Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 12. Simple linear regression. Ordinary least squares. (04 April)
Alla Tambovtseva

Regression models: idea

Goals of regression models

- To predict the values of one variable based on values of other variables.
- To describe the directed relationship between variables.

Example 1. We have a dataset on characteristics of films: film duration, film budget, film type and film rating. We construct a model, evaluate its coefficients and then become able to answer the following question: what is the predicted film rating if we know that it is 90 minutes long, it has budget of 2.5 million dollars and it is a drama?

Example 2. We have a dataset on characteristics of countries: GDP per capita, Control of corruption index and regime type. We construct a model, evaluate its coefficients and then become able to answer the following questions: how does the GDP per capita change when the Control of corruption index increases by one unit? what if we consider the regime type as well? how will the GDP per capita and the Control of corruption index be related if we will consider the interaction effect between the Control of corruption index and the regime type?

In the first case we are usually concerned about the predictive power of the model (goal common for machine learning) and in the second case we are more interested in the model coefficients, their statistical significance and other effects (goal common for economics and social sciences).

The idea of regression analysis is connected with the idea of correlation analysis. Both correlation coefficients and regression models describe the relationships between quantitative variables, however, in correlation analysis we study undirected relationships (no matter whether x affects y or vice versa) while in regression analysis we know which variable is dependent and which ones are independent.

NB. Neither correlation nor regression show cause-and-effect relationships. Although further we will operate a lot with terms like “dependent variable” or “linear dependence”, it is purely statistical term that has nothing in common with cause-and-effect dependence that can be detected in experiments or quasi-experiments.

Structure of regression models

Any regression model is an equation that has a dependent variable on its left-hand side and a set of independent variables on the right-hand side. Some examples:

- $y = a + b \cdot x$ (bivariate linear model);

- $y = a + b \cdot x + c \cdot z$ (multiple linear model);
- $y = a + b \cdot x^2$ (quadratic model).

An *independent variable* is a factor that influenses another variable. A *dependent variable* is a variable that is affected by other variable or variables.

Example 3. We have the following pair of variables: time spent on preparation for the test and the students' score for this test. Time spent on preparation for the test affects the score for this test (not vice versa, for sure), so `time` is the independent variable and `score` is the dependent variable.

As independent and dependent variables are usually called in different ways, here are some variants:

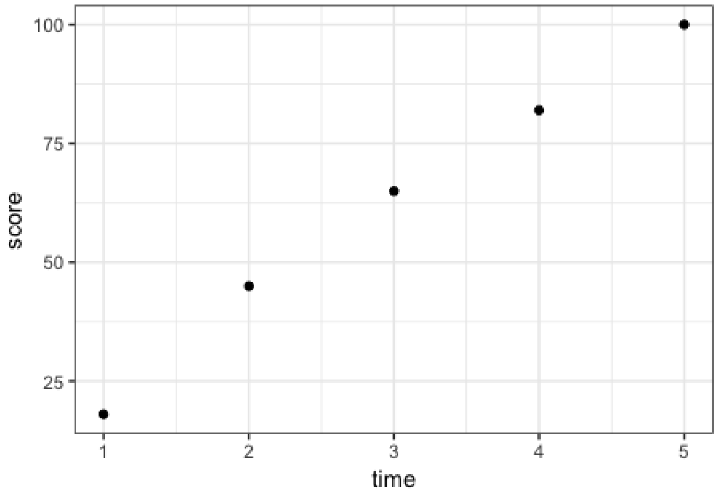
- independent variable = predictor = regressor = feature = factor;
- dependent variable = outcome = regressand = output.

Simple (bivariate) linear model

Consider a pair of variables that we discussed so far: time spent on preparation for the test and the students' score for this test. Imagine we have a toy dataset:

time (hours)	score
4	82
2	45
5	100
1	18
3	65

Let us make a scatter plot:



We can see that the score linearly depends on the time spent on preparation. This relationship can be described by a simple linear equation:

$$\hat{\text{score}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{time},$$

where the intercept $\hat{\beta}_0$ will show the average value of score when `time` equals to 0 and the slope $\hat{\beta}_1$ will show how `score` changes when `time` increases by one hour.

