PSY117
Statistická analýza dat v psychologii
Přednáška 6 - 2016

Vztahy mezi dvěma proměnnými II

Statistická predikce - lineární regrese

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

William Edwards Deming

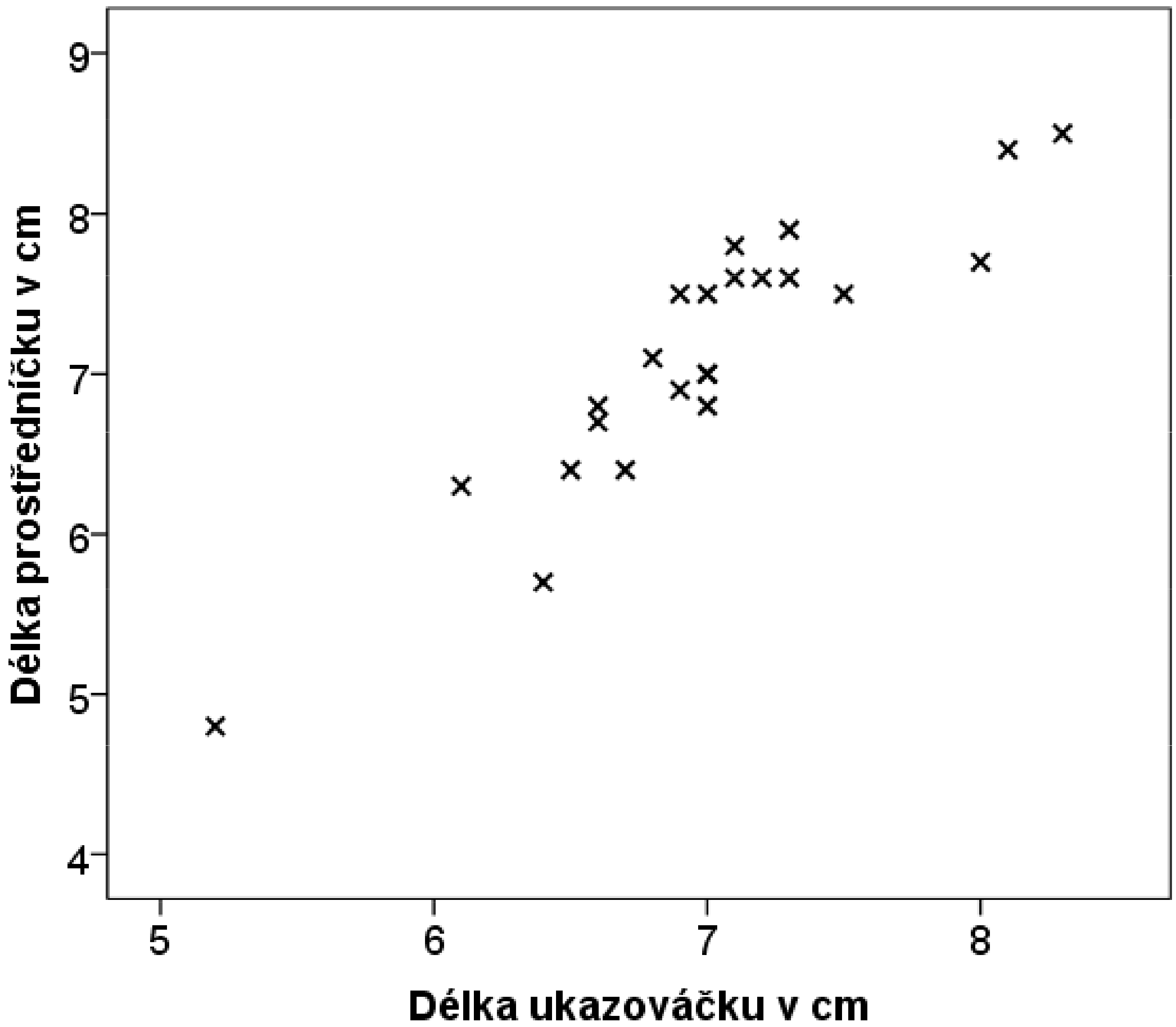
Statistická predikce

- 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?
  - délka vlasů: l = 31 cm
  - vzdělání: *vysokoškolské*
  - výsledek v testu paměti: z = 1,6
  - výsledek v jiném inteligenčním testu: IQ = 108

- **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

Predikce délky prostředníčku z ukazováčku



K predikci je třeba funkce

- fce = jak ze známé hodnoty X vypočítat tu neznámou Y : $\boldsymbol{Y = f(X)}$
  - různé fce: - stanovené výčtem

    - trigonometrické, exponenciální a logaritmické ...

