Tasks with Solutions for Introductory Econometrics

4. Multicollinearity

4.1 What is the problem of multicollinearity?

Solution:

A. Perfect multicollinearity

Perfect multicollinearity occurs when two or more independent variables have correlation coefficients of 1 or -1. An OLS estimation is then not possible because in this case columns of the observation matrix **X** (=variable vectors) are linearly dependent. The product matrix $\mathbf{X'X}$ is then singular, i.e. cannot be inverted, so that the OLS estimator $\hat{\boldsymbol{\beta}}$, i.e. the vector of the estimated regression coefficients, cannot be calculated. In practice, perfect multicollinearity only occurs when independent variables are linked by a definitory relationship.

B. Non-perfect multicollinearity

Independent variables are strongly correlated, but have no correlation coefficient of 1 or -1. In this case, the OLS estimator $\hat{\boldsymbol{\beta}}$ can be calculated. However, the variable values are not distributed over the entire x_1,x_2-level, e.g. with k=3 regressors (including the dummy variables), as in the case of an independence of the explanatory variables. Due to their dependence, they lie relatively close to a straight line in this plane. This makes the position of the additional regression level in the x_1,x_2,y space more uncertain, since it is already strongly influenced by small data changes. The greater uncertainty is also expressed by the fact that the standard errors of the estimated regression coefficients become larger with multi-linearity, so that their significance is more difficult to prove

4.2 Which interpretation problems will occur in econometric model estimates in presence of multicollinearity?

Solution:

● The jointly declared influence cannot be exactly separated into individual influences. Since the total influence is still correctly estimated, an overestimation of the influence of individual variables goes hand in hand with an underestimation of the influence of other variables.

● The distribution of the individual influences and thus the size of the regression coefficients is characterized by great uncertainty. Even small data changes can have a strong effect on the estimated regression coefficients.

● The high uncertainty is also shown by the fact that the standard error of the regression coefficients increases. The absolute values of the test quantities (t-values) tend to become smaller, so that the significance of individual regressors becomes more difficult to detect. Even if the null hypothesis of the F-test is rejected across the whole context, the results of the t-tests cannot be significant.

4.3 Outline the methods for detecting multicollinearity.

Solution:

● Simple correlation analysis:

The simple correlation coefficients measuring the relationship between the independent variables should not be greater than 0.8 in absolute terms.

Problem: Multicollinearity caused by the interdependence of multiple regressors may not be detected.

● Expanded correlation analysis:

The simple correlation coefficients between the regressors should not be greater than the multiple correlation coefficient, which measures the relationship between all x variables and the y variable.

Problem: see simple correlation analysis.

● Determinant of the product matrix **X'X** or the correlation matrix **R**:

The determinant of the product matrix **X'X** or the correlation matrix **R** of the exogenous variables is close to zero for multicollinearity.

Problem: There is no rule of thumb for which deviation from zero to assume multicollinearity. In addition, no information is obtained about which variables cause multicollinearity.

● Condition number of the **X'X**

The condition number of the product matrix **X'X**, which results from the root of the ratio of its largest and smallest eigenvalue, should not be greater than 30.
Problem: No information is obtained about which variables cause multicolinearity.

● Method of auxiliary regression:

Each of the k-1 "true" independent variables is explained by the other independent variables with a linear regression (= auxiliary regression). If the j-th x-variables are regressioned to the remaining x-variables, the determination coefficient R_j^2 is obtained.

- F-Test of determination coefficients R_j^2:

If one or more coefficients of determination R_j^2 are significant, multicollinearity is assumed.

- coefficient of tolerancet: $tol_j = 1 - R_j^2$

With multicollinearity, the tolerance coefficient should be close to 0. As a rule of thumb, a tolerance coefficient below 0.10 indicates multicollinearity.

- VIF coefficient (reciprocal of the tolerance coefficient):

The variance inflation index indicates the percentage by which the variance of the j-th estimated regression coefficient is "inflated" by multicollinearity. A value above 10 indicates multicollinearity.

