# PROBLÉMY PŘI LINEÁRNÍ REGRESI TERMOCHEMICKÝCH DAT

Petr Voňka[A], Jindřich Leitner[B]

[A]Ústav fyzikální chemie,
[B]Ústav inženýrství pevných látek
VŠCHT Praha, Technická 5, 16628 Praha 6

# KALSEM 2010, Hotel Skalský Dvůr

Statistické zpracování
termoanalytických dat

Petr Voňka a Jindřich Leitner
VŠCHT Praha

# KALSEM 2010, Hotel Skalský Dvůr
# Statistické zpracování termoanalytických dat

- Náhodné veličiny a rozdělení jejich pravděpodobnosti
- Náhodný výběr a jeho charakteristiky
- Odhad bodových a intervalových parametrů
- Testování hypotéz
- Zákon o šíření chyb
- **Lineární regrese**

# Špatně podmíněná soustava rovnic

$$8x + 53y = 61$$

$$3x + 20y = 23$$

$$x = \frac{\det(x)}{\det} = \frac{61 \times 20 - 53 \times 23}{8 \times 20 - 53 \times 3} = \frac{1}{1} = 1$$

$$x = y = 1$$

# Špatně podmíněná soustava rovnic

$$8x + 53y = 61$$
$$3x + 20y = 23$$

$$x = y = 1$$

$$7{,}951x + 53y = 61$$
$$3x + 20y = 23$$

$$x = 50 \quad y = -6{,}5$$

$$(8 + \varepsilon)x + 53y = 61$$
$$3x + 20y = 23$$

$$x = \frac{1}{1 + 20\varepsilon} \quad y = \frac{1 + 23\varepsilon}{1 + 20\varepsilon}$$

# Lineární regrese

Uvažujme náhodnou veličinu $Y$, která závisí na proměnné $x$

$$y = \sum_j a_j f_j(x) + e, \quad e \sim \mathrm{N}(0, \sigma^2)$$

$y$ ... závisle proměnná (**náhodná veličina**)
$x$ ... nezávisle proměnná (**není náhodná veličina**)
$f_j$ ... souřadnicová funkce
$e$ ... chyba (**náhodná veličina**)

## Více nezávisle proměnných

$$y = \sum_k \sum_j a_{jk} f_{jk}(x_k) + e, \quad e \sim \mathrm{N}(0, \sigma^2)$$

# Lineární regrese

Uvažujme $n$ experimentálních dvojic $(x_i, y_i), \quad i = 1, \ldots, n$

$$y_i = a_1 f_1(x_i) + a_2 f_2(x_i) + e_i \,, \;\; e_i \sim \mathrm{N}(0, \sigma^2) \,, \;\; i = 1, \ldots, n$$

# Metoda nejmenších čtverců

$$\Phi(a_1, a_2) = \sum_{i=1}^{n} \left( y_i - a_1 f_1(x_i) - a_2 f_2(x_i) \right)^2 \to \min$$

# Lineární regrese

$$\frac{\partial \Phi}{\partial a_1} = \frac{\partial \Phi}{\partial a_2} = 0$$

$$\mathbf{C}\,\vec{a} = \vec{b}$$

$$c_{jk} = c_{kj} = \sum_{i=1}^{n} f_j(x_i)\, f_k(x_i), \quad b_k = \sum_{i=1}^{n} y_i\, f_k(x_i), \quad j,k = 1,2$$

$$a_j = \hat{a}_j \pm 2\hat{\sigma}\sqrt{c^{jj}}, \quad \hat{\sigma} = \sqrt{\frac{\Phi(\hat{a}_1, \hat{a}_2)}{n-2}}$$

$$\hat{\sigma}^2 c^{jk} = \mathrm{Cov}(a_j, a_k), \quad \rho_{jk} = \frac{\mathrm{Cov}(a_j, a_k)}{\sqrt{\mathrm{Cov}(a_j, a_j)\,\mathrm{Cov}(a_k, a_k)}} = \frac{c^{jk}}{\sqrt{c^{jj} c^{kk}}}$$

