Matematické Metody v Ekonometrii 7.

Multicollinearity

Blanka Šedivá

KMA

zimní semestr 2016/2017

- One of the assumptions of the classical and normal regression models is that columns of $\boldsymbol{X}$ are linearly independent (i.e. $\boldsymbol{X}$ is a matrix with full rank).
- The so called multicollinearity is a high dependency of columns of the matrix $\boldsymbol{X}$ is almost singular and consequently it is problematic to find its inverse.
- Multicollinearity can be caused by adding polynomial terms or other regressor derived from already existing regressors.
- Another causes might be including too many variables in the model when some of them measures the same conceptual variable or wrong data collection procedure.

Gauss Markov theorem for normally distributed disturbances

Model $\boldsymbol{Y} \sim N\left(\boldsymbol{X}\beta, s^2\boldsymbol{I}\right)$,

(i) $\boldsymbol{b} \sim N\left(\beta; \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$;

(ii) $\dfrac{RSE}{\sigma^2} = \dfrac{s^2(n-p)}{\sigma^2} \sim \chi^2$ with $\nu = n - p$ degree of free (df)

(iii) $\boldsymbol{b}$ a RSE are independent

(iv) $\mathrm{E}\left(\boldsymbol{a}^T\boldsymbol{b}\right) = \boldsymbol{a}^T\beta$; where $\boldsymbol{a} = (a_0, a_1, \ldots, a_k)^T \neq \boldsymbol{0}$

(v) $\mathbf{Var}\left(\boldsymbol{a}^T\boldsymbol{b}\right) = s^2\boldsymbol{a}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{a}$;

(vi) $T = \dfrac{\boldsymbol{a}^T\boldsymbol{b} - \boldsymbol{a}^T\beta}{\sqrt{s^2\boldsymbol{a}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{a}}} \sim$ t-distribution with $\nu = n - p$ df.

The mean squared error (MSE) or mean squared deviation (MSD) of an estimator

- The MSE of an estimator $\hat{\theta}$ with respect to an unknown parameter θ is defined as

$$MSE(\hat{\theta}) = \mathrm{E}\left(\hat{\theta} - \theta\right)^2 = \mathbf{Var}\,\hat{\theta} + \left(\mathrm{E}\,\hat{\theta} - \theta\right)^2 = \mathbf{Var}\,\hat{\theta} + \mathrm{Bias}\left(\hat{\theta}, \theta\right)^2$$

- for model $\boldsymbol{Y} \sim \left(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}\right)$ are given

$$\mathrm{E}\left(\widehat{Y}^T \widehat{Y}\right) = \boldsymbol{Y}^T \cdot \boldsymbol{Y} + \sigma^2 \cdot \mathrm{rank}\left(\boldsymbol{X}\right)$$

and if $\boldsymbol{X}$ has linear independent columns

$$\mathrm{E}\left(\boldsymbol{b}^T \boldsymbol{b}\right) = \boldsymbol{\beta}^T \cdot \boldsymbol{\beta} + \sigma^2 \cdot \mathrm{tr}\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$$

Consequences of multicollinearity

- Small changes in $\boldsymbol{X}$ result in large changes in estimations of $\boldsymbol{\beta}$ (i.e. the OLS procedure is ill-conditioned).
- The standard errors of estimated coefficients tend to be large and therefore they often seem to be statistically insignificant despite a high value of R^2 and high significance of the whole model.
- The estimated coefficient can have wrong sign or unexpected values not corresponding with economical interpretation of the model.

Note: The important sign of multicollinearity is also a fact that increasing number of observation neither reduces standard errors of estimations nor helps to eliminate the other problems mentioned above.

Detection of multicollinearity - pair correlation coefficients

The basic approach is based on pair correlation coefficients between columns of $\boldsymbol{X}$

- Compute $\mu_{ij} = \mathrm{cor}\,(\boldsymbol{X}_i, \boldsymbol{X}_j)$, $i, j = 1, 2, \ldots, k$ for all pairs of columns.
- We can use several rules of thumb to test the multicollinearity. We say that the multicollinearity is present in our model if:
  - There exists $|\mu_{ij}| > 0.75$ (some literature suggest 0.8 or even 0.9).
  - There exists $|\mu_{ij}| \geq R^2$, where R^2 , the coefficient of determination of the regression model.
- This method is not very effective when the dependency is generated by three and more columns.

Detection of multicollinearity - auxiliary regressions

- Perform auxiliary regression of the column

$$X_j = X_1\alpha_1 + \ldots + X_{j-1}\alpha_{j-1} + X_{j+1}\alpha_{j+1} + \ldots + X_k\alpha_k + \varepsilon_j$$

- R_j^2 is the coefficient of determination of the corresponding auxiliary regression.
- Again, we can use several rules of thumb to test the multicollinearity. We say that the multicollinearity is present in our model if:
  - there exists $R_j^2 > R^2$
  - there exists $VIF_j = \frac{1}{1-R_j^2} > 10$ We call VIF_j variation inflation factor (of the regressor j) and it quantifies the severity of the multicollinearity's influence on the standard error of coefficient b_j The VIF_j are diagonal elements inverse matrix of correlations $diag(\mathrm{cor}\,(\boldsymbol{X})^{-1}) = VIF$.
  - The test statistics

$$F_j = \frac{R_j^2}{1-R_j^2} \cdot \frac{n-p}{p-1} \sim F_{\nu_1,\nu_2}$$

exceeds its critical value.

