PODSTATNÉ TESTY VÝZNAMNOSTI V KORELAČNÍ A REGRESNÍ ANALÝZE

- test významnosti korelačního koeficientu
- test významnosti modelu jako celku
- test významnosti jednotlivých regresních parametrů
- test shody lineárních regresních modelů

a mnoho dalších testů…..

TEST VÝZNAMNOSTI *R*

Test významnosti odpovídá, zda je korelace *R* mezi výběrovými proměnnými natolik silná, abychom ji mohli považovat za prokázanou i pro základní soubor ρ.

Pro párový R: $t_R = \dfrac{R \cdot \sqrt{n-2}}{\sqrt{1-R^2}}$ $t_{\alpha,n-2}$ *n* je počet hodnot výběru

Pro násobný R: $F_R = \dfrac{R^2(n-m)}{(1-R^2)(m-1)}$ $t_{\alpha,n-m}$ *m* je počet proměnných

Pro parciální R: $t_R = \dfrac{R \cdot \sqrt{n-k-2}}{\sqrt{1-R^2}}$ $t_{\alpha,n-k-2}$ *k* je počet „vyloučených“ proměnných

TEST VÝZNAMNOSTI REGRESNÍHO MODELU

Co vlastně testujeme?

$$Y = \mathbf{b_0} + \mathbf{b_1}x_1 + \mathbf{b_2}x_2 + \mathbf{b_3}x_3 + \ldots + \mathbf{b_m}x_m$$

Testujeme významnost odhadů jednotlivých parametrů: když je testovaný odhad parametru statisticky nevýznamný, pak jeho příslušná proměnná x_j **nepřispívá** ke zpřesnění odhadu závisle proměnné *y* a tato proměnná x_j je pak v modelu zbytečná.

Testujeme model jako celek: tj. zda příslušná kombinace všech nezávisle proměnných statisticky významně zpřesní odhad závisle proměnné *y* oproti použití pouhého průměru hodnot *y*.

TEST VÝZNAMNOSTI REGRESNÍCH PARAMETRŮ

H_0: $\beta_j = 0$, *tj. j-tý regresní parametr je nevýznamný*

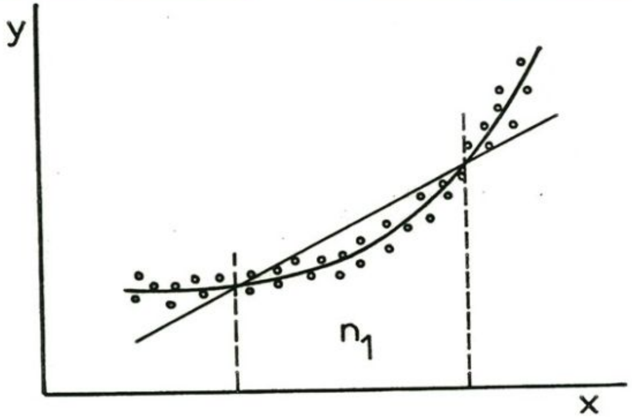
$$t = \frac{b_j - \beta_j}{s_b} \quad \text{pro } \beta_j = 0 \quad \Rightarrow \quad t = \frac{b_j}{s_b}$$

Pokud platí, že $|t| > t_{\alpha 2;n-m}$, potom je j-tý regresní parametr statisticky významný a příslušná proměnná musí zůstat v modelu.

Testy vhodnosti lineárního modelu

1. Test správnosti lineárního modelu $f(x, \beta) = X\beta$ dle Uttsové:

H_0: lineární model *vs.* H_A: nelineární model



RSC_1 regresí s využitím n_1 bodů,
RSC regresí s využitím všech n bodů.

Testační kritérium

$$F_U = \frac{(RSC - RSC_1)(n_1 - m)}{RSC_1 (n - n_1)}$$

Uttsová: volit $n_1 \approx n/2$ a body co nejblíže k těžišti

Test: Je-li $F_U < F_{1-\alpha}(n - n_1, n_1 - m)$, je H_0 přijata.

