

Simple Logistic Regression – One Continuous Independent Variable: Age

Our variable of interest, knowledge of neighbourhood policing, has two categories. As a result, we can model it using logistic regression, which requires a binary variable as the outcome. First, we can fit a logistic regression model with **neighpol1** as the dependent variable and **age** as the independent variable. However, before we begin, we should run exploratory bivariate analysis to get some answers about the relationship between **age** and **neighpol1**.

Because **age** is a continuous variable, we can run a two-sample t test to determine if there is a statistically significant difference in the mean ages for those who were aware of neighbourhood policing and those who were unaware. This, like all exploratory analysis, can help us determine whether or not it is worth fitting a logistic regression model for these variables. If the difference in mean age with respect to **neighpol1** is insignificant, running a logistic regression wouldn't be the best use of our time, as our results wouldn't be significant.

In addition to telling us if we're on the right track with our analysis, running this simple t test will also provide us with the frequencies of those who answered "Yes" and those who answered "No," which will allow us to calculate the percentage of the survey respondents who were unaware and those who were aware of neighbourhood policing.

Go to **Analyze, Compare Means**, and then **Independent-Samples T Test**.

Move **age** into the **Test Variables(s)** box and **neighpol1** into the **Grouping Variable** box.

Click on **Define Groups** and enter **1** in the **Group 1** box and **2** in the **Group 2** box (as 1=Yes and 2=No in **neighpol1** in our dataset).

Click **Continue**. Click **OK** to close the **Independent-Samples T Test** dialogue box.

Your SPSS output should look like this:

	Aware of Neighbourhood Policing Team in your local area - recoded	N	Mean	Std. Deviation	Std. Error Mean
Adult number 1 (respondent): Age	Yes	5001	52.03	17.750	.251
	No	6268	50.91	19.671	.248

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Adult number 1 (respondent): Age	Equal variances assumed	95.487	.000	3.148	11267	.002	1.125	.357	.424	1.825
	Equal variances not assumed			3.184	11098.025	.001	1.125	.353	.432	1.817

You can use the information in the t test output tables to answer the following questions:

What is the mean age for respondents who were aware of neighbourhood policing?

What is the mean age for respondents who were not aware of neighbourhood policing?

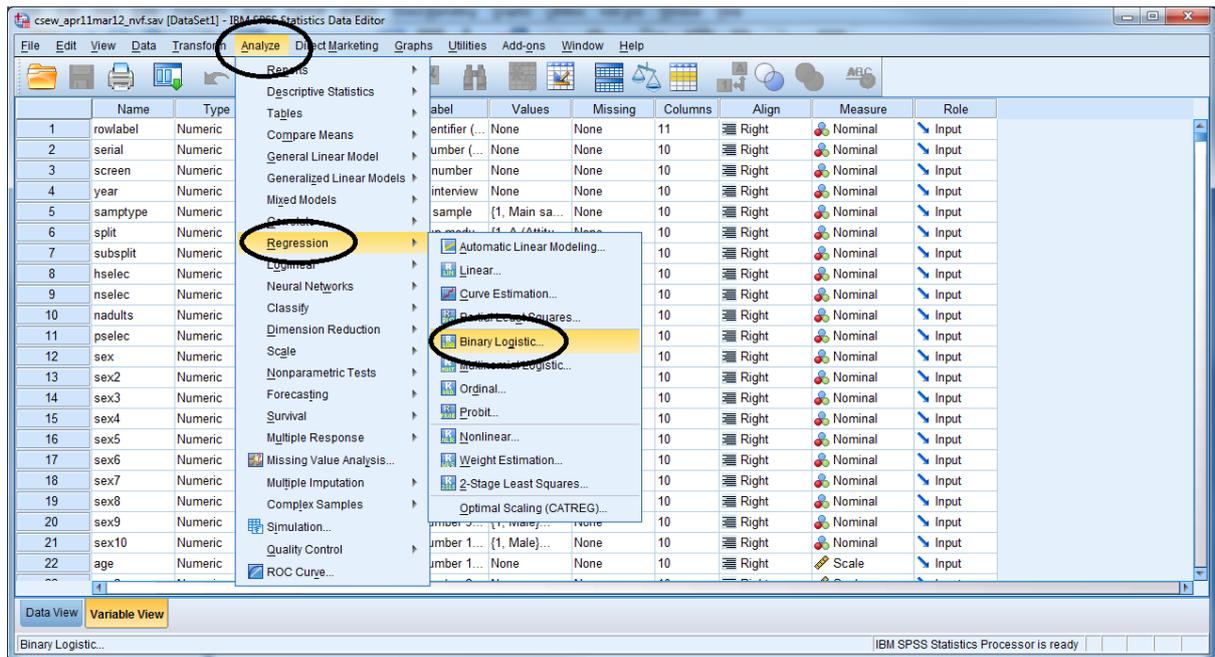
What percentage of respondents were not aware of neighbourhood policing?

