Simple Linear Regression – One Categorical Independent Variable with Several Categories

Does ethnicity influence police confidence score?

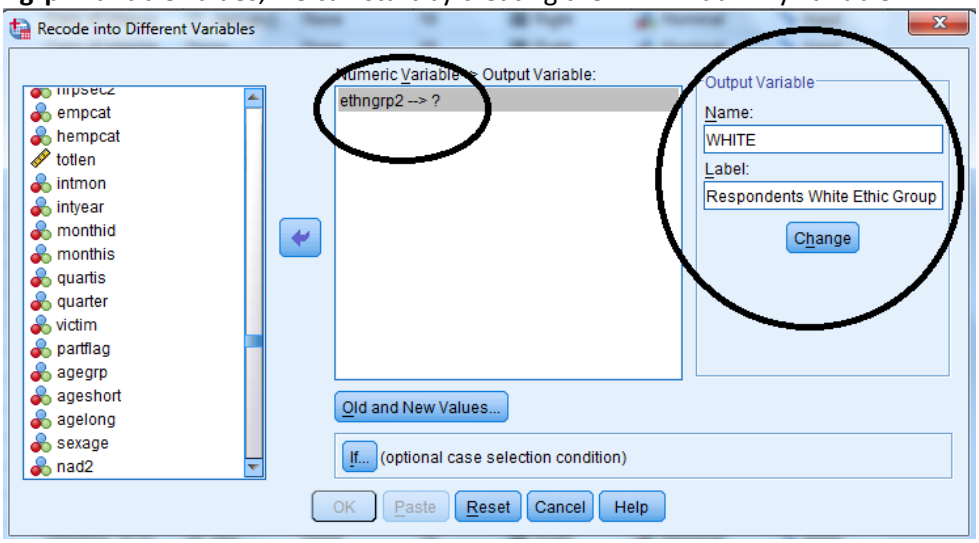
We've learned that variables with just two categories are called **binary** variables and are simple to use in regression. However, many variables have more than two categories. Suppose you want to see how the police confidence score of the respondents is related to their ethnic group. In this dataset, the variable **ethngrp2** has 5 categories (1= White, 2= Mixed, 3= Asian, 4= Black, and 5=Other). Much like with sex discussed in the previous page, the codes 1, 2, 3, 4, and 5 assigned to each ethnicity do not represent anything – the order is arbitrary. However, because linear regression assumes all independent variables are numerical, if we were to enter the variable **ethngrp2** into a linear regression model, the coded values of the five categories would be interpreted as numerical values of each category. As we found with **sex**, using **ethngrp2** in a linear regression without changing the coded values of the categories would give us results that would not make sense.

To avoid error, we're going to create dummy variables for **ethngrp2**. This is done in much the same way that we created the dummies for **sex**.