Solution of multicollinearity

We have several options what to do when we detect multicollinearity:

(i) Normalise or transform columns of $\boldsymbol{X}$ such that the multicollinearity is eliminated.

(ii) Select a submodel such that the regressors which cause multicollinearity are omitted.

(iii) Use transformed regressors which are linear combinations of the original regressors (the so called **principal component regression** (PCR)).

(iv) Use the so called **ridge regression**.

Ridge regression

- Consider a linear regression model $\boldsymbol{Y} \sim \left(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}\right)$, with diagnosed multicollinearity.
- Because multicollinearity causes the matrix $\boldsymbol{X}^T\boldsymbol{X}$ to be ill-conditioned, we use ridge regression estimator

$$\boldsymbol{b}_\delta = \left(\boldsymbol{X}^T\boldsymbol{X} + \delta \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T\boldsymbol{Y},$$

  where $\delta \geq 0$
- relation between $\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and $\boldsymbol{b}_\delta$ can be expressed as

$$\begin{aligned}\boldsymbol{b}_\delta &= \left(\boldsymbol{X}^T\boldsymbol{X} + \delta \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \\ &= \left(\boldsymbol{X}^T\boldsymbol{X} + \delta \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T\boldsymbol{X} \cdot \boldsymbol{b} = \\ &= \left[\boldsymbol{I} + \delta \cdot \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right]^{-1} \cdot \boldsymbol{b}\end{aligned}$$

Ridge regression - statistic properties of $\boldsymbol{b}_\delta$

- Obviously, ridge regression estimator is $\boldsymbol{b}_\delta$ biased which is not a desired property. However, it can be shown that under certain conditions the ridge regression estimator can be somewhat favourable
- but it can be shown that for $0 < \delta < 2\sigma^2 \cdot \frac{1}{\|\boldsymbol{\beta}\|^2}$ it holds that

$$MSE(\boldsymbol{b}) \geq MSE(\boldsymbol{b}_\delta)$$

- In practice, however, we do not know the real values of $\boldsymbol{\beta}$ and σ^2, so we are not able to directly determine the value of δ.
- Hence, the usual choice is $\delta_1 = k \cdot \frac{s^2}{\boldsymbol{b}^T\boldsymbol{b}} = k \cdot \frac{s^2}{\sum_{j=1}^n b_j^2}$, or we could employ values $\delta \in (0, \delta_{\max})$ and plot $\boldsymbol{b}_\delta$ to create the so called ridge trace. The desired value of δ is where $\boldsymbol{b}_\delta$ stabilise.

Choice of the submodel

- overspecification and underspecification of model
- including irrelevant variables in a regression model
  - „true model“ $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, $\mathrm{hod}(\boldsymbol{X}) = k$ and
  - „false“ selected model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \varepsilon_1$, $\mathrm{hod}([\boldsymbol{X}\boldsymbol{X}_2]) = k^* > k$.
  - the estimations $\boldsymbol{b}$ are unbias
  - but the variances of the OLS this estimators higher
  - there is also risk of multicollinearity
- misspecification of model
  - „true model“ $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, $\mathrm{rank}\,(\boldsymbol{X}) = k$ a $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \varepsilon$ and
  - „false“ selected model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$, $\mathrm{rank}\,(\boldsymbol{X}_1) = k_1 < k$
  - the estimations $\boldsymbol{b}$ are bias $\mathrm{E}\,(b_1) = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2$.

Model and submodel

- relationship between model $Y \sim N(\boldsymbol{X}\boldsymbol{\beta}; \sigma^2 \boldsymbol{I})$, $\mathrm{hod}(\boldsymbol{X}) = k$ and submodel with $\mathrm{rank}\,(\boldsymbol{X}) = k_0$
- the estimated parameters based on submodel we denote $\boldsymbol{b}_0$, s_0^2, $\boldsymbol{e}_0 = \widehat{\boldsymbol{y}_0} - \boldsymbol{y}$ and $RSE_0 = \boldsymbol{e}_0^T \boldsymbol{e}_0$,
- than it is hold

