6. Categorical Independent Variable

A categorical variable is a variable for which the possible outcomes are nameable characteristics, groups or treatments. Some examples are sex (male or female), highest educational degree attained (secondary school, university undergraduate, university postgraduate), blood pressure medication used (drug 1, drug 2, drug 3), etc.

We use **Dummy** (**Indicator**) variables to incorporate a categorical independent variable into a regression model. A dummy variable equals 1 when an observation is in a particular group and equals 0 when an observation is not in that group. An interaction between an indicator variable and a quantitative variable exists if the slope between the response and the quantitative variable depends upon the specific value present for the indicator variable.

A Simple Case

A simple regression model was used to show the relationship between the annual cardiac mortality rate per 100,000 for males averaged over the years 1958-1964 and the calcium concentration (in parts per million) in the drinking water supply for 61 large towns in England and Wales. (The higher the calcium content, the harder the water.)

$$y = \beta_1 + \beta_2 x + \varepsilon$$

Where y is the mortality rate and x is the calcium concentration.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.655[a]	.429	.419	143.029

a. Predictors: (Constant), calcium (ppm)

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	1676.356	29.298		57.217	.000
calcium (ppm)	-3.226	.485	-.655	-6.656	.000

a. Dependent Variable: mortality per 100,000

As we know, the water is generally harder in the South than the North.

Report

calcium (ppm)

North_South	Mean	N	Std. Deviation
North	30.40	35	26.134
South	69.77	26	40.361
Total	47.18	61	38.094

Perhaps the model would produce a clearer picture of the pattern of variation in male mortality, if the North/South difference could be built into the model. We can do this by creating a new dummy variable, *D*, with:

$$D = \begin{cases} 0, \text{ for town in the North} \\ 1, \text{ for town in the South} \end{cases}$$

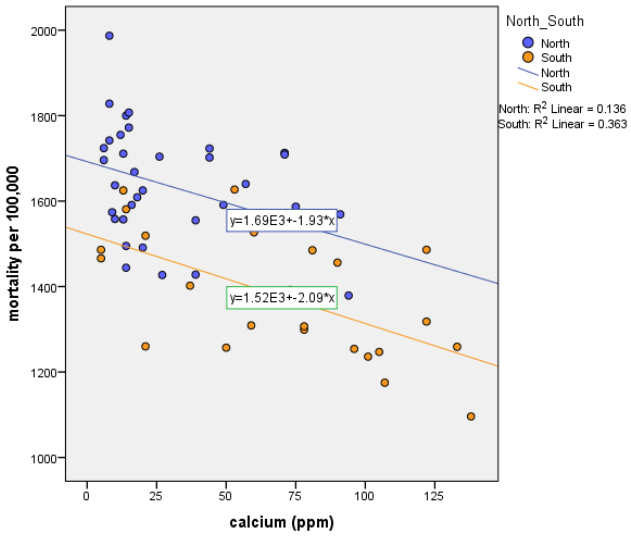
The new regression model is then:

$$y = \beta_1 + \beta_2 x + \beta_3 D + \varepsilon$$

What does this model mean for towns in the North?

What does this model mean for towns in the South?

How does this look graphically?



There seems to be two separate, almost parallel regression lines.

The overall regression is:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.769[a]	.591	.577	122.112

a. Predictors: (Constant), North_South, calcium (ppm)

b. Dependent Variable: mortality per 100,000

Coefficients[a]

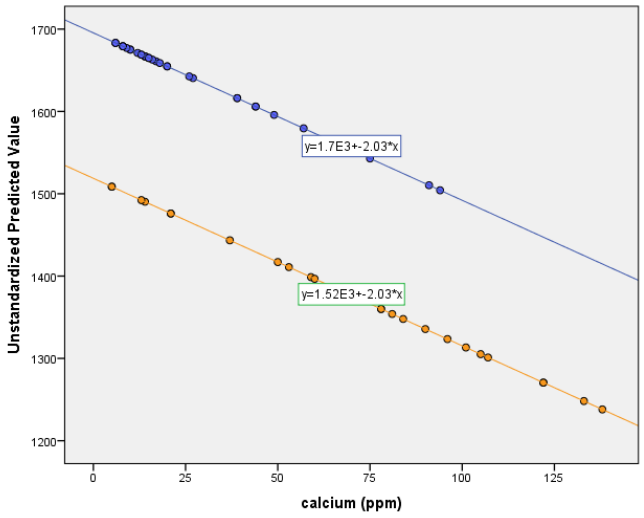
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1695.437	25.329		66.938	.000
	calcium (ppm)	-2.034	.483	-.413	-4.212	.000
	North_South	-176.711	36.891	-.470	-4.790	.000

