3 STATICKÉ MODELY

3.1. Korelačná a regresná analýza

⇒ Vhodná na **vyšetrovanie statických**, ale aj **dynamických systémov**.

⇒ Skúmame funkčný vzťah, podľa ktorého sa mení **závisle premenná y pri zmenách nezávislých veličín** $\mathbf{u} = (u_1, u_2, ..., u_n)^T$.

Predpokladajme, že skutočný funkčný vzťah má nasledovný tvar:

$$\boxed{y = F(\mathbf{u}, \boldsymbol{\theta}) + v}$$

t.j. predpokladáme, že **výsledok experimentu** $y \in R^1$, bude funkciou **systematicky pôsobiacich nezávislých veličín** $\mathbf{u} \in R^n$ (n-faktorový problém), a tiež od **náhodne pôsobiacich faktorov** $v \in R^1$. $\boldsymbol{\theta}$ je **vektor (neznámych) parametrov** a F je **funkcia štruktúry** identifikovaného systému.

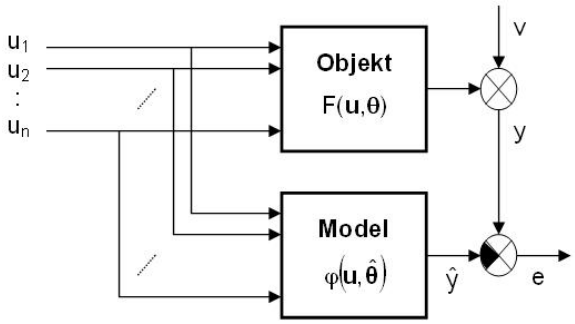
⇒ **Regresná analýza** sa zaoberá **určením štruktúry**, t.j. stanovením, ktoré nezávislé premenné a v akom charaktere väzieb do funkčnej závislosti vstupujú.

Korelačná analýza kvantifikuje **intenzitu vplyvov nezávislých premenných** na závislú premennú (**koeficienty**).

⇒ **Štruktúru** reálneho systému **presne poznať nebudeme**, preto budeme predpokladať, že je vyjadrená v tvare **mocninového radu**, tvoriaceho tzv. **teoretickú regresnú rovnicu**

$$\boxed{y = \theta_0 + \sum_{i=1}^{n} \theta_i u_i + \sum_{i=1}^{n} \sum_{j=1}^{i} \theta_{ij} u_i u_j + \ldots + v}$$

čo je funkcia **nelineárna voči nezávislým premenným** tvoriacim vektor $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T$ nazývaný **regresor**, ale **lineárna voči regresným koeficientom** $\theta_0, \theta_i, \theta_{ij}$.



⇒ **Model** tohto systému nech popisuje funkčná závislosť opäť **v tvare mocninového radu** s konečným stupňom s **bez náhodnej zložky**

$$\boxed{\hat{y} = \varphi(\mathbf{u}, \hat{\boldsymbol{\theta}})}$$

kde $\hat{\boldsymbol{\theta}}$ je **odhad vektora parametrov**

⇒ **n-faktorový polynomiálny model stupňa s**

Počet neznámych parametrov vrátane absolútneho člena pre model, ktorý má **n faktorov (= vstupov)** a je vyjadrený **polynómom stupňa s** je

$$1 + k = \binom{n+s}{s} \qquad \text{kde} \quad \binom{x}{y} = \left(\frac{x!}{y!\,(x-y)!} \right) \text{ je kombinačné číslo}$$

Pri väčšom počte faktorov n a so vzrastom stupňa polynómu s počet neznámych parametrov veľmi rýchle narastá, z praktického hľadiska úplne stačí uvažovať $1 \le s \le 3$, pričom:

- ak $s = 1$ – **lineárna regresia**

  $$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 u_1 + \ldots + \hat{\theta}_n u_n ,$$

- ak $s = 2$ – **kvadratická regresia**

  $$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 u_1 + \ldots + \hat{\theta}_n u_n + \hat{\theta}_{n+1} u_1^2 + \ldots + \hat{\theta}_{2n} u_n^2 + \hat{\theta}_{2n+1} u_1 u_2 + \ldots + \hat{\theta}_k u_{n-1} u_n ,$$

