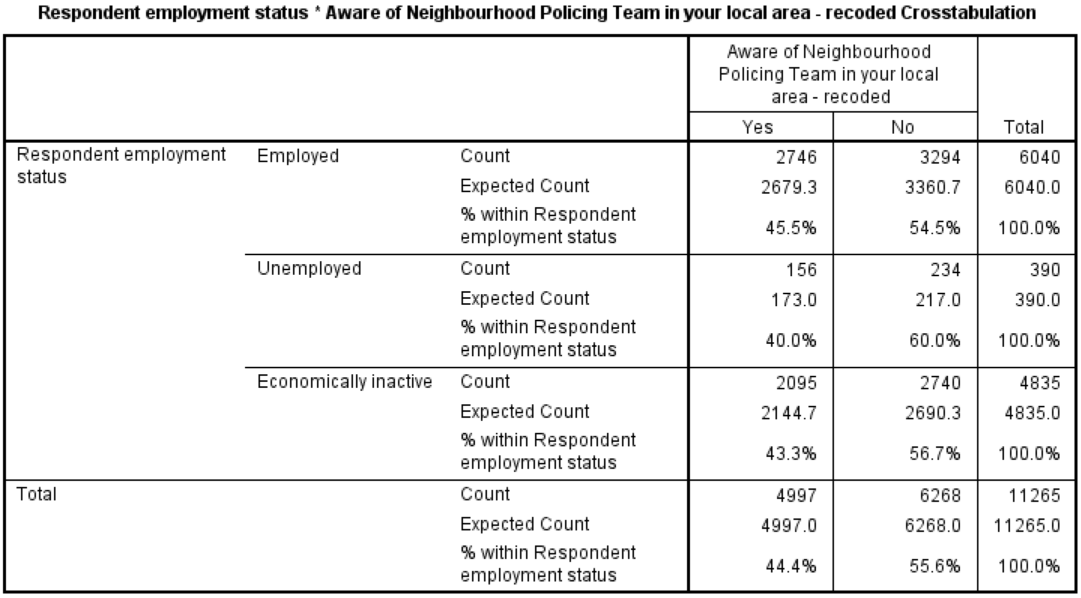
# *Simple Logistic Regression – One Categorical Independent Variable: Employment Status*

We've just run a simple logistic regression using **neighpol1** as a binary categorical dependent variable and **age** as a continuous independent variable. Suppose now we were interested to see if a respondent's employment status had any bearing on their awareness of neighbourhood policing. We may want to fit a logistic regression model using **neighpol1** as our dependent variable and **remploy**, respondent education level, as our independent variable to see if we can find a significant relationship between these two variables.

Just as we did at the beginning of our logistic regression investigation of **neighpol1** and **age**, we should run some exploratory analysis to determine if a relationship between these variables exists.

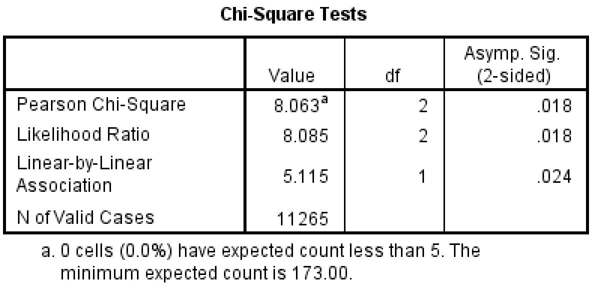
When our independent variable **age** was continuous, we used a t test to compare means. Now, our independent variable **remploy** is categorical, so we'll start by running crosstabulations. Select **Analyze**, **Descriptive Statistics**, and **Crosstabs**. Move **neighpol1** into the **Column(s)** box and **remploy** into the **Row(s)** box. Click the **Statistics** button and select **Chi-Square.** Click **Continue**. Because we are curious about **remploy**, we'd also like to see some row percentages. Click on **Cells**, and then under the **Percentages** header, select **Row**. Click **Continue**. Then, click **OK** to run the crosstabulation.

