4. Multicollinearity

4.1 The concept of multicollinearity

- Natural Science:
Experimentor can determine the values of the independent variables (independence, no correlation)

- Economics:
Values of the independent variables are observed (dependency, correlation)

Difficulty in economics:

Because of the fact that the independent variables are correlated, the measurement of their isolated influences on the dependent variable is hampered. A variable which is correlated with other variables includes in part the influence of the other variables. When the independent variables in a regression model are correlated with each other, we speak about **multicollinearity**.

In the event of multicollinearity, the total influence of the independent variables on the dependent variable can be accurately captured. But the attributability of the total influence is not unambiguously possible because of the mutual inter-dependencies of the independent variables.

Limiting case: Perfect Multicollinearity

Perfect multicollinearity, meaning a linear dependency between the regressors, would lead to a rank of the observation matrix **X** of less than k, implying that the product matrix **X'X** can not be inverted. In this extreme case the OLS estimator $\hat{\boldsymbol{\beta}}$ cannot be calculated.

Let us now consider the econometric model

$$(4.1) \qquad y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

with the pseudo variable x_{1t}= 1 and the two influencing variables x_2 and x_3. If the two regressors x_2 and x_3 are exactly linearly linked with each other,

$$(4.2) \qquad x_{3t} = \gamma x_{2t}$$

then the coefficients β_1, β_2, and β_3 can no longer be unambiguously determined because the inverse of the product matrix **X'X** does not exist. However, since a perfect interrelation between x_2 and x_3 is on hand, i.e. the correlation coefficient r_{23} is equal to 1, it is sufficient to draw on only one of the two exogenous variables to explain y. Substituting (4.2) in (4.1), we obtain the new regression equation

$$(4.3) \qquad y_t = \beta_1 x_{1t} + (\beta_2 + \gamma\beta_3) x_{2t} + u_t$$

which can be estimated by the OLS method without any difficulties. However, when performing such an estimation, only the total influence $\beta_2 + \gamma\beta_3$ of the two influencing factors x_2 and x_3 can be determined. It cannot be separated into the different individual influences of β_2 and β_3 respectively, so that one therefore describes both parameters as **not identifiable.**

From a technical estimation point of view, this means that instead of a regression plane only a regression line can be determined in three-dimensional space. If instead of (4.2) a linear dependency between the regressors x_2 and x_3 of the form