*Take a look at the significance levels in the **Independent Sample Test output** box. Is there a*

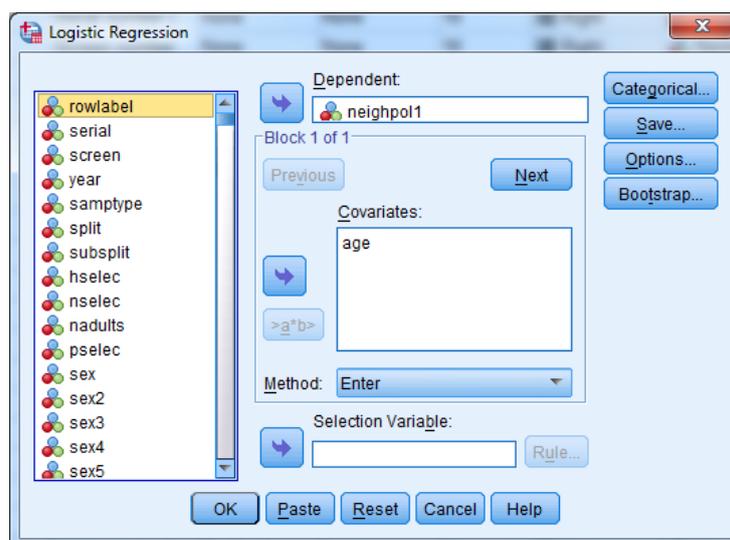
*significant difference between the mean ages for **neighpol1** respondents who were aware of neighbourhood policing and those who were not?*

Because we've just discovered that there is a significance difference between mean respondent ages, we know that there is a relationship between **neighpol1** and **age**. We can now continue on to fitting a logistic regression model to further explore this relationship.

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



Find our variable **neighpol1** from the variable list on the left of the dialogue box and move it the **Dependent** text box. Find the variable **age** and move it to the **Covariates** text box. Click **OK**.



You will now have several output tables open in the Output Viewer. Let's take a look at them.

The first table, called the **Case Processing Summary**, shows us that 11,269 cases were included in this logistic regression, and 34,762 are coded as Missing.

Unweighted Cases ^a		N	Percent
	Included in Analysis	11269	24.5
Selected Cases	Missing Cases	34762	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.

*Knowing what you do about **neighpol1**, why do you think there are so many cases coded as Missing?*

In our dataset, in the variable **neighpol1**, “Yes” is coded as “1” and “No” is coded as “2.” An answer of “No,” therefore, has been arbitrarily given a larger numeric code (as 2 is greater than 1). In logistic regression in SPSS, the variable category coded with the larger number (in this case, “No”) becomes the event for which our regression will predict odds. In other words, because the outcome “No” is coded as “2” in the dataset, the logistic regression will predict the odds of a respondent answering “No” to the question of whether or not they were aware of neighbourhood policing in their area. Because we are now predicting the odds of a respondent answering “No,” this answer becomes the success in our model, or “1.” An answer of “Yes” is a failure, or “0.” The **Dependent Variable Encoding** table below shows us that the Original Values of “Yes” and “No” (the answers to **neighpol1**) are coded as “0” and “1” in this analysis.

Original Value	Internal Value
Yes	0
No	1

You can change which category of your dependent variable is predicted by SPSS by simply recoding the values of the variable categories in the dataset. For example, if we wanted to predict the odds of a respondent answering “Yes” to **neighpol1**, we would recode “Yes” to “2” and “No” to “1,” therefore making “Yes” the category with the larger numeric code. (You can learn more about recoding the values in variables in the **Simple Linear Regression: One Binary Categorical Independent Variable** section of **Research Question 1**. The values for Male and Female in the variable **sex** are changed in exactly the same way you’d change the values of **neighpol1** if you wanted to change which category of your dependent variable is predicted by SPSS).

As you can see in your Output window, SPSS gives you many, many tables, most of which you won’t need to worry about. Here, we’ll highlight just the main points. The other tables become more useful as you do more in-depth analysis – we don’t have to worry about them now.

