Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 15. Different types of regressions. (23 May)
Alla Tambovtseva

Regressions with categorical (nominal) predictors

Not only numeric independent variables can be included in regression models. Including categorical predictors (provided it is done correctly, of course) does not make model quality worse. We can add variables with two values that can be easily converted into binary ones (0 and 1) or with more than two values that are usually split into a set of binary variables.

1. Binary variables

Examples: gender (male, female), membership in the European Union (yes, no).

Consider the following model:

$$\texttt{democracy} = \texttt{rule of law} + \texttt{postcom},$$

where `democracy` and `rule of law` are indices of democracy and rule of law, and `postcom` is a binary variable (1 – post-communist states, 0 – not post-communist states). And with coefficients estimated:

$$\texttt{democracy} = 3 + 1.5 \times \texttt{rule of law} - 0.8 \times \texttt{postcom}.$$

To see what the coefficient of `postcom` shows, let us look at two observations and predict the values of democracy using the equation above.

	rule of law	postcom
country 1	2	0
country 2	2	1

$$\texttt{democracy (country 1)} = 3 + 1.5 \times 2 - 0.8 \times 0 = 6.$$

$$\texttt{democracy (country 2)} = 3 + 1.5 \times 2 - 0.8 \times 1 = 5.2.$$

Now, we can easily get the **interpretation** of this coefficient:

- All else equal, the level of democracy in post-communist countries is less than in not post-communist countries by 0.8 on average.

The interpretation does not differ from those we had before, but it would be strange to say something like *when the political regime increases by one* since `postcom` is a categorical variable and the regime of a country does not change at once.

2. Categorical variables with more than two values

Examples: region, profession of a respondent, favourite party.

Consider the same model as in the previous section, but now take the region where a country is situated into account:

$$\texttt{democracy} \sim \texttt{rule of law} + \texttt{postcom} + \texttt{region}.$$

Suppose region takes six values: AF (Africa), AS (Asia), AU (Australia and Oceania), EU (Europe), SA (South America), NA (North America). It is clear that even if we encode these values with numbers, it will be useless to include this variable in the model *as is* because values still cannot be ordered.

So, we have to split this variable into a set of binary dummy variables:

country	region	AF	AS	AU	EU	NA	SA
Ghana	AF	1	0	0	0	0	0
China	AS	0	1	0	0	0	0
Australia	AU	0	0	1	0	0	0
France	EU	0	0	0	1	0	0
Canada	NA	0	0	0	0	1	0
Brazil	SA	0	0	0	0	0	1

And then, logically, we should add this set of dummy variables to the model. However, there is one difficulty: if we put all these variables in the model, we will get perfect multicollinearity. This will result in a serious problem – model coefficients will not be estimated at all. Why these variables are so correlated? Suppose we know that there are six mutually exclusive binary dummies and we know values of five of them for a particular country. We can easily guess the sixth value as there is only one 1 among them.

To avoid this problem we will keep one of the binary variables out of the model. This variable corresponds to a particular value of a categorical variable (take AF, for example) that is called **a base level**. In our case the base level is **a reference region** which we will compare other results with. If we take AF (Africa) as a base level, after estimation we will get the following model:

$$\texttt{democracy} = 2 + 1.8 \times \texttt{rule of law} - 0.6 \times \texttt{postcom} -$$

$$-0.15 \times \texttt{AS} + 0.2 \times \texttt{AU} + 0.5 \times \texttt{EU} + 0.15 \times \texttt{NA} + 0.1 \times \texttt{SA}.$$

Coefficients of binary dummy variables show how the democracy level in Asia, Europe, Australia, North America and South America differs from the level of democracy in Africa (base level, not included in the model).

Interpretation of some coefficients:

- All else equal, the level of democracy is less by 0.15 in Asian countries than in African ones (on average).
- All else equal, the level of democracy is higher by 0.5 in European countries than in African ones (on average).

Regressions with interaction effects

Consider the following model that describes how the price of a flat depends on its square, distance to the closest metro station (`metrdist`) and reachability by foot (`whip`):

$$\texttt{price} \sim \texttt{square} + \texttt{metrdist} + \texttt{walk}.$$

Imagine that now we want to take the following fact into account. The effect of distance to the closest metro station on price is not the same for flats in blocks that can be reached by foot and flats that cannot. In terms of modeling, we want to include **an interaction effect** of distance to the metro station and reachability by foot. We can add **an interaction term** to this model that is just a product of two predictors:

$$\texttt{price} \sim \texttt{square} + \texttt{metrdist} + \texttt{walk} + \texttt{metrdist} \times \texttt{walk}.$$

Note: all independent variables that are included in an interaction term should be included 'separately' as well (we cannot skip `walk`, for example, in the model above).

Let us again consider an example with two flats that differ only by one parameter `walk`. We have a model with already estimated coefficients:

$$\texttt{price} = 50 + 2 \times \texttt{square} - 0.1 \times \texttt{metrdist} + 0.3 \times \texttt{walk} + 0.4 \times \texttt{metrdist} \times \texttt{walk}.$$

Let's predict the prices for two flats using the equation above. The first one is not reachable by foot ($\texttt{walk} = 0$):

$$\texttt{price} = 50 + 2 \times \texttt{square} - 0.1 \times \texttt{metrdist} + 0.3 \times 0 + 0.4 \times 0 \times \texttt{metrdist} =$$

$$50 + 2 \times \texttt{square} - 0.1 \times \texttt{metrdist}.$$

The first one is reachable by foot ($\texttt{walk} = 1$):

$$\texttt{price} = 50 + 2 \times \texttt{square} - 0.1 \times \texttt{metrdist} + 0.3 \times 1 + 0.4 \times 1 \times \texttt{metrdist} =$$

$$50.3 + 2 \times \texttt{square} + 0.3 \times \texttt{metrdist}.$$

Now we see that the effect of `metrdist` is different for cases with `walk` $= 0$ and for cases with `walk` $= 1$ and that this difference is exactly the coefficient of an interaction term that is 0.4:

$$0.3 - (-0.1) = 0.4.$$

If we go further to some calculus and take partial derivatives of `price` with respect to `metrdist` (i.e. evaluate the 'pure' effect of `metrdist` on `price`), we will get the following:

$$\text{one unit change in price} = -0.1 + 0.4 \times \text{walk}.$$

So, if a flat is reachable by foot (walk = 1), the effect of the distance to the metro station on price is higher by 0.4 on average.

Anyway, the general **interpretation** of the coefficient of the interaction term is the following:

- If we have a model with interactions $y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_1 \times x_2$, where x_2 is binary, we can say that all else equal, the effect of x_1 on y is higher by β_3 for cases when $x_2 = 1$ (on average).

If we want to include some non-linear effects in the linear model, we can add an interaction of some independent variable with itself. For example, it is a common fact that age often has a non-linear effect on the salary: up to some point the increase in age corresponds to the increase in salary (more experience and skills), but after some value of age, the increase in age results in the decrease in salary (less flexibility, less concentration). So, in this case we can add $\texttt{age}^2$ to the model that will predict the salary of employees by age:

$$\texttt{salary} \sim \texttt{age} + \texttt{age}^2.$$