PSY117/454
Statistická analýza dat v psychologii
Přednáška 6

Vztahy mezi dvěma proměnnými II

Statistická predikce, modelování

Lineární regrese

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?
  - jeho/její pohlaví
  - 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 **modelů** vztahu mezi predikovanou proměnnou a jejími koreláty.

AJ: statistical prediction, estimate, predicted value, line slope

Složitý model

POHLAVÍ

VZDĚLÁNÍ

IQ TEST

PAMĚŤ

IQ TEST 2

K predikci je třeba funkce

- $Y = f(X)$
- funkce je „návod“, jak ze známé hodnoty (X) vypočítat tu neznámou (Y)
  - jsou různé funkce…
    - stanovené výčtem
    - trigonometrické, exponenciální 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) proměnnou Y, pomocí proměnné 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

$$Y' = a + bX + e$$

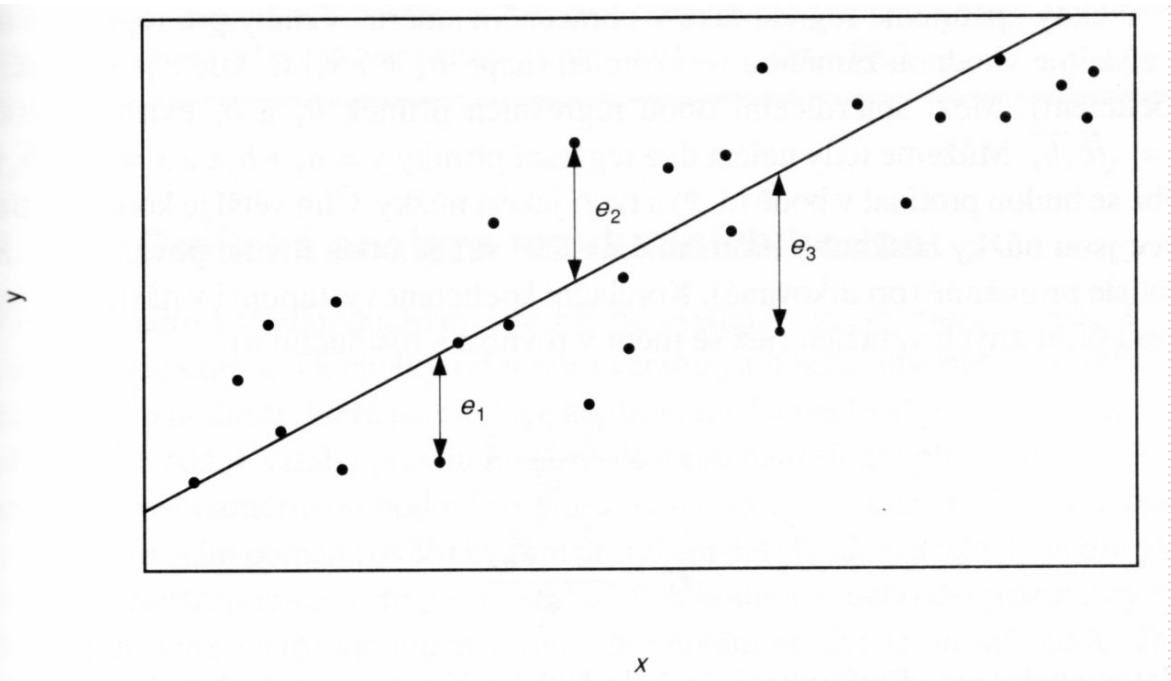
b – směrnice

a – průsečík

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 $b = r_{xy}$

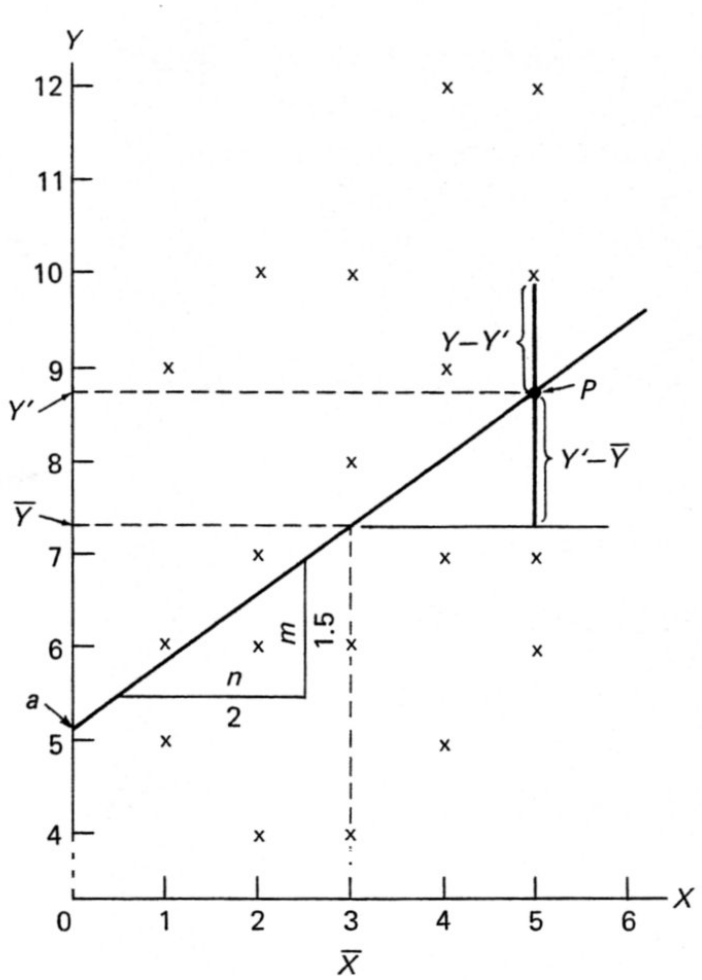
AJ: slope, intercept, least squares (estimation), regression coefficents (a,b)

Lineární regrese II. – úspěšnost predikce

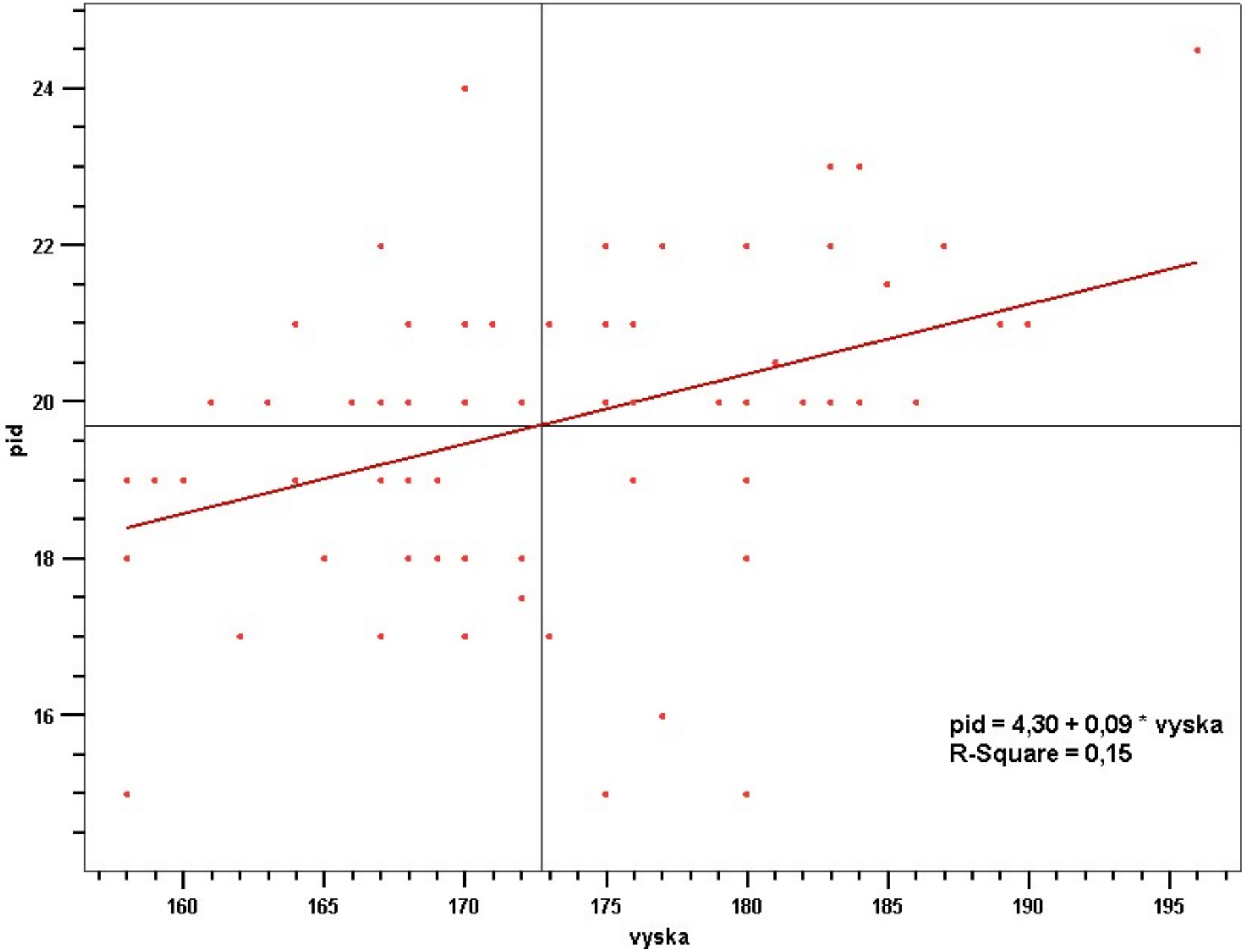
$$s^2_{reg} = \frac{\sum(\bar{Y} - Y')^2}{n-1} \qquad s^2_{res} = \frac{\sum(Y - Y')^2}{n-1}$$