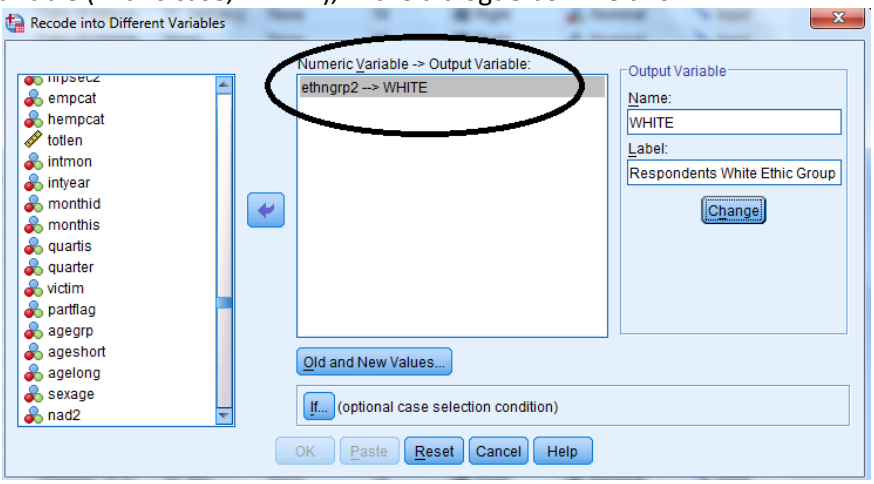
Dummy Variables

Remember that a dummy variable is a variable created to assign numerical value to levels of categorical variables. Each dummy variable represents one category of the explanatory variable and is coded with 1 if the case falls in that category and with 0 if not. For example, in the dummy variable for **Mixed** ethnicity, all cases in which the young person's ethnicity is **Mixed** will be coded as 1 and all other cases are coded as 0. In the dummy variable for **White**, all cases in which the young person's ethnicity is **White** will be coded as 1 and all other cases are coded as 0. The same will be done in the **Black**, in the **Asian**, and in the **Other** ethnicity dummy variables. This allows us to enter in the ethnicity values as numerical.

To begin creating our five dummy variables (one for each of the categories in **ethngrp2**), select **Transform** and then **Recode into Different Variables**. Find **ethngrp2** in the variable list on the left and move it to the **Numeric Variable → Output Variable** text box. Under the **Output Variable** header, type in the name and label of the first dummy variable you want to create. Because 1=White in the **ethngrp2** variable values, we can start by creating the **WHITE** dummy variable.

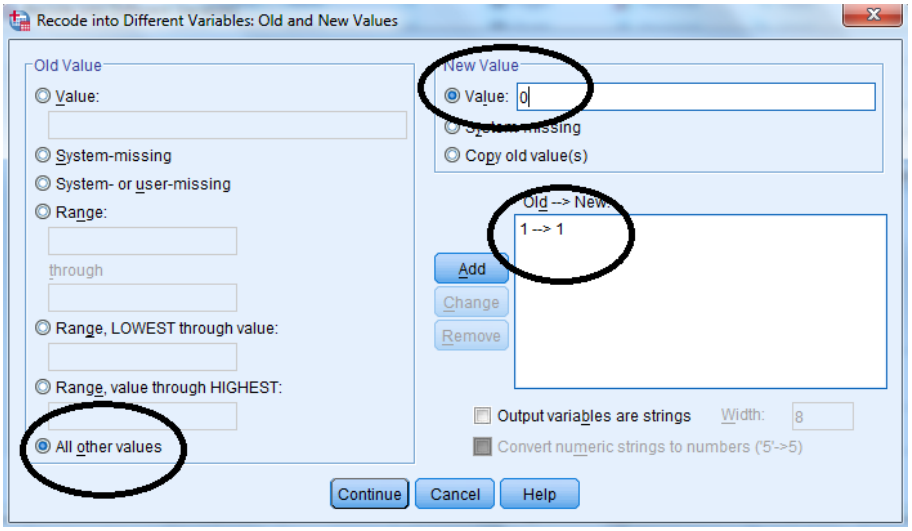


When you've finished entering the Output variable name and label, click **Change**. You should see your output variable (in this case, **WHITE**), in the dialogue box like this:

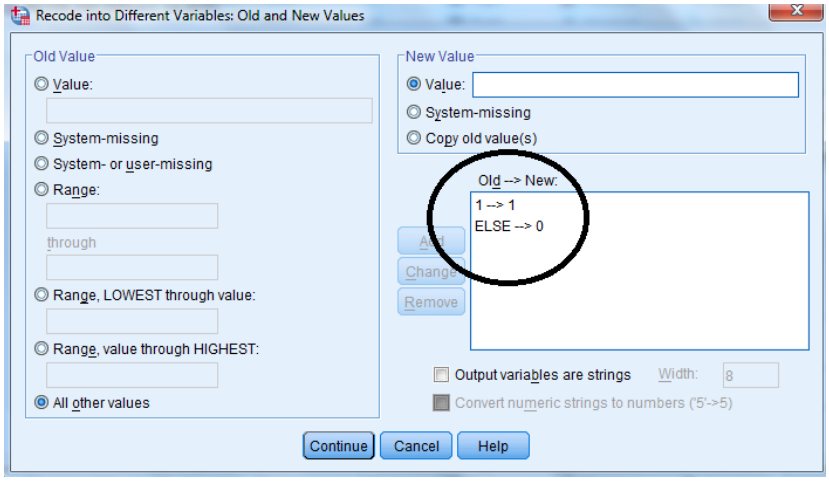


Next, click **Old and New Values**.