# Multikolinearita

Vliv „skoro existující“ lineární vazby mezi souřadnicovými funkcemi nebo proměnnými *x*

$$y = a_1 f_1(x) + a_2 f_2(x)$$

$$a_1 f_1(x) + a_2 f_2(x) \approx 0, \quad x \in (x_1, x_2)$$

$$f_2(x) \approx \beta f_1(x)$$

**Takto NE !**

$$y = a_1 + a_2 x + a_3 x^2 + a_4 \ln x$$

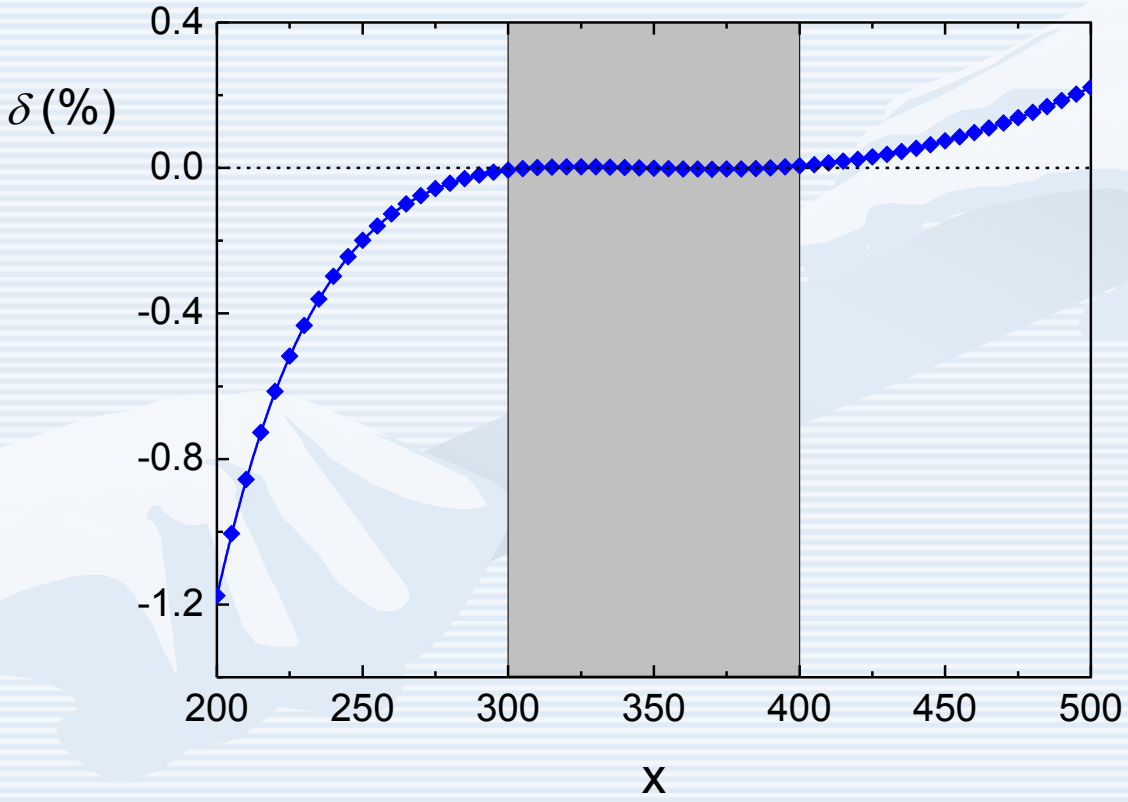
$$\ln x \approx \beta_1 + \beta_2 x + \beta_3 x^2$$

# Multikolinearita

$$\ln x \approx \beta_1 + \beta_2 x + \beta_3 x^2, \quad x \in (300, 400)$$

$$\beta_1 = 4,348 \pm 0,012, \quad \beta_2 = (5,759 \pm 0,069) \times 10^{-3}, \quad \beta_3 = -(4,127 \pm 0,098) \times 10^{-6}$$