$$x_{3t} = \delta + \gamma \cdot x_{2t} \tag{4.4}$$

would exist, the estimable regression equation would then be

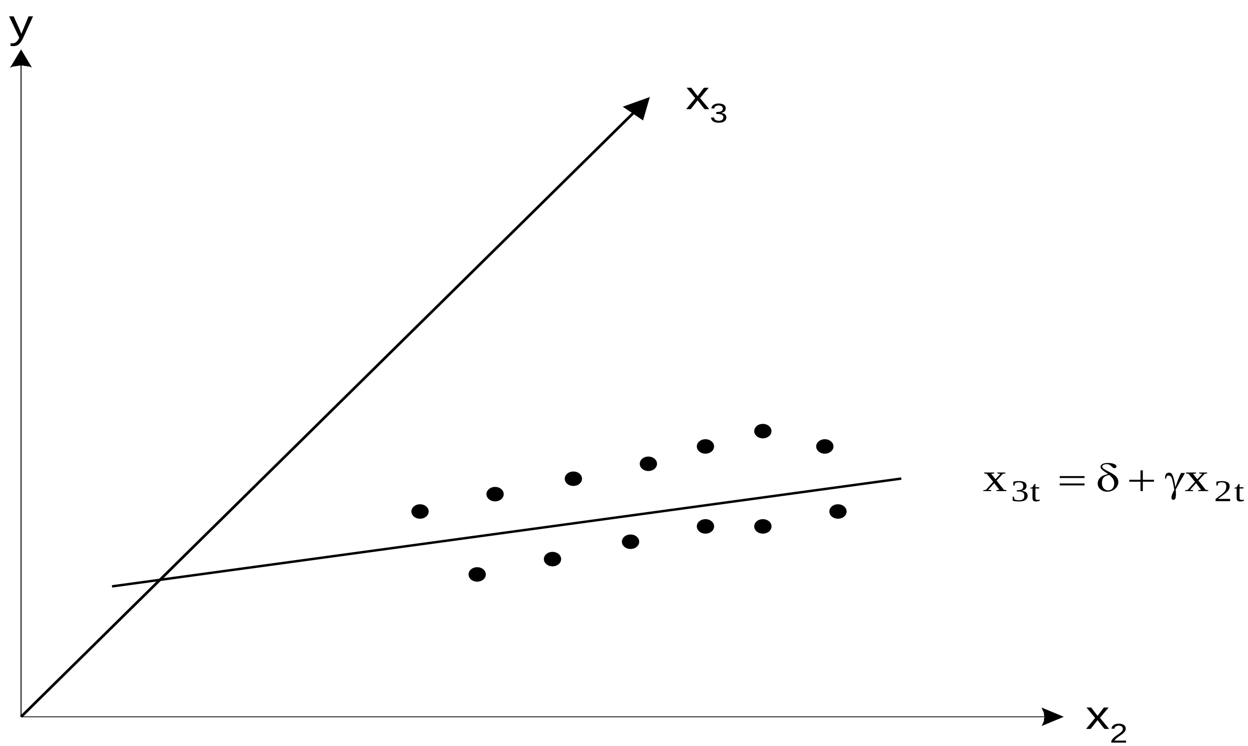
$$y_t = (\beta_1 + \delta\beta_3)x_{1t} + (\beta_2 + \gamma\beta_3)x_{2t} + u_t \tag{4.5}$$

In this case, the intercept β_1 would also no longer be identifiable.

Occurrence of perfect multicollinearity:
- because of definitional interrelations
- because of erroneous definition of dummy variables

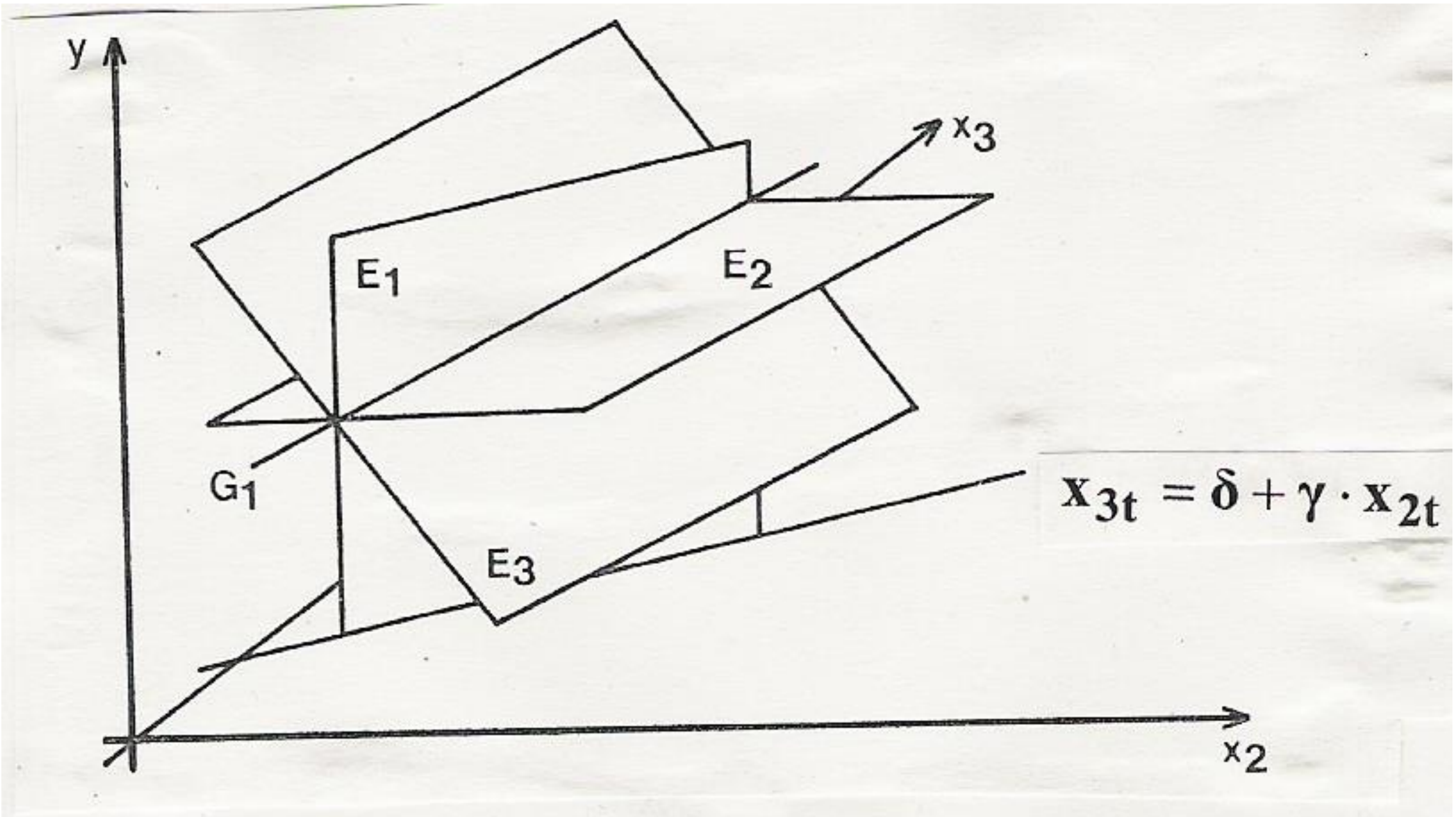
In **econometric practice**, with a correct model specification, perfect multicollinearity does normally not occur, but **imperfect multicollinearity** is still possible, emerging from high, but different from +1 or -1, correlations between the independent variables.