Because 1=White in **ethngrp2**, enter **1** under the **Old Value** header and **1** under the **New Value** header. Click **Add**. You should see **1→1** in the **Old → New** text box. Now, because in this dummy variable we want all the other values to be 0, click **All other values** under the **Old Value** header and enter **0** under the **New Value** header.



Click **Add**.



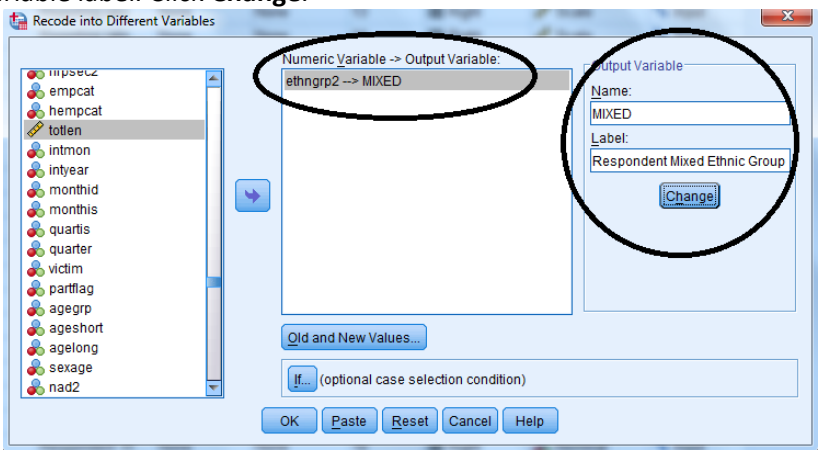
Click **Continue**, and then **OK** in the original **Recode into Different Variables** dialogue box.

To check that you have successfully created a dummy variable called **WHITE**, scroll down to the end of the variable list in **Variable View**. **WHITE** should be the last variable in the list, as it is the latest variable to be created.

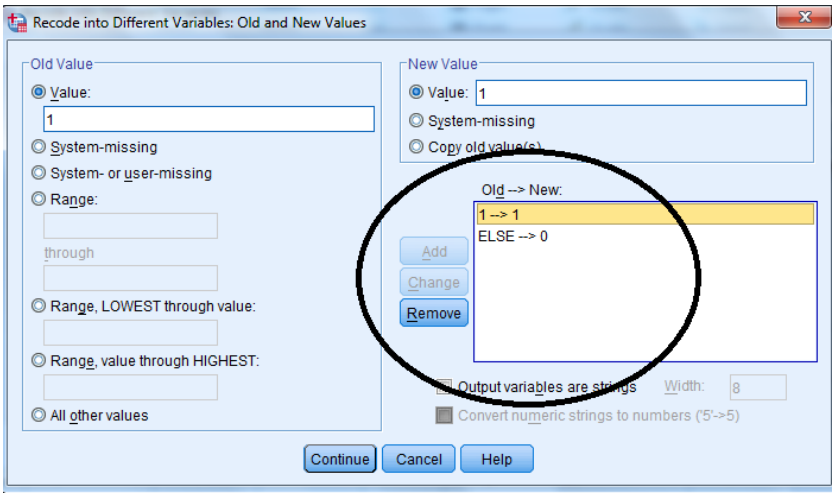
Let's recode one more dummy variable together. Go back to **Transform** and **Recode into Different Variables**.

All of our previous information will still be entered. Click to highlight **ethngrp2 → WHITE** and then click the blue arrow to remove this from the **Numeric Variable → Output Variable** text box.