$$s^2_y = \frac{\sum(Y - \bar{Y})^2}{n-1}$$

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



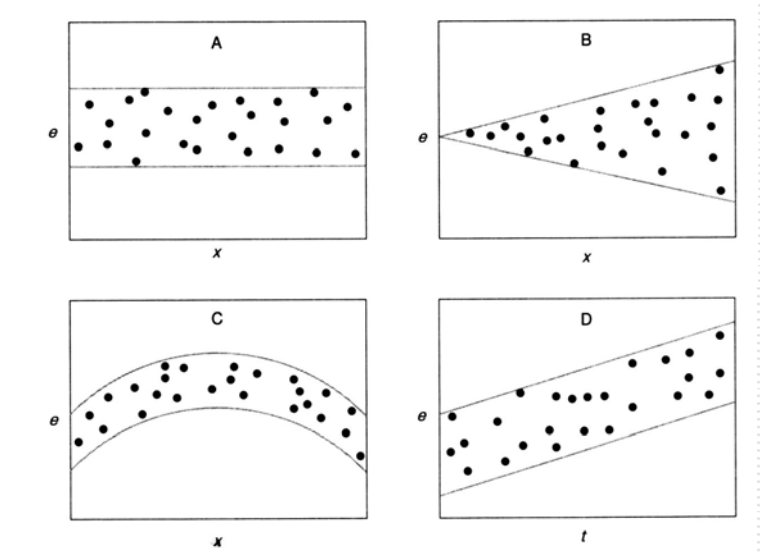
AJ: regression and residual variance (sum of squares), explained variance, model fit with the data, coefficient of determination (R square)


24
22
20
18
16
pid
160
165
170
175
180
185
190
195
vyska
pid = 4,30 + 0,09 * vyska
R-Square = 0,15

Lineární regrese IV. – předpoklady, platnost

Předpoklady oprávněnosti použití lineárně-regresního modelu

- 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.
  - nelineární regrese je spojena s mnoha obtížemi

AJ: multiple regression, logistic regression, nonlinear regression

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 lineárního vztahu mezi proměnnými.
- Model se vždy liší od skutečných dat
  - díky zjednodušení
  - díky chybě měření
- Míra shody modelu s daty je ukazatelem vhodnosti modelu.

- Hendl: kapitoly 7.3 – 7.3.2, 7.3.6, pro absolventy metodologie i 7.4