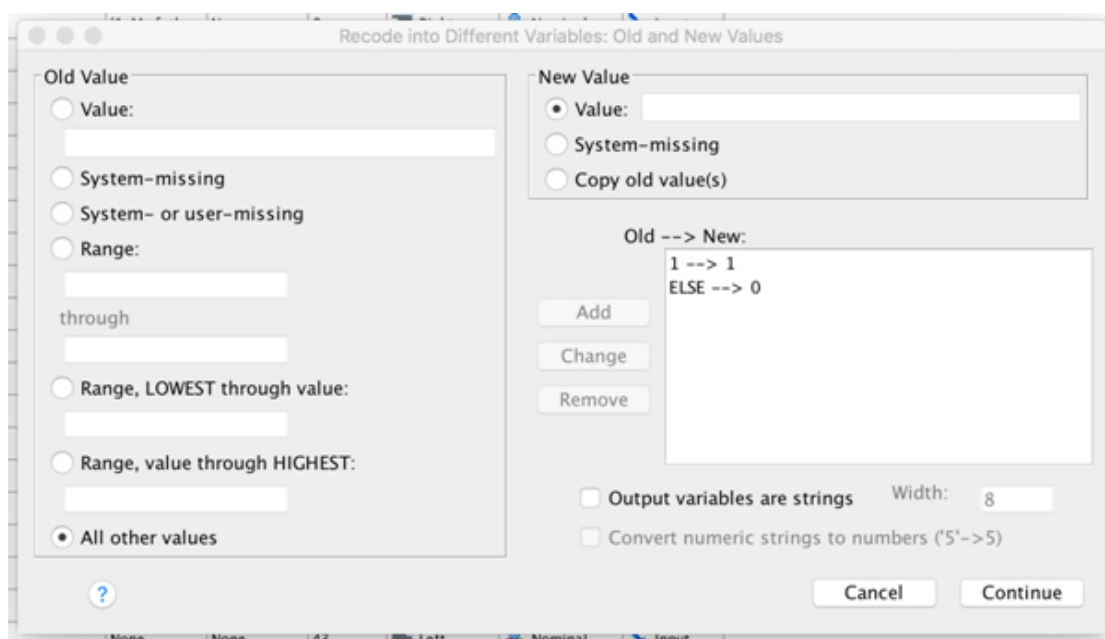
1. Select 'Continue', then on the window, enter a name and label for thenew dummy variable. Call it something intuitive, e.g. 'Under18' (to account for the fact that we have coded it so Under 18s = 1, and all other groups= 0).
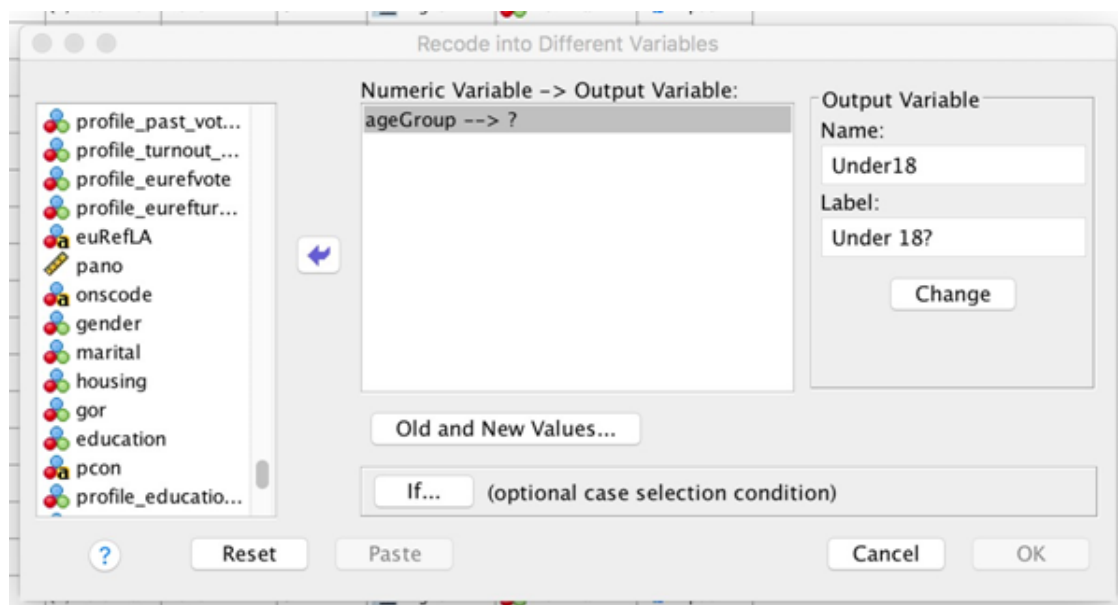