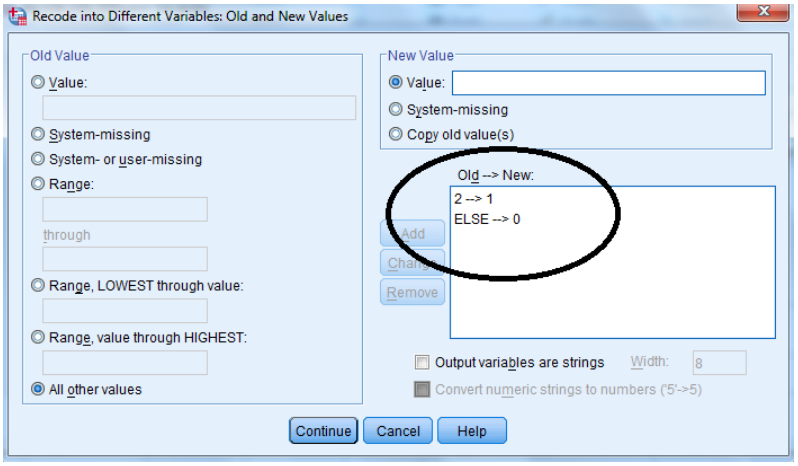
Find **ethngrp2** again and move it to the **Numeric Variable → Output Variable** box. Under the **Output Variable** header, enter in **MIXED** as the output variable name and **Respondent Mixed Ethnic Group** as the output variable label. Click **Change**.



Click **Old and New Values**. All of our previous work will be saved here and we no longer need it. Highlight the commands in the **Old → New** box and click **Remove**.



Because 2=Mixed, enter **2** under the **Old Value** header and **1** under the **New Value** header (as in this dummy variable, we want Mixed to have a value of 1). Click **Add**. Under the **Old Value** header, select **All other values** and enter **0** under the **New Value** header. Click **Add**.



Click **Continue** and then **OK** in the original **Recode into Different Variables** dialogue box.

Repeat steps 9-14 for the **ASIAN**, **BLACK**, and **OTHER** ethnicity categories. Remember when entering these following categories, you must use their corresponding values when recoding: 3= Asian, 4= Black, and 5=Other. For example, when recoding **ethngrp2** into the **ASIAN** dummy variable, you will use **3** as the **Old Value** and **1** as the **New Value** for **ASIAN**, while recoding all other values to **0**. You'll use **4** as the **Old Value** and **1** as the **New Value** for **BLACK**, while recoding all other values to **0**. And, finally, you'll use **5** as the **Old Value** and **1** as the **New Value** for **OTHER**, while recoding all other values to **0**.

When you are finished, you should have five new dummy variables at the end of your variable list in **Variable View**.

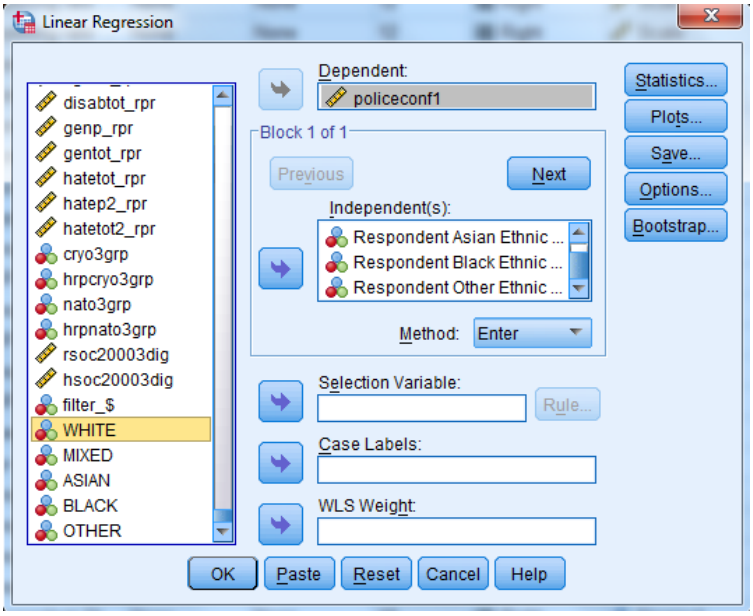
2327	WHITE	Numeric	8	2	Respondent W...
2328	MIXED	Numeric	8	2	Respondent Mi...
2329	ASIAN	Numeric	8	2	Respondent As...
2330	BLACK	Numeric	8	2	Respondent Bl...
2331	OTHER	Numeric	8	2	Respondent Ot...
2332					
2333					

Now we're ready to fit a linear regression model for this categorical data! This does seem very long winded, and it is, but this is the process you need to go through each time you have a categorical variable with more than two categories and are performing linear regression.

