PSY117/454
Statistická analýza dat v psychologii
Přednáška 6

Vztahy mezi dvěma proměnnými III

Statistická predikce, modelování

Lineární regrese, parciální korelace

The only useful action for a statistician is to make predictions, and thus to provide basis for action.

William Edwards Deming

Statistická predikce

- Jaký výsledek v inteligenčním testu lze nejspíše očekávat od náhodně přišedšího, víme-li, že test má přibližně normální rozložení s průměrem 100 a směrodatnou odchylkou 15 ?
- Jaká informace by nám pomohla zpřesnit náš odhad?
  - délka vlasů
  - vzdělání
  - výsledek v testu paměti
  - výsledek v jiném inteligenčním testu

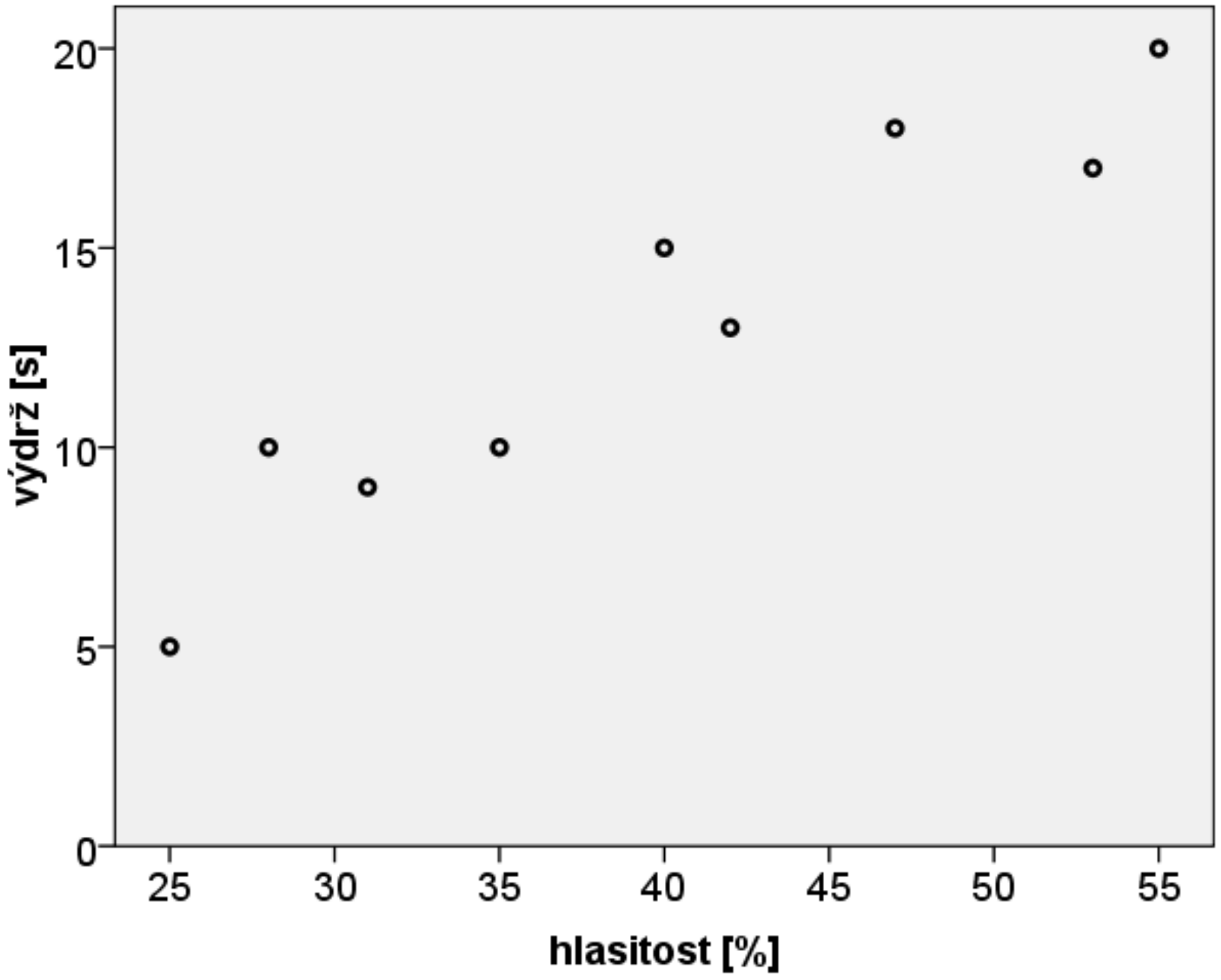
- **Statistická predikce** je předpovídání (kvalifikované odhadování) nejpravděpodobnější hodnoty proměnné z údajů, které již známe, a to pomocí **modelu vztahu** mezi predikovanou proměnnou a jejími **koreláty**.