3. Find the new dummy variable at the bottom of the main variable list window, and create the value labels as appropriate for its values. E.g.:
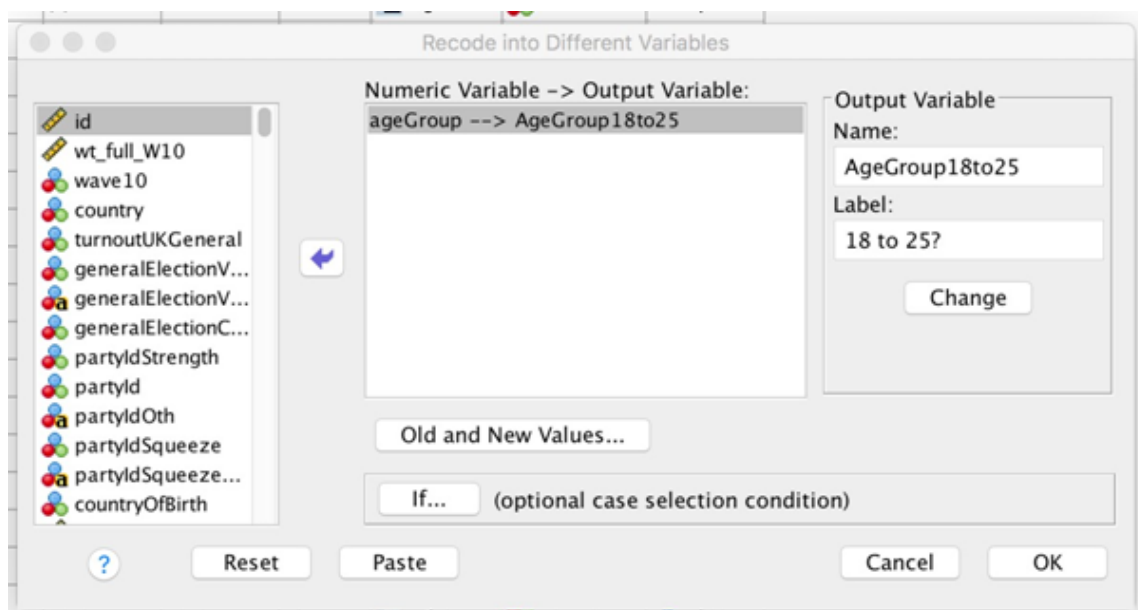