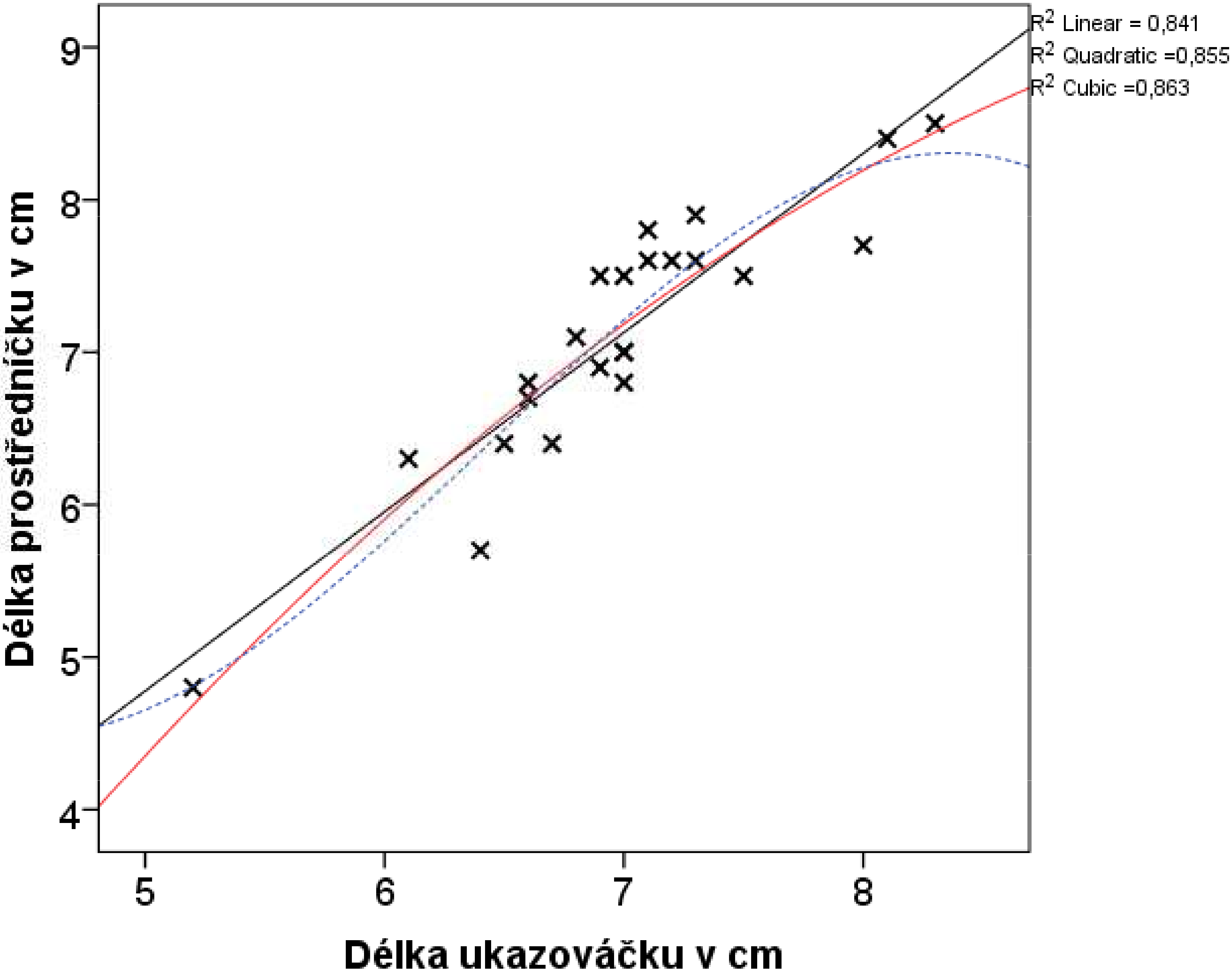
    - polynomické: $\boxed{\text{lineární: } Y = bX + a}$ (rovná čára ... Pearsonova *r*)

    kvadratické: $Y = cX^2 + bX + a$ (jedna zatáčka)

Ve statistice...

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

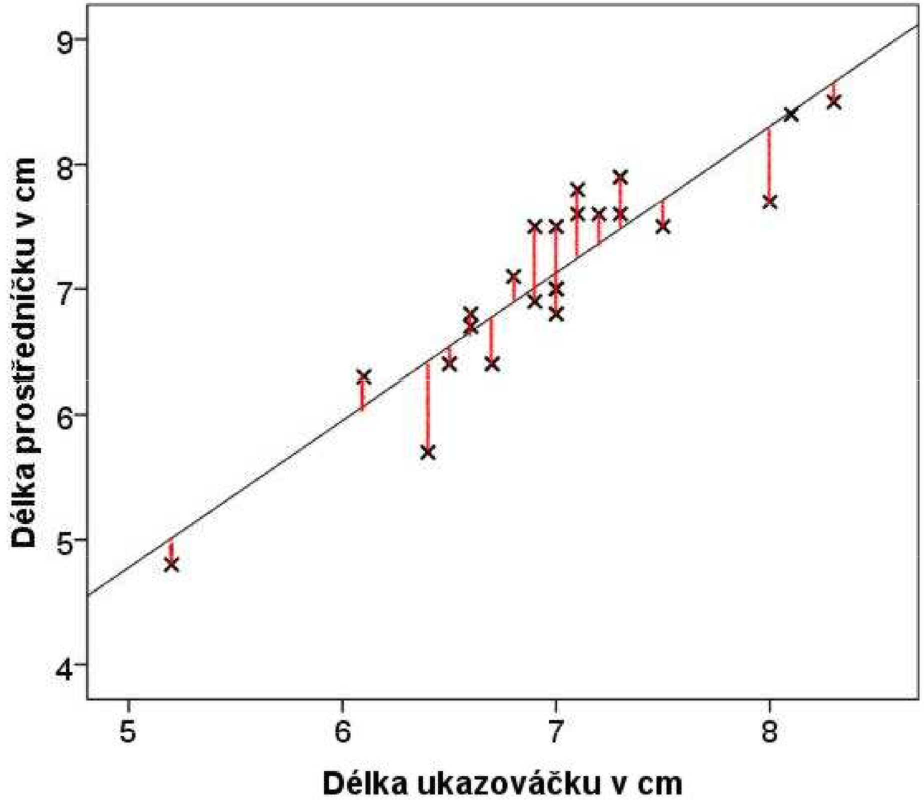
AJ: function, polynomial, linear, quadratic, estimation, modelling, estimate *n.*, regression, residual *n.*, predictor, sources of variability(variance), dependent and independent variable



R2 Linear = 0,841
R2 Quadratic =0,855
R2 Cubic =0,863
Délka prostředníčku v cm
Délka ukazováčku v cm
4
5
6
7
8
9

Lineární regrese I. – odhad přímou úměrou

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



$$Y' = a + bX$$

b ... směrnice

a ... průsečík

$(Y' - m_y) = b(X - m_x)$

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

Nejlepší přímka?