- ak $s = 3$ – **kubická regresia**

$$\hat{y} = \hat{\theta}_0 + \sum_{i=1}^{n} \hat{\theta}_i u_i + \sum_{i=1}^{n}\sum_{j=1}^{i} \hat{\theta}_{ij} u_i u_j + \sum_{i=1}^{n}\sum_{j=}^{i}\sum_{k=1}^{j} \hat{\theta}_{ijk} u_i u_j u_k ,$$

$\Rightarrow$ Pri jednofaktorových modeloch môže byť aj **polynomiálna** závislosť

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 u + \hat{\theta}_2 u^2 \ldots + \hat{\theta}_k u^k$$

Na model **lineárny** voči regresným koeficientom môžeme **pretransformovať** aj niektoré **nelineárne závislosti**, napr.

$$z = c.x_1^{\hat{\theta}_1}.x_2^{\hat{\theta}_2} \ldots x_n^{\hat{\theta}_n} \quad \rightarrow \quad \hat{y} = \hat{\theta}_0 + \hat{\theta}_1 u_1 + \ldots + \hat{\theta}_n u_n$$

$$\text{kde} \quad \hat{y} = \ln z \qquad \hat{\theta}_0 = \ln c \qquad u_i = \ln x_i \qquad i = 1,2,\ldots,n$$

alebo

$$z = c.e^{\left(\hat{\theta}_1 u_1 + \ldots + \hat{\theta}_n u_n\right)} \quad \rightarrow \quad \hat{y} = \hat{\theta}_0 + \hat{\theta}_1 u_1 + \ldots + \hat{\theta}_n u_n$$

$$\text{kde} \quad \hat{y} = \ln z \qquad \hat{\theta}_0 = \ln c$$

$\Rightarrow$ Vo **všeobecnosti** možno **všetky** uvedené **modely vyjadriť v tvare lineárnej regresie**

$$\boxed{\hat{y} = \hat{\theta}_0 f_0(\mathbf{u}) + \hat{\theta}_1 f_1(\mathbf{u}) + \hat{\theta}_2 f_2(\mathbf{u}) \ldots + \hat{\theta}_k f_k(\mathbf{u})}$$

$$\boxed{\hat{y} = \hat{\boldsymbol{\theta}}^T \mathbf{f}(\mathbf{u})}$$

$\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_k\right)^T$ je **vektor odhadovaných parametrov**

$\mathbf{f}(\mathbf{u}) = (f_0(\mathbf{u}), f_1(\mathbf{u}), \ldots, f_k(\mathbf{u}))^T$ je **regresný vektor**

Poznámka: všetky vektory sú stĺpcové, preto sú transponované, keď sú zapísané ako riadkové.

$\Rightarrow$ **Predpokladajme, že uskutočníme celkom** $N >> 1 + k$ **meraní**, pri každom z nich budeme určovať hodnotu výstupu zo sústavy y_i aj hodnoty zložiek vektora $\mathbf{f}(\mathbf{u}_i)$

$$\mathbf{f}(\mathbf{u}_i) = (f_0(\mathbf{u}), f_1(\mathbf{u}), \ldots, f_k(\mathbf{u}))^T = (1, h_{1i}, h_{2i}, \ldots, h_{ki})^T = \mathbf{h}_i \qquad i = 1, \ldots, N$$

Potom v i-tom experimente vyjadríme hodnotu odchýlky medzi nameraným výstupom a výstupom z modelu

$$e_i = y_i - \hat{y}_i = y_i - \hat{\boldsymbol{\theta}}\mathbf{h}_i$$

čím sa dostaneme k **preurčenému systému rovníc (t.j. máme viac rovníc ako neznámych)**

$$\boxed{\begin{aligned} e_1 &= y_1 - \hat{\theta}_0 - \hat{\theta}_1 h_{11} - \hat{\theta}_2 h_{21} - \ldots - \hat{\theta}_k h_{k1} \\ e_2 &= y_2 - \hat{\theta}_0 - \hat{\theta}_1 h_{12} - \hat{\theta}_2 h_{22} - \ldots - \hat{\theta}_k h_{k2} \\ &\vdots \\ e_N &= y_N - \hat{\theta}_0 - \hat{\theta}_1 h_{1N} - \hat{\theta}_2 h_{2N} - \ldots - \hat{\theta}_k h_{kN} \end{aligned}}$$