Your output should look like this:



**Respondent employment status * Aware of Neighbourhood Policing Team in your local area - recoded Crosstabulation**

| | | | Aware of Neighbourhood Policing Team in your local area - recoded | | |
|---|---|---|---|---|---|
| | | | Yes | No | Total |
| Respondent employment status | Employed | Count | 2746 | 3294 | 6040 |
| | | Expected Count | 2679.3 | 3360.7 | 6040.0 |
| | | % within Respondent employment status | 45.5% | 54.5% | 100.0% |
| | Unemployed | Count | 156 | 234 | 390 |
| | | Expected Count | 173.0 | 217.0 | 390.0 |
| | | % within Respondent employment status | 40.0% | 60.0% | 100.0% |
| | Economically inactive | Count | 2095 | 2740 | 4835 |
| | | Expected Count | 2144.7 | 2690.3 | 4835.0 |
| | | % within Respondent employment status | 43.3% | 56.7% | 100.0% |
| Total | | Count | 4997 | 6268 | 11265 |
| | | Expected Count | 4997.0 | 6268.0 | 11265.0 |
| | | % within Respondent employment status | 44.4% | 55.6% | 100.0% |



**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 8.063[a] | 2 | .018 |
| Likelihood Ratio | 8.085 | 2 | .018 |
| Linear-by-Linear Association | 5.115 | 1 | .024 |
| N of Valid Cases | 11265 | | |

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 173.00.

*How many unemployed people were aware of neighbourhood policing?*

*How many economically inactive people were not aware of neighbourhood policing?*

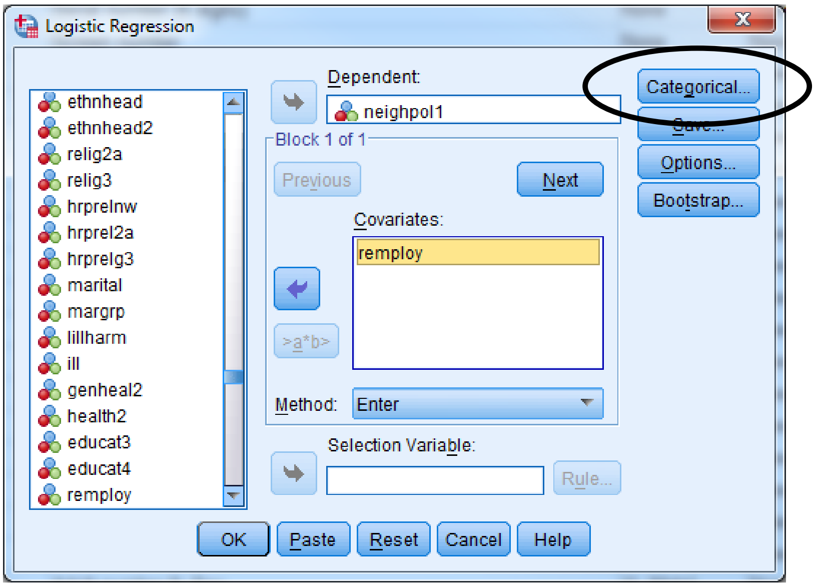
*What percentage of employed people were aware of neighbourhood policing?*

*Is there a significant relationship between* ***neighpol1*** *and* ***remploy****? How can you tell?*

Now we can fit our logistic regression model using **neighpol1** as the dependent variable and **remploy** as the independent variable.

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

Move **neighpol1** to the **Dependent** text box. Move **remploy** to the **Covariates** text box. Because **remploy** is a categorical variable, we have to tell SPSS to create dummy variables for each of the categories. (SPSS will do this for us in logistic regression – unlike in linear regression, when we had to create the dummies ourselves.) To tell SPSS that **remploy** is a categorical variable, click **Categorical** in the upper right corner of the **Logistic Regression** text box.

