Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 14. Multiple linear regression. (25 April)
Alla Tambovtseva

Model diagnostics

Usually problems with linear regression models occur when Gauss-Markov conditions, basic assumptions of OLS models, are not satisfied. Plus, some inconsistencies might arise when there are outliers in our data. So, potential problems are:

- patterns in residuals, especially non-linear ones;
- multicollinearity;
- heteroskedasticity;
- influential observations.

Let us discuss all these problems in detail.

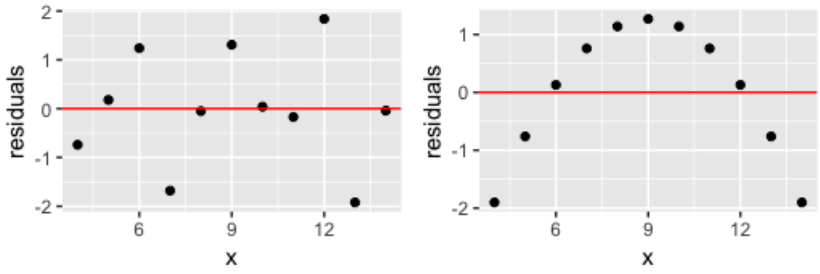
Patterns in residuals

Why it is bad?

As it stems from Gauss-Markov theorem, OLS estimates of coefficients are good if residuals of a model are scattered randomly. If this condition is violated, so if there are certain patterns in residuals, there is no guarantee that OLS estimates are unbiased and reliable. Moreover, definitely non-linear patterns can serve as evidence that it is incorrect to fit a linear model on our data (we should try a quadratic one, for example).

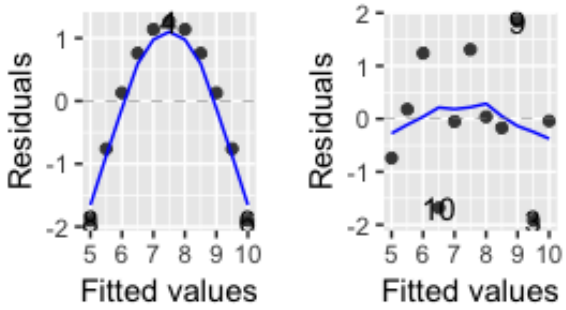
How to detect?

Plot different scatterplots with residuals. Firstly, we can create a series of scatterplots **residuals vs independent variable**. Let us compare two situations: an ideal one (with randomly scattered residuals) and a bad one (with a quadratic pattern in residuals).



In the first case there are no problems in the model, in the second case we face serious problems with model specification. Here we had better consider a different type of model, not a linear OLS regression, but a quadratic one (or at least with some quadratic terms).

Secondly, we can search for patterns in a different type of graph, a scatterplot **predicted values vs residuals**.



In the picture on the left we see such a graph that shows a certain quadratic pattern, in the picture on the right there is a case without any patterns.

The interpretation is the same: if there are some non-linear patterns, it is bad, if not, no problems are detected.

How to deal with it?

Think about the specification of the model. Considering a linear model with quadratic terms might help or taking a completely different type of regression with a different fit method.

Multicollinearity

Multicollinearity: high correlation between independent variables.

Why it is bad?

Leads to extremely high standard errors of coefficients (so called *inflated standard errors*) that causes non-stable results. Besides, it can result in insignificance of coefficients that should be statistically significiant and would be if multicollinearity was absent.

How to detect?

First, we can look at the **correlation matrix** for all independent variables in a model and check whether there are variables that are highly correlated (with correlation coefficients greater than 0.8). If yes, it might be a sign for multicollinearity.

Secondly, we can become suspicious when the R^2 of the model is very high (0.9 and higher), but at the same time there are **a lot of insignificant coefficients** in the model.

Thirdly, we can compute the value of the **Variance Inflation Factor (VIF)** that is

supposed to be a measure of the multicollinearity. It is calculated for each independent variable j as follows:

$$VIF = \frac{1}{1 - R_j^2},$$

where R_j^2 is `R-squared` from the model where j-th variable is taken as a dependent variable and other independent variables from the original model are taken as independent ones.

Example 1. The original model is:

$$\texttt{y} \sim \texttt{x1 + x2 + x3}$$

We want to compute VIFs for every independent variable in RStudio. So, R automatically runs three models:

$$\texttt{x1} \sim \texttt{x2 + x3}$$
$$\texttt{x2} \sim \texttt{x1 + x3}$$
$$\texttt{x3} \sim \texttt{x1 + x2}$$

takes `R-squared` from every model (e.g. 0.8, 0.7 and 0.6) and calculates VIFs:

- VIF (x1) = $\frac{1}{1-0.8} = 5$
- VIF (x2) = $\frac{1}{1-0.7} = 3.33$
- VIF (x2) = $\frac{1}{1-0.6} = 2.5$.

Values of VIF greater than 10 are considered very high, so variables with such VIFs are problematic. Probably, they describe the same thing, so some information is included in the model several times. If we want to get a single measure, we can calculate the **mean VIF** averaging by all independent variables. Usually it is supposed that the mean VIF greater than 5 signals about multicollinearity.