2. Test kvadratického členu

H_0: $\beta_2 = 0$ (test významnosti β_2)

$$E(y/x) = \beta_1 x + \beta_2 x^2 + \beta_3$$

RSC_K pro kvadratický model
RSC_L pro lineární model

Testační kritérium linearity

$$F_L = \frac{(RSC_L - RSC_K)(n - 3)}{RSC_K (1)}$$

Test: Je-li $F_L < F_{1-\alpha}(1, n - 3)$, je H_0 přijata.

3. Linearita testem všech charakteristik

a) Střední kvadratická chyba predikce

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

kde $b_{(i)}$ je odhad, určený ze všech bodů kromě i-tého
Platí vztah

$$MEP = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{e}_i^2}{(1 - H_{ii})^2}$$

pro velké n jsou prvky $H_{ii} \approx 0$ a MEP = RSC/n.

b) Predikovaný koeficient determinace

$$\hat{R}_P^{\,2} = 1 - \frac{n\ \text{MEP}}{\sum_{i=1}^{n} y_i^2 - n\bar{y}}$$

c) Akaikovo informační kritérium

$$\text{AIC} = n \ln\left(\frac{\text{RSC}}{n}\right) + 2m$$

nejvhodnější model má AIC minimální

HODNOCENÍ KVALITY REGRESNÍHO MODELU

Střední kvadratická chyba predikce (MEP)

$$\text{MEP} = \frac{1}{n}\sum_{i=1}^{n} \frac{e_i^2}{(1 - H_{ii})^2}$$

e_i^2 čtverec reziduí modelu
H_{ii} *i*-tý diagonální prvek projekční matice H

Akaikovo informační kritérium (AIC)

$$\text{AIC} = n \cdot \ln\left(\frac{\text{RSC}}{n}\right) + 2m$$

RSC reziduální součet čtverců
m počet parametrů

Čím je AIC (MEP) menší, tím je model vhodnější.

Příklad 6.16 *Výběr ze tří polynomických regresních modelů*
Regresní analýzou dat vyšetřete, zda místo kvadratického modelu by lépe vyhovoval polynom třetího nebo pátého stupně.

Řešení:

a) Polynom 3. stupně:
MEP, $\hat{R}_P^2$ a AIC indikují polynom třetího stupně jako nejvhodnější

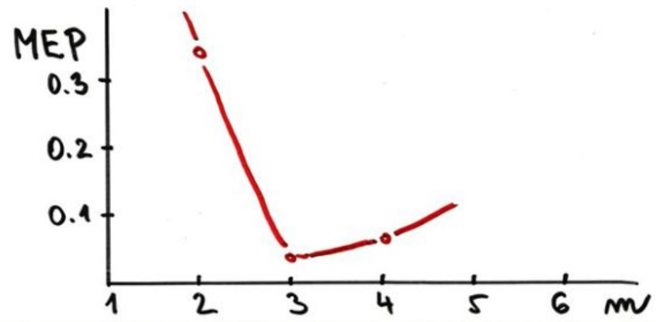
$$\hat{y}_P = 860.2\ (\pm\ 85.17) - 5.057\ (\pm\ 0.485)\ x + 9.77 \cdot 10^{-3}\ (\pm\ 9.19 \cdot 10^{-4})\ x^2 - 6.146 \cdot 10^{-6}\ (\pm\ 5.78 \cdot 10^{-7})\ x^3$$