$$\sigma = 1,84 \times 10^{-4}$$



# Multikolinearita

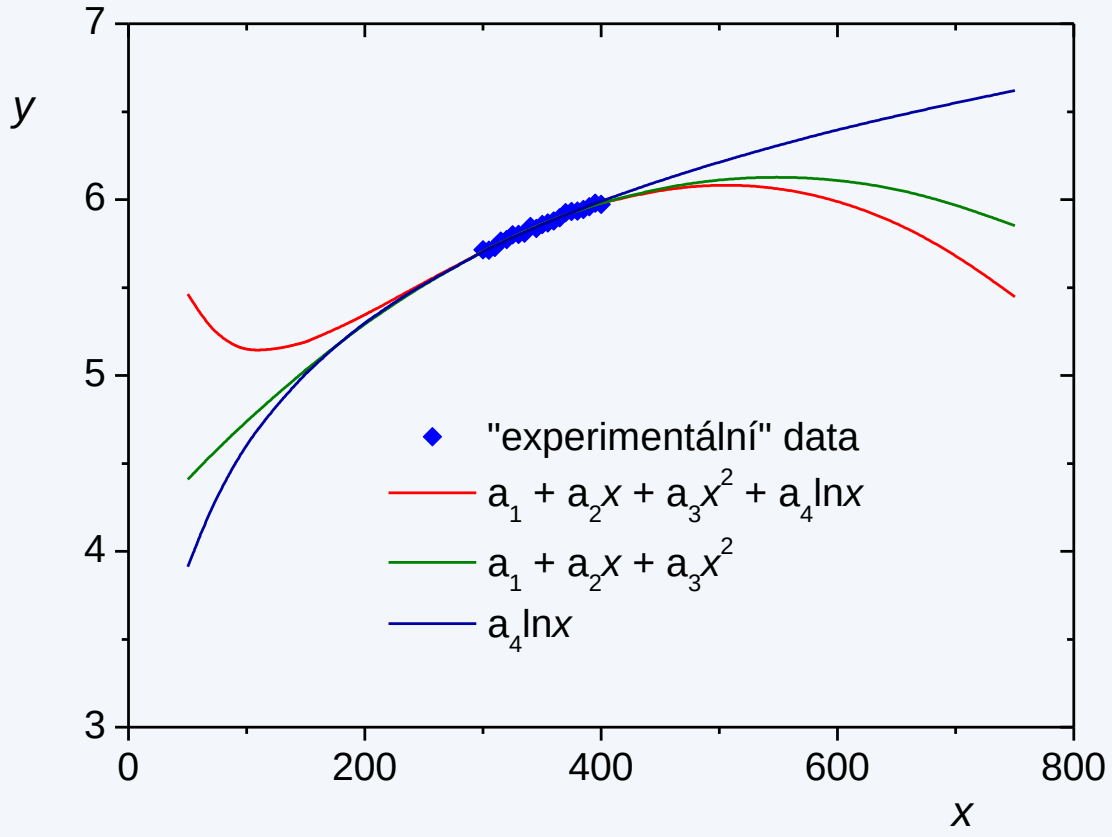
$$y_i = \ln x_i + \delta x_i, \quad \delta x_i \sim \mathrm{N}(0;1\times10^{-4})$$

$$y_i = a_1 + a_2 x_i + a_3 x_i^2 + a_4 \ln x_i$$

| | | | |
|---|---|---|---|
| $a_1$ | | 4,064±0,583 | 10,457±102,651 |
| $a_2$ | | (7,612±3,348)×10$^{-3}$ | (16,096±135,888)×10$^{-3}$ |
| $a_3$ | | (-6,941±4,780)×10$^{-6}$ | (-1,302±9,740)×10$^{5}$ |
| $A_4$ | ≈ 1 | | -1,474±23,605 |
| $\sigma$ | 0,0095 | 0,0089 | 0,0092 |
| $y$(100)/$y$(600) | 4,605/6,397 | 4,738/6,115 | 5,147/5,999 |

# Multikolinearita

$$y = a_1 + a_2 x + a_3 x^2 + a_4 \ln x$$



# Příklad 1

$$G = -a_1 T \ln T + a_2 + a_3 T$$

$$H = G + TS = G - T\left(\frac{\partial G}{\partial T}\right)_p = a_1 T + a_2$$

$$C_p = \left(\frac{\partial H}{\partial T}\right)_p = a_1$$

$$f_1(T) = -T \ln T \quad f_2(T) = 1 \quad f_3(T) = T$$

Modelové hodnoty: $a_1 = 1, \; a_2 = 100, \; a_3 = 5$

$$G = -T \ln T + 100 + 5T$$

# Příklad 1

Simulovaná exp. data:

$$G_{\mathrm{exp}} = G + \delta G$$

$\delta G$ … generátor sady náhodných čísel N(μ = 0; σ = 0,4) (Compaq Fortran)