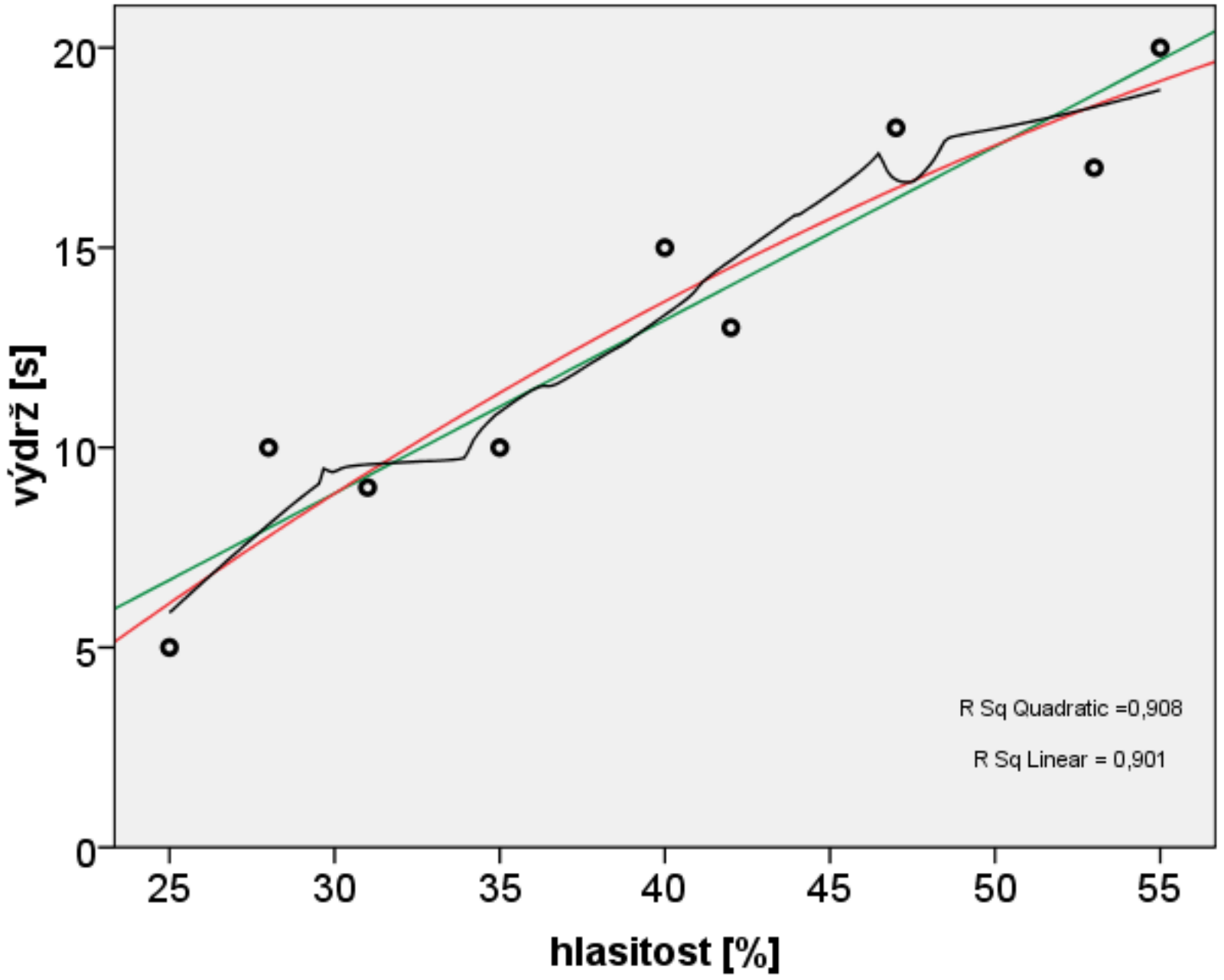
AJ: statistical prediction, estimate, predicted value

Dhodobá adaptace sluchu

Lidé, kteří poslouchají osobní přehrávač na vysokou **hlasitost** [% z maxima přehrávače], **vydrží** nepříjemný hlasitý zvuk déle?

hlasitost [%]	výdrž [s]
25	5
31	9
55	20
42	13
47	18
53	17
40	15
35	10
28	10


20
15
10
5
0
výdrž [s]
25
30
35
40
45
50
55
hlasitost [%]


20
15
10
5
0
výdrž [s]
25
30
35
40
45
50
55
hlasitost [%]
R Sq Quadratic =0,908
R Sq Linear = 0,901

K predikci je třeba funkce

- fce = jak ze známé hodnoty X vypočítat tu neznámou Y : $\boldsymbol{Y = f(X)}$
  - různé fce: - stanovené výčtem
    - trigonometrické, exponnciální a logaritmické ...
    - polynomické: lineární: $Y = bX + a$ (rovná čára)
      kvadratické: $Y = cX^2 + bX + a$ (jedna zatáčka)

Ve statistice...

