Matematické Metody v Ekonometrii 3.

Multiple Regression Analysis - matrix notation

Blanka Šedivá

KMA

zimní semestr 2016/2017

Matrix notation of linear regression models

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \\ \vdots \\ \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \\ \vdots & \vdots & \ddots & \vdots \\ \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \vdots \\ \\ \varepsilon_n \end{bmatrix},$$

where

- $\boldsymbol{Y} = [Y_1, Y_2, \dots, Y_n]^T$ is column vector of regressand, dependent variable, observation of dependent variable;
- $\boldsymbol{X}$ matrix $n \times (k+1)$ of explanatory variables, regressors, number of columns p and $p = k+1$; (DATA MATRIX)
- $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_p]^T$ is column vector $p = k+1$ of unknown parameters (coefficients);
- $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]^T$ column vector of stochastic disturbances.

Assumptions of the classical matrix notation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(P1) $\mathrm{E}\,(\boldsymbol{\varepsilon}) = \mathbf{0}$;

(P2) $\mathbf{Var}\,(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$; homoscedasticity, non-correlated

(P3) $rank(\boldsymbol{X}) = p = k + 1$ ... no exact linear relationship between columns, $\boldsymbol{X}$ is an matrix with rank p

(P4) $\boldsymbol{X}$ may be fixed or random, but it is generated by a mechanism that is unrelated to $\boldsymbol{\varepsilon}$
$\mathrm{E}\,(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbf{0}$ and $\mathbf{Var}\,(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$;

Assumptions of the normal classical matrix notation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(N1) Normal distribution $\boldsymbol{\varepsilon} \sim N_n\left(0; \sigma^2 \boldsymbol{I}_n\right)$;

(N2) $rank(\boldsymbol{X}) = p$ ... no exact linear relationship between columns, $\boldsymbol{X}$ is an matrix with rank p

(N3) $\boldsymbol{X}$ may be fixed or random, but it is generated by a mechanism that is unrelated to $\boldsymbol{\varepsilon}$

The least square coefficient vector

- Minimazing the sum of square $Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$
- For $\boldsymbol{b}$ to minimize the sum of squared residuals, it must solve the first order condition

$$\left.\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\beta}}\right|_b = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{b} - \boldsymbol{Y}) = \boldsymbol{0}$$

- Least square normal equations $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{Y}$
- If the inverse of $\boldsymbol{X}^T\boldsymbol{X}$ exists, then the solution is $\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$;
- For this solution to minimize the sum of square $\left.\frac{\partial^2 \boldsymbol{Q}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right|_b = 2\boldsymbol{X}^T\boldsymbol{X}$ must be a positive definitive matrix

Numerical Methods for Least Squares Problems

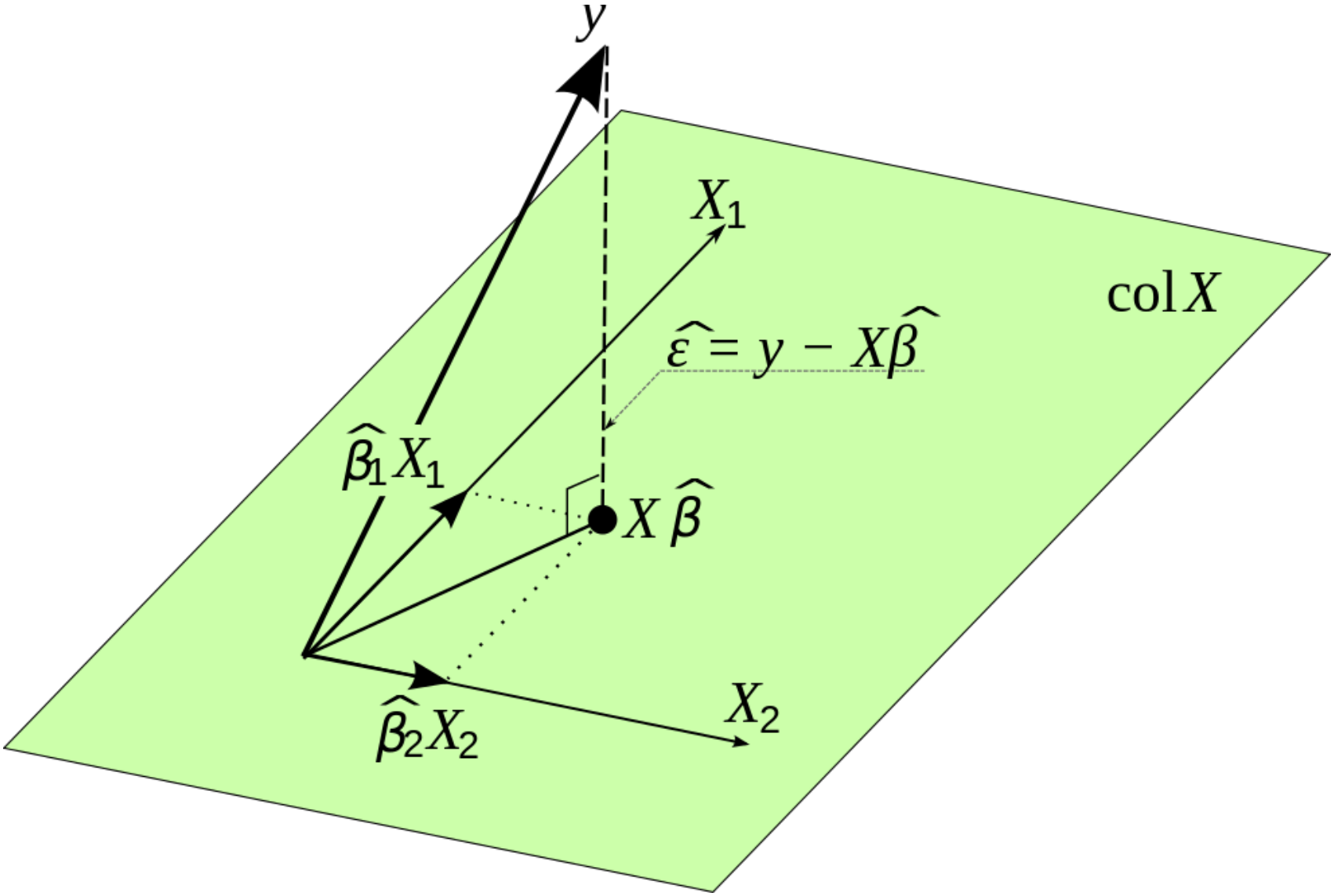
- QR Decomposition
- Consider a matrix $\boldsymbol{X}_{n\times p}$ with $n \geq p$ and $rank(X) = p$ then there exists $\boldsymbol{Q}$ and $\boldsymbol{R}$