Figure: Imperfect multicollinearity



The estimate of the regression plane in the space spanned by the variables y, x_2 and x_3 is uncertain. Because with increasing proximity of the projections of the observed values to the straight line, the regression plane, which adjusts optimally to the three-dimensional cloud of points, translates more and more into a regression line within the three-dimensional space. When all projected observations lie on the straight line (4.4), the location of the regression plane is completely indeterminate. By contrast, the more the observations disperse within the x_2, x_3-plane (= stronger tendency towards independence), the wider is the basis for determining a regression plane in three-dimensional space.

Fig .: Perfect multicollinearity



4.2 Consequences of multicollinearity

In the event of multicollinearity - resulting from the statistical interdependencies between the exogenous variables - the OLS–estimator $\hat{\boldsymbol{\beta}}$ can still be calculated. However it can be shown that the variances of the estimated regression coefficients will be large, involving a loss of precision (accuracy) of the estimates.

To show this on the basis of the regression model (4.1), we use for the sake of simplicity the deviations of the observed values from their means:

(4.6) $$\tilde{y}_t = \beta_2 \tilde{x}_{2t} + \beta_3 \tilde{x}_{3t} + u_t$$

with $\tilde{y}_t = y_t - \bar{y},\ \tilde{x}_{2t} = x_{2t} - \bar{x}_2$ und $\tilde{x}_{3t} = x_{3t} - \bar{x}_3$

