PSY117
Statistická analýza dat v psychologii
Přednáška 6 - 2018

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

Dvě základní otázky predikce

1. Jakou hodnotu predikovat?
   - Stanovení modelu
     - výběr z mnoha „šablon“
     - stanovení parametrů modelu
   - Použití modelu k predikci
2. S jakou přesností predikujeme?
   - Chyby ve volbě modelu
   - Chyby ve stanovení parametrů modelu
   - Chyby implikované modelem

1. Stanovení modelu

Pokud víme (ze zkušenosti, z výzkumu, z teorie…), že….

…pokud související proměnná (*X*) má hodnotu *x*…

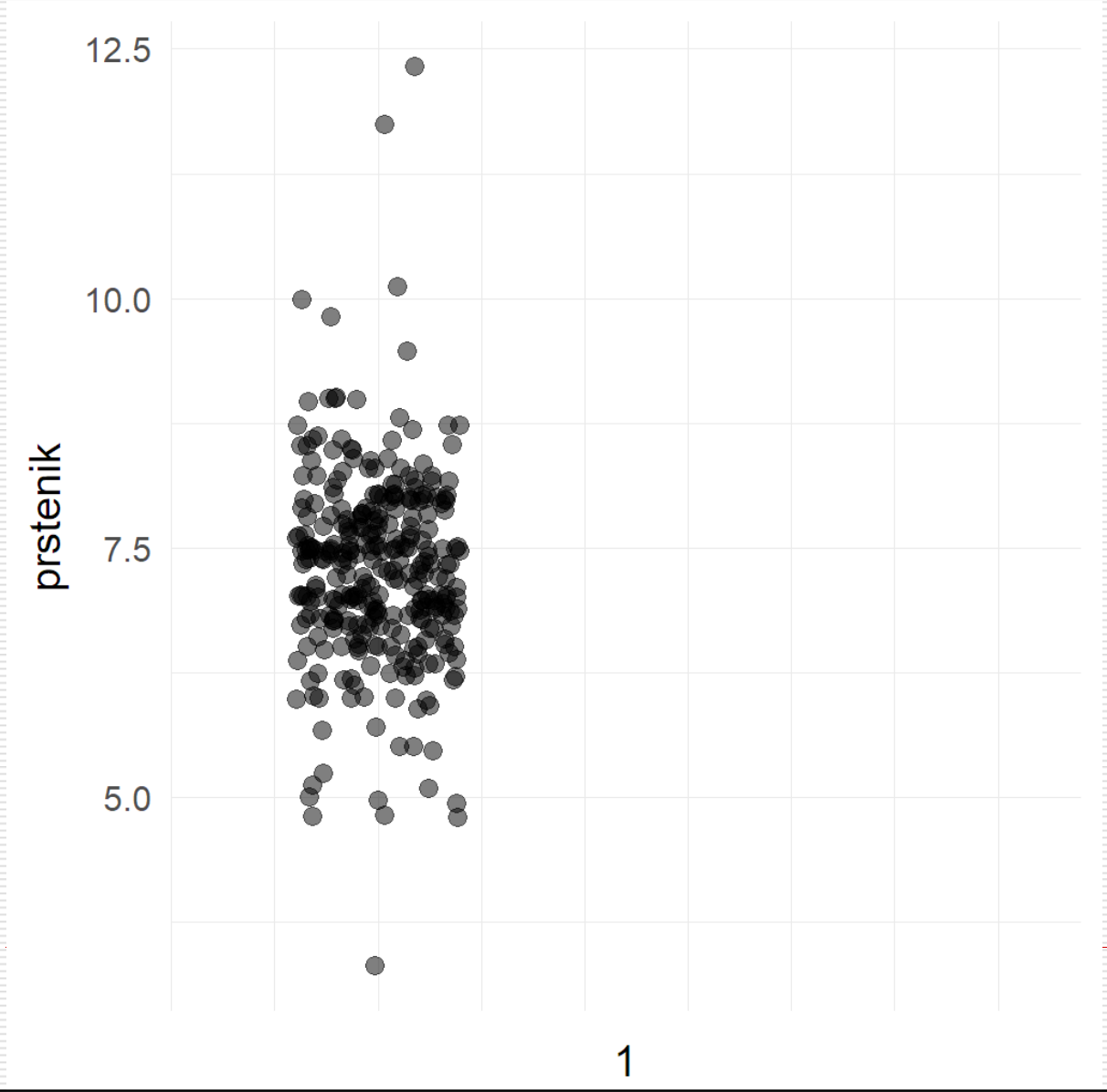
…tak predikovaná proměnná (*Y*) může nabývat…
…omezeného rozpětí hodnot | jen určitých hodnot | jen určité hodnoty…

…budeme predikovat…

…střední hodnotu těchto možných hodnot.