- tuto funkci odhadujeme (modelujeme)
  - Jak dobře dokážeme vyjádřit (=predikovat) Y pomocí X a funkce f ?
- říkáme výsledku výpočtu **odhad** (Y') a stanovení té funkce říkáme **regrese**
- regrese Y na X : $Y' = f(X) + e$,kde $e = Y - Y'$ (1)
  - e je reziduální hodnota (reziduum), Y je závislá p., X je prediktor (nezáv.)
  - e představuje všechny ostatní zdroje variability vyjma X

AJ: function, polynomial, linear, quadratic, estimation, modelling, estimate *n.*, regression, residual *n.*, predictor, sources of variablity(variance), dependent and independent variable

Lineární regrese I. - odhad

Je-li Pearsonova korelace dobrým popisem vztahu mezi dvěma proměnnými, lze popsat vztah mezi nimi lineární funkcí

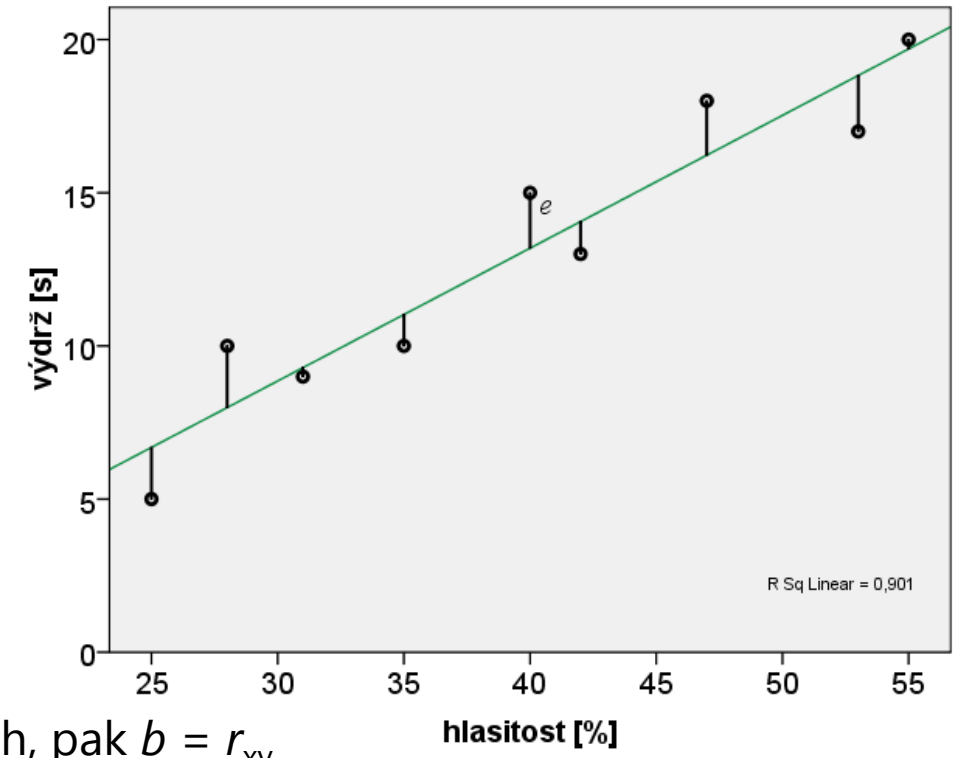
$$Y' = a + bX$$

b – směrnice

a – průsečík

$Y = Y' + e$

$Y = a + bX + e$



Odhad metodou nejmenších čtverců

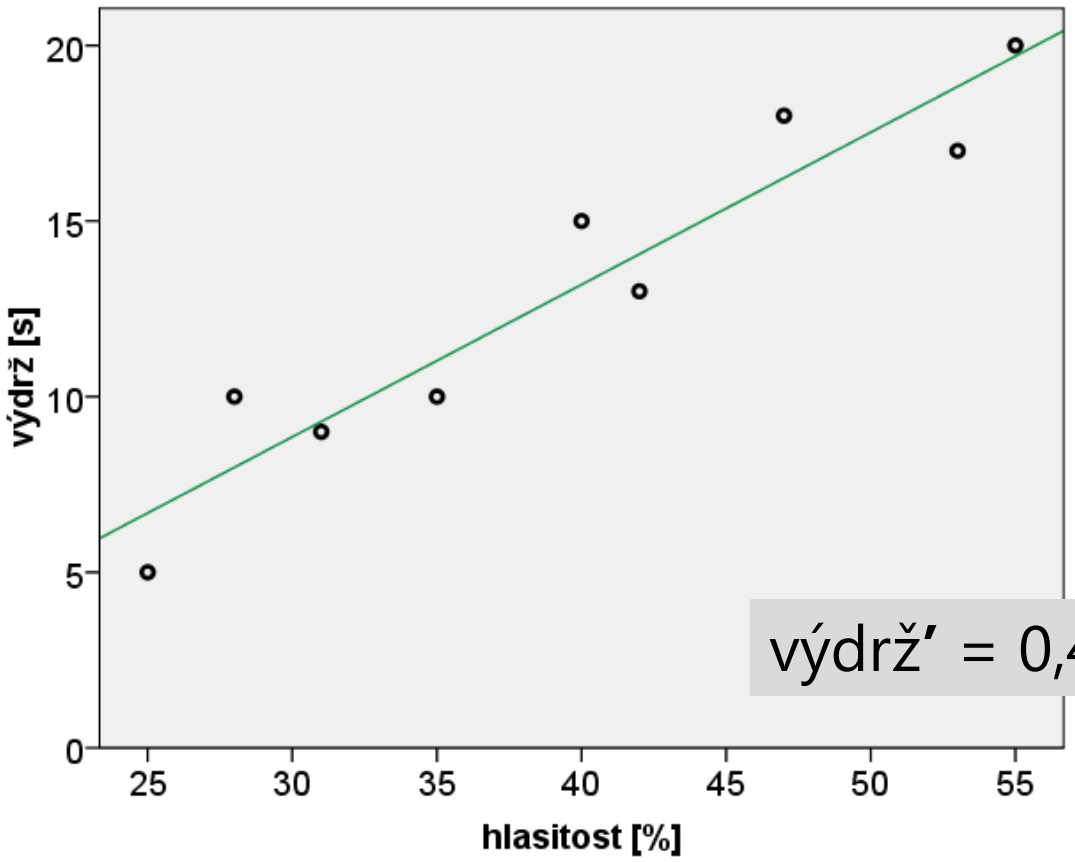
$$b = r_{xy}(s_y/s_x)$$

$$a = m_y - bm_x$$

Jsou-li X a Y vyjádřeny v z-skórech, pak $b = r_{xy}$

AJ: slope, intercept, least squares (estimation), regression coefficents (a,b)