$$\boldsymbol{X}_{(n\times p)} = \boldsymbol{Q}_{(n\times r)}\boldsymbol{R}_{(r\times p)}$$

where $\boldsymbol{R}$ is upper triangular matrix and $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$ is orthogonal matrix

$$\begin{aligned}
\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} &= \boldsymbol{X}^T\boldsymbol{y} \\
(\boldsymbol{Q}\boldsymbol{R})^T\boldsymbol{Q}\boldsymbol{R}\boldsymbol{b} &= (\boldsymbol{Q}\boldsymbol{R})^T\boldsymbol{y} \\
\boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R}\boldsymbol{b} &= \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{y} \\
\boldsymbol{R}\boldsymbol{b} &= \boldsymbol{Q}^T\boldsymbol{y} \\
\boldsymbol{b} &= \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{y}
\end{aligned}$$

Geometric interpretation of least squares solutions



https://en.wikipedia.org/wiki/Ordinary_least_squares

Projection matrix

For model $\boldsymbol{Y} \sim \left(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}\right)$ we define projection matrix by

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

Properties of projection matrix

- $\boldsymbol{H}_{n\times n}$
- $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$
- $\boldsymbol{H}^2 = \boldsymbol{H}$,;
- $\mathrm{tr}(\boldsymbol{H}) = \sum \mathrm{h_{ii}} = \mathrm{p}$;
- $\mathrm{rank}(\boldsymbol{H}) = \mathrm{p}$;
- $\boldsymbol{H}^T = \boldsymbol{H}$;
- $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$;
- $\mathbf{Var}\left(\widehat{\boldsymbol{Y}}\right) = \sigma^2\boldsymbol{H}$.

Residual matrix

For model $\boldsymbol{Y} \sim \left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right)$ we define residual matrix

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}.$$

Properties of residual matrix

- $\boldsymbol{e} = \boldsymbol{M}\boldsymbol{\varepsilon}$;
- $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$;
- $\boldsymbol{M}^2 = \boldsymbol{M}$, ;
- $\boldsymbol{M}^T = \boldsymbol{M}$;
- $\mathrm{rank}(\boldsymbol{M}) = \mathrm{n} - \mathrm{p}$;
- $\boldsymbol{M}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{M} = \boldsymbol{0}$
- $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y} + \boldsymbol{M}\boldsymbol{Y} = \textit{projection} + \textit{residual}$
- $\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{Y} + \boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Y} = \widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}} + \boldsymbol{e}^T\boldsymbol{e}$

SW support of regression

- EXCEL
  - add trend line to plot
  - slope(y;x)
  - intercept(y;x)
  - forecast(newx;y;x)
  - lintrend(y;[x];[newx];[b])
  - linregrese(y;[x];[b];[stat]) (matrix function)
  - nástroje/analýza dat/regrese
- MATLAB
  - $b = inv(X' * X) * X' * y$
  - $b = (X' * X) \setminus X' * y$
  - [b,bint,r,rint,stats] = regress(y,X)
  - stats = regstats(y,x,model,whichstats) Statistics and Machine Learning Toolbox
  - [p,S,mu] = polyfit(x,y,n)