Simple Linear Regression – One Binary Categorical Independent Variable

Does sex influence mean GCSE score?

In order to answer the question posed above, we want to run a linear regression of **s1gcseptsnew** against **s1gender**, which is a binary categorical variable with two possible values. (If you check the **Values** cell in the **s1gender** row in **Variable View**, you can see that the categories in this sex variable are labelled as 1= Male and 2= Female). However, before we begin our linear regression, we need to recode the values of **Male** and **Female**. Why do we need to do this?

The codes 1 and 2 are assigned to each gender simply to represent which place each category occupies in the variable **s1gender**. However, linear regression assumes that the numerical amounts in all independent, or explanatory, variables are meaningful data points. So, if we were to enter the variable **s1gender** into a linear regression model, the coded values of the two gender categories would be interpreted as the numerical values of each category. This would provide us with results that would not make sense, because for example, the sex Female does not have a value of 2.

We can avoid this error in analysis by creating **dummy variables**.

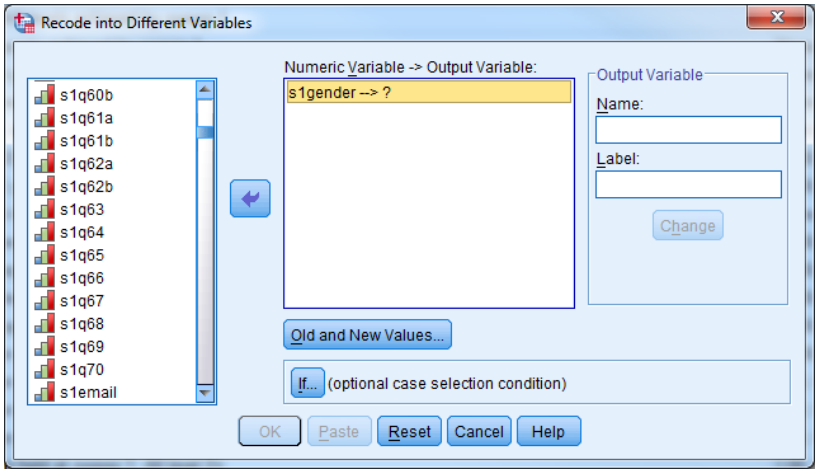
Dummy Variables

A dummy variable is a variable created to assign functional numerical values 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 a dummy variable for Female, all cases in which the respondent is female are coded as 1 and all other cases, in which the respondent is Male, are coded as 0. This allows us to enter in the sex values as numerical. (These numbers are just indicators.)

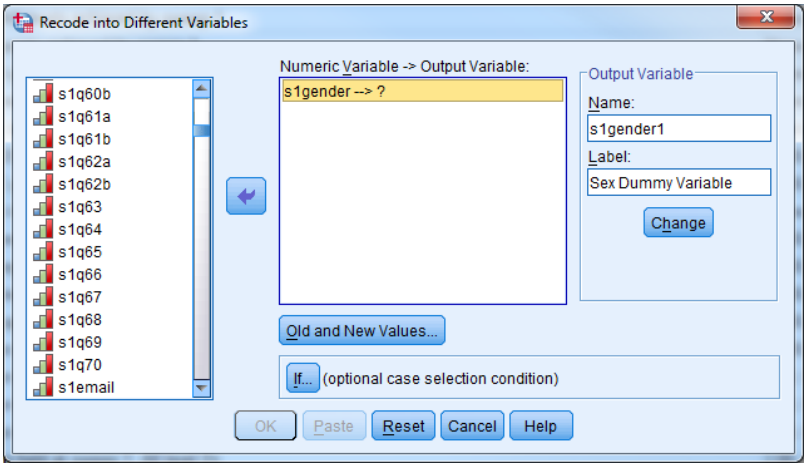
Because our sex variable only has two categories, turning it into a dummy variable could be as simple as recoding the values of Male and Female from 1=Male and 2=Female to 0=Male and 1=Female. (We will see later that creating dummy variables for categorical variables with multiple levels takes just a little more work.) However, it's good practice to create a new variable altogether when you are creating dummy variables. This way, if you make an error while building the dummy variables, you haven't altered your original variable and can always start again.