Before we begin: when we fit our model in SPSS, we need to select one dummy variable as the baseline category (the category against which we compare all the other categories). In this example, we will use **WHITE** as the baseline category. As the **WHITE** variable is now our baseline, we don't have to include it the linear regression model. We will, however, need to include all of the other dummy variables for ethnicity in the model. Basically, this means we are comparing all the ethnicities to the WHITE ethnicity.

To perform simple linear regression, select **Analyze**, **Regression**, and then **Linear…**

In the dialogue box that appears, move **policeconf1** to the **Dependent** box and **MIXED**, **ASIAN**, **BLACK**, and **OTHER** to the **Independent(s)** box.



Click **OK**.

You should get the following output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.050[a]	.003	.002	4.31089

a. Predictors: (Constant), Respondent Other Ethnic Group, Respondent Mixed Ethnic Group, Respondent Black Ethnic Group, Respondent Asian Ethnic Group

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2019.134	4	504.784	27.163	.000[b]
	Residual	791651.841	42599	18.584		
	Total	793670.976	42603			

a. Dependent Variable: I have confidence in the police

b. Predictors: (Constant), Respondent Other Ethnic Group, Respondent Mixed Ethnic Group, Respondent Black Ethnic Group, Respondent Asian Ethnic Group

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13.550	.022		622.371	.000
	Respondent Mixed Ethnic Group	1.067	.251	.021	4.258	.000

Respondent Asian Ethnic Group	-.839	.108	-.038	-7.771	.000
Respondent Black Ethnic Group	.517	.145	.017	3.571	.000
Respondent Other Ethnic Group	-.740	.188	-.019	-3.930	.000

a. Dependent Variable: I have confidence in the police

We can use our SPSS results to write out the fitted regression equation for this model and use it to predict values of **policeconf1** for given certain values of **ethngrp2**. In this case, **WHITE** is our baseline, and therefore the **Constant** coefficient value of 13.550 represents the predicted police confidence score of a respondent in that category. Remember that the dummy variables used in this regression model are coded as Mixed=1, Asian=1, Black=1, and Other=1. This means that we will enter in 1 as the value for X in the regression equation

Y = a + bX

The predicted scores are as follows:

policeconf1 = 13.550 + (1.067 x 1) = 14.617 (Mixed)
policeconf1 = 13.550 + (-0.839 x 1) = 12.711 (Asian)
policeconf1 = 13.550 + (0.517 x 1) = 14.067 (Black)
policeconf1 = 13.550 + (-0.740 x 1) = 12.810 (Other)

So, on average, Mixed respondents report a police confidence score that is 1.067 points ***higher*** than White respondents.

*On average, Asian respondents report a police confidence score that is how many points **lower** than White respondents?*

*On average, Black respondents report a police confidence score that is how many points **higher** than White respondents?*

*On average, respondents in the Other ethnic group category report a police confidence score that is how many points **lower** than White respondents?*

Remember we can use the r^2 statistic (which is calculated in the **Model summary** output table) to gauge how much variation in the dependent variable is explained by the independent variable. In our ethnicity example, the r^2 is low at .003, or 0.3%. Only 0.3% of the variation in police confidence score is explained by ethnicity.

Check our calculations again by using the **Compare Means** function, just as we did for sex on the previous page.

*Are the mean police confidence scores we calculated using linear regression the same as those SPSS calculated using the **Compare Means** function?*

Take a look at the significance levels for this ethnicity linear regression. Are these results statistically significant?

Summary

Here, we've used linear regression to determine the statistical significance of police confidence scores in people from various ethnic backgrounds. We've created dummy variables in order to use our ethnicity variable, a categorical variable with several categories, in this regression. We've learned that there is, in fact, a statistically significant relationship between police confidence score and ethnicity, and we've predicted police confidence scores using the ethnicity coefficients presented to us in the linear regression. Now, we may want to see how our predicted scores change if we run a linear regression using both sex and ethnicity as independent variables.

*****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!**