Diagnostics

How to deal with linear regression

Chuvakin Sergey

«School of Advanced Studies»

January 11, 2021

Outline

- Why?
- Efficient Sample
- Multicollinearity
- Linear dependency
- Homoscedasticity
- Exogeneity
- Model Selection

why

Diagnosticts helps to understand if regression model is efficient, robust and unbiased. We can use and interpret just in case our coefficients are unbaised.

Efficient Sample

Sample is implied to be representative. First of all it assumed to be random, but in case other technique is chosen, it should be proven.

Multicollinearity

- Collinearity is a linear association between two explanatory variables
- Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly correlated with each other
- For more information click here

You can also use Variance Inflation Factor (VIF).

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

where R_j^2 is the coefficient of determination of a regression of explanatory variable j on all the other explanatory variables.
VIF should be less than 4.

Multicollinearity

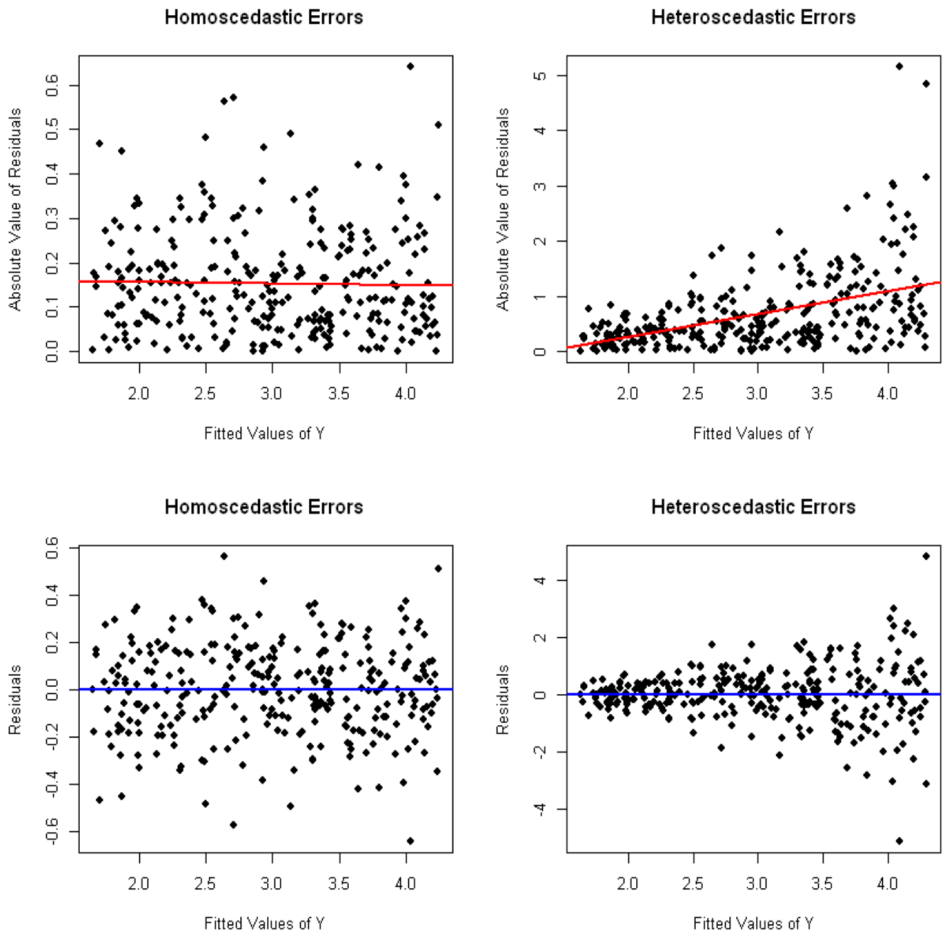
	Blood pressure	Age	Weight	Body surface area	Duration of hypertension	Pulse
Age	0.66					
Weight	0.95	0.41				
Body surface area	0.87	0.38	0.88			
Duration of hypertension	0.29	0.34	0.20	0.13		
Pulse	0.72	0.62	0.66	0.47	0.40	
Stress	0.16	0.37	0.03	0.02	0.31	0.51

Linear dependency

By default true relationship should resamble linear. Make scatter plot to ensure it. But, unfortunately, there is a really little chance that you face it. In these cases use some function to transform your X in regression formula. For example add a quadratic (second-order) polynomial of X if you see a U-shape relationship between Y and X (cubic, or third-order, if a Sigma-shape link). You can use any function are known, the only condition - is should satisfy visual relationship.

Homoscedasticity

Heteroscedasticity



Homoscedasticity

- The term «heteroscedasticity» refers to the case of non-constant error variance.
- Standard errors obtained under heteroskedastictity are generally incorrect

Except for visually representativness it also could be captured via Breusch-Pagan test. In R, this test can be performed by the function ncvTest from the car package or the function bptest from the lmtest package. Significant tests indicate heteroskedastisity. Use p-values to decide.

Exogeneity

- Exogeneity means that each X variable does not depend on the dependent variable Y, ratherY depends on theXs and on e
- Since Y depends on e, this means that the Xs are assumed to be independent of Y hence e
- required because if the «independent variables» are not independent of e and Y, then the estimated regression coefficients are not consistentif we use the OLS estimating equations
- X is exogenous if $Corr(X, e) = 0$
- X is endogenous if $Corr(X, e) \neq 0$
- If OLS is to be unbiased and consistent, requires that X is exogenous.

Exogeneity

- Simultaneous equations bias
- Omitted variables bias
- Errors-in-variables
- Regression model (time series) includes a lagged dependent variable and the error term is serially correlated.

Model Selection

How to choose model between several?

- R^2 should be close to 1
- AIC – Akaike Information Criterion, should be less
- BIC – Bayesian Information Criterion, should be less

What else?

- Think about standartization
- Remove extreme values
- Normally distributed errors