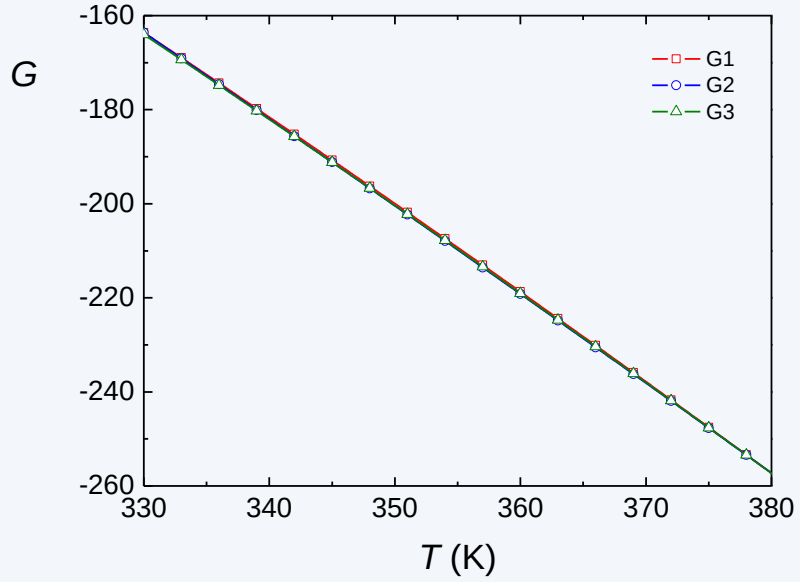
| $T$(K) | $-G_{\mathrm{m}}$ | exp_1 | exp_2 | exp_3 |
|---|---|---|---|---|
| 330 | 163,70 | 163,37 | 163,31 | 163,86 |
| 333 | 169,11 | 169,65 | 168,62 | 168,90 |
| 336 | 174,55 | 174,76 | 175,21 | 174,02 |
| | ..... | | | |
| 351 | 202,14 | 202,45 | 201,97 | 202,10 |
| 354 | 207,73 | 207,33 | 208,51 | 206,81 |
| 357 | 213,35 | 213,06 | 213,41 | 213,13 |
| | ..... | | | |
| 375 | 247,60 | 247,79 | 247,54 | 247,41 |
| 378 | 253,39 | 253,64 | 253,43 | 253,00 |
| 381 | 259,21 | 259,08 | 259,88 | 259,18 |

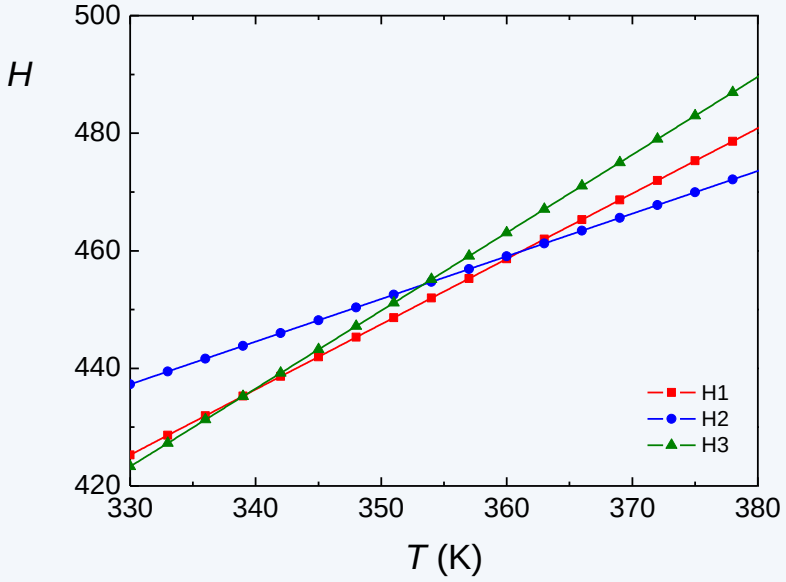
# Příklad 1

$$\Phi = \sum_{i=1}^{n} \left[ G_i - a_1 \left( -T_i \ln T_i \right) - a_2 - a_3 T_i \right]^2$$