a. Dependent Variable: mortality per 100,000

Mean mortality for North ($D = 0$): $\hat{y} = 1695.437 - 2.034x$

Mean mortality for South ($D = 1$): $\hat{y} = (1695.437 - 176.711) - 2.034x = 1518.726 - 2.034x$

Difference: North – South = $-b_3 = 176.711$



Independent Variables with Three Categories

Consider the following example.

Example 6.1

Suppose we are analysing data for a clinical trial to compare the effectiveness of three different medications used to treat depression. Nine participants are randomly divided into three groups of 3 patients and each group is assigned a different medication. The response variable is the measure of the effectiveness of the treatment. In addition to the treatment variables, another predictor will be age, x_2. We are examining three different treatments so we can define the following three dummy variables for the treatment:

D_3 = 1 if patient used treatment 1, 0 otherwise;

D_4 = 1 if patient used treatment 2, 0 otherwise;

D_5 = 1 if patient used treatment 3, 0 otherwise.

The regression model, therefore, is:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_5 D_5 + \varepsilon$$

On the surface, it seems that the model is well-defined. But what is the problem with it?

Patients	Treatments	Intercept x_1	x_2	D_3	D_4	D_5
1	1	1	52	1	0	0
2	3	1	35	0	0	1
3	1	1	65	1	0	0
4	2	1	57	0	1	0
5	2	1	71	0	1	0
6	3	1	66	0	0	1
7	1	1	42	1	0	0
8	2	1	53	0	1	0
9	3	1	39	0	0	1

Matrix $\boldsymbol{X}$

It is clear that matrix $\boldsymbol{X}$ is singular, as $D_3 + D_4 + D_5 = x_1$, and not invertible. We recognise this problem as **perfect multicollinearity** and is called **Dummy Variable Trap**.

One solution (there are others) for avoiding this difficulty is the "*leave one out*" method. The method has the general rule that whenever a categorical predictor variable has *m* categories, we should only use (*m*-1) of them to describe the differences among the *m* categories. For the overall fit of the model, it does not matter which set of (*m*-1) dummy variables we use. The choice of which (*m*-1) indicators we use, however, does affect the interpretation of the coefficients that multiply the specific indicators in the model.

Coefficient Interpretations

In the example above with three treatments, we might leave out the third indicator resulting in the following model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 D_3 + \beta_4 D_4 + \varepsilon$$

If patient received treatment 1, $D_3 = 1$ and $D_4 = 0$.

$$y = (\beta_1 + \beta_3) + \beta_2 x_2 + \varepsilon$$

If patient received treatment 2, $D_3 = 0$ and $D_4 = 1$.

$$y = (\beta_1 + \beta_4) + \beta_2 x_2 + \varepsilon$$

If patient received treatment 3, $D_3 = 0$ and $D_4 = 0$.

$$y = \beta_1 + \beta_2 x_2 + \varepsilon$$

Compare the three equations to each other. The only difference between the equations for treatments 1 and 3 is the coefficient β_3. The only difference between the equations for treatments 2 and 3 is the coefficient β_4. This leads to the following interpretations for the coefficients:

- β_3 = difference in mean response for treatment 1 versus treatment 3, assuming the same age.
- β_4 = difference in mean response for treatment 2 versus treatment 3, assuming the same age.

Note: In general, a coefficient of a dummy variable in the model measures the difference between the category defined by the dummy variable included the model and the category defined by the dummy variable that was left out (*reference category*).