Sheffield Methods Institute.
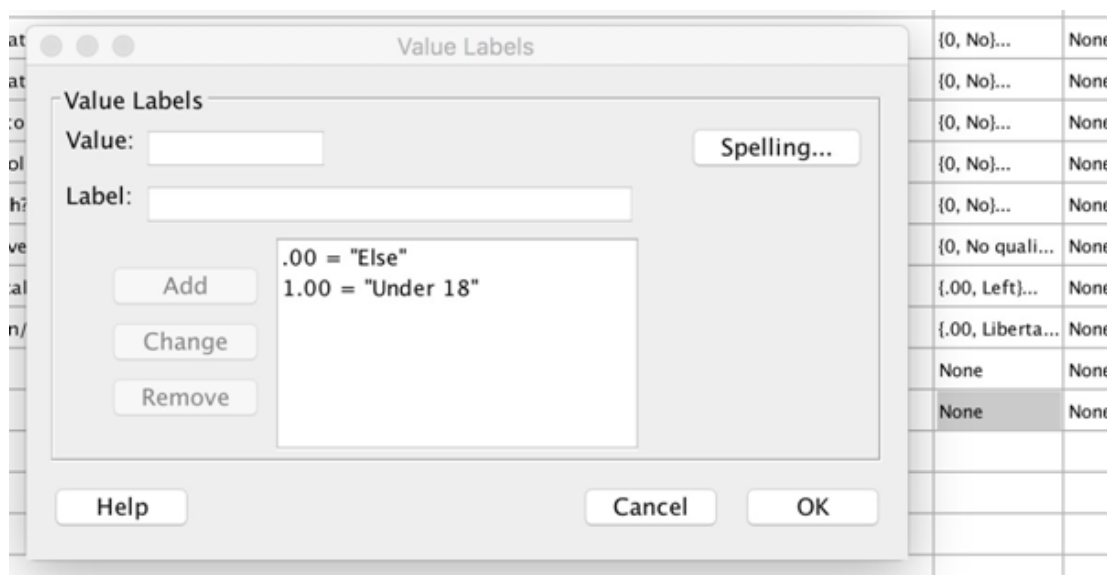
4. Repeat the above for every value of the original Age Group variable, but coded respectively for each category. E.g. 18-25 = 1, Else = 0; then 26-35 = 1, Else = 0; and so on.
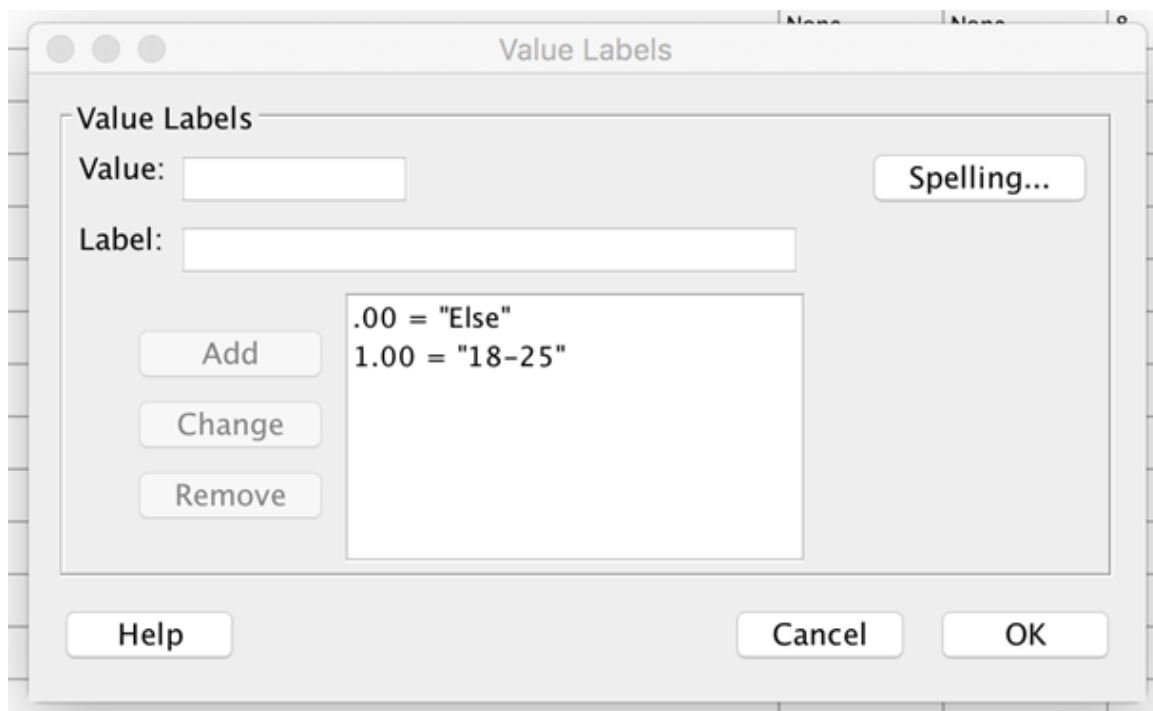
E.g. for 18-25:

Recode so that 18-25s are '1', and all other values are '0':
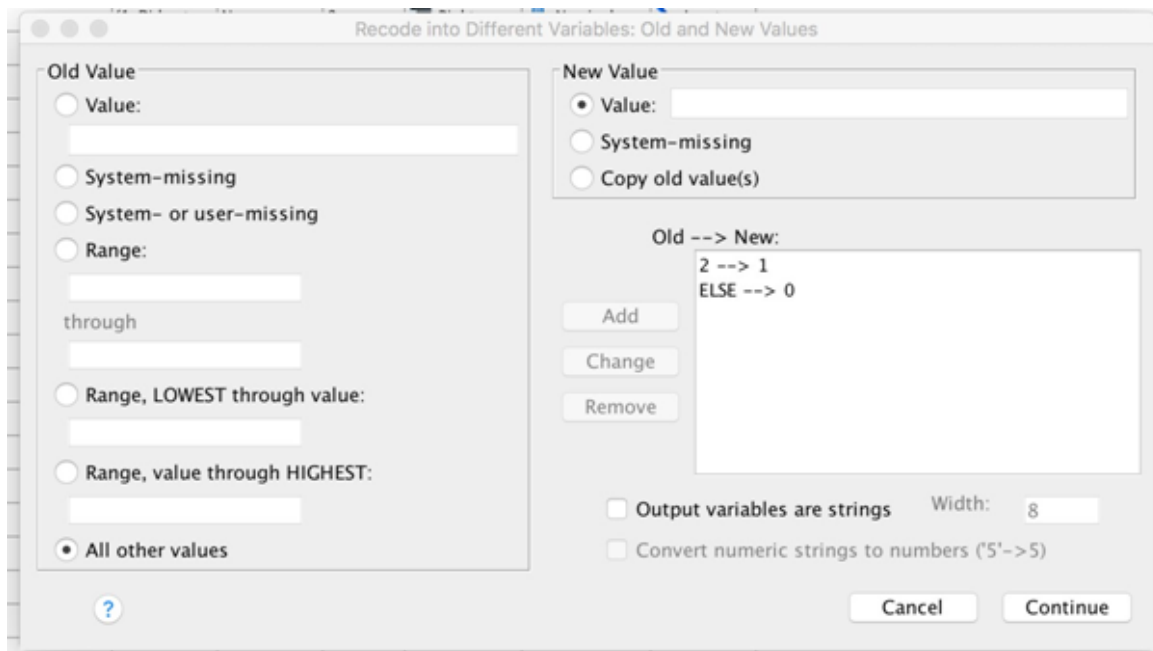
Set the name and label to something intuitive:

Sheffield
Methods
Institute.

Go to the new AgeGroup18to25 variable and change the corresponding Value Labels:

...and repeat for the remaining categories.

5. Once all of the categories have a respective dummy variable, you can then enter these into the Regression. Note: when you do this, you enter all of the variables EXCEPT ONE. This is important, because the one you leave out is the category that you compare the other dummy variables' coefficients to. E.g. if you enter the dummy variables for all the age group categories except the Under 18s, then the coefficients that they produce are telling you the difference between that age group compared to Under 18s.

(This YouTube video walks through this interpretation: https://www.youtube.com/watch?v=AJ88dheUieY in a *relatively* simple way).

Sheffield
Methods
Institute.