Odhad metodou **nejmenších čtverců**

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

$$a = m_y - bm_x$$

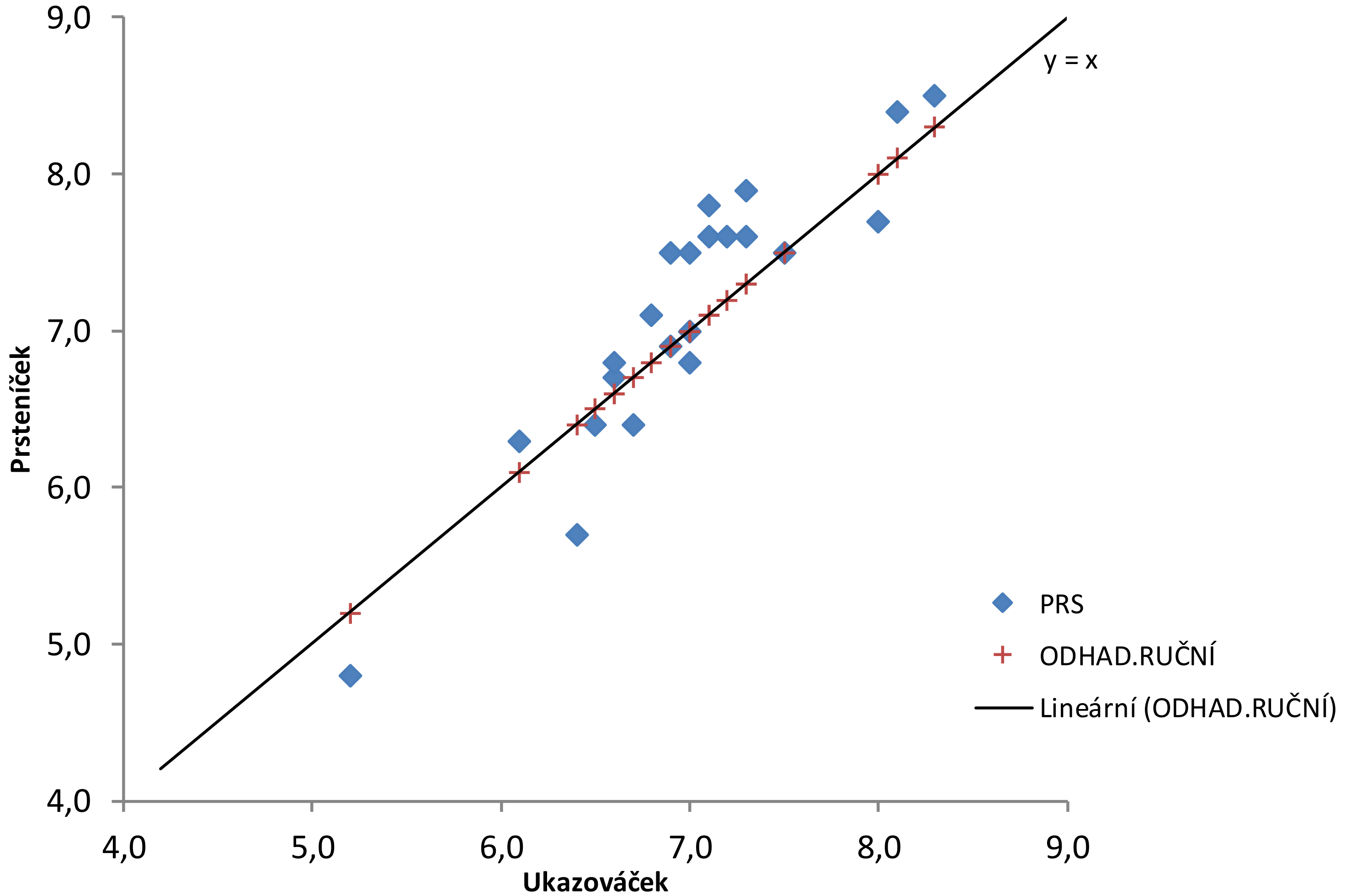
Jsou-li X a Y vyjádřeny v z-skórech, pak $b = r_{xy}$

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

Nastavme si parametry regresní přímky ručně

průsečík	0
směrnice	1

X	Y	
UKA	PRS	ODHAD. RUČNÍ
6,5	6,4	6,5
7,0	7,0	7,0
7,5	7,5	7,5
5,2	4,8	5,2
6,6	6,7	6,6
6,6	6,8	6,6
7,0	7,0	7,0
7,1	7,8	7,1
6,8	7,1	6,8
7,0	6,8	7,0
6,9	6,9	6,9
6,7	6,4	6,7
6,4	5,7	6,4
7,1	7,6	7,1
6,9	7,5	6,9
7,2	7,6	7,2
6,1	6,3	6,1
7,3	7,6	7,3
7,0	7,5	7,0
8,3	8,5	8,3
7,3	7,9	7,3
8,0	7,7	8,0
8,1	8,4	8,1