Tab. II. Výsledky lineární regrese

| sada | $a_1$ | $a_2$ | $a_3$ | $\hat{\sigma}$ | $H_m(280)$ | $H_m(260)$ |
|---|---|---|---|---|---|---|
| 1 | 1,11148 ± 0,63 | 58,494 ± 224 | 5,7713 ± 4,34 | 0,41 | 369,71 | 347,48 |
| 2 | 0,72587 ± 0,73 | 197,74 ± 259 | 3,1144 ± 5,02 | 0,47 | 400,99 | 386,47 |
| 3 | 1,32581 ± 0,71 | -14,207 ± 253 | 7,2358 ± 4,89 | 0,46 | 357,02 | 330,50 |



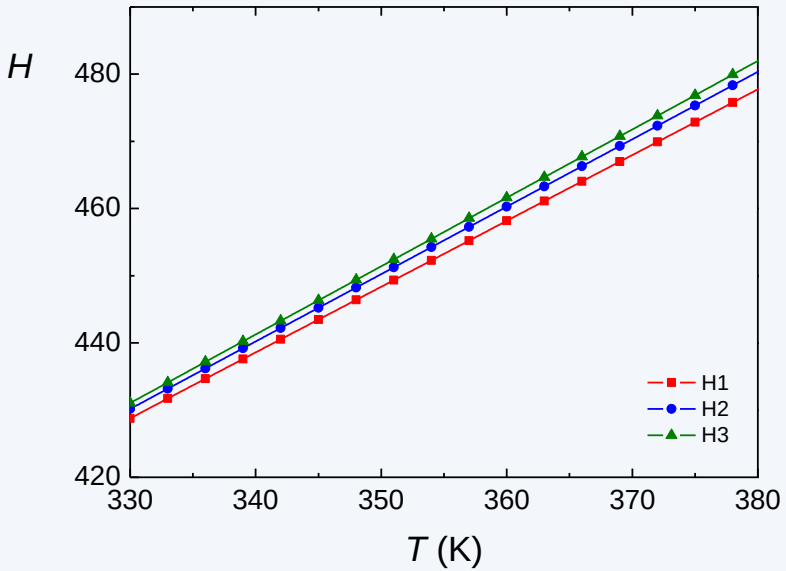



# Příklad 1

$$\Phi = \sum_{i=1}^{n}\left[ G_i - a_1\left(-T_i \ln T_i\right) - a_2 - a_3 T_i \right]^2 + \sum_{i=1}^{L}\left[ H_i - a_1 T_i - a_2 \right]^2$$

Tab. III. Výsledky simultánní korelace

| sada | $a_1$ | $a_2$ | $a_3$ | $\hat{\sigma}$ | $H_m(280)$ | $H_m(260)$ |
|---|---|---|---|---|---|---|
| 1 | $0,97941 \pm 0,05$ | $105,577 \pm 13$ | $4,8631 \pm 0,31$ | $0,38$ | $379,81$ | $360,22$ |
| 2 | $1,00308 \pm 0,06$ | $99,168 \pm 15$ | $5,0202 \pm 0,37$ | $0,45$ | $380,03$ | $359,97$ |
| 3 | $1,01797 \pm 0,06$ | $95,129 \pm 15$ | $5,1199 \pm 0,37$ | $0,44$ | $380,16$ | $359,80$ |



# Příklad 2

Vyhodnocení teplotní závislost $C_{\mathrm{pm}}$ pro FeAs z měření relativních entalpií $H_{\mathrm{m}}(T)$-$H_{\mathrm{m}}(T_0)$