How to deal with it?

- Exclude some variables (usually with highest VIFs) so as to get rid of pairs of strongly correlated variables.
- Use PCA (principal component analysis) to get a smaller set of uncorrelated variables and include it in the model instead of a larger set of original highly correlated variables. We will discuss PCA later.

Heteroskedasticity

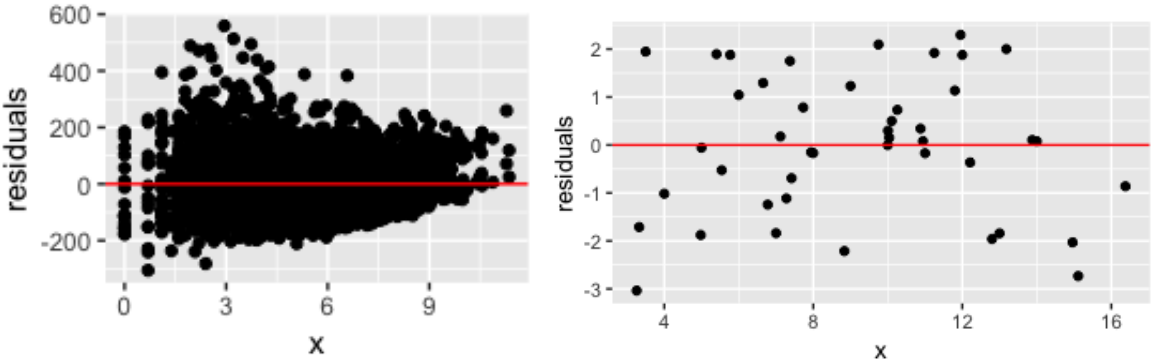
Heteroskedasticity: constant variance of residuals.

Why it is bad?

Leads to unstable estimates of coefficients and estimates that might be inefficient.

How to detect?

Firstly, we can draw conclusions from the same scatteplot as one used for detecting patterns of residuals, **residuals vs independent variable**. Consider two cases:



The left picture corresponds to the model with heteroskedasticity: the spread of residuals is not constant when values of x change. If we look how points are scattered around the line $y = 0$, we will see that they are widely spread when x is less than 3, but then the magnitude of the spread decreases as x increases. The right picture corresponds to the model with no heteroskedasticity: residuals are scattered randomly and, in general, they remain equally distant from the line $y = 0$ as x gets higher.

Secondly, we can use a formal **Breush-Pagan test** that tests the hypothesis about the absence of hetereroskedasticity (i.e. homoskedasticity). However, this test is conservative, very strict, so if we reject the null hypothesis about homoskedasticity, it does not necessarily mean that it exists and seriously affects the estimates. It is better to base final conclusions on the data features and visual analysis.

How to deal with it?

- Nothing if graphs show approximately the same spread of residuals for different values of independent variables and if there are no substantial reasons for heteroskedasticity.
- Use heteroskedasticity consistent standard errors of regression coefficients (sometimes called robust, but term 'robust' is less concrete) instead of ordinary standard errors.

Inflential observations

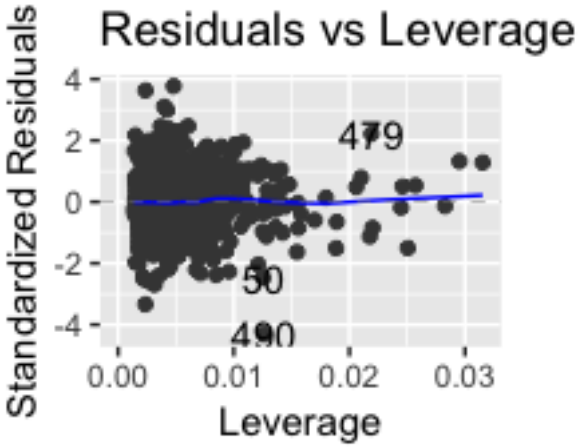
Why it is bad?

The presence of influential observations leads to unstable results because such observations can "pull" a regression line towards themselves and, hence, significantly change the estimates of regression coefficients. In other words, values of coefficients estimated on the full data set can be dramatically different from their values estimated with influential observations excluded (they can even change the sign).

How to detect?

Firstly, calculate **Cook's distance**, a measure of influence, for every observation. The calculation is pretty difficult, so let us skip it. The observation (point) is supposed to be influential if its Cook's distance is greater than 1.

Secondly, Plot a graph **residuals vs leverage**. Residuals of the model are calculated as usual, but they are scaled (standardized). The leverage is calculated for every observation i, it is just the difference between the predicted value of the dependent variable computed on the given data set and its value computed on the data set with i-th observation excluded.
The graph **residuals vs leverage** is the following:



Points that are non-typical are distant from the line $y = 0$ (as residuals are standardized, their typical values lie between -2 and 2), and points with high leverage (level of influence) are on the right in this graph. Influential points are those that are non-typical and with high leverage at the same time. R automatically puts their numbers (row names in a data set) so as we can see which observations are suspicious. Here it is 479. Points 490 and 50 can be safely kept since the values of leverage are now very high.

How to deal with it?

Exclude influential observations from analysis.