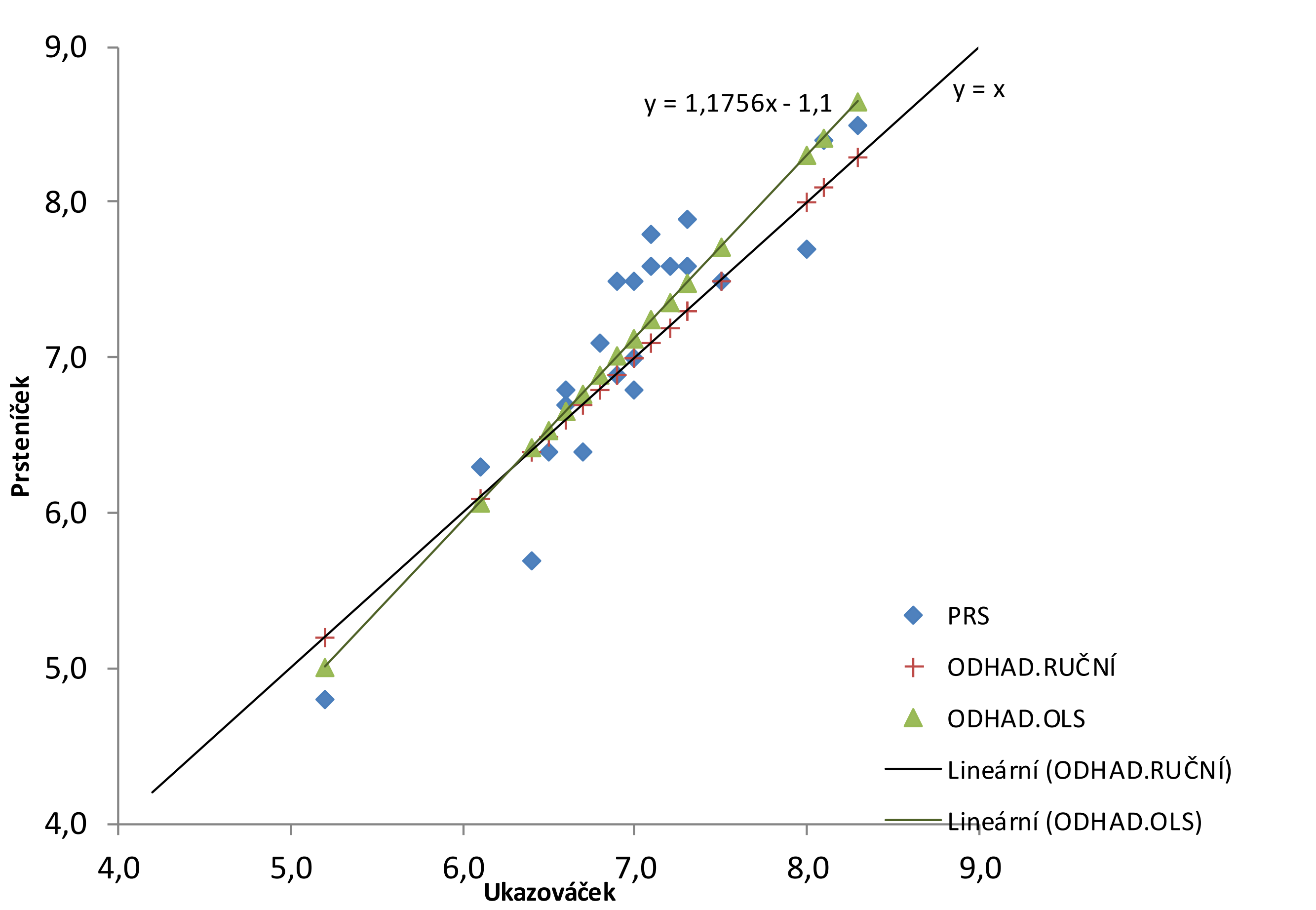


Jak stanovit „nejlepší přímku“?

- Více možných kritérií
- Kritérium nejmenších čtverců
  - Snažíme se minimalizovat sumu čtverců reziduí

Nastavme si parametry regresní přímky ručně						průsečík	0	-1,10 OLS
						směrnice	1	1,18

X UKA	Y PRS	ODHAD. RUČNÍ	ODHAD. OLS	e.RUČNÍ	e.OLS
6,5	6,4	6,5	6,54	0,10	0,14
7,0	7,0	7,0	7,13	0,00	0,13
7,5	7,5	7,5	7,72	0,00	0,22
5,2	4,8	5,2	5,01	0,40	0,21
6,6	6,7	6,6	6,66	-0,10	-0,04
6,6	6,8	6,6	6,66	-0,20	-0,14
7,0	7,0	7,0	7,13	0,00	0,13
7,1	7,8	7,1	7,25	-0,70	-0,55
6,8	7,1	6,8	6,89	-0,30	-0,21
7,0	6,8	7,0	7,13	0,20	0,33
6,9	6,9	6,9	7,01	0,00	0,11
6,7	6,4	6,7	6,78	0,30	0,38
6,4	5,7	6,4	6,42	0,70	0,72
7,1	7,6	7,1	7,25	-0,50	-0,35
6,9	7,5	6,9	7,01	-0,60	-0,49
7,2	7,6	7,2	7,36	-0,40	-0,24
6,1	6,3	6,1	6,07	-0,20	-0,23
7,3	7,6	7,3	7,48	-0,30	-0,12
7,0	7,5	7,0	7,13	-0,50	-0,37
8,3	8,5	8,3	8,66	-0,20	0,16
7,3	7,9	7,3	7,48	-0,60	-0,42
8,0	7,7	8,0	8,30	0,30	0,60
8,1	8,4	8,1	8,42	-0,30	0,02
			SUMA:	-2,90	0,00
			SUMA.ČT:	3,15	2,49



