Simple Linear Regression – One Categorical Independent Variable with Several Categories

Does ethnicity influence total GCSE 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 total GCSE score of the respondents is related to their ethnic group. In the YCS dataset, the variable **s1eth2** has five categories (1=White, 2=Black, 3=Asian, 4=Mixed, and 5=Other). Much like with sex discussed above, 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 **s1eth2** 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 **s1gender**, using **s1eth2** 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 **s1eth2**. This is done in much the same way that we created the dummies for **sex**.

However, before we begin, let's check the value labels for **s1eth2**, to make sure this variable is ready for analysis. Find **s1eth2** in the **Variable View** window. Click on the **Values** cell in the **s1eth2** row to view the values assigned to our variable. You should see that in addition to the five ethnicity categories, there is a category called "Not answered," which is labelled as "-9.00." Because we are interested in the influence of ethnicity on GCSE score, these unanswered cases are effectively missing data, so we can code "-9.00" as missing. To do so, just click to open the **Missing** cell, select **Discrete** missing value, and enter "-9.00" into the text box. Click **OK**.

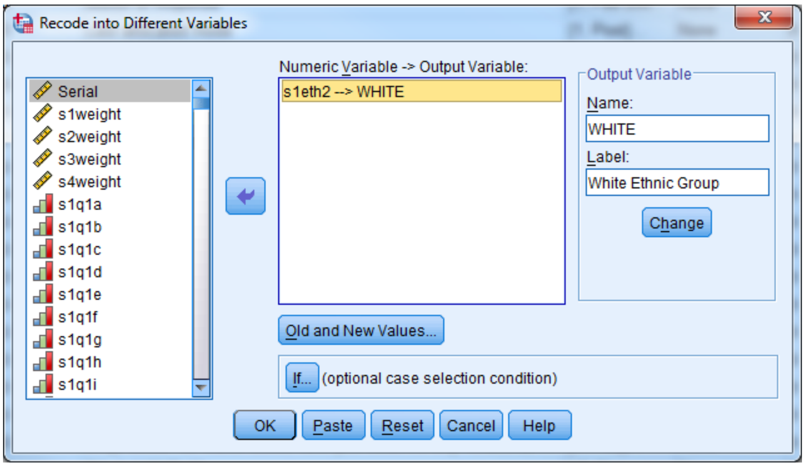
Now that we've edited **s1eth2**, we're ready to build our dummy variables and conduct our linear regression analysis!

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 1 if the case falls in that category and zero 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 **s1eth2**), select **Transform** and then **Recode into Different Variables**. Find **s1eth2** 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 **s1eth2** 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:



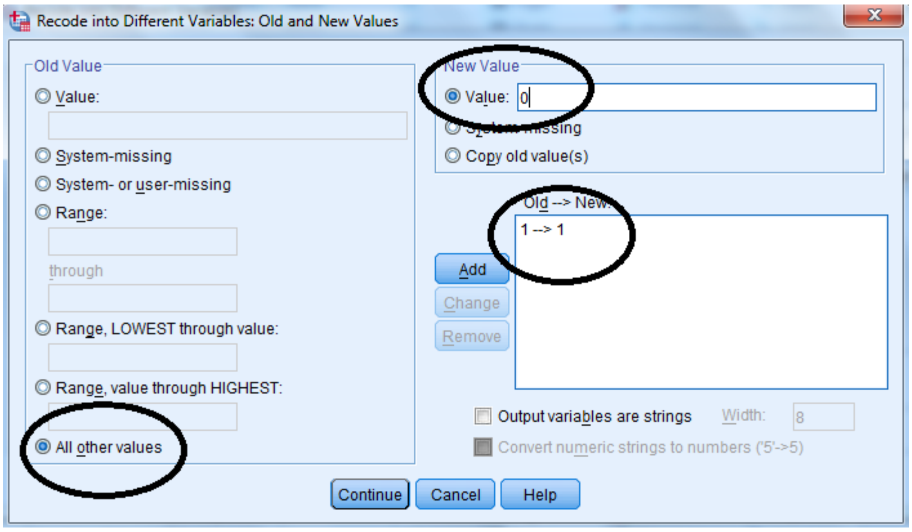
Numeric Variable -> Output Variable: s1eth2 --> WHITE

Name: WHITE

Label: White Ethnic Group

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

Because 1=White in **s1eth2**, 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.

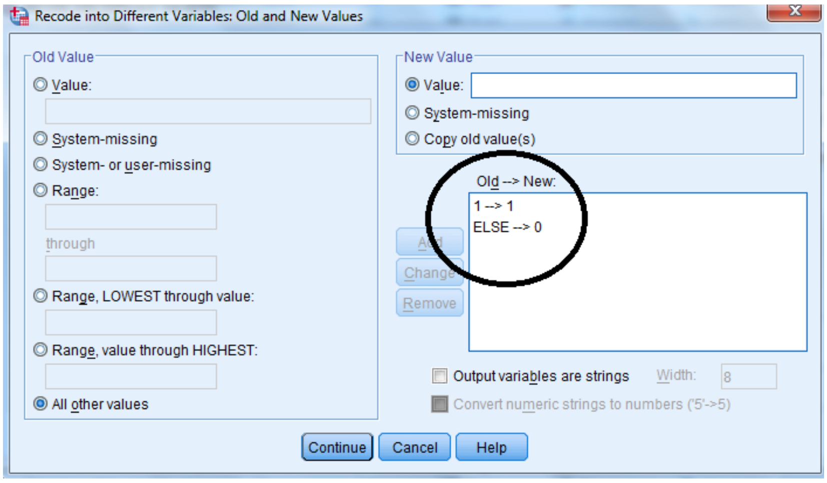

Recode into Different Variables: Old and New Values
Old Value
○ Value:
○ System-missing
○ System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
New Value
○ System-missing
○ Copy old value(s)
Add
Change
Remove
□ Output variables are strings
Width: 8
Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help

● All other values

New Value: ● Value: 0

Old --> New:
1 --> 1

Click **Add**.