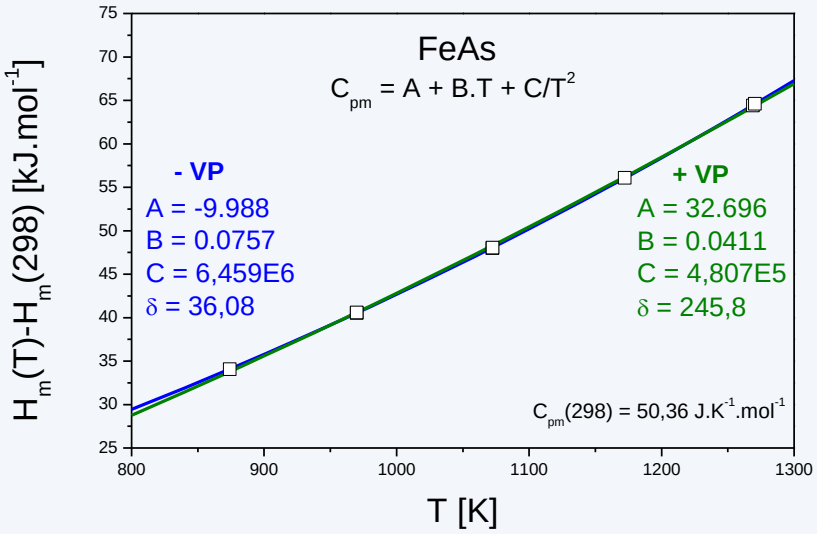
$$C_{p\mathrm{m}} = A + B \cdot T + \frac{C}{T^2}$$

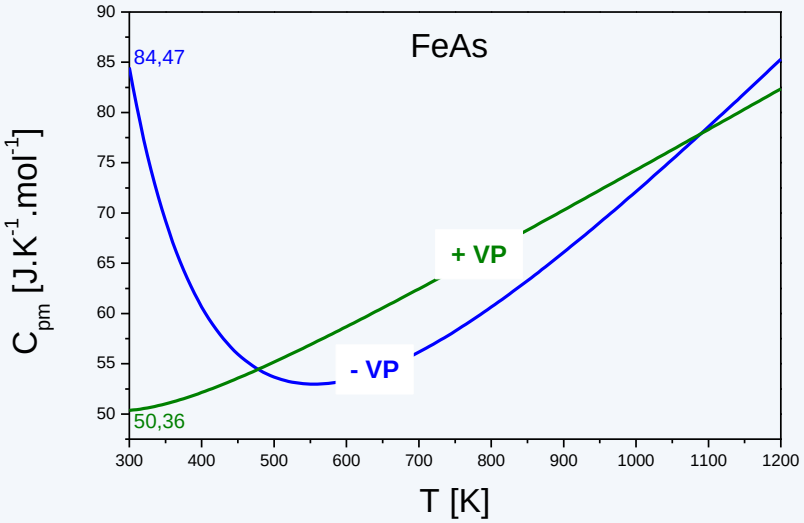
$$H_{\mathrm{m}}(T) - H_{\mathrm{m}}(T_0) = A\left(T - T_0\right) + \frac{B}{2}\left(T^2 - T_0^2\right) - C\left(\frac{1}{T} - \frac{1}{T_0}\right)$$

# Příklad 2

$$\Phi = \sum_{i=1}^{n} w_i \left[ H_{\mathrm{m}}(T_i) - H_{\mathrm{m}}(T_{0,i}) - A\left(T_i - T_{0,i}\right) - \frac{B}{2}\left(T_i^2 - T_{0,i}^2\right) + C\left(\frac{1}{T_i} - \frac{1}{T_{0,i}}\right) \right]^2$$

$$\Phi_{\mathrm{VP}} = \Phi + \lambda \left( C_{p\mathrm{m}}(298) - A - 298B - \frac{C}{298^2} \right)$$



