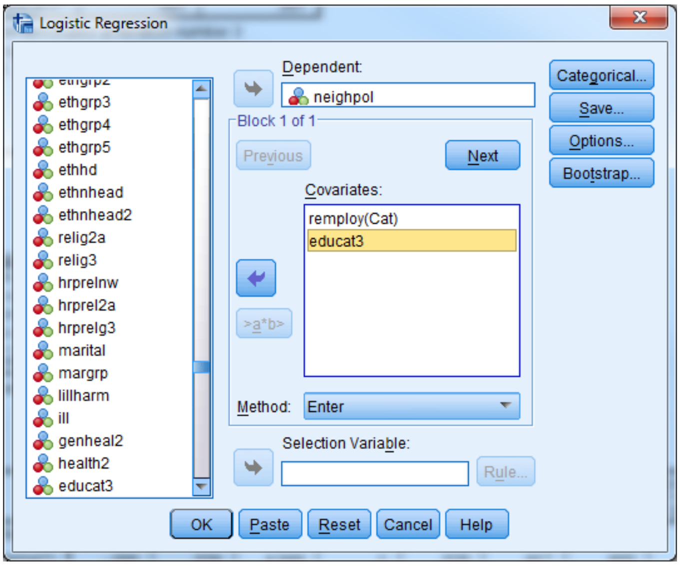
Multiple Logistic Regression – Two Categorical Independent Variables: Employment Status and Education Level

We've seen in the previous section that employment status has a significant influence on the odds of neighbourhood policing awareness. We might also be interested in the relationship between educational attainment and awareness. However, because respondent employment status may be informed by respondent educational qualification, it is best if we fit another logistic regression model, using **neighpol1** as the dependent variable and entering both **remploy** and **educat3**, a second categorical independent variable measuring respondent education level, as the independent variables at the same time.

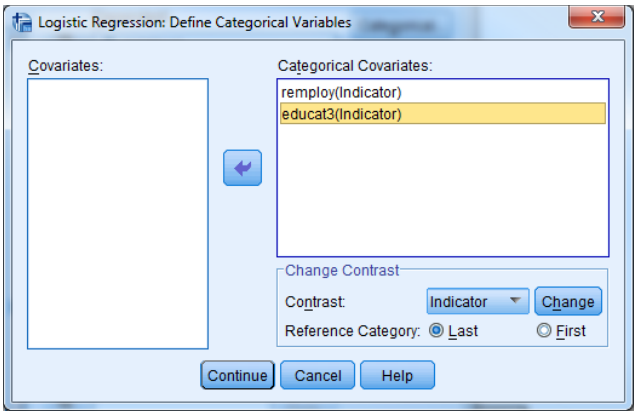
Select **Analyze**, **Regression**, and then **Binary Logistic**.

Make sure that **neighpol1** is in the **Dependent** box and **remploy(Cat)** is in the **Covariates** box. Find **educat3** in the variable list on the left and add it to the **Covariates** box.



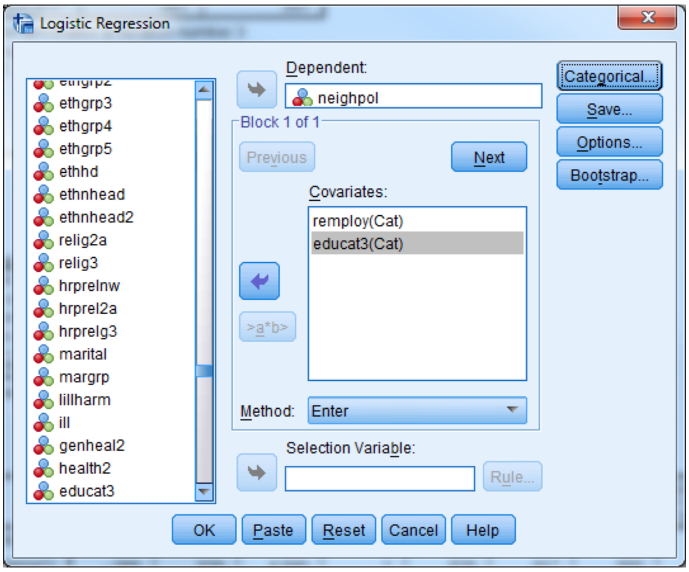
Because **educat3** is another categorical variable, we need to have SPSS create dummy variables. Click on **Categorical** in the upper right corner of the **Logistic Regression** dialogue box.

Move **educat3** to the **Categorical Covariates** box on the right.



Click **Continue**.

You should now see both **remploy(Cat)** and **educat3(Cat)** in the **Logistic Regression** dialogue box. Click **OK**.



Now we can look over the output of our new logistic regression model.

Case Processing Summary

Unweighted Cases[a]		N	Percent
	Included in Analysis	11251	24.4
Selected Cases	Missing Cases	34780	75.6
	Total	46031	100.0
Unselected Cases		0	.0
Total		46031	100.0

a. If weight is in effect, see classification table for the total number of cases.