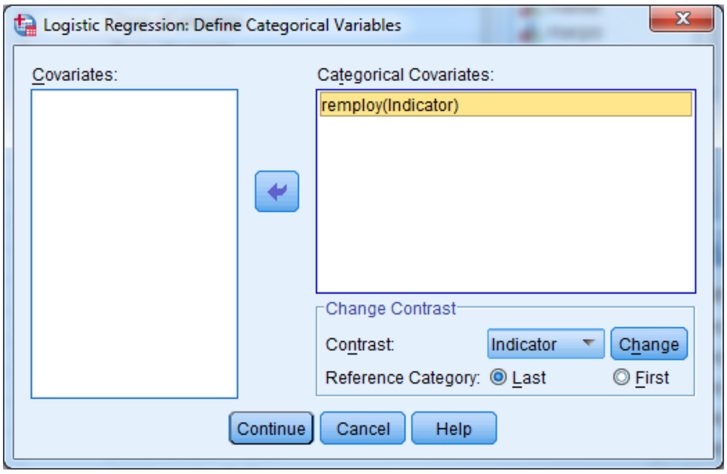


Dependent: neighpol1

Covariates: remploy

Method: Enter

Move **remploy** from the **Covariates** text box on the left to the **Categorical Covariates** text box on the right. Click **Continue**.



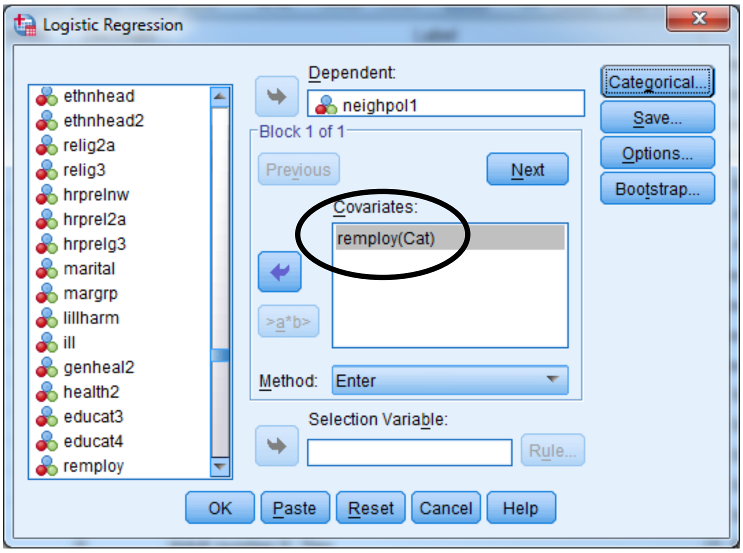
Logistic Regression: Define Categorical Variables

Covariates:

Categorical Covariates:
remploy(Indicator)

Change Contrast
Contrast: Indicator
Reference Category: ● Last ○ First

The original **Logistic Regression** dialogue box should now have **remploy(Cat)** in the **Covariates** text box. Click **OK**.



Logistic Regression

ethnhead
ethnhead2
relig2a
relig3
hrprelnw
hrprel2a
hrprelg3
marital
margrp
lillharm
ill
genheal2
health2
educat3
educat4
remploy

Dependent: neighpol1

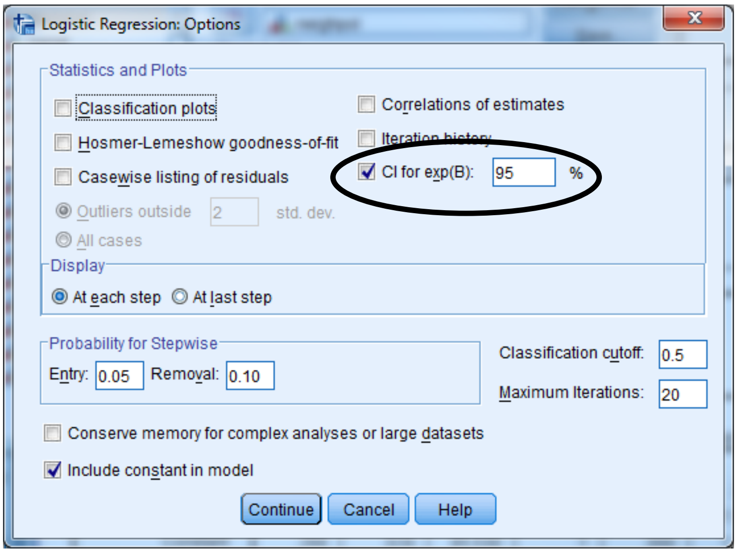
Block 1 of 1

Covariates:
remploy(Cat)

Method: Enter

Selection Variable:

We can also have SPSS calculate confidence intervals for **remploy** for us. In the **Logistic Regression** dialogue box you should have open, click **Options**. Under **Statistics and Plots**, select **CI for exp(B)**. This should already be set at 95%.