4.4 What problems do the methods of variable suppression and the method of differences entail for overcoming multicollinearity?

Solution:

a) Method of variable suppression

Of the regressors that are highly correlated, only one regressor is included in the regression model. The following problems are associated with this:

- The user must determine which regressors are to be excluded. The exclusion is often difficult to justify in terms of content and the procedure is therefore subjective.

- By excluding one or more regressors, the disturbance variable can contain classifications. This would mean that the model assumptions would no longer be fulfilled.

b) Method of taking differences

Multicollinearity may be due to the fact that exogenous variables are subject to trends. In a linear trend, multicollinearity can be excluded by simple difference formation. The following problems are to be mentioned here:

- Limited application possibilities: The method can only be used if the correlated exogenous variables are afflicted with a linear trend.

- The level relationship between the variables is lost, which can be of interest in the sense of economic theory.
- By limiting the range of values, the formation of differences tends to lead to a reduction in the dispersion of the regressors and thus to an increase in the standard error of the estimated regression coefficients. The test variable of the t-test (empirical t-value) then remains low.
- An observation value is lost due to the formation of differences, which can be a problem especially with annual data.
- If the interference terms were previously uncorrelated, they are autocorrelated due to the difference formation. The BLUE property of the OLS estimator is thus lost.

4.5 The correlation matrix of the explanatory variables of the Energy Model III is given by

	GASPR	DHC	DIINC
GASP	1.000	0.749	-0.848
DHC	0.749	1.000	-0.589
DIINC	-0.848	-0.589	1.000

a) Assess the extent of multicollinearity on the basis of an advanced simple correlation analysis!
Note: The coefficient of determination in Energy Model III is 0.796.

Solution:

Variabels: X_2: GASPR, X_3: DHC, X_4: DIINC

Multiple correlation coefficient: $R_{yx2x3x4} = \sqrt{0,796} = 0,892$

$|r_{23}| = 0,749 < 0,892$ → No indication of multicollinearity by X_2 and X_3

$|r_{24}| = 0,848 < 0,892$ → No indication of multicollinearity by X_2 and X_4

$|r_{34}| = 0,589 < 0,892$ → No indication of multicollinearity by X_3 and X_4

b) Assess the extent of multicollinearity with the aid of the determinant of the correlation matrix of the exogenous variables!

Solution:

Correlation matrix:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{23} & r_{24} \\ r_{32} & 1 & r_{34} \\ r_{42} & r_{43} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0{,}7489 & -0{,}8476 \\ 0{,}7489 & 1 & -0{,}5889 \\ -0{,}8476 & -0{,}5889 & 1 \end{pmatrix}$$

By a 3×3 matrix the determinant can be determined most quickly with the Sarrusian rule.

Determinant of the Correlation Matrix (with Enhancement)

$$|\mathbf{R}| = \begin{vmatrix} 1 & 0{,}7489 & -0{,}8476 & 1 & 0{,}7489 \\ 0{,}7489 & 1 & -0{,}5889 & 0{,}7489 & 1 \\ -0{,}8476 & -0{,}5889 & 1 & -0{,}8476 & -0{,}5889 \end{vmatrix}$$

Determinant:

$$|\mathbf{R}| = 1 \cdot 1 \cdot 1 + 0{,}7489 \cdot (-0{,}5889) \cdot (-0{,}8476) + (-0{,}8476) \cdot 0{,}7489 \cdot (-0{,}5889)$$

$$- [(-0{,}8476) \cdot 1 \cdot (-0{,}8476) + (-0{,}5889) \cdot (-0{,}5889) \cdot 1 + 1 \cdot 0{,}7489 \cdot 0{,}7489]$$

$$= 1 + 0{,}3738 + 0{,}3738 - (0{,}7184 + 0{,}3468 + 0{,}5609) = 1{,}7476 - 1{,}6261 = 0{,}1215$$

Interpretation:
Since the determinant deviates significantly from zero, there is no multicollinearity.

c) What is the correlation between $\hat{\beta}_2$ and $\hat{\beta}_4$? What does it mean in terms of the values of the two regression coefficients and their significance?