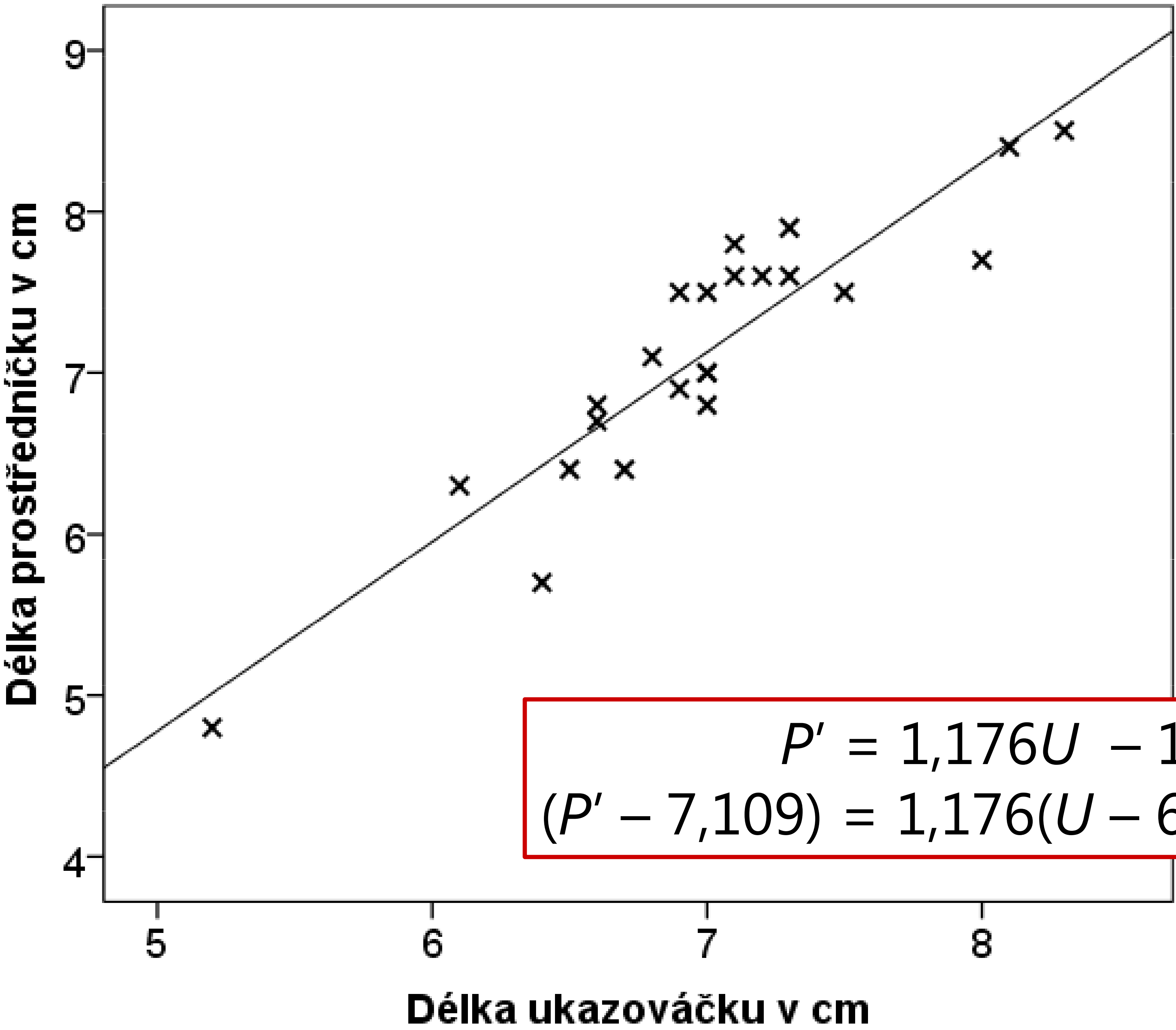
Řešení metodou nejmenších čtverců

$Y' = a + bX$: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 $b = r_{xy}$
- Přímka prochází m_x a m_y
- Průměr Y a Y' je stejný
- Součet reziduí je nulový, součet reziduí umocněných na druhou nejmenší možný