To begin, select **Transform** and **Recode into Different Variables**.

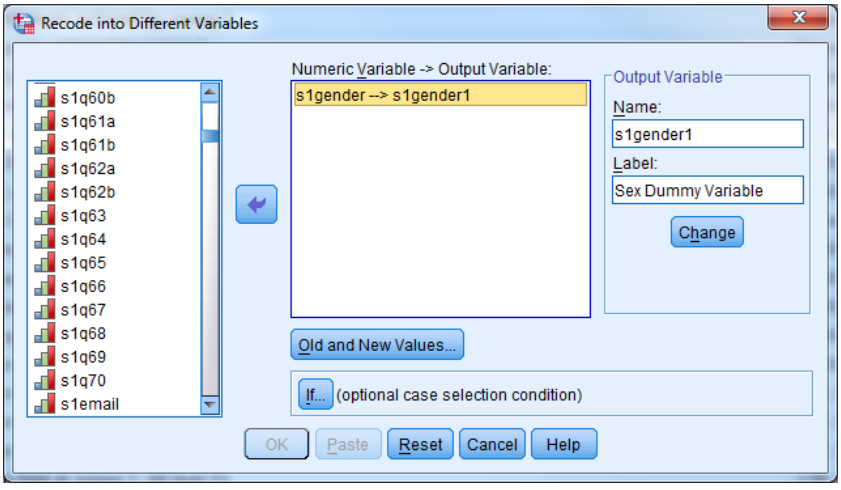
Find our variable **s1gender** in the variable list on the left and move it to the **Numeric Variables** text box.



Next, under the **Output Variable** header on the left, enter in the name and label for the new sex variable we're creating. We've chosen to call this new variable **s1gender1** and label it **Sex Dummy Variable**.



Click **Change**, to move your new output variable into the **Numeric Variable -> Output Variable** text box in the centre of the dialogue box.



Then, select **Old and New Values**.

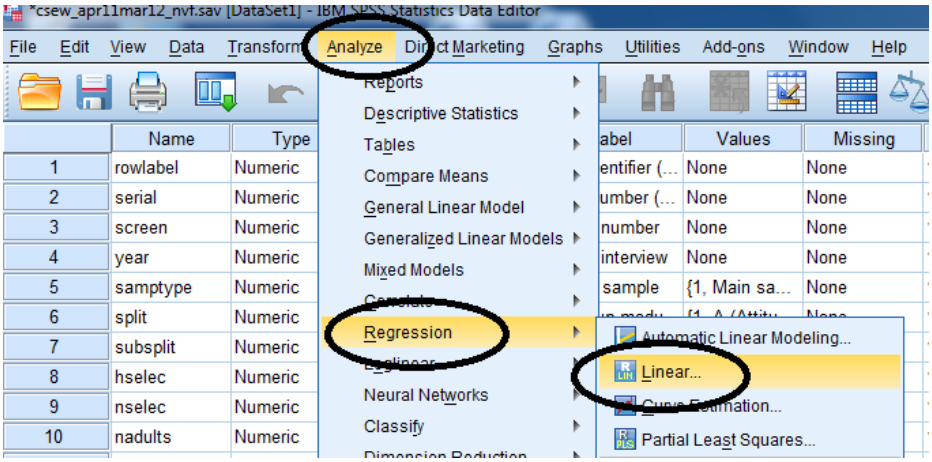
Enter **1** under the **Old Value** header and **0** under the **New Value** header. Click **Add**. You should see **1 → 0** in the **Old → New** text box. Now enter **2** under the **Old Value** header and **1** under the **New Value** header.

Click **Add**, and then **Continue**.

Finally, click **OK** in the original **Recode into Same Variables** dialogue box. You have successfully recoded the values for Male and Female in the **s1gender** variable!

Now, let's run our first linear regression, exploring the relationship between **s1gcseptsew** and **s1gender**.

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



Find **s1gcseptsnew** in the variable list on the left and move it to the **Dependent** box at the top of the dialogue box. Find **s1gender1** (our dummy variable) in the variable list and move it to the **Independent(s)** box in the centre of the dialogue box. Click **OK**.

