



Univerzita Palackého  
v Olomouci

# Lineární regrese I

## Chemometrie I (ACH/CHEX1)

(c) David MILDE, 2023



Univerzita Palackého  
v Olomouci

## Úvod

### – DRUHY ZÁVISLOSTÍ DVOU PROMĚNNÝCH:

- **FUNKČNÍ VZTAH:** dvě závisle proměnné, tj. určité hodnotě  $x$  odpovídá jediná hodnota  $y$ .
- **KORELACE:** dvě nezávislé proměnné. Hodnotí se síla vzájemného vztahu popisovaná např. korelačním koeficientem  $R$ .
- **REGRESE:** vztah nezávislé (náhodné) proměnné  $x$  a závisle proměnné  $y$ , které má určité rozdělení pravděpodobnosti.

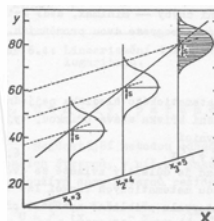
- Statistickou analýzou pomocí lineární regrese tedy objasňujeme vztah mezi závisle proměnnou  $y$  a nezávisle proměnnou (nezávisle proměnnými)  $x$ .

- Výsledkem je regresní model (LRM):

- $y_i = f(x_i, \mathbf{b}) + \varepsilon_i$

- $b_i$  – regresní koeficienty,  $\varepsilon$  – náhodná chyba

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

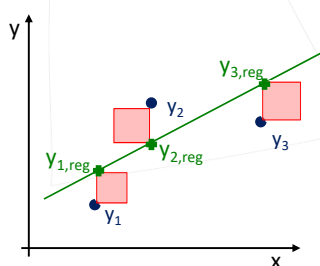


Lineární regrese



## Metoda nejmenších čtverců (MNČ)

- MNČ je v regresní analýze nejpoužívanější pro výpočet odhadů parametrů  $b_i$  do regresního modelu. Např. pro přímku:  $y = b_0 + b_1x$ .
- Legendre a Gauss navrhli, aby se od každého bodu vedla ve svislém směru úsečka až k uvažované přímce. Tato úsečka se bere jako strana čtverce. Řekne se, že přímka je tím lepší, čím menší součet čtverců vytváří.
- Máme-li proložit přímku více než 2 body, řešíme tzv. přeuredený systém – více rovnic než neznámých.



RSC:  $\sum \square = \min.$

Tento součet čtverců je účelová funkce U:

$$U(b_i) = \sum_{i=1}^n (y_i - b_0 - b_1x)^2 = \min$$
$$\sum_{i=1}^n (y_i - y_{i,reg})^2 = \min$$

reziduum:  $e_i = y_i - y_{i,reg}$

RSC – reziduální součet čtverců

$$y_{1,reg} = b_0 + b_1x_1$$

$$y_{2,reg} = b_0 + b_1x_2$$

$$y_{3,reg} = b_0 + b_1x_3$$

Tuto soustavu rovnic lze řešit pouze pro určitou podmínku a tou je požadavek na nejmenší čtverce.



## MNČ – výpočet regresních koeficientů $b_i$

- Pro určení regresních koeficientů z účelové funkce použijeme podmínku nejmenších čtverců. V případě přímky:

$$\sum_{i=1}^n (y_i - b_0 - b_1x)^2 = \text{minimum}$$

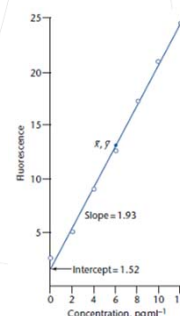
- Rovnici derivujeme podle obou koeficientů ( $b_0$  a  $b_1$ ). Získáme soustavu dvou tzv. normálních rovnic a jejím řešením koeficienty  $b_0$  a  $b_1$ .

$$nb_0 + (\sum_{i=1}^n x_i)b_1 = \sum_{i=1}^n y_i$$
$$(\sum_{i=1}^n x_i)b_0 + (\sum_{i=1}^n x_i^2)b_1 = \sum_{i=1}^n x_i y_i$$

