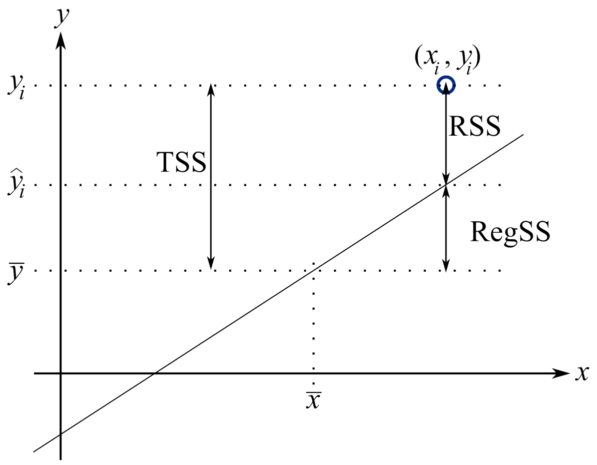
Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 13. Multiple linear regression. (18 April)
Alla Tambovtseva

More on squares and R^2

Last time we discussed the idea of the least squares method (OLS) and concluded that the linear model is fitted in a way that minimizes the sum of residuals squared. However, it is not the only type of squares that is considered in models. There are three types of squares:

- *Explained sum of squares*: corresponds to the variance of the dependent variable y that is explained by the model:
$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$
- *Residual sum of squares*: corresponds to the variance of the dependent variable y that is not explained by the model:
$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$
- *Total sum of squares*: total variance of the dependent variable y:
$$TSS = ESS + RSS = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

We can visualize all these values (RegSS is ESS here [1]):



[1] Taken from: https://learnche.org/pid/least-squares-modelling/least-squares-model-analysis.

Hence, as we defined R^2 as a share of the variance of y explained by our model, it can be obtained in the following way:

$$R^2 = \frac{ESS}{TSS} = \frac{ESS}{ESS + RSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

That is why R^2 is often considered as a measure of model quality.

In the R output for regression models we also see the adjusted R^2 (`Adjusted R-squared`). Unlike a simple R^2, it poses some penalties for the number of independent variables in the model:

$$R^2_{adj} = R^2 \times \frac{n-1}{n-k-1},$$

where n is the number of observations and k is the number of independent variables.

Adjusted R^2 is always less than simple R^2. If our model is too complex (includes too many independent variables for the given number of observations), the difference between R^2 and R^2_{adj} will be high. As it follows from some empirical rules in econometrics, if we want to get a good model, for one independent variable we should have at least 10 observations. For instance, if we have 100 observations in our data set, we should run a linear model with no more than 10 independent variables.

Multiple linear regression model

Suppose we have a model that predicts the price of the flat (in million rubles) based on its square (in sq.meters) and its distance from the closest metro station (in meters):

$$\hat{\text{price}} = 2.5 + 0.4 \times \text{livesp} - 0.05 \times \text{distance}.$$

This model cannot be plotted as a line because in this case $\hat{\text{price}}$ is the function of two variables `livesp` and `distance`, and functions of two variables describe surfaces, not lines (see here, for example). Thus, so as to keep approximately the same logic of interpretation as we had in a paired linear regression, we should fix the values of some variable and make price dependent only on one variable.

Imagine we compare two flats with the same living space of 60 meters^2, but with different distance from the metro station: 100 and 101 meters. The price of the first flat is $2.5 + 0.4 \times 60 - 0.05 \times 100 = 21.5$, the price of the second one is $2.5 + 0.4 \times 60 - 0.05 \times 101 = 21.45$. These two prices are different exactly by 0.05, the value of the coefficient of distance in the equation above. Hence, we can suggest the following interpretation of model coefficients:

- All else equal (*ceteris paribus*), with one square meter increase in `square` the price of a flat increases by 0.4 on average.
- All else equal (*ceteris paribus*), with one meter increase in `distance` the price of a flat decreases by 0.05 on average.

Or, generally:

- If we have a model $y = \beta_0 + \beta_1 \times x + \beta_2 \times z$, we can say, for example, that all else equal (*ceteris paribus*) as x increases by one, y changes by β_1 on average (increases or decreases depending on the sign).

Assessing model quality

OLS estimates of coefficients in a linear model are supposed to be the best linear unbiased estimators (BLUE) if a linear model satisfies the following **Gauss-Markov conditions** that stem from the **Gauss-Markov Theorem**:

1. Linear relationships between independent variables and a dependent one.
2. No or low *multicollinearity*: no high correlation between independent variables.
3. No systematic patterns of residuals: $E(\varepsilon) = 0$.
4. No *heteroskedasticity*: $Var(\varepsilon) = const$, constant variance of residuals.
5. No correlation between residuals that correspond to different observations: $Corr(\varepsilon_i, \varepsilon_j) = 0$, where $i \neq j$ are numbers of observations.

The normality of residuals is not included in these conditions, but if residuals of a model are distributed normally, OLS estimates of coefficients are the best not only among linear estimators, but among all possible estimators, so they are called BUE – Best Unbiased Estimators.