$m_P = 7{,}109$
$s_P = 0{,}843$

$m_U = 6{,}983$
$s_U = 0{,}658$

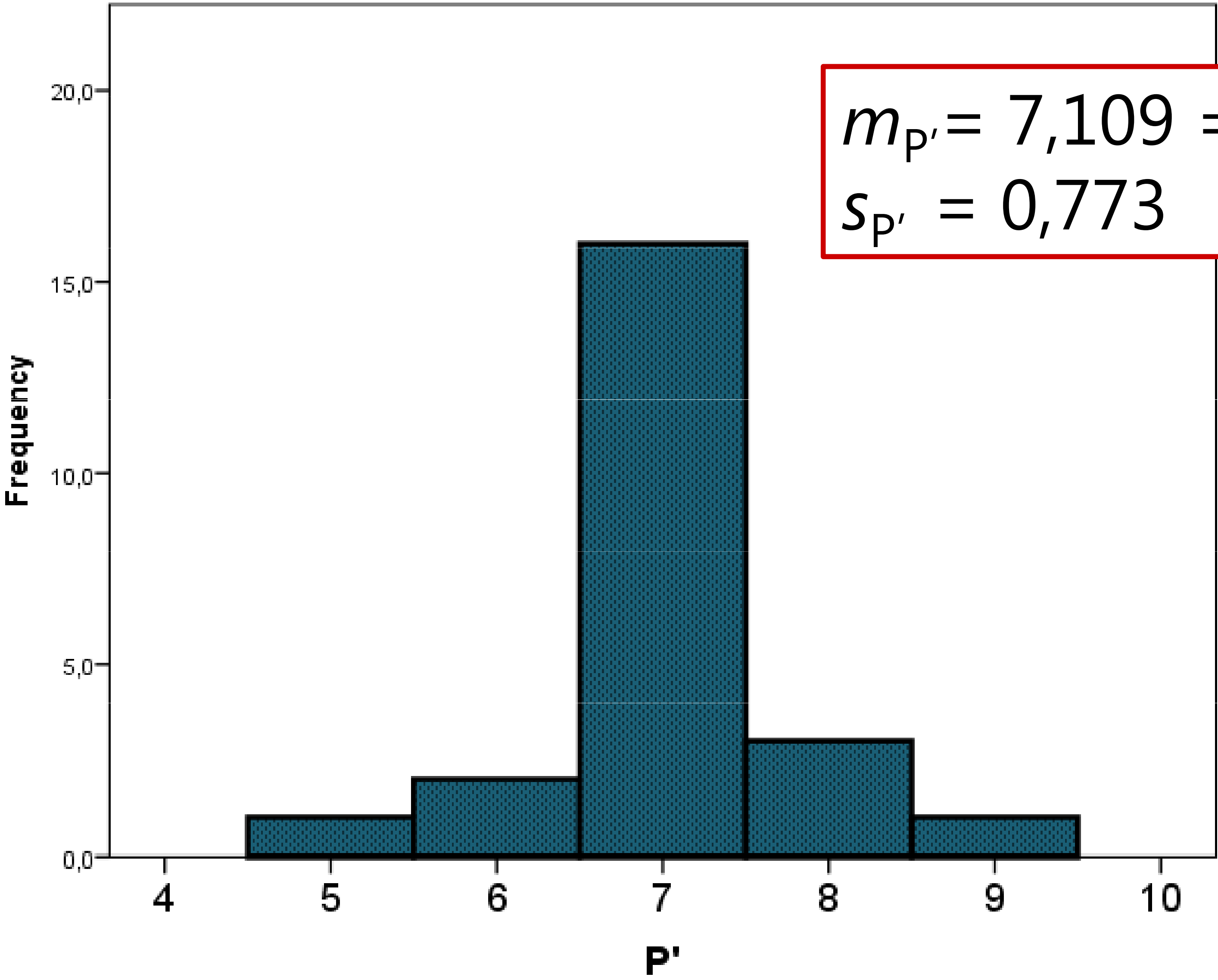
$r_{PU} = 0{,}917$

$$P' = 1{,}176U - 1{,}100$$
$$(P' - 7{,}109) = 1{,}176(U - 6{,}983)$$

Predikované hodnoty

U	P	P'
6,5	6,4	6,5413
7	7	7,1291
7,5	7,5	7,7169
5,2	4,8	5,0130
6,6	6,7	6,6589
6,6	6,8	6,6589
7	7	7,1291
6,8	?	