Using

$$\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{x}_{21} & \tilde{x}_{31} \\ \tilde{x}_{22} & \tilde{x}_{32} \\ \vdots & \vdots \\ \tilde{x}_{2n} & \tilde{x}_{3n} \end{bmatrix} \qquad \tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \begin{bmatrix} \sum \tilde{x}_{2t}^2 & \sum \tilde{x}_{2t}\tilde{x}_{3t} \\ \sum \tilde{x}_{2t}\tilde{x}_{3t} & \sum \tilde{x}_{3t}^2 \end{bmatrix} \qquad \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{adj} = \begin{bmatrix} \sum \tilde{x}_{3t}^2 & -\sum \tilde{x}_{2t}\tilde{x}_{3t} \\ -\sum \tilde{x}_{2t}\tilde{x}_{3t} & \sum \tilde{x}_{2t}^2 \end{bmatrix}$$

$$\left|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right| = \sum \tilde{x}_{2t}^2 \sum \tilde{x}_{3t}^2 - \left(\sum \tilde{x}_{2t}\tilde{x}_{3t}\right)^2$$

one obtains for the OLS-estimator

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \qquad \text{with} \quad \tilde{\mathbf{y}} = (\tilde{y}_1 \tilde{y}_2 \ldots \tilde{y}_n)$$

the variance-covariance matrix

$$\text{(4.7)} \qquad \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} = \frac{\sigma^2}{\sum\tilde{x}_{2t}^2\sum\tilde{x}_{3t}^2 - \left(\sum\tilde{x}_{2t}\tilde{x}_{3t}\right)^2}\begin{bmatrix}\sum\tilde{x}_{3t}^2 & -\sum\tilde{x}_{2t}\tilde{x}_{3t} \\ -\sum\tilde{x}_{2t}\tilde{x}_{3t} & \sum\tilde{x}_{2t}^2\end{bmatrix}.$$

Following this the variances of the estimated regression coefficients $\hat{\beta}_2$ *and* $\hat{\beta}_3$ are given by:

$$\mathrm{Var}\left(\hat{\beta}_2\right) = \frac{\sigma^2\sum\tilde{x}_{3t}^2}{\sum\tilde{x}_{2t}^2\sum\tilde{x}_{3t}^2 - \left(\sum\tilde{x}_{2t}\tilde{x}_{3t}\right)^2} \quad \text{and} \quad \mathrm{Var}\left(\hat{\beta}_3\right) = \frac{\sigma^2\sum\tilde{x}_{2t}^2}{\sum\tilde{x}_{2t}^2\sum\tilde{x}_{3t}^2 - \left(\sum\tilde{x}_{2t}\tilde{x}_{3t}\right)^2}$$

One achieves a suitable interpretation of multicolliniarity when one inserts in these expressions the correlation coefficient r_{23} between x_2 and x_3 which is defined by

$$r_{23} = \frac{\sum\tilde{x}_{2t}\tilde{x}_{3t}}{\sqrt{\sum\tilde{x}_{2t}^2\sum\tilde{x}_{3t}^2}}$$

After squaring it follows herefrom

$$\left(\sum \tilde{x}_{2t}\tilde{x}_{3t}\right)^2 = r_{23}^2\left(\sum \tilde{x}_{2t}^2 \sum \tilde{x}_{3t}^2\right)$$

so that the variances of $\hat{\beta}_1$ *and* $\hat{\beta}_2$ are presentable in the form of

(4.8) $$\mathrm{Var}\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\left(1-r_{23}^2\right)\sum \tilde{x}_{2t}^2}$$

and

(4.9) $$\mathrm{Var}\left(\hat{\beta}_3\right) = \frac{\sigma^2}{\left(1-r_{23}^2\right)\sum \tilde{x}_{3t}^2}$$

In the event of multicollinear regressors x_2 and x_3, r_{23} tends towards 1 so that the denominator of (4.8) and (4.9) will become small and the variances of the estimated regression coefficients will become large. Significance of a single regression coefficient is then more difficult to prove. The overall modell may also in this case be statistically significant even though the estimated regression coefficients are insignificant.

In addition, it can be shown that the correlation between the regression coefficients $\hat{\beta}_2$ *and* $\hat{\beta}_3$ amounts to

$$r_{\hat{\beta}_2,\hat{\beta}_3} = -r_{23}$$

This means that an overestimation of $\hat{\beta}_2$ is accompanied by an underestimation of $\hat{\beta}_3$ and vice versa.