Neighpol1 has again been coded with “Yes” as “0” and “No” as “1,” meaning that yet again, we will be predicting the odds of being *unaware* of neighbourhood policing.

Dependent Variable Encoding

Original Value	Internal Value
Yes	0
No	1

In the **Categorical Variables Codings** table below, you can see that our two categorical variables, **remploy** and **educat3**, have been recoded into dummy variables. In **educat3**, the baseline variable is **Other**. You can tell this because **Other** has not been coded as “1” in any of the **Parameter Code** columns. In **remploy**, **Economically Inactive** is again the baseline variable. We will need this information when we want to analyse the odds ratios.

		Frequency	Parameter coding			
			(1)	(2)	(3)	(4)
Respondent education (5 categories)	None	2698	1.000	.000	.000	.000
	O level/GCSE	2235	.000	1.000	.000	.000
	Apprenticeship or A/AS level	2014	.000	.000	1.000	.000
	Degree or diploma	3833	.000	.000	.000	1.000
	Other	471	.000	.000	.000	.000
Respondent employment status	Employed	6032	1.000	.000		
	Unemployed	390	.000	1.000		
	Economically inactive	4829	.000	.000		

Block 0

Again, we've left out the output tables for Block 0. We won't need them in our analysis.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	138.663	6	.000
	Block	138.663	6	.000
	Model	138.663	6	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	15314.646[a]	.012	.016

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	remploy			6.556	2	.038			
	remploy(1)	.087	.042	4.264	1	.039	1.091	1.004	1.185
	remploy(2)	.209	.109	3.691	1	.055	1.232	.996	1.524
	educat3			128.363	4	.000			
	educat3(1)	.234	.102	5.222	1	.022	1.264	1.034	1.544
	educat3(2)	-.144	.103	1.960	1	.161	.866	.707	1.059
	educat3(3)	-.200	.104	3.688	1	.055	.819	.667	1.004

	educat3(4)	-.385	.100	14.846	1	.000	.681	.560	.828
	Constant	.316	.096	10.859	1	.001	1.371		

a. Variable(s) entered on step 1: remploy, educat3.

Does this final model have a better fit than the previous two logistic regression models we created? Looking at the output in the **Model Summary** table, we can see that the Cox & Snell r^2 has risen from 0.001, its value in both of our previous logistic regressions, to 0.012 in this multiple logistic regression (meaning that 1.2% of the variation in neighbourhood policing awareness can be explained by this model). Therefore, this model has a better fit than our previous two simple logistic regression models.

Examining the **Block 1** output, we can see what (if anything) has changed in the predicted odds of being employed and unaware of neighbourhood policing, now that the model controls for education level. Remember that in this model, "Economically Inactive" was selected as our baseline comparison dummy variable and is called **remploy** in our model outputs. Because **remploy (1)** (with a p-value of .039) is a significant predictor of the odds of neighbourhood policing awareness, we can use the odds ratio information provided for us in the **Exp(B)** column to say that a respondent who is employed has odds of being unaware of neighbourhood policing that are 1.091 of the odds of someone who is economically inactive. We can compare that result to 0.917, the odds ratio for an employed respondent being unaware of neighbourhood policing we found in our previous logistic regression. Because the odds ratio for employed respondents is now greater than 1, this model predicts that the employed are now *less* likely than the economically inactive to know about neighbourhood policing. An odds ratio less than 1 means that the odds of an event occurring are lower in that category than the odds of the event occurring in the baseline comparison variable. An odds ratio more than 1 means that the odds of an event occurring are higher in that category than the odds of the event occurring in the baseline comparison variable.

This change in neighbourhood policing awareness in people who are employed is reflected in the **B** (or log-odds) column of the **Variables in the Equation** table. In our previous simple logistic regression of **neighpol1** and **remploy**, the **B** coefficient for **remploy(1)** (or employed respondents) was -0.086, meaning that the odds of an employed person being unaware of neighbourhood policing were lower than those of an economically inactive person. Now, however, the **B** coefficient for **remploy(1)** is 0.087, meaning that in this multiple logistic regression that controls for education level, the odds of an employed person being unaware of neighbourhood policing are now *higher* than those of an economically inactive person.

It can be quite confusing that the relationship between variables changes from a positive one (i.e. employed people are more likely to be aware) to a negative one (i.e. employed people are less likely to be aware) when we run a model twice with different variables. However, this just means that when we take out the effect of education, the relationship between neighbourhood policing and employment is the other way around. We are now calculating the relationship between employment and awareness of community policing while *controlling for* the effect of education.

Summary

Here, you've run a multiple logistic regression using neighpol1 as a binary categorical dependent variable and both educat3 and remploy as categorical independent variables. Using the output of this multiple logistic regression, you predicted the odds of a survey respondent being employed and unaware of neighbourhood policing, much like you did in the previous logistic regression including only neighpol1 and remploy. You were able determine how these predicted odds

changed after you added educat3 to the model and controlled for the influence of respondent education.

*****Note: as we are making changes to a dataset we'll continue using for the rest of this section, please make sure to save your changes before you close down SPSS. This will save you having to repeat sections you've already completed!**