Change in the Slope

So far we have assumed that the categorical variables affect the intercept only. The same technique can be used to incorporate changes in slope coefficients. To demonstrate this we consider a simple example of one numeric ad one dummy variable, the extension to more complex cases is however straightforward.

- Change in the intercept: $y = \beta_1 + \beta_2 x + \beta_3 D + \varepsilon$

$$\begin{cases} y = \beta_1 + \beta_2 x + \varepsilon, & D = 0 \\ y = (\beta_1 + \beta_3) + \beta_2 x + \varepsilon, & D = 1 \end{cases}$$

- Change in the slope: $y = \beta_1 + \beta_2 x + \beta_4 Dx + \varepsilon$

$$\begin{cases} y = \beta_1 + \beta_2 x + \varepsilon, & D = 0 \\ y = \beta_1 + (\beta_2 + \beta_4) x + \varepsilon, & D = 1 \end{cases}$$

- Changes in both: $y = \beta_1 + \beta_2 x + \beta_3 D + \beta_4 Dx + \varepsilon$

$$\begin{cases} y = \beta_1 + \beta_2 x + \varepsilon, & D = 0 \\ y = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) x + \varepsilon, & D = 1 \end{cases}$$

This procedure is sometimes called testing for **Structural Stability** of the regression.

Interactions

Consider the linear regression model with two independent variables:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

As we discussed in the interpretations of coefficients, the effect of x_2 on y is the same, whatever the value of x_3 and vice versa.

The effects of x_2 and x_3 on the response are called *main effects*.

The assumption that the effect of x_2 on response is the same no matter what value x_3 takes is a strong assumption and may not be appropriate in all cases. The assumption can be avoided by adding an *interaction* term to the model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 x_3 + \varepsilon$$

Interactions can occur between any types of independent variables, a numeric variable and a categorical variable, two numeric variables and between two categorical variables.

Note that when there is an interaction between two categorical variables, if the first variable has *m* categories and the second has *p* categories, the interaction will provide information about how the $m \times p$ two-way classification of individuals influences mean response.

In addition to the main effects terms, the interaction model will include terms for the $(m-1)(p-1)$ dummy variables representing the interaction effects.

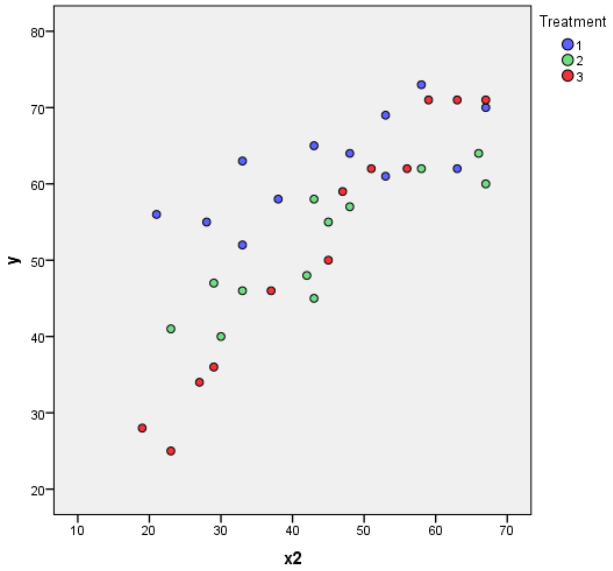
Note: While interaction between numeric variables can mathematically be interpreted as a form of non-linearity, the other kinds of interaction have more informative interpretations.

Example 6.2

Consider the model described in Example 6.1-but this time for 36 individuals:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 D_3 + \beta_4 D_4 + \varepsilon$$

The scatter plot of the data with treatment effectiveness against age is:



A (second-order) multiple regression model with interaction terms is:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_{23} x_2 D_3 + \beta_{24} x_2 D_4 + \varepsilon$$

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	6.211	3.350		1.854	.074
x2	1.033	.072	1.218	14.288	.000
D3	41.304	5.085	1.591	8.124	.000
D4	22.707	5.091	.874	4.460	.000
x2D3	-.703	.109	-1.298	-6.451	.000
x2D4	-.510	.110	-.922	-4.617	.000

a. Dependent Variable: y

Hence, the estimated line is:

$$\hat{y} = 6.211 + 1.033x_2 + 41.304D_3 + 22.707D_4 - 0.703x_2D_3 - 0.510x_2D_4$$