Lineární regrese II. – příklad



m_h=39,6

s_h = 10,7

m_v=13,0

s_v = 4,9

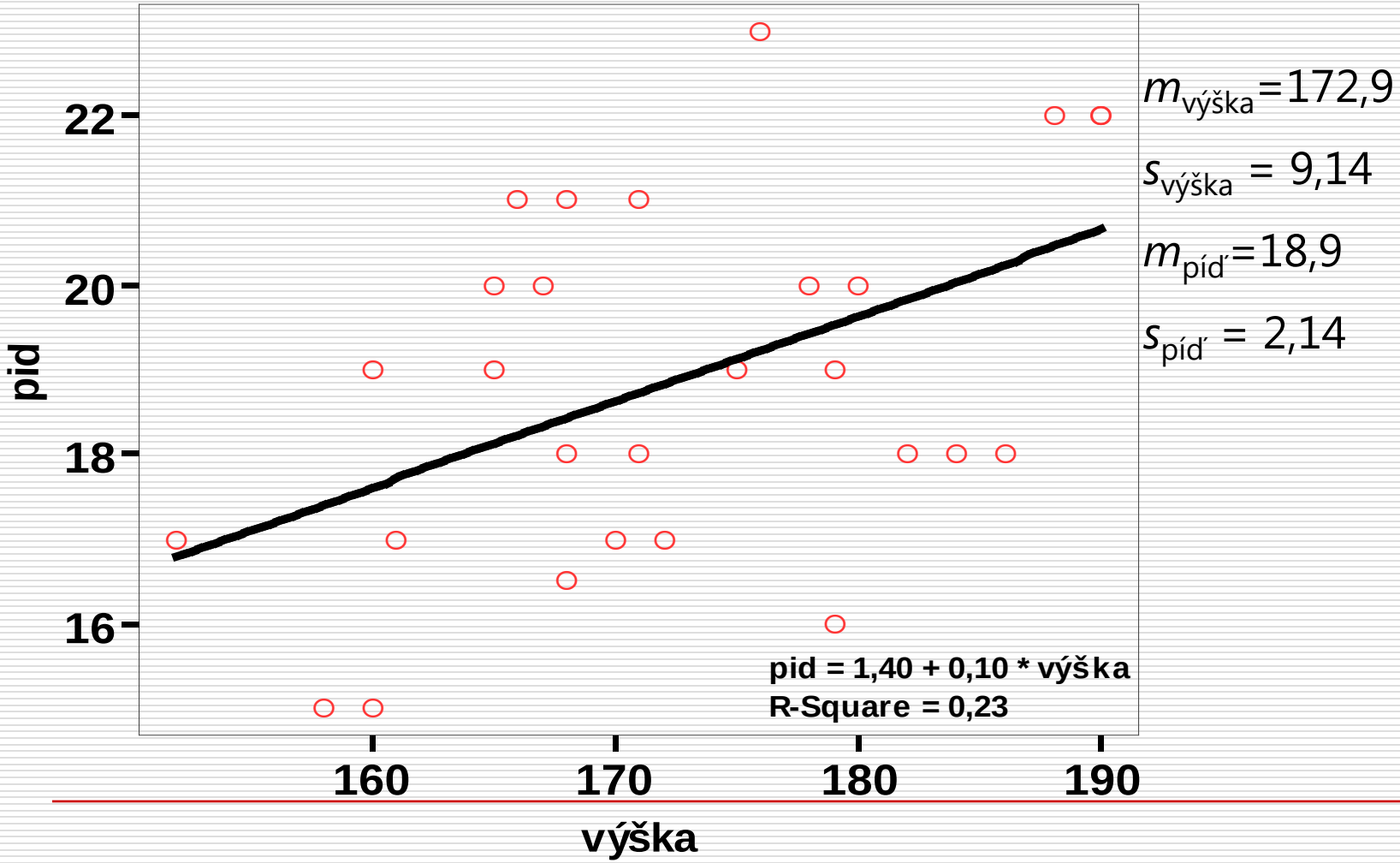
r = 0,95

výdrž' = 0,43.hlasitost – 4,15

Predikované hodnoty a rezidua

hlasitost [%]	výdrž [s]	výdrž' [s]	reziduum [s]
25	5	6,69	-1,69
31	9	9,29	-0,29
55	20	19,70	0,30
42	13	14,06	-1,06
47	18	16,23	1,77
53	17	18,83	-1,83