Rozložení predikovaných hodnot



$m_{P'} = 7{,}109 = m_P$

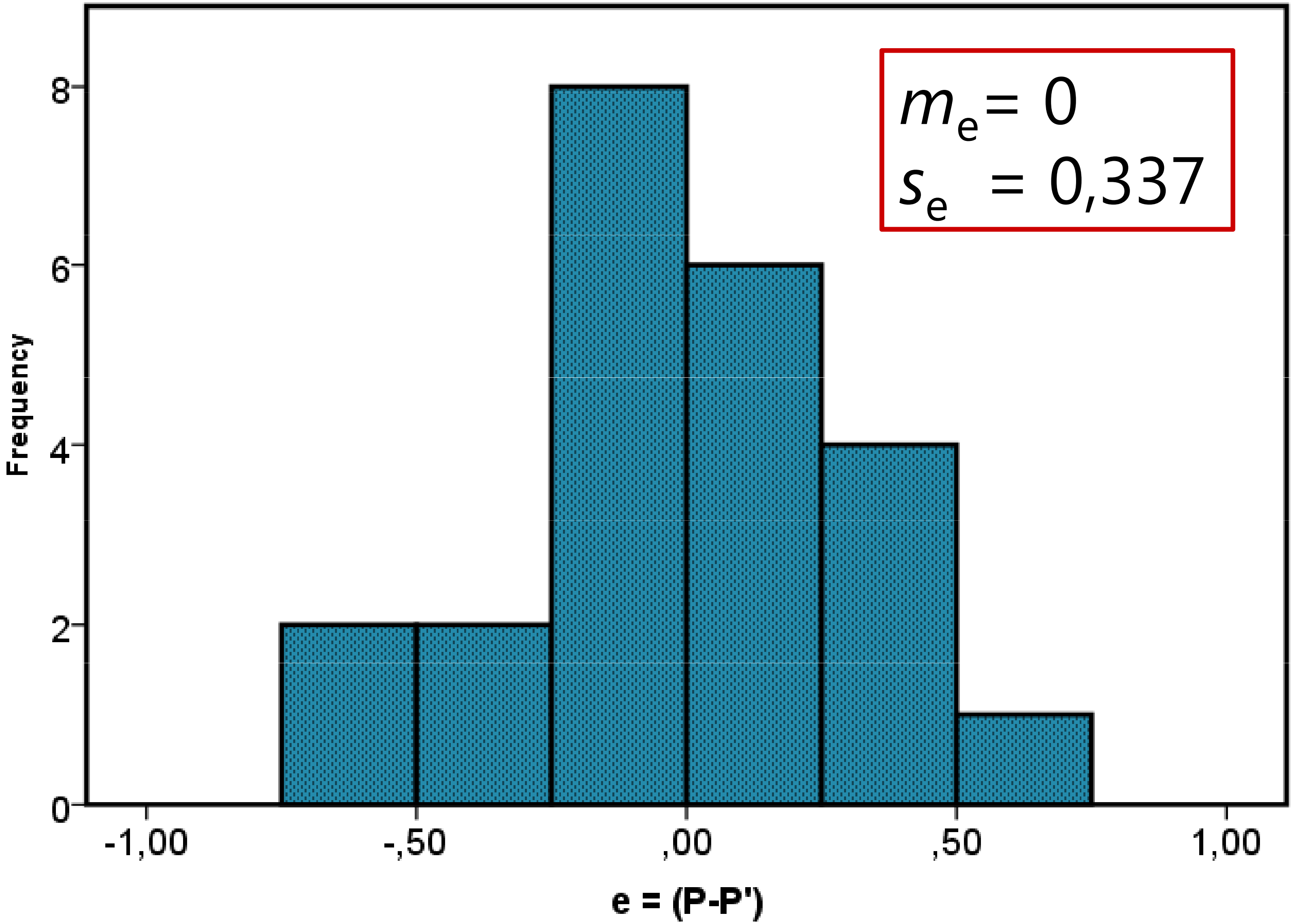
$s_{P'} = 0{,}773$

Lineární regrese II. – úspěšnost predikce

- Jak *dobré* jsou takto predikované hodnoty?
- Dobré ≈ přesné ≈ s co nejmenšími rezidui
  - Odhad metodou nejmenších čtverců
- Jak velká jsou rezidua?

U	P	P'	*e* = (P-P')
6,5	6,4	6,5413	-0,1413
7	7	7,1291	-0,1291
7,5	7,5	7,7169	-0,2169
5,2	4,8	5,0130	-0,2130
6,6	6,7	6,6589	0,0411
6,6	6,8	6,6589	0,1411
7	7	7,1291	-0,1291

Rozložení reziduí



Přesnost predikce

- **s_e** vyjadřuje míru chyby při individuální predikci způsobenou nedokonalou těsností lineárního vztahu
  - vzhledem k normálnímu rozložení reziduí je pravděpodobnost určitých intervalů reziduí dána kvantily normálního rozložení (standardizovaného **s_e**)
  - Např. 68% reziduí délky prsteníčků <|0,337| neboli pravděpodobnost, že se při odhadu délky prsteníčku mýlíme o 0,337 a méně je přibližně 68%

- *Zatím nezohledňujeme nejistotu predikce způsobenou tím, že jsme parametry regresní přímky pouze odhadovali z (malého) vzorku*
- *Také nezohledňujeme to, že chyby odhadu jsou v extrémech X vyšší než okolo průměru X* *(viz Hendl, s. 285 s chybou)*