4.3 Detection of multicollinearity

Techniques for detecting multicollinearity

- Simple and extended correlation analysis

- Determinant

- Condition number

- Auxiliary regressions

Simple and extended correlation analysis

- Simple correlation analysis

Detection of multicollinearity on the basis of the pairwise correlation coefficient r_{23}, r_{24}..., $r_{k-1, k}$,

Rule of thumb: Multicollinearity if $r_{jk} > 0.8$

The method is sufficient for the simple regression analysis but does not reveal interdependencies between several influencing factors.

- Extended correlation analysis

Assessment of multicollinearity on the basis of the multiple correlation coefficient $R_{yx2, \ldots xk}$:

Rule of thumb: Multicollinearity if $r_{jk} > R_{yx2, \ldots xk}$

No lifting (repeal) of the restrictions applying to the simple correlation analysis, but objectification of the critical value

- The determinant of the product matrix X'X

Multicollinearity means that two or more regressors are approximatively linearly dependent. If the columns of the observation matrix **X** are approximately linearly dependent, the product matrix **X'X** becomes almost singular. Although the inverse ($\mathbf{X'X}$)$^{-1}$ required for the estimation of $\hat{\boldsymbol{\beta}}$ can still be calculated, with an increasing degree of linear dependence the estimation becomes however unstable.

With perfect multicollinearity the determinant of the product matrix **X'X** is equal to 0, so that a value of |**X'X**| close to 0 can be taken to indicate multicollinearity:

$$|\mathbf{X'X}| \approx 0 \rightarrow \text{Multikollinearität}$$

[The same applies to the correlation matrix **R** that is obtained from the standardized observation matrix **Z**: **R** = **Z'Z** / n.]

- Condition number of the product matrix X'X

The condition number C indicates how close a matrix is to a singular matrix.

$$C = \sqrt{\lambda_{max} / \lambda_{min}}$$

λ_{Max}: Largest eigenvalue of **X ' X**. λ_{min}: Smallest eigenvalue of **X ' X**

With an increasing degree of multicollinearity the ratio $\lambda_{Max} / \lambda_{min}$ rises

Rule of thumb: Multicollinearity if C> 30

- Approach of auxiliary regressions

The approach of the auxiliary regressions implies that for each of the k-1 (real) exogenous variables, a regression on the remaining k-2 regressors is carried out:

$$\begin{array}{lcllll}
\hat{x}_{2t} & = & \beta_1^2 x_{1t} & + \beta_3^2 x_{3t} & + \cdots & + \beta_k^2 x_{kt} \\
\hat{x}_{3t} & = & \beta_1^3 x_{1t} & + \beta_2^3 x_{2t} & + \cdots & + \beta_k^3 x_{kt} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\hat{x}_{kt} & = & \beta_1^k x_{1t} & + \beta_2^k x_{2t} & + \cdots & + \beta_{k-1}^k x_{k-1,t} \,.
\end{array}$$

For each of these auxiliary regressions the coefficient of determination R_j^2 is calculated, wherein the index j indicates that x_j acts as the dependent variable. Multicollinearity can be diagnosed on the basis of a

- F-test or
- Tolerance coefficients and variance inflation factors.

a) F-test of the auxiliary regressions

A high R_j^2 indicates a linear statistical dependence of the variable x_j with the other regressors. The significance of such a relationship can be examined with an F-test, whereby its test statistic is given by:

$$F_j = \frac{R_j^2 / (k-2)}{(1-R_j^2)/(n-k+1)}, j = 2,3,\ldots,k \tag{4.10}$$

Test decision:

$F_j > F_{k-2;n-k+1;1-\alpha}$ → Significant linear statistical relationship between x_j and the other regressors

In the case of a significant test result, an existence of multicollinearity is ascertained. The approach identifies at the same time those independent variables through which multicollinearity is generated.