40	15	13,19	1,81
35	10	11,02	-1,02
28	10	7,99	2,01

Lineární regrese II. – příklad



$m_{výška}$=172,9

$s_{výška}$ = 9,14

$m_{píď}$=18,9

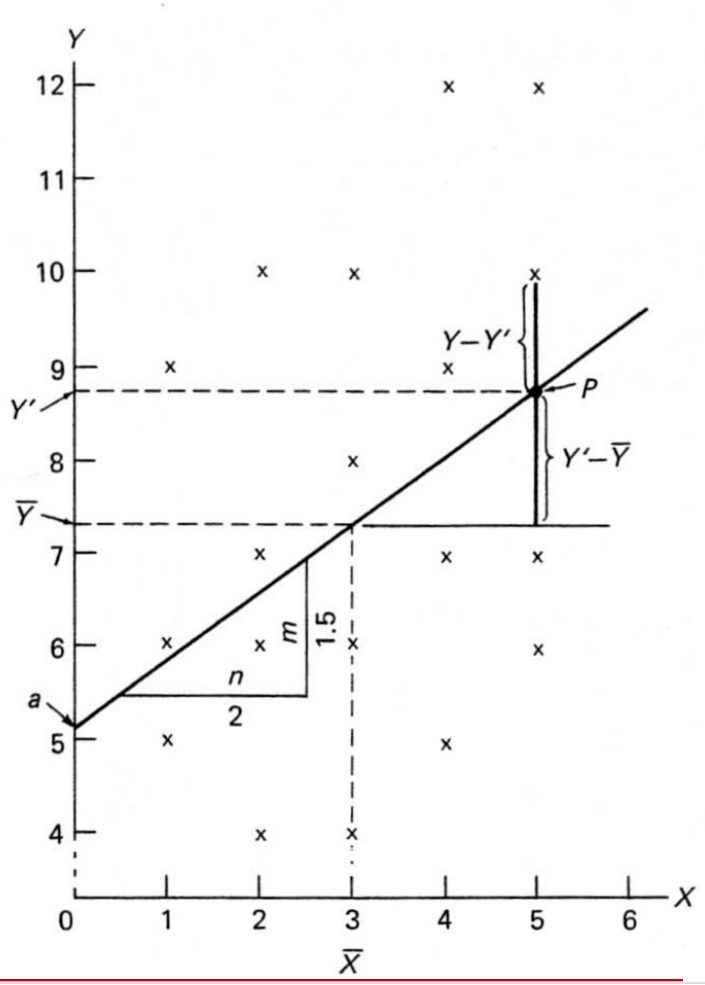
$s_{píď}$ = 2,14

Lineární regrese III. – úspěšnost predikce

$$s_{reg}^2 = \frac{\sum (m_y - Y')^2}{n-1} \qquad s_{res}^2 = \frac{\sum (Y - Y')^2}{n-1}$$