Rozložení predikovaných hodnot a reziduí

$m_P = 7{,}109$

$s_P = 0{,}843$

$m_{P'} = 7{,}109 = m_P$ $\quad + \quad$ $m_e = 0$

$s_{P'} = 0{,}773$ $\qquad\qquad$ $s_e = 0{,}337$

Rozložení predikovaných hodnot a reziduí

m_{P}=7,109

s^{2}_{P}=0,711

$m_{P'}$= 7,109 = m_{P}

$s^{2}_{P'}$ = 0,598

+

m_{e}= 0

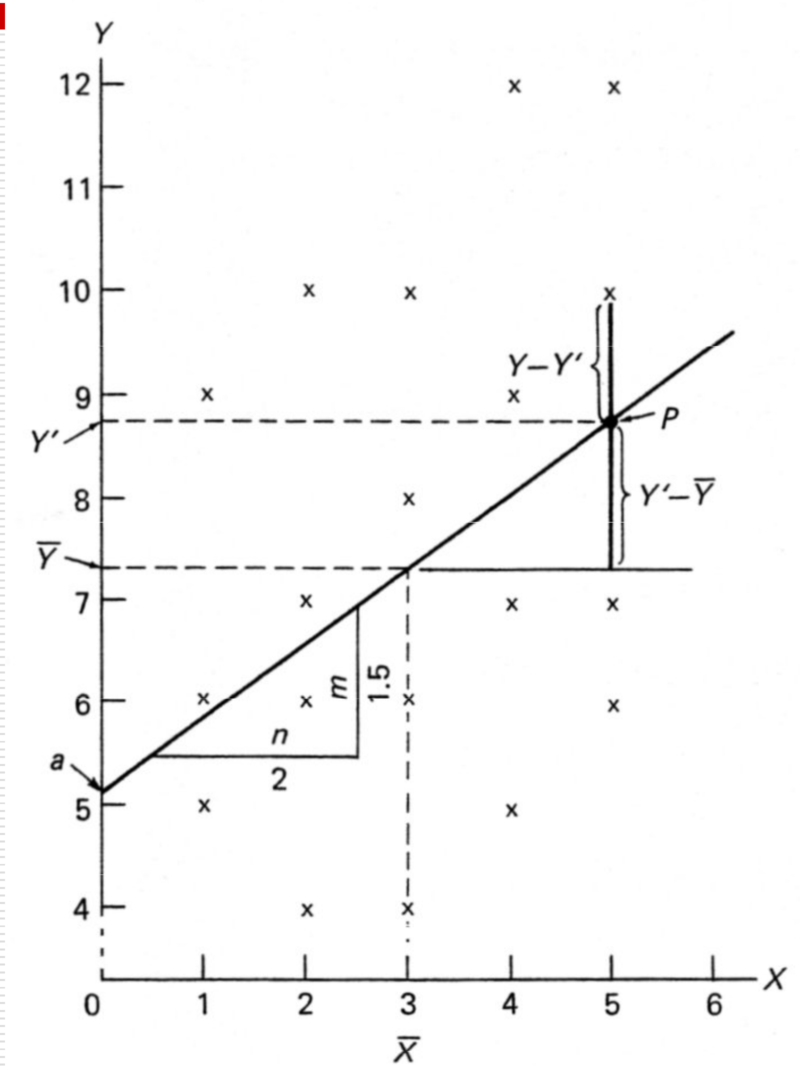
s^{2}_{e} = 0,113

Lineární regrese II. – úspěšnost predikce

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

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

- $s_y^2 = s_{reg}^2 + s_{res}^2$ $(ss_y = ss_{res} + ss_{reg})$
- $R^2 = s_{reg}^2 / s_y^2$ … $s_{res}^2 = s_y^2(1-R^2)$
- Koeficient determinace (R^2)
  - Podíl vysvětleného rozptylu
  - Je ukazatelem kvality, úspěšnosti regrese
  - Vyjadřuje shodu modelu s daty
- **Pro jednoduchou lin. regr. platí $R^2 = r^2$**



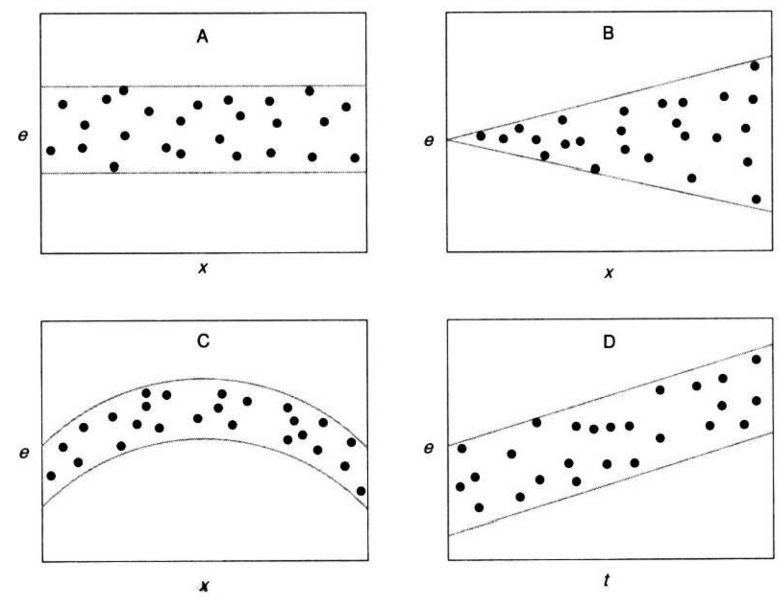
AJ: regression and residual variance (sum of squares), explained variance, model fit with the data, coefficient of determination (R square)

Pozn. Zde uvedené vzorce jsou pro s^2_{res}. Pro populační parametr,tj. nejlepší odhad z výběrových dat σ^2_{res} počítáme ss_{res}/ (n-2).

Lineární regrese III. – předpoklady, platnost

Předpoklady oprávněnosti použití lineárně-regresního modelu

- jako u Pearsonovy korelace
- <u>konceptuální</u> předpoklady:
  - vztah je ve skutečnosti lineární
  - *X* je jediným zdrojem *Y*
- rezidua mají normální rozložení s průměrem 0 a SD=s_{res}
- 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,

Použití (lineární) regrese

- Prozkoumání (lineárního) vztahu mezi proměnnými (místo korelace)
  - analyticko-konceptuální využití
  - středem zájmu je *b*

- Predikce
  - praktické využití
  - středem zájmu je odhad a jeho chyba

Predikce Y pro nového jedince

- Dosazením do regresní rovnice získáme odhad Y′
- Jak přesný?
  - Rezidua (=chyby odhadu) mají podle *předpokladů* LR normální rozložení s $m=0$ a $s=s_{res}$
  - 95% chyb odhadu se tak bude přibližně mezi $-2s_{res}$ a $+2s_{res}$
- Přesněji, jak přesný?
  - s_{res} je „průměrná" chyba. Čím dále je *X* od průměru, tím jsou chyby větší.
  - Parametry regrese (*a* a *b*) stanovujeme s chybou. Ta závisí hlavně na *N*.
  - Pak $s_{Y'} = s_{res}\sqrt{1+\frac{1}{N}+\frac{z_X^2}{N-1}}$ a rozložení chyb je *t* s *N*-2 st.v.

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 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
  - … opatrně zvážíme, zda se pustit do nelineární regrese

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 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

- Hendl: kapitoly 7.3 – 7.3.2, 7.3.6, 7.4