- $b_0$  nazýváme úsek (aj. intercept) a  $b_1$  je směrnice (aj. slope)

$$b_1 = \frac{n \sum_{i=1}^n (x_i y_i) - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

$$b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n}$$





## MNČ – intervaly spolehlivosti

✳ Výběrové odhady směrodatných odchylek:

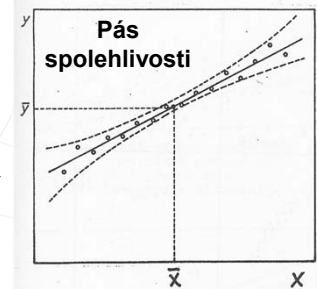
$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - y_{i,reg})^2}{n-2}}$$
$$s_{b0} = \sqrt{s_y \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}, \quad s_{b1} = \sqrt{\frac{s_y^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- IS regresních koeficientů  $b_i$ :

$$L_{1,2} = b_i \pm s_{bi} \cdot t_{(1-\frac{\alpha}{2}, n-m)}$$

$n$  je počet hodnot a  $m$  je počet parametrů  $b_i$

- IS pro regresní hodnoty  $y_{i,reg}$ : vyjadřuje se graficky pomocí pásu spolehlivosti, který je nejužší v místě  $x_i = \bar{x}$  a oběma směry od tohoto bodu se rozšiřuje.



## Testování hypotéz v lineární regresi



## Testování významnosti v regresi

### - Test významnosti regresních koeficientů $b_i$ :

- Je-li parametr nevýznamný, znamená to, že je roven 0. Pokud  $b_0 = 0$  v rovnici  $y = b_0 + b_1x$ , přímka prochází počátkem a regresní model se zjednoduší na  $y = b_1x$ .

$$H_0: b_i = 0; H_1: b_i \neq 0 \quad t_i = \frac{b_i}{s_{b_i}}$$

- $t_i$  v absolutní hodnotě porovnáváme s kritickou hodnotou Studentova rozdělení  $t_{krit(1-\alpha/2; n-m)}$ , kde  $n$  je počet hodnot a  $m$  je počet parametrů  $b_i$ . Je-li  $|t_i| < t_{krit}$ , přijímáme  $H_0$ .

### - F test významnosti regrese:

- Test významnosti všech regresních koeficientů  $b_i$  kromě absolutního členu ( $b_0$ ), označován jako test významnosti  $R^2$

$$H_0: R^2 = 0; H_1: R^2 \neq 0 \quad F_R = \frac{R^2 \cdot (n-m)}{(1-R^2) \cdot (m-1)}$$

- $F_R$  porovnáváme s kritickou hodnotou F-rozdělení  $F_{krit(m-1, n-m)}$ . Je-li  $F_R < F_{krit}$ , přijímáme  $H_0$ .



## F testy v LR

### - Test linearity:

- Např.: test linearity (tj. vhodnosti přímkového regresního modelu), který je založený na volbě mezi přímkou ( $y = b_0 + b_1x$ ) a parabolou ( $y = b_0 + b_1x + b_2x^2$ ).

$$H_0: b_2 = 0, H_1: b_2 \neq 0 \quad F_L = \frac{(RSC_L - RSC_K) \cdot (n-3)}{RSC_K}$$

- $RSC_L$  a  $RSC_K$  je reziduální součet čtverců pro lineární a kvadratickou závislost.

- $F_L$  porovnáváme s  $F_{krit(1, n-3)}$ . Pokud je  $F_L < F_{krit}$ , přijmeme  $H_0$ , je závislost lineární. Pokud zamítneme  $H_0$  a přijmeme  $H_1$ , je závislost kvadratická.

### - Chowův test shody 2 lineárních modelů:

- Testuje shodu parametrů  $b$  pro regresní model A a model B:

- $y_A = b_{0A} + b_{1A}x + \varepsilon_A$ , který má  $RSC_A$ ;  $y_B = b_{0B} + b_{1B}x + \varepsilon_B$ , který má  $RSC_B$

$$H_0: b_A = b_B; H_1: b_A \neq b_B \quad F_C = \frac{(RSC - RSC_A - RSC_B) \cdot (n-2m)}{(RSC_A + RSC_B) \cdot m}$$

- Za předpokladu homoskedasticity obou výběrů se statistika  $F_C$  porovnává s  $F_{krit(m, n-2m)}$ . V případě heteroskedasticity je nutné použít jiný způsob určení stupňů volnosti, který zde není uveden.  $H_0$  přijmeme, pokud  $F_C < F_{krit}$ .