$$s_y^2 = \frac{\sum (Y - m_y)^2}{n-1}$$

- $s_y^2 = s_{reg}^2 + s_{res}^2$ $(ss_y = ss_{res} + ss_{reg})$
- $R^2 = s_{reg}^2 / s_y^2$
- Koeficient determinace (R^2)
  - Podíl vysvětleného rozptylu
  - Je ukazatelem kvality, úspěšnosti regrese
  - Vyjadřuje shodu modelu s daty
- Pro jednoduchou lin. regr. platí $R^2 = r^2$

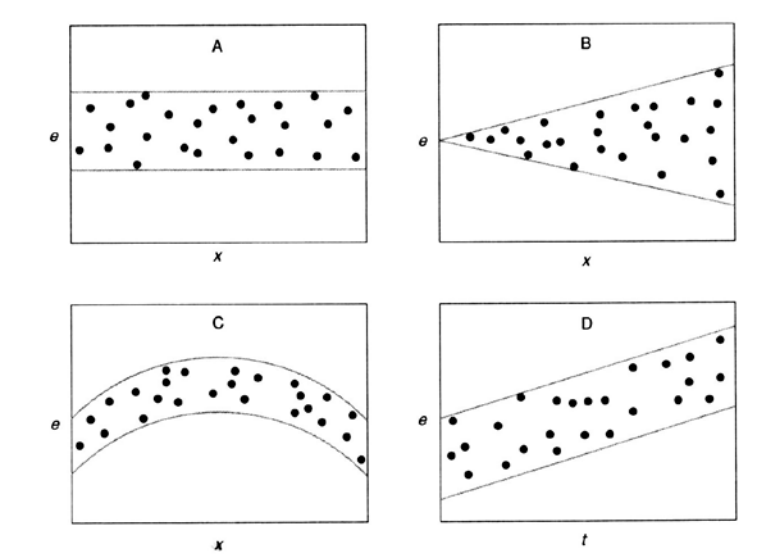


AJ: regression and residual variance (sum of squares), explained variance, model fit with the data, coefficient of determination (R square)

Lineární regrese IV. – předpoklady, platnost

Předpoklady oprávněnosti použití lineárně-regresního modelu

- jako u Pearsonovy korelace
- konceptuální předpoklad: vztah je ve skutečnosti lineární
- rezidua mají normální rozložení s průměrem 0
- homoskedascita
  - =rozptyl reziduí (chyb odhadu) se s rostoucím *X* nemění



- Platnost modelu je omezena daty, z nichž byl získán, a teorií.
  - Extrapolace, neoprávněná extrapolace (≈jako generalizace nad rámec empirických dat)
  - Pozor na odlehlé hodnoty – jako u všech ostatních momentových statistik

AJ: assumptions of the linear regression model, residuals normally distributed, homoscedascity,

Další druhy regrese

Zde je prezentovaná pouze jednoduchá lineární regrese, tj. s jednou závislou a jednou nezávislou proměnnou. Potřeb a možností je více.

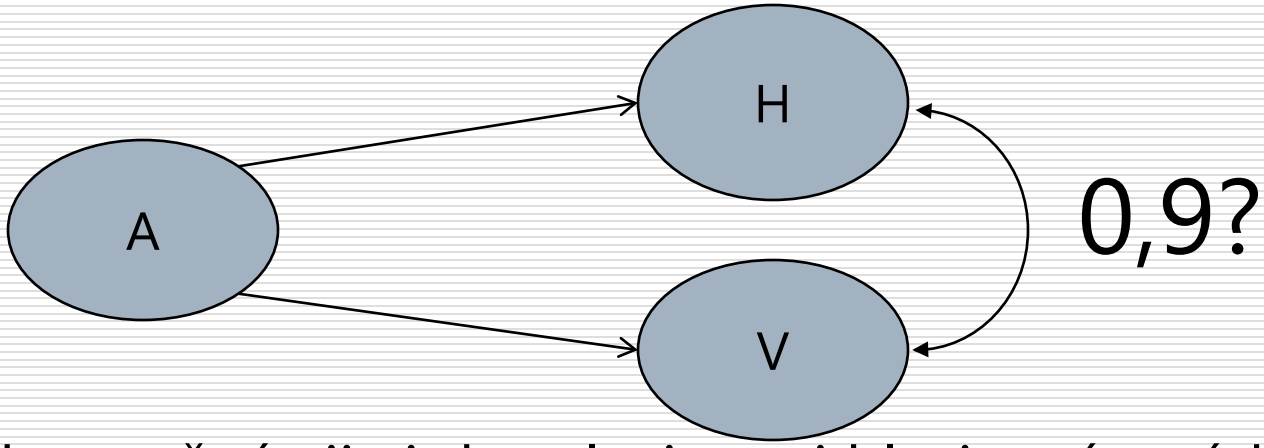
- mnohočetná (mnohonásobná) lineární regrese
  - $Y = a + b_1X_1 + b_2X_2 + \ldots + b_mX_m$
  - komplikují ji vztahy mezi nezávislými proměnnými - prediktory
- logistická regrese
  - pokud je závislá dichotomie, nominální proměnná
  - predikuje se tak pravděpodobnost jednotlivých hodnot závislé

