Multicollinearity

Madari Zoltán

Contents

Repeat

1. Model assumptions

Multicollinearity

1. Multicollinearity
2. Measuring multicollinearity
3. VIF indicator

Solution

1. Generally
2. Principal component analysis

Model assumptions

1. **Linearity**
2. **No exact multicollinearity**
3. **Strong or strict exogeneity**
4. **Homoscedasticity**
5. **No autocorrelation**

In case of IID

1. $f(X) = \beta_0 + \beta_1 X + \varepsilon$

2. The data matrix has full column rank.

   Variables cannot be written as a linear combination of each other

3. $\mathbb{E}(\varepsilon_i | X_i) = 0$

   The errors are independent of the explanatory variables

4. $\mathbb{D}^2(\varepsilon_i | X) = \sigma^2$

   The standard deviation of errors for different observations is constant

5. The errors for different observations are uncorrelated

http://real.mtak.hu/42944/1/dim.2016.07.pdf

Multicollinearity

Explanatory variables are correlated with each other

Higher standard errors → higher p-value → unreliable t statistics

Structural

- Specification problem
- Using interaction or quadratic term

Due to the nature of the data

- Data is naturally correlated
- E.g. age and working experience

Measuring multicollinearity

Measuring the relation between explanatory variables

$$x_i = \beta_0 + \beta_1 x_1 \ldots + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} \ldots + \beta_k x_k$$

R_i^2: what percentage of the variance of the given explanatory variable is explained by the other explanatory variables

Tolerance

$$Tol(i) = 1 - R_i^2 = 1 - R^2_{x_i|x_1,x_2 \ldots x_{i-1},x_{i+1},\ldots x_k}$$

The larger R_i^2 the smaller tolerance → try to measure the „new" information contain of the given variable

The smaller the value of the tolerance indicator, the greater the risk of multicollinearity

VIF indicator – variance inflation factor

$$VIF_i = \frac{1}{1 - R_i^2} = \frac{1}{Tol(i)}$$

How many times does the inclusion of given independent variable increase the estimated variance of the given parameter

VIF (i) = 1 – the theoretical minimum → The i-th explanatory variable is independent from the other explanatory variables

VIF (i) = 2: the variance of the given parameter is doubled due to multicollineartity compared to the no multicollinearity case

	High multicollinearity	
R_i^2	≥ 80%	≥ 90%
Tol	≤ 0,2	≤ 0,1
VIF	≤ 5	≤ 10

Solving multicollinearity

- **Omitting highly correlated** independent varibles from the model
  - ❖Questionable – What will be the limit? Omitted variable bias?
- **Dimension reduction** procedures
  - ❖**„Merge” strongly correlated** variables into fewer number of variables
  - ❖Aim: preserving as much information as possible

Principal component analysis

Aim: reduce the number of variables in such a way that you lose as little information as possible

Measuring information content: **with variance** – the larger standard deviation of an observed property, the more information it contains

Variance depends on **unit of measure**→ the data must be **standardized**

$$\frac{x_i - \bar{x_i}}{\delta} \qquad E(x) = 0, D(x) = 1$$

Principal component: linearly independent variables that contain the information content of the previously correlated data

We keep the variables with the highest variance

The mathematics of principal component analysis

1. Create the dataset – only continous variables
2. Standardization or normalization
3. Prepare covariance matrix (R)
4. Calculate the eigenvectors and eigenvalues of the covariance matrix
    1. Eigenvalues : $\lambda_1 \geq \ldots \lambda_k \geq \ldots \lambda_p > 0$
    2. Eigenvectors (unit length): a_1, a_k, a_p
    3. Matrix of eigenvalues and eigenvectors: $L = \text{diag}(\lambda_i)$, $A = a_i$ vectors as the column of matrix
    4. Connection: $\mathbf{R = A \Lambda A^T}$
5. Determine the PCs

Path analysis

The explanatory variables are correlated in some cases

They have effect on the dependent variable through each other

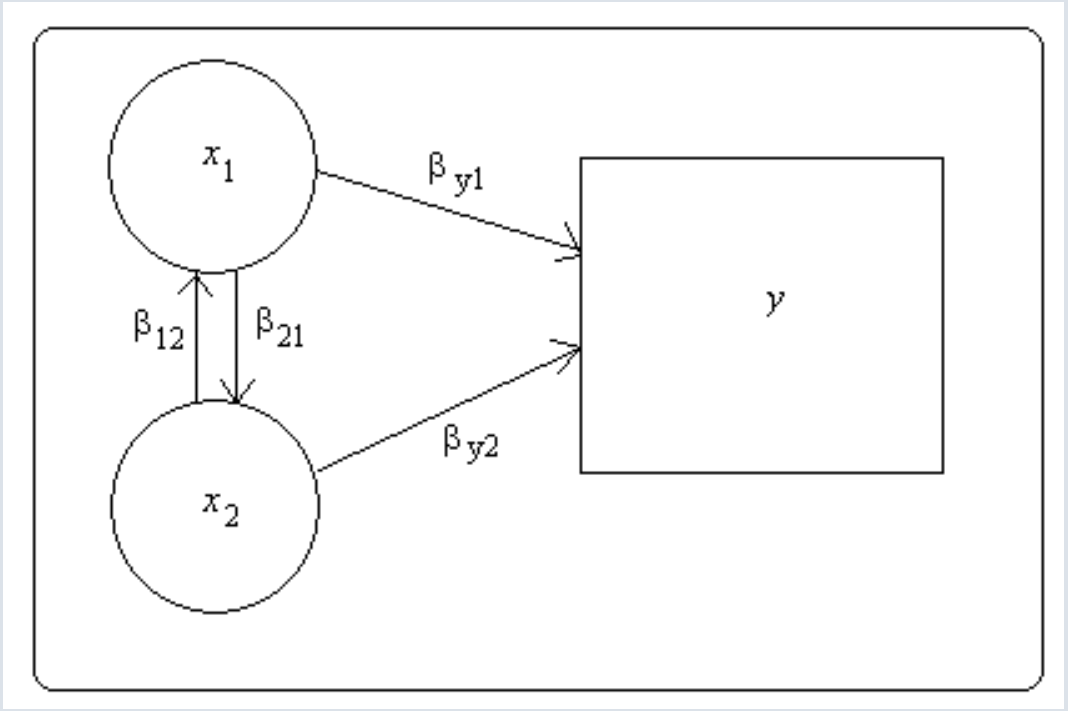
We would like to identify the direct and indirect effects

$$y = \beta_{y0} + \beta_{y1}x_1 + \beta_{y2}x_2 + \varepsilon$$

$$x_2 = \beta_{20} + \beta_{21}x_1 + \varepsilon$$

$$\beta_1 = \beta_{y1} + \beta_{21} * \beta_{y2}$$

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$



Forrás: https://mersz.hu/dokumentum/m366stkett__66


Thank you for your attention!