You should have the following output in the SPSS Output window:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	S1 Gender[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	.092[a]	.009	.008	125.25303

a. Predictors: (Constant), S1 Gender

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1855299.614	1	1855299.614	118.260	.000[b]
	Residual	215996794.476	13768	15688.320		
	Total	217852094.090	13769			

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

b. Predictors: (Constant), S1 Gender

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	381.233	1.611		236.627	.000
	S1 Gender	23.390	2.151	.092	10.875	.000

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

We can see in the **Coefficients** table above that the relationship between sex and GCSE score is significant, as the p-value is 0.000, well below the $p < 0.05$ threshold. Now, we can use the SPSS results above to write out a fitted regression equation for this model and use it to predict values of GCSE scores for given certain values of **s1gender**. We can calculate the mean GCSE score for boys and girls using the following regression equation:

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

where **Y** is equal to our dependent variable and **X** is equal to our independent variable.
Into this equation, we will substitute **a** and **b** with the statistics provided in the **Coefficients** SPSS output table, with **a** being the **constant coefficient** and **b** being the **coefficient associated with s1gender** (our explanatory variable).

In this example, our equation should look like this:

s1gcseptsnew = 381.233 + (23.390 x s1gender)

Since **s1gender** takes on the value of 1 for female students and 0 for male students, the predicted scores are as follows:

s1gcseptsnew = 381.233 + (23.390 x 1) = 404.623 (Females)
s1gcseptsnew = 381.233 + (23.390 x 0) = 381.233 (Males)

So, according to our linear regression, on average, female students earned higher total GCSE scores than male students.

Note: If you've been following through the previous pages of this section, these numbers may look familiar to you. (Hint: they are the means we calculated while running mean comparisons in bivariate analysis. Calculating the mean scores using simple linear regression, with just one independent variable, was effectively the same function as comparing the means. As we'll see later,

multiple linear regression allows the means of many variables to be considered and compared at the same time, while reporting on the significance of the differences.)

Rather than just accepting these results, we can now gauge how much of the variation in **s1gcseptsnew** is explained by **s1gender**. To do this, we can simply use the r^2 statistic which is already calculated for you in the **Model summary** output table above. In this example, the r^2 is very low at 0.009. This shows that only 0.9% of the variation in GCSE is explained by sex (0.009 X 100 gives us the percentage). This suggests that other factors are affecting a young person's total GCSE score.

Determining the Significance of the Independent Variable

What is the significance of sex as a predictor of total GCSE score?

Our sample of data has shown us that, on average, female students earn total GCSE scores that are 23.390 points higher than male students. We want to know if this is a statistically significant effect in the population from which the sample was taken. To do this, we carry out a hypothesis test to determine whether or not **b** (the coefficient for females) is different from zero in the population. If the coefficient could be zero, there is no statistically significant difference between males and females.

SPSS calculates a t statistic and a corresponding p-value for each of the coefficients in the model. These can be seen in the **Coefficients** output table. A t statistic is a measure of how likely it is that the coefficient is not equal to zero. It is calculated by dividing the **coefficient** by the **standard error**. If the standard error is small relative to the coefficient (making the t statistic relatively large), the coefficient is likely different from zero in the population.

The p-value is in the column labelled **Sig.** As in all hypothesis tests, if the p-value is less than 0.05, then the variable is significant at the 5% level. That is, we have evidence to reject the null and conclude that **b** is different from zero.

In this example, we can see in the **Coefficients** output table that t = 10.875 with a corresponding p-value of 0.000. This means that the chances of the difference between males and females that we have calculated is actually happening due to luck is very small indeed. Therefore, we have evidence to reject the null and conclude that **s1gender** is a significant predictor of **s1gcseptsnew** in the population.

Summary

You've just used linear regression to study the relationship between our continuous dependent variable s1gcseptsnew and gender1, a categorical independent variable with just two categories. Using linear regression, you were able to predict GCSE scores for men and women. What if you wanted to fit a linear regression model using police confidence score and something like ethnicity, a categorical independent variable with more than two categories? The next page will take you through how to run a simple linear regression with a categorical independent variable with several categories.

*****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!**