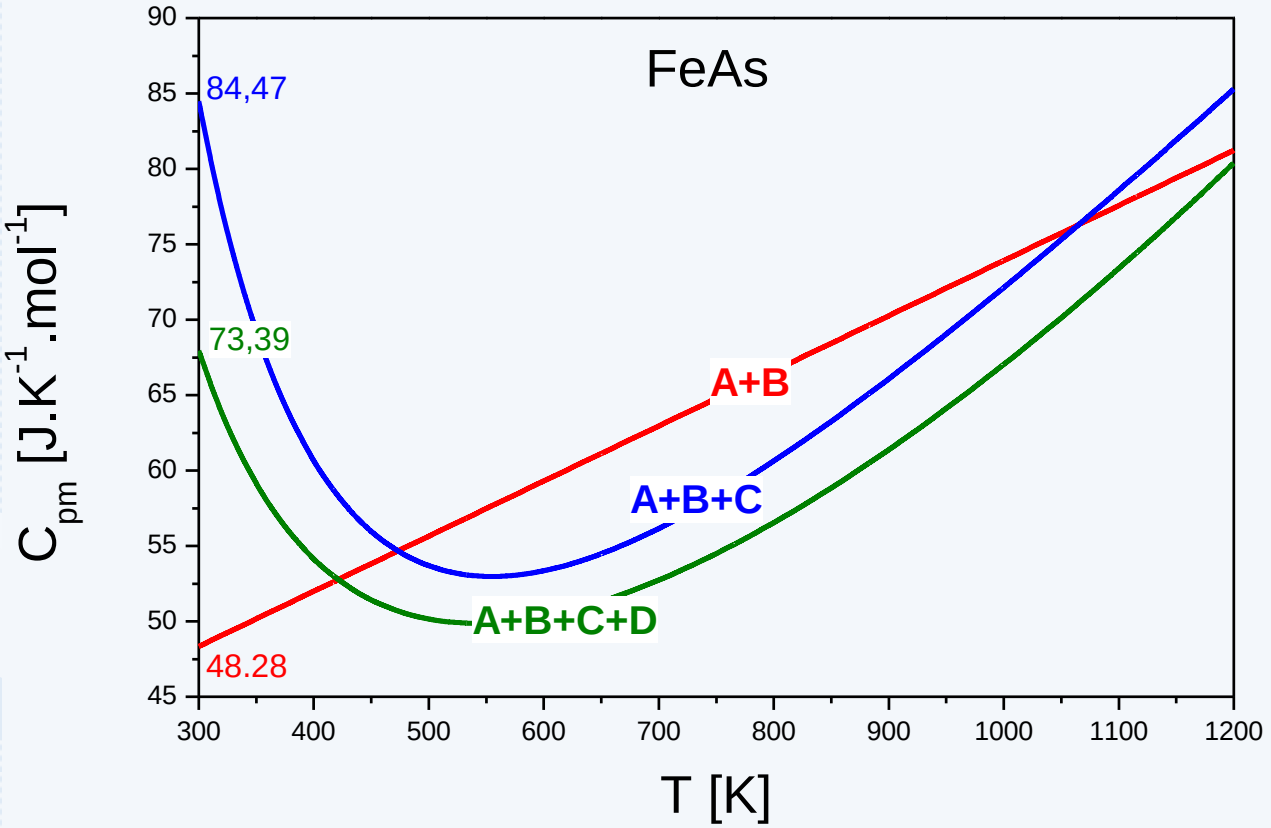
# Příklad 2

$$C_{pm} = A + B \cdot T + \frac{C}{T^2} + D \cdot T^2$$

| | | | |
|---|---|---|---|
| *A* | 37,381±2,066 | -9,523 ±4,861 | 25,307±88,436 |
| *B* | (36,54±3,03)×$10^{-3}$ | (75,28±4,03)×$10^{-3}$ | (18,93±142,92)×$10^{-3}$ |
| *C* | | (6,395±0,662)×$10^{6}$ | (3,581±7,166)×$10^{6}$ |
| *D* | | | (2,423±6,141)×$10^{-5}$ |
| σ | 0,179 | 0,026 | 0,027 |
| $C_{pm}$(298) | 48,28 | 84,87 | 73,39 |

# Příklad 2

$$C_{pm} = A + B \cdot T + \frac{C}{T^2} + D \cdot T^2$$



# Abychom z našich výsledků měli radost …

- ☺ Souřadnicové funkce pro regresi vybírejme uváženě.
- ☺ Kde je to možné provádějme simultánní regresi relevantních dat (nutnost vážit !!!).
- ☺ Buďme ostražití při extrapolaci mimo rozsah vstupních hodnot *x*.
- ☺ Nebojme se statistiky, neboť její správně používaný aparát nám řadu „problémů“ může predikovat či následně vysvětlit.

# Děkuji za Vaši pozornost