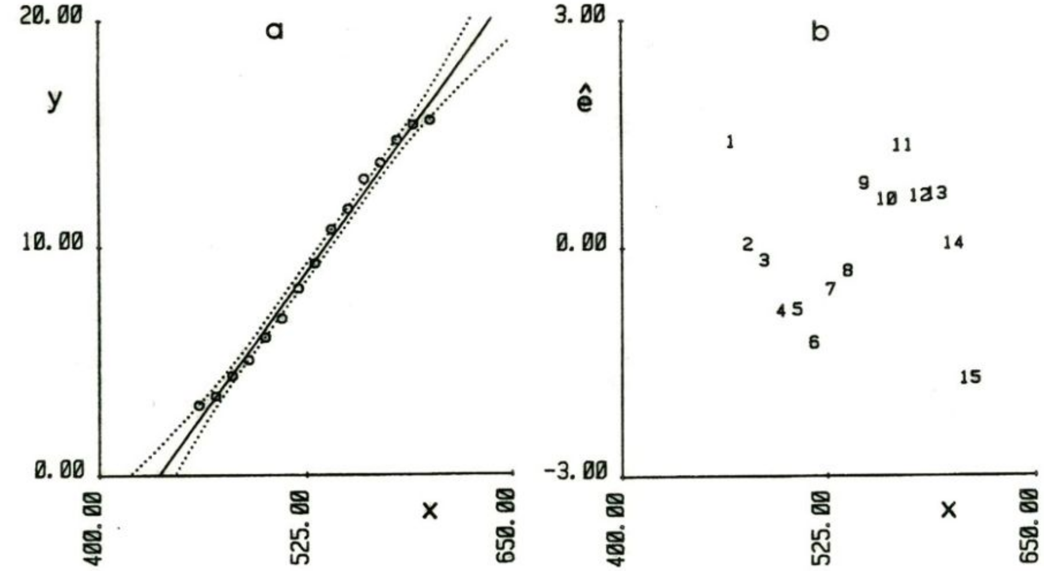
přičemž odhady všech tří parametrů vycházejí statisticky významné.

b) Polynom 5. stupně:
všechny parametry β kromě β_3 vycházejí statisticky nevýznamné, protože se zde projevuje multikolinearita.

Tabulka 6.3 Rozlišení stupně regresního polynomu statistikami MEP, $\hat{R}_P^2$, $\hat{R}^2$ a AIC

Stupeň polynomu	MEP	$\hat{R}_P^2$	$\hat{R}^2$	AIC
m = 2	0.3502	0.9905	0.9915	-21.65
3	0.0283	0.9992	0.9992	-56.02
5	0.0613	0.9983	0.9997	-55.04





Úloha L6.02 *Závislost výšky píku kyseliny kyanurové na koncentraci želatiny*

Při stanovení kyseliny kyanurové metodou diferenční pulsní polarografie byl sledován vliv přítomnosti povrchově aktivních látek. (1) Určete stupeň polynomu *m* závislosti výšky píku kyseliny kyanurové *y* na koncentraci želatiny *x*. (2) Které z kritérií, *MEP* nebo *AIC*, má lepší rozlišovací schopnost při určení stupně polynomu? (3) Pokuste se snížit multikolinearitu. Jak se indikuje multikolinearita v datech a jakou modifikaci MNČ je potom nutno použít k získání nejlepších odhadů neznámých parametrů **β** a výstavby regresního modelu?

Data: Koncentrace *x* [mg/l], výška píku *y* [mm]:

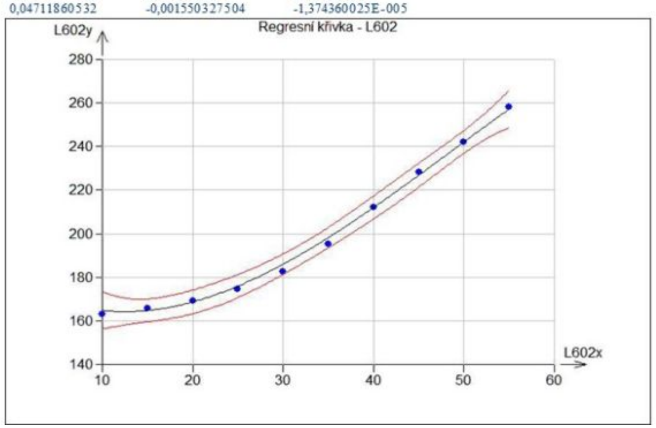
x	*y*
10	163.2
55	258.2

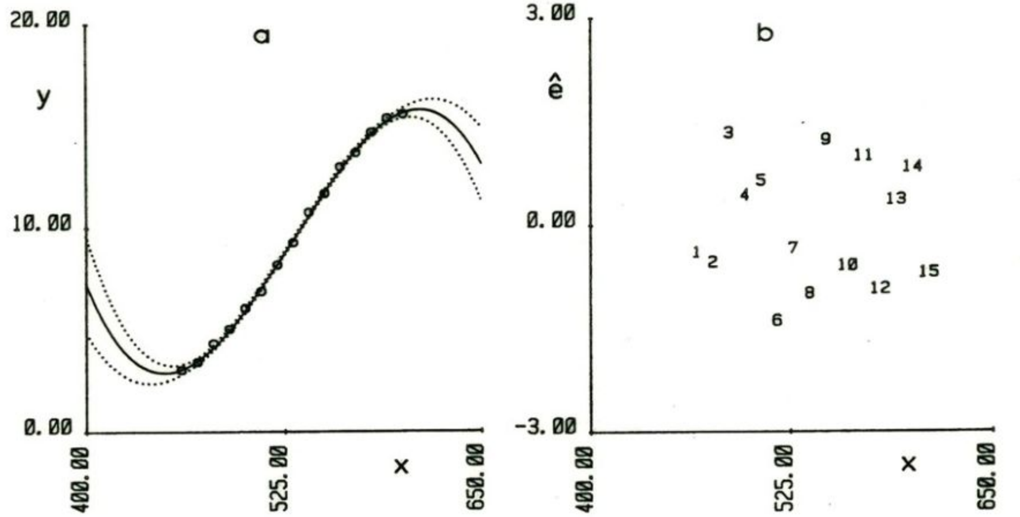
Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	179,0372815	7,930198308	Významný	4,943150438E-007	159,6327853	198,4417778
L602x	-2,4888912	0,915025471	Významný	0,03464167574	-4,727877869	-0,249904531
L602x^2	0,1141248515	0,03084462348	Významný	0,01008847791	0,03865077678	0,1895989262
L602x^3	-0,000782035	0,00031398431	Významný	0,04711860532	-0,001550327504	-1,374360025E-005

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R :	0,9986434468
Koeficient determinace R^2 :	0,9972887338
Predikovaný korelační koeficient Rp :	0,9707849726
Střední kvdratická chyba predikce MEP :	15,49028891
Akaikeho informační kritérium :	18,48707943





<u>Závěr:</u> Je patrné, že některé statistiky pro vystižení linearity modelu nebo vhodnosti specifikace modelu selhávají.

Úloha L6.10 *Závislost teploty tuhnutí nitrobenzenu na obsahu vody*

Byla změřena závislost bodu tuhnutí nitrobenzenu *y* na obsahu vody *x*. (1) Rozhodněte, zda je daná závislost lépe vystižena lineárním nebo kvadratickým modelem. (2) Užijte testu kvadratického členu. (3) Mají nalezené odhady parametrů statistický význam?

Data: Obsah vody *x* [hm. %], bod tuhnutí *y* [EC]:

x	*y*
0.041	5.57
...	...
0.38	5.25

Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	5,727095246	0,04398713614	Významný	9,989698224E-007	5,587108547	5,867081945
L610x	-4,246051891	1,009151765	Významný	0,02451525638	-7,457623196	-1,034480585
L610x^2	17,52172931	5,680999542	Nevýznamný	0,05395516144	-0,5577466946	35,60120531
L610x^3	-26,02187677	9,073748545	Nevýznamný	0,0641646745	-54,8985943	2,854840765

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R :	0,9951621992
Koeficient determinace R^2 :	0,9903478027
Predikovaný korelační koeficient Rp :	0,8534457695
Střední kvdratická chyba predikce MEP :	0,0013830323
Akaikeho informační kritérium :	-52,5456134