- Není-li vztah lineární
  - snažíme se transformovat proměnné tak, aby byl lineární.
  - dělíme vzorek na podskupiny, v nichž vztah za lineární považovat lze

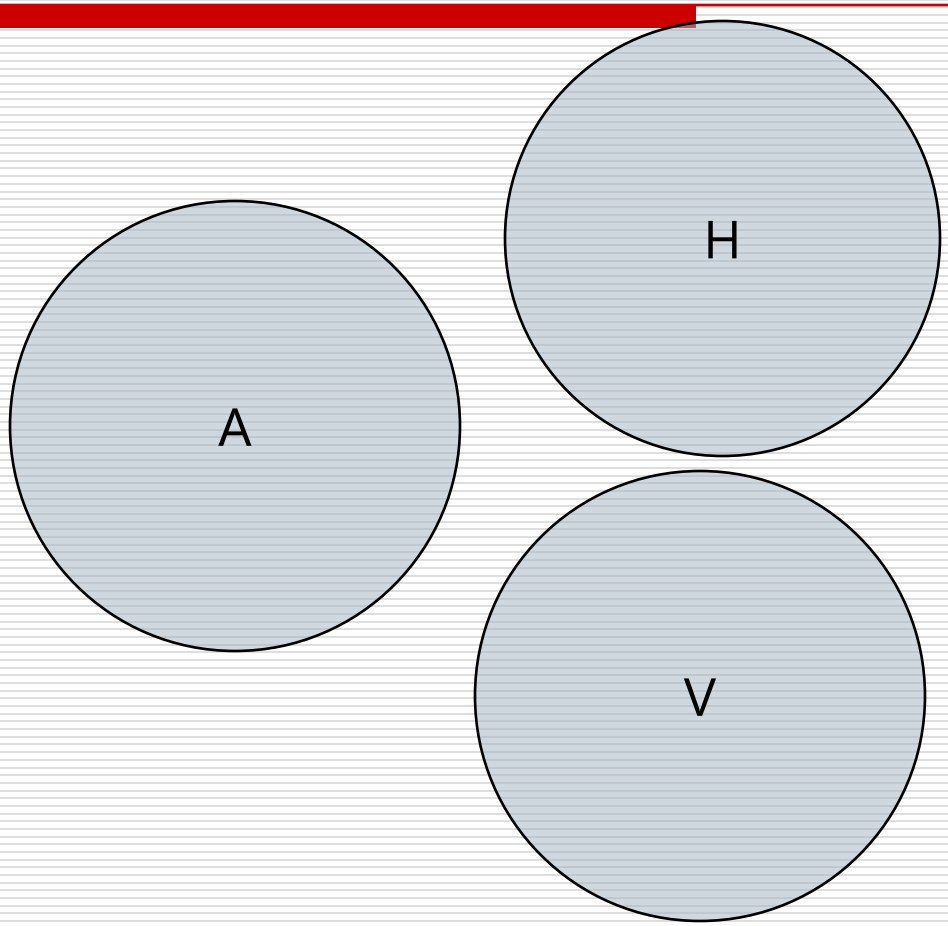
AJ: multiple regression, logistic regression, nonlinear regression

Vztah mezi třemi proměnnými
Parciální a semiparciální korelace

Zjistili jsme, že účastníci našeho experimentu se nám opili. To nám vadí, protože opilost snižuje citlivost na podněty a zvyšuje obě naše proměnné.



Bylo by možné zjistit korelaci mezi hlasitostí a výdrží, bez vlivu alkoholu?


H
A
V

Jak ale rozdělovat ty rozptyly?