alebo v maticovom tvare

$$\boxed{\mathbf{e} = \mathbf{y} - \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{e}\left(\hat{\boldsymbol{\theta}}\right)} \quad \text{kde} \quad \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (e_1, e_2, \ldots, e_N)^T$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \qquad \mathbf{H} = \begin{pmatrix} 1 & h_{11} & h_{21} & \ldots & h_{k1} \\ 1 & h_{12} & h_{22} & \ldots & h_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & h_{1N} & h_{2N} & \ldots & h_{kN} \end{pmatrix} = \begin{pmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \vdots \\ \mathbf{h}_N^T \end{pmatrix}$$

⇒ **Vektor neznámych parametrov** $\hat{\boldsymbol{\theta}}$ z tohto preurčeného systému rovníc budeme určovať **minimalizáciou súčtu štvorcov odchýlok (metóda najmenších štvorcov, MNŠ –** ako prvý ju použil Gauss**,** publikoval ju v r. 1795**)**

$$Q\left(\hat{\boldsymbol{\theta}}\right)=\frac{1}{2}\sum_{i=1}^{N}e_i^2=\frac{1}{2}\mathbf{e}\left(\hat{\boldsymbol{\theta}}\right)^T\mathbf{e}\left(\hat{\boldsymbol{\theta}}\right)=\frac{1}{2}\left(\mathbf{y}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{y}-\mathbf{H}\hat{\boldsymbol{\theta}}\right)$$ **účelová funkcia**

Optimálnu hodnotu $\hat{\boldsymbol{\theta}}^*$, ktorú budeme považovať za **odhad hľadaných parametrov**, určíme z **nulovej hodnoty gradientu účelovej funkcie**

$$\nabla_{\hat{\boldsymbol{\theta}}}Q\left(\hat{\boldsymbol{\theta}}\right)_{|\hat{\boldsymbol{\theta}}^*}=\mathbf{H}^T\left(\mathbf{H}\hat{\boldsymbol{\theta}}^*-\mathbf{y}\right)=0$$

$$\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}}^*=\mathbf{H}^T\mathbf{y} \quad \Rightarrow \quad \hat{\boldsymbol{\theta}}^*=\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}$$ **Gaussov vzorec (1809)**

Aby tento vektor zabezpečil **minimum účelovej funkcie**, musí byť matica $\mathbf{R}=\mathbf{H}^T\mathbf{H}$ **pozitívne definitná**, čo je splnené, pretože súčin nenulovej matice s jej transpozíciou je vždy symetrická a pozitívne definitná matica

$$\nabla_{\hat{\boldsymbol{\theta}}}\nabla_{\hat{\boldsymbol{\theta}}}Q\left(\hat{\boldsymbol{\theta}}\right)_{|\hat{\boldsymbol{\theta}}^*}=\mathbf{H}^T\mathbf{H}=\mathbf{R}$$ **informačná matica**

$$\mathbf{P}=\mathbf{R}^{-1}=\left(\mathbf{H}^T\mathbf{H}\right)^{-1}$$ **disperzná matica**

$$\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}}^*=\mathbf{H}^T\mathbf{y}$$ **normálna sústava rovníc**

Príklad → priklady_regresia.pdf + cvičenia

⇒ **MATLAB**

1. Curve Fitting Toolbox

Umožňuje **spracovanie údajov** (vyhladenie, odstránenie niektorých vzoriek), **aproximáciu funkciou** a **porovnanie výsledkov** graficky aj numericky

Možnosti použitia:

- GUI – spúšťa sa príkazom *cftool*
- Matlab – príkazy

  fit(x,y,FT)

2. Basic Fitting GUI

v okne **Figure** vybrať **Tools – Basic Fitting**

⇒ **Vlastnosti odhadu**:

Predpokladajme, že:

1. **hod(H)=1+k ≤ N** (počet pozorovaní je väčší ako počet neznámych parametrov modelu)
2. **parazitný šum spĺňa**:

   stredná hodnota šumu sa **rovná nule**, t.j. vylučujeme výskyt jednosmernej zložky aj nestacionárnych šumov

   $$E\{v_i\}=0 \qquad i=1,2,\ldots,N$$

   jednotlivé **zložky** šumov sú navzájom **nekorelované**

$$E\{v_i v_j\} = 0 \qquad i \neq j$$

disperzia šumov je **konštantná**

$$E\{v_i^2\} = \sigma_i^2 = \sigma^2 \qquad i = 1,2,\ldots,N$$

potom **kovariančná matica má tvar**

$$\mathrm{cov}(\mathbf{v}) = E\{\mathbf{v}\mathbf{v}^T\} = \sigma^2\mathbf{I}$$

Gaussova – Markovova veta: Ak platia predpoklady 1. a 2., potom $\hat{\boldsymbol{\theta}}^* = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$ je výdatný nevychýlený odhad (BLUE) .

BLUE (Best Linear Unbiased Estimator)

- **lineárny vzhľadom k dátam**
- **nevychýlený** (unbiased)
- **výdatný** (best, minimum variance)

a) Linearita

$$\hat{\boldsymbol{\theta}}^* = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$$

odhad je **lineárny vzhľadom k nameraným výstupom**

b) Nevychýlenosť odhadu parametrov

Pre výstup reálneho systému platí $\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{v}$

$$\hat{\boldsymbol{\theta}}^* = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T(\mathbf{H}\boldsymbol{\theta} + \mathbf{v}) = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} + (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{v} = \boldsymbol{\theta} + (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{v}$$

$$E\{\hat{\boldsymbol{\theta}}^*\} = \boldsymbol{\theta} + E\{(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{v}\}$$

$$\Downarrow$$

$$\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{v} = \mathbf{P}\mathbf{H}^T\mathbf{v}$$

Keďže $E\{\mathbf{v}\} = \mathbf{0}$, potom platí $\boxed{E\{\hat{\boldsymbol{\theta}}^*\} = \boldsymbol{\theta}}$.

Ak sa parazitný šum v riadi rozdelením $N(0,\sigma^2)$, potom vektor $\hat{\boldsymbol{\theta}}^*$ **je nevychýleným odhadom** vektora $\boldsymbol{\theta}$.

c) Výdatnosť odhadu parametrov

Kovariančná matica odhadu je

$$\mathrm{cov}\{\hat{\boldsymbol{\theta}}^*\} = E\{(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta})^T\} = E\{\mathbf{P}\mathbf{H}^T\mathbf{v}\mathbf{v}^T\mathbf{H}\mathbf{P}\} = \mathbf{P}\mathbf{H}^T E\{\mathbf{v}\mathbf{v}^T\}\mathbf{H}\mathbf{P} = \sigma^2\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{P} = \sigma^2\mathbf{P}\mathbf{P}^{-1}\mathbf{P} = \sigma^2\mathbf{P}$$

$$\boxed{\mathrm{cov}(\hat{\boldsymbol{\theta}}^*) = \sigma^2\mathbf{P}}$$

čo znamená, že **výdatnosť odhadu** sa **zvyšuje** so **vzrastom hladiny užitočných signálov u** a naopak **klesá so zväčšovaním disperzie šumu** σ^2.

Tiež $\hat{y}$ je **nevychýleným** odhadom y, pričom jeho kovariancia je

$$\boxed{\mathrm{cov}(\hat{y}) = \sigma^2\mathbf{H}^T\mathbf{P}\mathbf{H}}$$

Problém **zabezpečenia výdatnosti hľadaných odhadov** je teda spojený s otázkou **vhodného modelu** pre aproximáciu určovanej funkčnej závislosti a tiež so starostlivou **prípravou experimentu**.

Experiment by mal byť zostavený tak, aby

- prinášal **maximum užitočnej informácie**
- bol realizovaný s **minimálnymi nákladmi**
- bol **vyhodnocovaný** metódami **matematickej štatistiky**.

Objektivita a presnosť modelu sa zvyšuje **priamo úmerne s počtom experimentov**, avšak **ekonomická stránka** identifikácie smeruje k **menšiemu počtu experimentov**.