So we have the following equations (regimes):

- If patient received treatment 1, $D_3 = 1$ and $D_4 = 0$.

  $$y = (\beta_1 + \beta_3) + (\beta_2 + \beta_{23})x_2 + \varepsilon \qquad \hat{y} = 47.515 + 0.330x_2$$

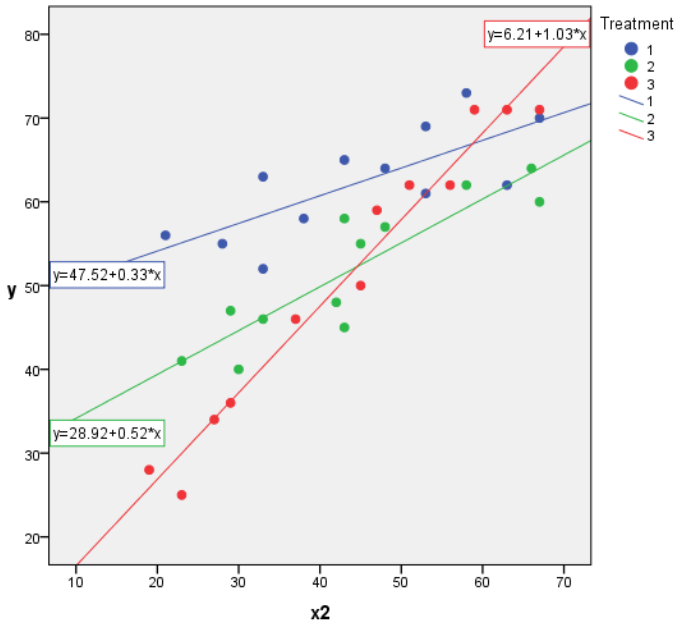
- If patient received treatment 2, $D_3 = 0$ and $D_4 = 1$.

  $$y = (\beta_1 + \beta_4) + (\beta_2 + \beta_{24})x_2 + \varepsilon \qquad \hat{y} = 28.918 + 0.523x_2$$

- If patient received treatment 3, $D_3 = 0$ and $D_4 = 0$.

  $$y = \beta_1 + \beta_2 x_2 + \varepsilon \qquad \hat{y} = 6.211 + 1.0330x_2$$

Plotting these 3 equations, we obtain:



The estimated slopes tell us:

- For patients *in this study* receiving treatment 1, the effectiveness of the treatment is predicted to increase 0.33 units for every additional year in age.
- For patients *in this study* receiving treatment 2, the effectiveness of the treatment is predicted to increase 0.52 units for every additional year in age.
- For patients *in this study* receiving treatment 3, the effectiveness of the treatment is predicted to increase 1.03 units for every additional year in age.

In short, the effect of age on the predicted treatment effectiveness depends on the treatment given. That is, age appears to **interact** with treatment in its impact on treatment effectiveness. The interaction is exhibited graphically by the intersecting lines.

Practical Week 6

The data file **logdrink.sav** records the weekly alcohol intake in units for a sample of 16-21 year olds who had taken some alcohol in the previous week. The number of units consumed has an extremely skewed distribution and a log transform is advisable:

Logdrink is the logarithm of the weekly units consumed.

Sex =1 for men and 0 for women. This variable is already in a form suitable for using as a dummy variable.

Age records the individual's age in years.

Set up a new variable to represent the interaction between sex and age.

Write down the regression model for the regression of **Logdrink** on **Age**, **Sex** and include the interaction.

Now write down the model for men. What are the slope and intercept in terms of the regression parameters?

What do the slope and intercept represent in terms of alcohol consumption?

And for women, what are the slope and intercept for women?

Now carry out the regression.

Does age significantly affect mean alcohol consumption?

Does sex have a statistically significant effect?

Is the interaction statistically significant?

Now plot the fitted regression lines, using sex to determine the markers. Comment on the results.