$$F_0 = \frac{(RSE_0 - RSE)\,/\,(k - k_0)}{RSE/\,(n-k)} \sim F_{\nu_1,\nu_2},$$

where $\quad \nu_1 = k - k_0$ and $\nu_2 = n - k$

Choice of the model - critters

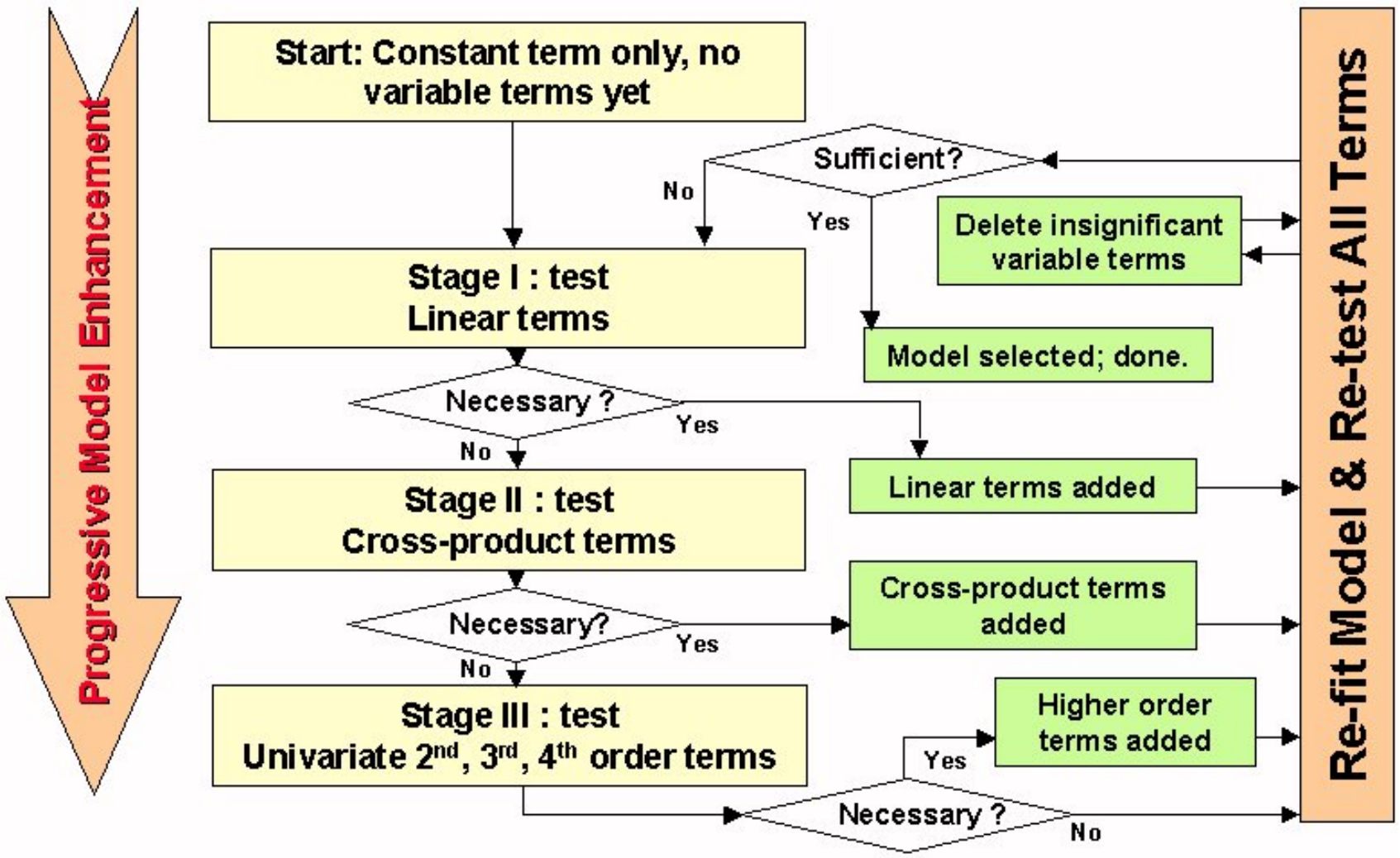
There are many different critters derived for the choice of the model:

- Minimization of the residual sum of squares $RSE = \boldsymbol{e}^T\boldsymbol{e}$
- Maximization of the coefficient of determination $R^2 = 1 - \frac{\boldsymbol{e}^T\boldsymbol{e}}{\boldsymbol{y}^T\boldsymbol{y}}$
- Maximization of the adjusted coefficient of determination $R^2_{adj} = 1 - \frac{(\boldsymbol{e}^T\boldsymbol{e})/(n-k)}{(\boldsymbol{y}^T\boldsymbol{y})/(n-1)}$
- Minimization of the residual variance $s^2 = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-k}$
- Maximization of Mallows C_k

$$C_k = \frac{RSS_0}{s^2} + 2 \cdot k_0 - n,$$

- Minimization of an information criterion such as
  - $AIC = \ln s^2 + 2k$ ... (Akaike)
  - $A = s^2\left(1 + k \cdot n^{-1/4}\right)$ ... (Anděl a kol.)
  - $SR = \ln s^2 + k \cdot \frac{\ln n}{n}$ ... (Swarz,Rissanen)
  - $HQ = \ln s^2 + 2 \cdot c \cdot k \cdot \frac{\ln(\ln n)}{n}$, $c = 2$ or 3 ... (Hannan,Quinn)

Stepwise regression

- **Forward selection**, which involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until none improves the model.
- **Backward elimination**, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable (if any) that improves the model the most by being deleted, and repeating this process until no further improvement is possible.
- **Bidirectional elimination**, a combination of the above, testing at each step for variables to be included or excluded.

Stepwise regression