Univerzita Palackého  
v Olomouci

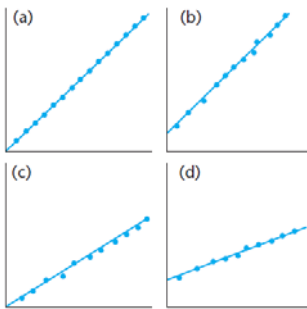
## Aplikace lineární regrese



Univerzita Palackého  
v Olomouci

### Porovnání dvou metod pomocí LR

- Rozšíření párového testu, někdy nepřesně nazývané validace pomocí LR.
- Výsledky 1. metody se vynesou na osu x, výsledky 2. metody na osu y a body se proloží přímkou pomocí MNČ.
- V případě shody výsledků obou metod získáme regresní model  $y = x$ . Jestli  $b_0 = 0$  a  $b_1 = 1$  ověřujeme pomocí IS obou regresních parametrů.



- (a) shoda metod,  $b_0 = 0$ ,  $b_1 = 1$ ,  $R = 1$   
(b) vyšší/nížší výsledky jedné metody,  $b_0 \neq 0$ ,  $b_1 = 1$   
(c) systematická chyba jedné metody,  $b_0 = 0$ ,  $b_1 \neq 1$   
(d) kombinace (b) a (c)



## Další aplikace LR

- Kalibrace
- Polynomická regrese
  - $y = b_0 + b_1x + b_2x^2 + \dots + b_nx^n$
- Vícerozměrná lineární regrese
  - $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$



## Rozlišení lineární a nelineární regrese

- Pro lineární regresi platí následující podmínka:

$$g_j = \frac{\delta f(\mathbf{x}, \mathbf{b})}{\delta b_i} = \text{konst.}$$

- Pokud alespoň pro jeden  $b_i$  je parciální derivace  $g_j$  funkcí, tak jde o nelineární regresi!
- Př. lineárních regresních modelů:
  - $y = b_0 + b_1x - b_2x^2$  tj. polynom;  $y = b_0 + (b_1/x)$  tj. hyperbola
- Př. nelineárních regresních modelů:
  - $y = b_0 \cdot x^{b_1}$ ;  $y = b_0 + b_1 \cdot \exp(b_2x)$
- Nelineární regrese je náplní ACH/CHE2.



Univerzita Palackého  
v Olomouci

## Předpoklady pro použití MNČ, postupy odhadu $b_i$ při porušení předpokladů MNČ



Univerzita Palackého  
v Olomouci

### Předpoklady pro použití MNČ

1. Regresní koeficienty ( $b_i$ ) mohou nabývat libovolných hodnot. (V praxi existují omezení fyzikálního smyslu.)
2. Regresní model je lineární v parametrech. (Pomocí LR lze řešit např. i polynomicke závislosti.)
3. Žádné dva sloupce matice  $\mathbf{X}$  nejsou kolineární (rovnoběžné) vektory. V datech matice  $\mathbf{X}$  se nevyskytuje multikolinearita, matice  $\mathbf{X}^T\mathbf{X}$  je dobře podmíněná.
4. Náhodné chyby  $\varepsilon_i$  mají nulovou střední hodnotu.
5. Náhodné chyby  $\varepsilon_i$  mají konstantní rozptyl (homoskedasticita).
6. Hodnoty závisle proměnné jsou získány nezávislým měřením (nekorelované chyby  $\varepsilon_{ij}$ ).
7. Nezávisle proměnné veličiny nejsou zatíženy systematickými či náhodnými chybami nebo jsou náhodné chyby na ose  $x$  zanedbatelné vůči náhodným chybám na ose  $y$ .