Solution:

Correlation coefficient of the regression coefficients $\hat{\beta}_2$ and $\hat{\beta}_4$:

$$r_{\hat{\beta}_2 \hat{\beta}_4} = -r_{24} = -(-0{,}848) = 0{,}848$$

Situation 1:

The higher β_2 is estimated, the higher β_4 is estimated due to multicollinearity (rule of thumb: $|r_{24}| > 0{,}8$):

$\hat{\beta}_2 \uparrow$ (overestimation) → $\hat{\beta}_4 \uparrow$ (overestimation)

While the inflation of the variance estimators for Var($\hat{\beta}_2$) and Var($\hat{\beta}_4$) makes it more difficult to prove the significance of the regression coefficients $\hat{\beta}_2$ and $\hat{\beta}_4$, their overestimation has the opposite effect. This reduces the "inflation effect", but does not necessarily completely eliminate it.

Situation 2:

The lower β_2 is estimated, the lower β_4 is estimated due to multicollinearity (rule of thumb: $|r_{24}| > 0{,}8$

$\hat{\beta}_2 \downarrow$ (underestimation) → $\hat{\beta}_4 \downarrow$ (underestimation)

In this case, the regression coefficients $\hat{\beta}_2$ and $\hat{\beta}_4$ not only the inflation of their estimated variances, but also the negative correlation between X_2 and X_4 could become insignificant.

4.6 The application of the auxiliary regression method yields the following estimated results in Energy Model III:

$$\hat{GASP}_t = 0{,}912 + 0{,}807 \cdot FERNWP_t - 0{,}00050 \cdot VEINK_t, \quad R_2^2 = 0{,}825$$

$$\hat{FERNWP}_t = 0{,}554 + 0{,}420 \cdot GASP_t + 0{,}00006 \cdot VEINK_t, \quad R_3^2 = 0{,}572$$

$$\hat{VEINK}_t = 2244{,}304 - 1162{,}332 \cdot GASP_t + 276{,}269 \cdot FERNWP_t. \; R_4^2 = 0{,}723.$$

Make a diagnosis of multicollinearity using tolerance coefficients and variance inflation factors (VIF)!

Solution:

1. Auxiliary regression: GASPR (X2) is explained by DHC (X3) and DIINC (X4)

Coefficient of tolerance:

$tol_2 = 1 - R_2^2 = 1 - 0{,}825 = 0{,}175 > 0{,}10$ → no serious multicollinearity

Variance inflation factor:

$VIF_2 = 1 / tol_2 = 1 / 0{,}175 = 5{,}714$ → Inflating the variance of $\hat{\beta}_2$ by a factor of 5.714 compared to the case of a missing correlation between X_2 with X_3 and X_4.

2. Auxiliary regression: DHC (X3) is explained by GASPR (X2) and DIINC (X4).

Coefficient of tolerance:

$tol_3 = 1 - R_3^2 = 1 - 0{,}572 = 0{,}428 > 0{,}10$ → no serious multicollinearity

Variance inflation factor:

$VIF_3 = 1 / tol_3 = 1 / 0{,}428 = 2{,}336$ → Inflating the variance of $\hat{\beta}_3$ by a factor of 2,336 compared to the case of a missing correlation between X_3 with X_2 and X_4.

3. Auxiliary regression: DIINC (X4) is explained by GASPR (X2) and DHC (X3)

Coefficient of tolerance:

$tol_4 = 1 - R_4^2 = 1 - 0{,}723 = 0{,}277 > 0{,}10$ → no serious multicollinearity

variance inflation factor:

$VIF_4 = 1 / tol_4 = 1 / 0{,}277 = 3{,}610$ → Inflating the variance of $\hat{\beta}_4$ by a factor of 3.610 compared to the case of a missing correlation between X_4 with X_2 and X_3.