[Auxiliary regressions can be conducted in the same manner for subsets of the exogenous variables with appropriately modifying the test statistic.]

b) Tolerance coefficients and variance inflation factors

With a more descriptively oriented approach to the use of auxiliary regressions, multicollinearity can be determined on the basis of tolerance coefficients and variance inflation factors. The **tolerance coefficients** tol_j are defined by:

(4.11) $tol_j = 1 - R_j^2$, $j = 2,3, \ldots , k$

Here low values of the tolerance coefficients indicate multicollinearity.

<u>Rule of thumb</u>: Multicollinearity if $tol_j < 0.10$

Using the tolerance coefficients tol_j, **variance inflation factors** can be formed:

(4.12) $$VIF_j = \frac{1}{tol_j}$$. $j = 2,3, \ldots , k$

from which the multicollinearity induced inflation of the variances of the estimated regression coefficients emerge (by rule of thumb: multicollinearity if $VIF_j > 10$). More specifically, the indicator VIF_j indicates the increase of the variance $Var(\hat{\beta}_j)$, which in turn can be traced back to the variable x_j not being uncorrelated with the other exogenous variables of the regression model, butforming a collinear relationship with these, as measured by the coefficient of determination R_j^2.

Example:

With the money demand function (2.28), the logarithm of money demand (ln m) is explained by the, on a logarithmic scale measured, exogenous variables income (ln y) and interest rate (ln r). We now deploy the previously described approaches to the detection of multicollinearity:

- **Simple and expanded correlation analysis**

  Correlation coefficient r_{23} between ln y and ln r: -0.689

  $|r_{23}| = 0.689 < 0.8$ → no serious multicollinearity

  $|r_{23}| = 0.689 < R_{yx2x3} = 0.984$ → no serious multicollinearity

- **Determinant of the product matrix X ' X**

$$\underset{3x3}{|\mathbf{X'X}|} = \det\begin{bmatrix} n & \Sigma \ln y_t & \Sigma \ln r_t \\ \Sigma \ln y_t & \Sigma (\ln y_t)^2 & \Sigma \ln y_t \cdot \ln r_t \\ \Sigma \ln r_t & \Sigma \ln r_t \cdot \ln y_t & \Sigma (\ln r_t)^2 \end{bmatrix}$$

$$= \det\begin{bmatrix} 19{,}000 & 145{,}986 & 17{,}779 \\ 145{,}986 & 1121{,}937 & 135{,}657 \\ 17{,}779 & 135{,}657 & 23{,}963 \end{bmatrix} = 18{,}776 >> 0 \rightarrow \text{no serious multicoll.}$$

- **Determinant of the correlation matrix R**

$$|\mathbf{R}| = \det\begin{bmatrix} 1 & 0{,}689 \\ 0{,}689 & 1 \end{bmatrix} = 1 - 0{,}689^2 = 0{,}5253$$ >> 0 → no serious multicollinearity

- **Method of auxiliary regressions**

In the approach using auxiliary regressions running *one* regression is sufficient, since in terms of determination it does not matter whether ln y is regressed on ln r or whether ln r is regressed on ln y.

Auxiliary regression: $\widehat{\ln y} = 7{,}804 - 0{,}129 \cdot \ln r$ $\qquad R_2^2 = 0.475$

- F test

Test statistic (n = 19, k = 3):

$$F_2 = \frac{R_2^2/(k-2)}{(1-R_2^2)/(n-k+1)} = \frac{0{,}475/(3-2)}{(1-0{,}475)/(19-3+1)} = \frac{0{,}475}{0{,}525/17} = 15{,}381$$

Critical value (α= 0.05): $F_{1;\,17;\,0.95} = 4.43$

Testing decision:
$F_2 = 15.381 > F_{1;\,18;\,0.95} = 4.43$ → Significant statistical linear relationship between x_2 and x_3

→ Indication of multicollinearity

- Tolerance coefficient and variance inflation factor

Tolerance coefficient:

$tol_2 = 1 - R_2^2 = 1 - 0.475 = 0.525 > 0.10 \rightarrow$ no serious multicollinearity

Variance inflation factor:

$$VIF_2 = \frac{1}{tol_2} = \frac{1}{0{,}525} = 1{,}904$$

Compared to a situation of uncorrelated regressors, due to the weak linear statistical dependence between x_2 and x_3, the variance of the OLS–estimator $\hat{\beta}_2$ (and hence also of that of the OLS–estimator $\hat{\beta}_3$) is inflated by the factor 1,904. This effect of reducing the precision of the estimate has to be accepted, if the existence of multicollinearity is neglected.