Regrese dělí proměnnou na sdílený rozptyl a reziduální rozptyl….

Parciální korelace $r_{HV.A}$

- Uděláme regresi výdrže na alkohol – reziduum výdrže bez alkoholu
- Uděláme regresi hlasitosti na alkohol – reziduum hlasitosti bez alkoholu
- Korelace dvou reziduí je PARCIÁLNÍ KORELACE

$$r_{HV.A} = \frac{r_{HV} - r_{HA}r_{VA}}{\sqrt{1-r_{HA}^2}\sqrt{1-r_{VA}^2}}$$

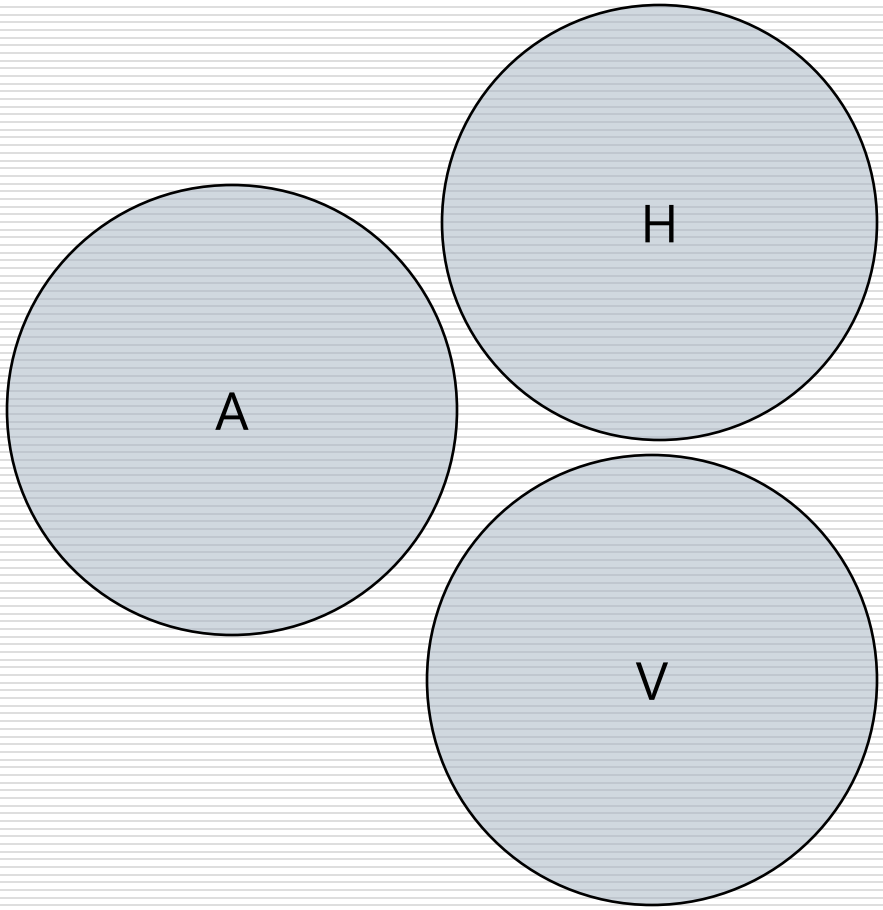
Semiparciální korelace $r_{V(H.A)}$

- Korelace rezidua (V.A) se závislou proměnnou (H)

$$r_{H(V.A)} = \frac{r_{HV} - r_{HA}r_{VA}}{\sqrt{1-r_{VA}^2}}$$

AJ: partial correlation, part (semi-partial) correlation

Korelace mezi hlastiostí a výdrží , **kontrolujeme-li statisticky*** alkohol je…



	hlasitost	vydrz	alkohol
hlasitost	1,000	,949**	,864**
vydrz	,949**	1,000	,902**
alkohol	,864**	,902**	1,000

$r_{HV.A} = 0{,}78$

* Též, „pokud by alkohol byl konstantní“

Shrnutí

- Pro praktické účely (predikce/odhad) je korelace málo, je třeba uvažovat o funkčním vztahu mezi proměnnými.
- Vztah můžeme znát analyticky nebo ho zkoušet modelovat.
- Lineární regrese je model lineár. vztahu mezi proměnnými.
- Model se vždy liší od skutečných dat
  - díky zjednodušení
  - díky chybě měření
- Míra shody modelu s daty je ukazatelem vhodnosti modelu.
  - U lineární regrese R^2 – podíl vysvětleného rozptylu
- Vliv nežádoucích třetích proměnných lze někdy eliminovat použitím parciální nebo semiparciální korelace.

- Hendl: kapitoly 7.3 – 7.3.2, 7.3.6, 7.4