Suppose we estimated a model in R and got $\hat{\beta}_0 = 1.7$ and $\hat{\beta}_1 = 20.1$. Hence,

$$\hat{\text{score}} = 1.7 + 20.1 \times \text{time},$$

and we can interpret it in the following way:

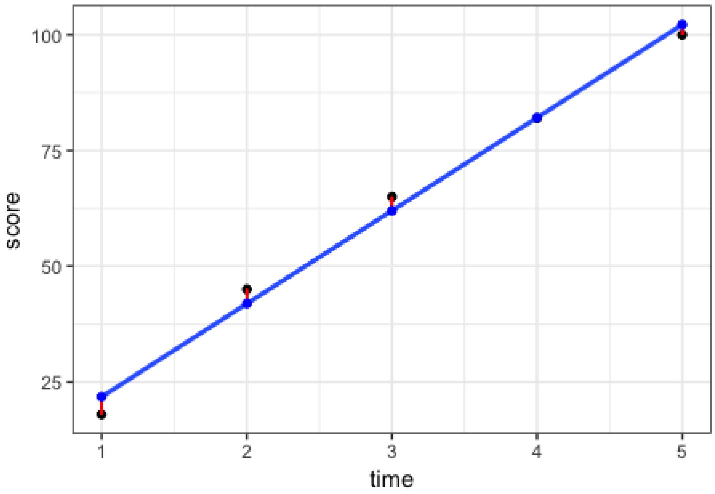
- The expected value of the score for the test if a student did not prepare for this test at all (`time=0`) is 1.7.
- If the time spent on preparation increases by one hour, the score for the test increases by 20.1 on average.

So, we got an equation that explains us the relationship between `time` and `score` and that can be used to predict the score of a student if we know only the time he or she spent on preparation. For instance, if he prepared for 2 hours, his predicted score for this test is $1.7 + 20.1 \times 2 = 41.9$.

Ordinary least squares method

A logical question can arise: how these coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$ were estimated? The method that is commonly used for estimating coefficients in linear models is called *ordinary least squares* method or *OLS*. It is not the only method that can be applied; $\hat{\beta}_0$ and $\hat{\beta}_1$ could be obtained by maximum likelihood estimation (MLE) or weighted least squares (WLS) and so on. This method is widely used since it is simple and efficient.

So as to understand this method, let us look at the scatterplot with regression line added:



Points that are below or above the regression line correspond to “real” values of `time` and `score`, ones we have in our table. Points that are on the regression line correspond to predicted values of `score`, ones that are expected based on our model. Our goal is to fit such a model that will result in the smallest error of prediction. In other words, while drawing a line defined by some slope and intercept, we want to make the differences between real values of `score` and predicted values of `score` (marked as short red lines) as low as possible. Let us write this idea in a more formal way. First, let us introduce some notations:

- x are values of `time` in our data and x_i is time spent by the i-th student;
- y are values of `score` in our data and y_i is a score of the i-th student;
- $\hat{y}$ are values of `score` predicted by our model and $\hat{y_i}$ is a predicted score of the i-th student;
- $\varepsilon = y - \hat{y}$ are residuals of our model, the errors of prediction and ε_i is a residual for the i-th student.

Now we can sum up all the squared residuals (squared so as not to get 0 while adding positive and negative numbers) and say that we need such $\hat{\beta}_0$ and $\hat{\beta}_1$ that

$$\sum_i \varepsilon_i^2 = \sum_i (y_i - \hat{y_i})^2 \to \min .$$

If substitute $\hat{y_i}$ with the line equation $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 \times x_i$ we will get the function that depends on $\hat{\beta}_0$ and $\hat{\beta}_1$. To find values of these coefficients where this function reaches its minimum, we have to calculate derivatives, solve a system of equations and do a lot of technical work. We will skip these details and just write the results. Finally,

$$\hat{\beta}_0 = \bar{x} \text{ and } \hat{\beta}_1 = r \times \frac{s_y}{s_x},$$

where $\bar{x}$ is the average of x, s_y is the sample standard deviation of y and s_x is the sample standard deviation of x.

Some notes on model quality

A basic measure of a linear model quality is R^2 (R-squared) that is also called the *determination coefficient*. We will discuss this measure in more detail later, but for a simple linear model, with one independent variable x, R^2 is simply the Pearson’s correlation coefficient squared, so:

$$R^2 = r^2 = [corr(x, y)]^2.$$

As for interpretation, R^2 stands for the share of variance of the dependent variable explained by our model. If express the same idea in an informal way, it shows the share of “reality” explained by our model. As you can guess, R^2 takes values from 0 to 1, and values close to 1 correspond to good models. There is no certain agreement which values of R^2 are acceptable. For social sciences values of R^2 about 0.6 and higher are supposed to be enough for good models.