Logistic Regression: Options

Statistics and Plots
□ Classification plots
□ Hosmer-Lemeshow goodness-of-fit
□ Casewise listing of residuals
○ Outliers outside 2 std. dev.
○ All cases
□ Correlations of estimates
□ Iteration history
☑ CI for exp(B): 95 %

Display
● At each step ○ At last step

Probability for Stepwise
Entry: 0.05 Removal: 0.10

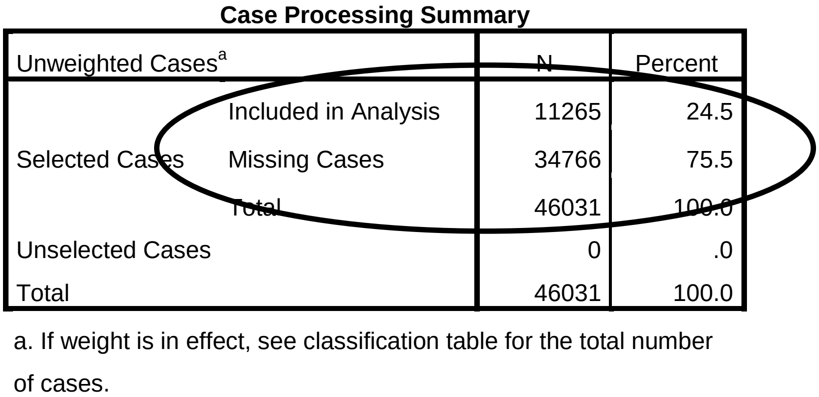
Classification cutoff: 0.5
Maximum Iterations: 20

□ Conserve memory for complex analyses or large datasets
☑ Include constant in model

Click **Continue** and then **OK** in the original **Logistic Regression** dialogue box.

Now we can examine the output.

You can see in the **Case Processing Summary** that again, we're only analysing about one quarter of the survey respondents, because our dependent variable **neighpol1** was only asked in Module A.

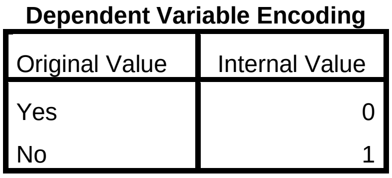


**Case Processing Summary**

| Unweighted Cases[a] | | N | Percent |
|---|---|---|---|
| Selected Cases | Included in Analysis | 11265 | 24.5 |
| | Missing Cases | 34766 | 75.5 |
| | 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.

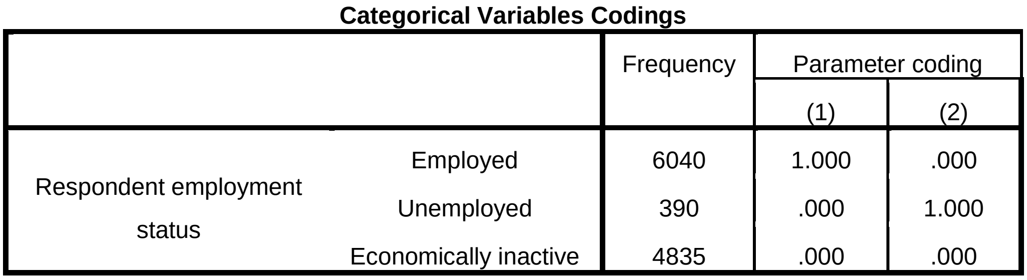
In the **Dependent Variable Encoding** table, you can see that awareness of neighbourhood policing ("Yes") is coded as 0 and being unaware of neighbourhood policing ("No") is coded as 1. Again, just like in the simple logistic regression we performed on the previous page, we will be predicting the odds of being unaware of neighbourhood policing in this logistic regression.



