Let's Start

Multicollinearity and its impact on estimates of model parameters

Pylypiva Oleksandra, Katrich Constantine, Koval Oleg

Multicollinearity

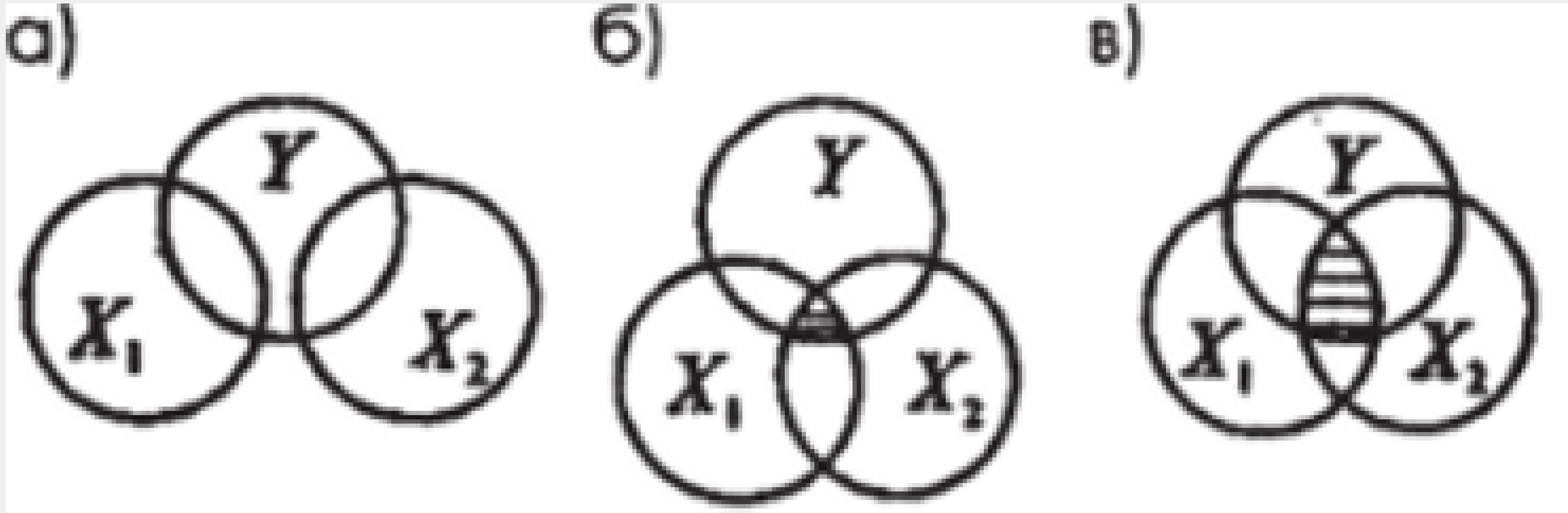
Multicollinearity is a phenomenon in which there is a dense linear relationship between two or more factor variables of the model or, in other words, factor variables have a high degree of correlation

The term "multicollinearity" means that in a multivariate regression model, two or more independent variables, factors are interconnected by a linear relationship or, in other words, have a high degree of correlation.

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1m} \\ r_{21} & 1 & r_{23} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & r_{m3} & \cdots & 1 \end{pmatrix}$$

г)

Graphically, this situation is presented in the form of a pie chart:



- Subset 1 - describes the individual effect of factor x1.
- Subset 2 - describes the individual effect of factor x2.
- Subset 3 - the combined effect of both factors on the variable y, which can not be separated. Graphically describes the situation of collinearity.

The inevitability of multicollinearity

- Multicollinearity is a normal phenomenon.
- Almost any model contains multicollinearity.
- We do not pay attention to multicollinearity until the appearance of obvious symptoms.
- Only excessively strong connections get in the way.

Methods for determining multicollinearity

The most complete study of multicollinearity is carried out according to the Farrar-Glober algorithm, which uses three types of statistical criteria and allows to detect multicollinearity:

- the whole array of independent variables (criterion 2)
- each independent variable with all others (F-criterion)
- each pair of independent variables (t-test)

The Farrar-Glober algorithm
Seven steps

Step 1:

1

Normalize variables x1, x2, ..., xm of ecometric model, calculating new (normalized) values of variables xy

$$x_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{n}\sigma_j}$$

Step 2:

Based on the matrix X *, the elements of which are normalized independent variables Xij *, calculate the correlation matrix (matrix of moments of the normalized system of normal equations):

$$R = X^{*T} X^{*} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{pmatrix}$$

Step 3:

Determine | R | - the determinant of the correlation matrix R; calculate the experimental value of criterion 2:

$$\chi^2_{exp} = -[n - 1 - \frac{1}{6}(2m + 5)] ln|R|$$

Compare the value of x^2 with the table at degrees 1 / 2m (m-1) of freedom and the significance of α

4

Step 4:

Determine the matrix of errors:

$$C = R^{-1} = \begin{pmatrix} c_{11} & c_{22} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & c_{mm} \end{pmatrix}$$

Step 5:

Calculate F-criteria:

$$F = \frac{(c_{kk} - 1)(n - m)}{m - 1}$$

Calculate the coefficients of determination for each variable:

$$R_k^2 = 1 - \frac{1}{c_{kk}}$$

Step 6:

Find the partial correlation coefficients that characterize the density of the relationship between two variables, provided that other variables do not affect this relationship (the existence of paired multicollinearity):

$$r_{kj} = \frac{-c_{kj}}{\sqrt{c_{kk}c_{jj}}}$$

Step 7:

Calculate t-criteria:

$$t_{kj} = |r_{kj}| \frac{\sqrt{n-m}}{\sqrt{1-r_{kj}^2}}$$

The values of the t_kj criteria are compared with the tabular ones at n-m degrees of freedom and significance levels α; if t_kj> ttable, then there is multicollinearity between the independent variables xk and xi.

Conclusions of the method:

There may be a linear relationship between independent variables, but it may not be a phenomenon of multicollinearity

If F_exp> F_table, then x_k depends on all other independent variables and it is necessary to decide whether to exclude it from the list of variables

If t_kj> t_table, then xk and xi are closely related

Analyzing the F- and t-criteria, we conclude which of the variables should be excluded from the model (of course, if possible for economic, logical and theoretical reasons)

If following paragraphs. 2-4, we did not achieve the goal, ie did not eliminate multicollinearity, the estimation of the model parameters should be calculated using another method

Conclusion

Multicollinearity and its impact on estimates of model parameters

End

Thank you

Do you have any questions?