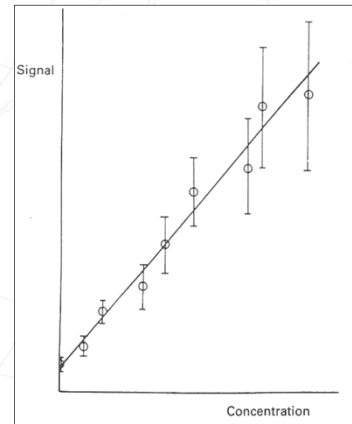
## Metody při porušení MNČ

### – METODA VÁŽENÝCH NEJMENŠÍCH ČTVERCŮ:

- Aplikujeme ji v případě, kdy náhodné chyby nemají konstantní rozptyl, tedy byla-li prokázána heteroskedasticita v datech (porušení předpokladu 5).
- Heteroskedasticita se často vyskytuje u kalibračních modelů, které se používají přes více koncentračních řádů.
- Do rovnice pro výpočet účelové funkce se vkládá váhový koeficient  $w_i$ .

$$U(b) = \sum_i \left( w_i y_i - \sum_j w_i x_{ij} b_j \right)^2, \quad w_i \text{ je např. } \frac{1}{y} \text{ či } \frac{1}{y^2}$$

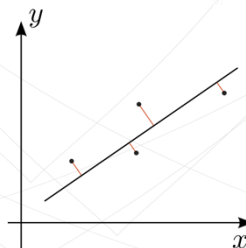
- Posouzení homoskedasticity či heteroskedasticity dat – viz prezentace CHEX1-07-LR-II.



## Metody při porušení MNČ

### – ORTOGONÁLNÍ REGRESE:

- Používá se, kdy je zatížena chybami i nezávisle proměnná  $x$  (porušení předpokladu 7).
- Obdoba MNČ, zde se však minimalizuje suma čtverců kolmých vzdáleností experimentálních bodů od přímky.



### – METODA KOREKCE HODNOSTI:

- Používá se v případě prokázané multikolinearity v datech (porušení předpokladu 3). Obvykle se aplikuje při řešení polynomické regrese u vyšších stupňů polynomu.
- Detaily v prezentaci CHEX1-07-LR-II.



## Metody při porušení MNČ

### – ROBUSTNÍ REGRESE:

- V případě porušení předpokladu č. 6, který vede k odchylkám od normality chyb závisle proměnné nebo v případě přítomnosti OB, které nechceme vyloučit je vhodným řešením robustní regrese.
- Místo kritéria „MNČ“ použijeme robustního kritéria, které je jak na porušení předpokladu normality, tak na OB málo citlivé.
- Z velké skupiny metod robustní regrese zmíníme pouze  **$L_p$  regresi:  $U = \sum |e_i|^p$** 
  - **$p = 1$ : robustní mediánová regrese, tedy metodu nejmenších absolutních odchylek vhodnou pro data s často se vyskytujícími odlehlými hodnotami na obou stranách nebo s rozdělením podobným Laplaceovu,  $U = \sum_{i=1}^n |y_i - b_0 - b_1 x_i|$**
  - **$p = 2$ : odpovídá metodě nejmenších čtverců,**
  - **$p \approx 5-10$  odpovídá metodě nejmenší maximální chyby (minimax),**
  - **$1 \leq p < 2$ : vykazují robustnost vůči odlehlým hodnotám.**



## Metody při porušení MNČ

### – METODA STEPWISE ALL:

- Slouží jako pomůcka k sestavení dobrého modelu na základě dat i bez předběžné informace o možných vztazích mezi proměnnými.
- Vypočítá regrese se všemi možnými kombinacemi vybraných nezávisle proměnných v regresním modelu. Pro každou regresi vypočítá tři kritéria kvality regrese: F-kritérium (FIS), Akaikeho informační kritérium (AIC) a střední kvadratickou chybu predikce (MEP).
- Výstup metody se ukládá do protokolu a rovněž do zvláštního datového listu s názvem StepAll, který se při výpočtu vytvoří. Tento datový list slouží ke snadné identifikaci nejlepších modelů.
- K výpočtu jednotlivých regresí je použita vždy klasická metoda nejmenších čtverců.
- *Budeme používat u složitých polynomičtých regresních modelů.*