**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| Yes | 0 |
| No | 1 |

This **Categorical Variables Codings** table shows us the frequencies of respondent employment. In addition, it also tells us that the three categories of **remploy** have been recoded in our logistic regression as dummy variables. In logistic regression, just as in linear regression, we are comparing groups to each other. In order to make a comparison, one group has to be omitted from the comparison to serve as the baseline. In our logistic regression, "Economically inactive" has been selected as the baseline (or constant) dummy variable to which we will compare the predictions for "Employed" and "Unemployed." Therefore, "Economically inactive" won't be included in our model. (You can see that in the table below it isn't coded with a "1" in any case, because it is the baseline, comparison category and has not been added to the model.) You can change the category to be used as the baseline to either the first or last categories – this is done where you specify that the variable is categorical.



**Categorical Variables Codings**

| | | Frequency | Parameter coding | |
|---|---|---|---|---|
| | | | (1) | (2) |
| Respondent employment status | Employed | 6040 | 1.000 | .000 |
| | Unemployed | 390 | .000 | 1.000 |
| | Economically inactive | 4835 | .000 | .000 |

## Block 0

As we're not going to use any of the information provided for us in Block 0, the output has been left out of this worksheet. If you'd like to work through some of the information provided for you in Block 0, you can use the interpretation provided for the **neighpol1** and **age** logistic regression model we did on the previous page.

## Block 1: Method = Enter

Remember that the **Omibus Tests of Model Coefficients** output table shows the results of a chi-square test to determine whether or not employment has a significant influence on neighbourhood policing awareness. The Chi-square has produced a p-value of .018, making our employment status model significant at the 5% level.

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 8.085 | 2 | .018 |
| | Block | 8.085 | 2 | .018 |
| | Model | 8.085 | 2 | .018 |

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 15464.811[a] | .001 | .001 |

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 | | | 8.054 | 2 | .018 | | | |
| | remploy(1) | -.086 | .039 | 4.949 | 1 | .026 | .917 | .850 | .990 |
| | remploy(2) | .137 | .107 | 1.630 | 1 | .202 | 1.147 | .929 | 1.415 |
| | Constant | .268 | .029 | 85.530 | 1 | .000 | 1.308 | | |

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

Take a look at the **Variables in the Equation** output table above. Let's first look at the significance levels. **Remploy (1),** or "Employed," has a p-value of .026, making it significant at the p < .05 level. **Remploy (2),** or "Unemployed," on the other hand, has a p-value of .202, telling us that those who are in this category are no different in their awareness than the baseline category of economically inactive.

*If we were to fit this model again, and wanted to use **remploy**, we may be tempted to remove **remploy (2)** from the model, as it is not significant. However, we can't do this. Why?*

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 .026) is a significant predictor of the odds of neighbourhood policing, 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 0.917 of the odds of someone who is economically inactive. This means that the employed are more 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.

*A respondent who is unemployed has odds of being unaware of neighbourhood policing that are _____________ of the odds of a respondent who is economically inactive. This means that the unemployed are ________ likely than the economically inactive to know about neighbourhood policing.*

In addition, SPSS has calculated confidence intervals for us. Remember that confidence intervals allow us to extend out analyses from the sample in our data to the population as a whole. We can say, with 95% confidence, that for the entire population of England, employed people have odds of being unaware of neighbourhood policing that are 0.850 to 0.990 the odds of people who are economically inactive.

***Summary***

***First, you used a chi square test test to determine whether or not a statistically significant relationship existed between our categorical independent variable remploy and our categorical dependent variable neighpol1. Then, using simple logistic regression, you predicted the odds of a survey respondent being unaware of neighbourhood policing with regard to their employment status. Finally, using the odds ratios provided by SPSS in the Exp(B) column of the Variables in the Equation output table, you were able to interpret the odds of employed respondents being unaware of neighbourhood policing.***

*****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!**