4.4 Remedying multicollinearity

While the problem of multicolliniarity for **forecasting purposes** is not necessarily a troublesome issue, since the focus is on the overall effect of the explanatory variables on the variable to be forecasted, it is often desirable to eliminate or at least reduce multicollinearity in **econometric structural models.**

Methods for remedying multicollinearity

- **Extending of the database**

  By extending the database, a greater range of variation of the data is to be expected, which tends to reduce multicollinearity.

  Cross-sectional data:
  Increasing sample size may perhaps be possible for cross-sectional units such as consumers, households, enterprises, yet not for spatial units such as regional labor markets or administrative regions.

  Time series data:
  Inclusion of data from further back in the past carry with it the risk of structural breaks.

  Panel data:
  By combining cross section with time series data, the spread in the variation of the data is often considerably increased, thereby mitigating substantially the problem of multicollinearity.

- **External information**

Suppose the demand for a product group, in dependence of income and the price level, is investigated econometrically by using time-series data. As income is often correlated with the price level multicollinearity is to be expected. Now, if the magnitude of the influence of income on the demand for the product group is known, for example from a cross–sectional survey, it is possible to revert to this **external information** to econometrically estimate the separate influence of the price level on demand.

Problem:
Within the framework of a **time series analysis** a **long-term response coefficient** of income with respect to demand could be looked for, while from the **cross section analysis** only a short-term reaction coefficient can be determined.

- **Method of variable suppression**

If, for example, two variables are highly correlated, one of the two variables captures already most of the influence of the other variable, so suppressing one variable by and large does not reduce explanatory power. Multicollinearity will thereby not only be reduced, but even completely eliminated.

Problem:
Misspecification of the regression equation (error term captures the systematic)

- **Method of differences**

Using the **method of differences** it is assumed that multicollinearity in econometric models shows up in a trend in the exogenous variables. If an approximately linear trend can be presumed, it can simply be eliminated by taking differences. For two exogenous variables (without pseudo variable) the regression equation for the period t is given by:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

If one hereof subtracts the regression equation for the previous period t-1,

$$y_{t-1} = \beta_1 + \beta_2 x_{2,t-1} + \beta_3 x_{3,t-1} + u_{t-1}$$

we obtain the regression model

(4.13) $$\Delta y_t = \beta_2 \Delta x_{2t} + \beta_3 \Delta x_{3t} + \Delta u_t$$

whith e.g. $\Delta y_t = y_t - y_{t-1}$. To the extent that we succeed in eliminating the trend from the exogenous variables x_2 and x_3, multicollinearity will be reduced compared to the original model with the level variables.

Problem:
- Reduction of the dispersion of the regressors by taking differences;
- Autocorrelation of disturbance variables - increasing the disturbance variance

- **Principal component regression**

The principal component regression is a multivariate method based on principal component analysis. With the principal component analysis the exogenous variables of a regression model are transformed into uncorrelated artificial variables (= principal components). The principal components arise from linear combinations of the regressors. Frequently just a few principal components explain most of the variance of the regressors.

Inserting these uncorrelated principal components instead of the observable exogenous variables into the regression equation, the corresponding regression coefficients can be estimated in a clear-cut manner. If the principal components are economically interpretable, the economist can just turn his interest to the reaction coefficients. In general, however, the interest will continue to focus on the effects of the original variables. In this case, the estimated regression coefficients of the principal components have to be transformed back.

<u>Problem:</u>
- The estimated regression coefficients will be affected by the number of employed principal components.
- The transformation back involves a weighting problem which is likewise also reflected in the estimated regression coefficients.