The output tables in **Block 0: Beginning Block** show neighbourhood policing awareness predictions before the addition of our independent variable **age** into the model. **Block 0** shows us the odds of a respondent being unaware of neighbourhood policing without the influence of age.

Block 0: Beginning Block

In the **Variables in the Equation** table, we can see the odds of being unaware of neighbourhood policing. These odds ratios are presented as the **Exp(B)** output in this table. Here we see that without the addition of age, the odds that a respondent would not be aware of neighbourhood policing are 1.253 greater than that a respondent would be aware of neighbourhood policing.

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	.226	.019	141.848	1	.000	1.253

In the **Variables not in the Equation** table, we see the predicted significance for the variable **age**. This table will predict age being significant and improving the fit of the model if $p < 0.05$. Before we move on to the logistic regression that includes **age**, take a look at the information provided for us here.

We can see that the predicted p-value for the **age** in this model is 0.002.

	Score	df	Sig.
Step 0 Variables age	9.901	1	.002
Overall Statistics	9.901	1	.002

*What do you think the addition of our explanatory variable will do to our model? Will it improve the fit of this logistic regression? Will we be better able to predict odds when we add **age** to our model? Why or why not?*

Let's move on to **Block 1: Method = Enter**, and see what changes (if any) our independent variable has on the predicted odds of a respondent not being aware of neighbourhood policing.

Block 1: Method = Enter

The **Omnibus Tests of Model Coefficients** table shows us the results of a chi-square test. This hypothesis test examines whether or not there is a statistically significant impact of age on the prediction of neighbourhood policing awareness. In order to accept that age has statistically significant influence on awareness of neighbourhood policing, the p-value must be less than 0.05.

	Chi-square	df	Sig.
Step 1 Step	9.903	1	.002

Block	9.903	1	.002
Model	9.903	1	.002

Based on the results above, do you think age has a statistically significant impact on the prediction of neighbourhood policing awareness?

We use the **Cox & Snell r^2** statistic calculated in the **Model Summary** output table below to gauge how much of the variation in neighbourhood policing awareness is explained by this model, and therefore how well our model fits our data. In this example, the r^2 is very low at 0.001. This shows that only 0.1% of the variation in neighbourhood policing awareness is explained by age. This indicates that other factors are affecting a respondent's awareness of neighbourhood policing.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	15469.494 ^a	.001	.001

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

Is our logistic regression model a good one? How can you tell?

In the Variables in the Equation output table, you can see that the p-value for **age** in this regression is $p=0.002$, meaning this variable does have a statistically significant influence on respondent awareness of neighbourhood policing.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a age	-.003	.001	9.895	1	.002	.997
Constant	.389	.055	49.548	1	.000	1.475

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

As you can see, actually running the logistic regression is not a problem – as long as you remember to put the binary outcome variable in the correct box in SPSS, it is difficult to go wrong! However, the interpretation of the results is a bit trickier – and the interpretation is what you are really interested in.

As noted above, we can easily see that the relationship between the age of the respondent and awareness of neighbourhood policing is significant (as shown by the p-value being less than 0.05). But what is this relationship? For every year older someone is, the log-odds of someone not being

aware decreases by 0.003 units. This does not mean much in terms of interpretation, which is unfortunate, because logistic regression actually conducts the analysis on the log odds. A better way of interpreting this is by using the odds ratio – which is included in the **Exp(B)** column, the final column of the table. In our example here, the odds ratio is 0.997. Because **age** is a continuous variable, we can say that with every one year increase in age, the odds of being unaware of neighbourhood policing are multiplied by 0.997. Because 0.997 is less than 1, any odds being multiplied by 0.997 will decrease. Therefore, as age increases, the odds of being unaware of neighbourhood policing decrease. For additional year, the odds reduce by 3%, meaning that people become more aware as they are getting older.

Logistic regression results can also be interpreted as probabilities – but we will not do this here.

This difference by **age** is subtly reflected in the mean ages for **neighpol1** respondents we calculated before running this logistic regression. The mean age for respondents who answered “Yes” was 52.03 years old. The mean age for respondents who answered “No” was 50.91.

Summary

First, you used an independent-sample (or two-sample) t test to determine whether or not a statistically significant relationship existed between our continuous independent variable age 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 age. You've learned that the results of a logistic regression are presented first as log-odds, but that those results often cause problems in interpretation. While the results of a logistic regression model can also be interpreted as probability, a favoured way of describing the results is to use the odds ratio provided by SPSS in the Exp(B) column of the Variables in the Equation output table.

*****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!