Recode into Different Variables: Old and New Values
Old Value
○ Value:
○ System-missing
○ System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
New Value
● Value:
○ System-missing
○ Copy old value(s)
Add
Change
Remove
□ Output variables are strings
Width: 8
Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help

● All other values

Old --> New:
1 --> 1
ELSE --> 0

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.

Repeat the above steps for the **BLACK, ASIAN**, **MIXED**, and **OTHER** ethnicity categories. However, remember when entering these following categories, you must use their corresponding values when recoding: 2=Black, 3=Asian, 4=Mixed, and 5=Other. For example, when recoding **s1eth2** into the **BLACK** dummy variable, you'll use **2** as the **Old Value** and **1** as the **New Value** for **BLACK**, and recode all other values to **0**. When creating 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 **MIXED**, 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've finished, you should have five new dummy variables at the end of your variable list.

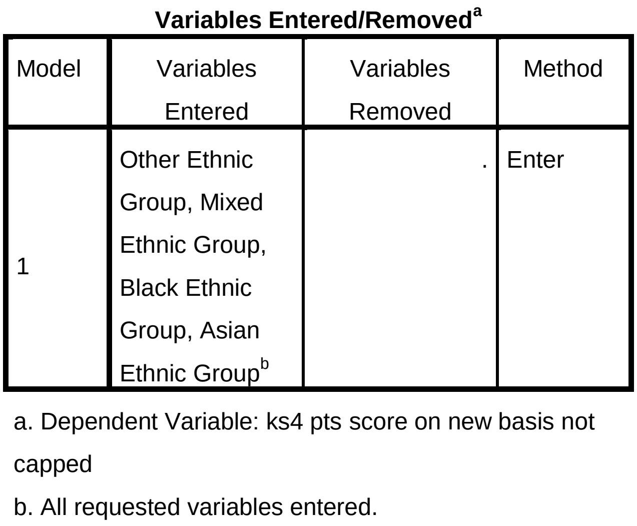
Now we're ready to fit a linear regression model for this categorical data! This does seem very long winded, and it is, but this is the process you need to go through each time you are conducting linear regression with a categorical variable with more than two categories.

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 that we are comparing all the ethnicities to the WHITE ethnicities.

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

In the dialogue box that appears, move **s1gcseptsnew** to the **Dependent** box and **MIXED**, **ASIAN**, **BLACK**, and **OTHER** to the **Independent(s)** box. Click **OK**.

You should get the following output:



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Other Ethnic Group, Mixed Ethnic Group, Black Ethnic Group, Asian Ethnic Group[b]	.	Enter

a. Dependent Variable: ks4 pts score on new basis not capped

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.066[a]	.004	.004	125.53087

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

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	943245.567	4	235811.392	14.965	.000[b]
1	Residual	216908848.523	13765	15757.998		
	Total	217852094.090	13769			

a. Dependent Variable: ks4 pts score on new basis not capped

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	394.550	1.156		341.322	.000
	Black Ethnic Group	-45.857	6.663	-.059	-6.883	.000
1	Asian Ethnic Group	8.000	3.799	.018	2.106	.035
	Mixed Ethnic Group	4.974	7.211	.006	.690	.490
	Other Ethnic Group	30.367	12.798	.020	2.373	.018

a. Dependent Variable: ks4 pts score on new basis not capped

Take a look at the p-values calculated for our ethnicity dummy variables. Which of them are statistically significant at the $p < 0.05$ level?

All of the dummy variables are statistically significant except for the Mixed ethnic group variable. The p-value in that case is 0.490, much higher than the $p < 0.05$ threshold. This means that any predicted values we may be able to calculate for the Mixed ethnic group won't be significant.

We can use our SPSS results to write out the fitted regression equation for this model and use it to predict values of **s1gcseptsnew** for given certain values of **s1eth2**. 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 when calculating each of their predicted values, we will enter in 1 as the value for X in the regression equation

$$\mathbf{Y = a + bX}$$

The predicted scores are as follows:

s1gcseptsnew = 394.550 + (-45.857 x 1) = 340.693 (Black Ethnic Group)
s1gcseptsnew = 394.550 + (8.000 x 1) = 402.55 (Asian Ethnic Group)
s1gcseptsnew = 394.550 + (4.974 x 1) = 399.524 (Mixed Ethnic Group)
s1gcseptsnew = 394.550 + (30.367 x 1) = 424.917 (Other Ethnic Group)

Looking at the predicted total GCSE scores above, we can see that on average, Asian young people earn GCSE scores that are 8 points higher than White students (as WHITE is our baseline category).

On average, Black students earn GCSE scores that are how many points ***lower*** *than White respondents?*

On average, respondents in the Other ethnic group category earn GCSE scores that are how many points ***higher*** *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 .004, or 0.4%. Only 0.4% of the variation in total GCSE score is explained by ethnicity.

Summary

Here, we've used linear regression to determine the statistical significance of GCSE 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 GCSE score and ethnicity, and we've predicted GCSE 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!**