Např. $M(Y|X=c)$, $Md(Y|X=c)$, $Mo(Y|X=c)$

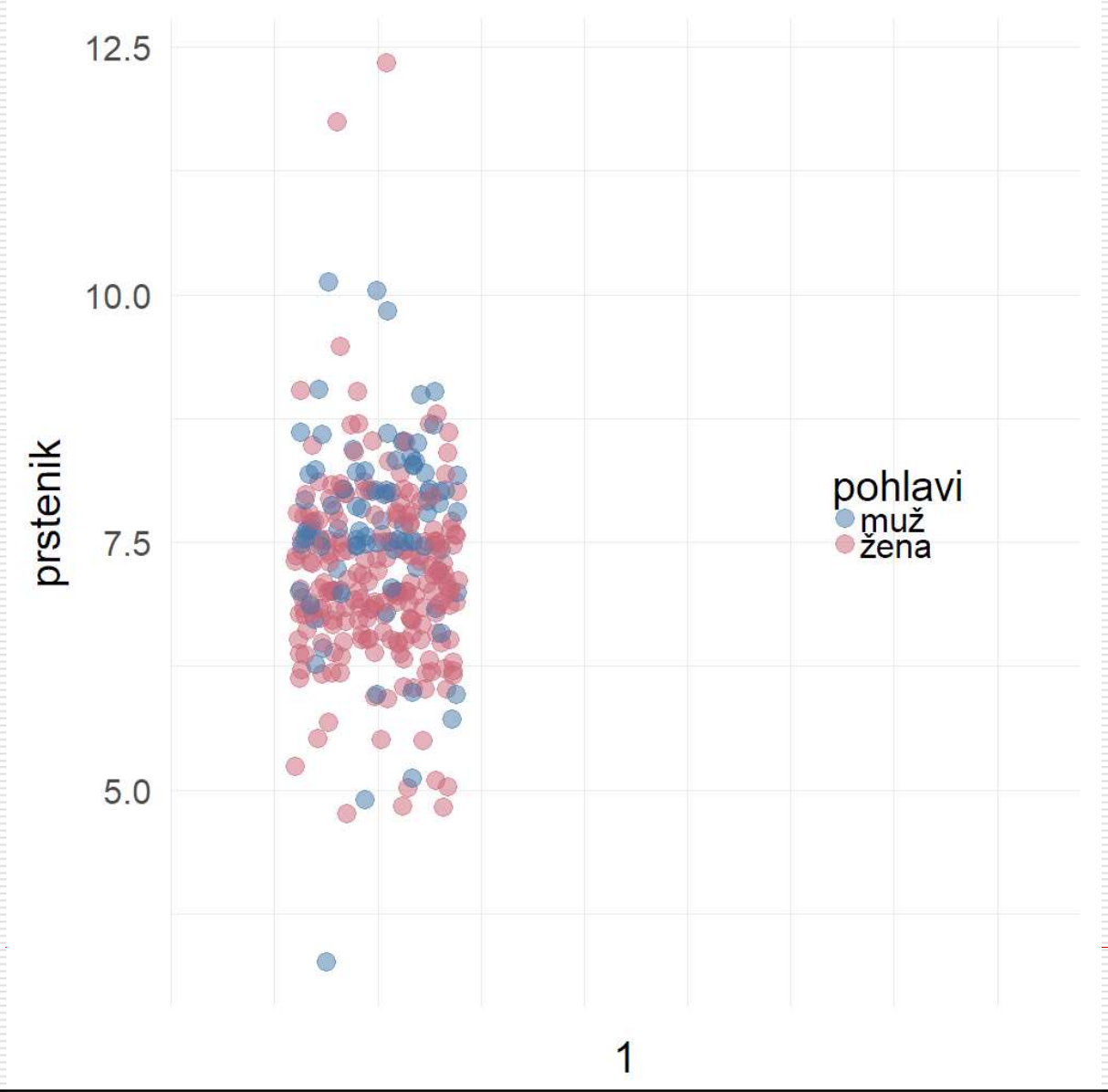
Predikce délky prsteníku z pohlaví



Výchozí znalost:
data o 312 lidech.

Predikce:
Jak dlouhý prsteník udělat muži, který si přišel nechat vyrobit jeho náhradu po úraze?

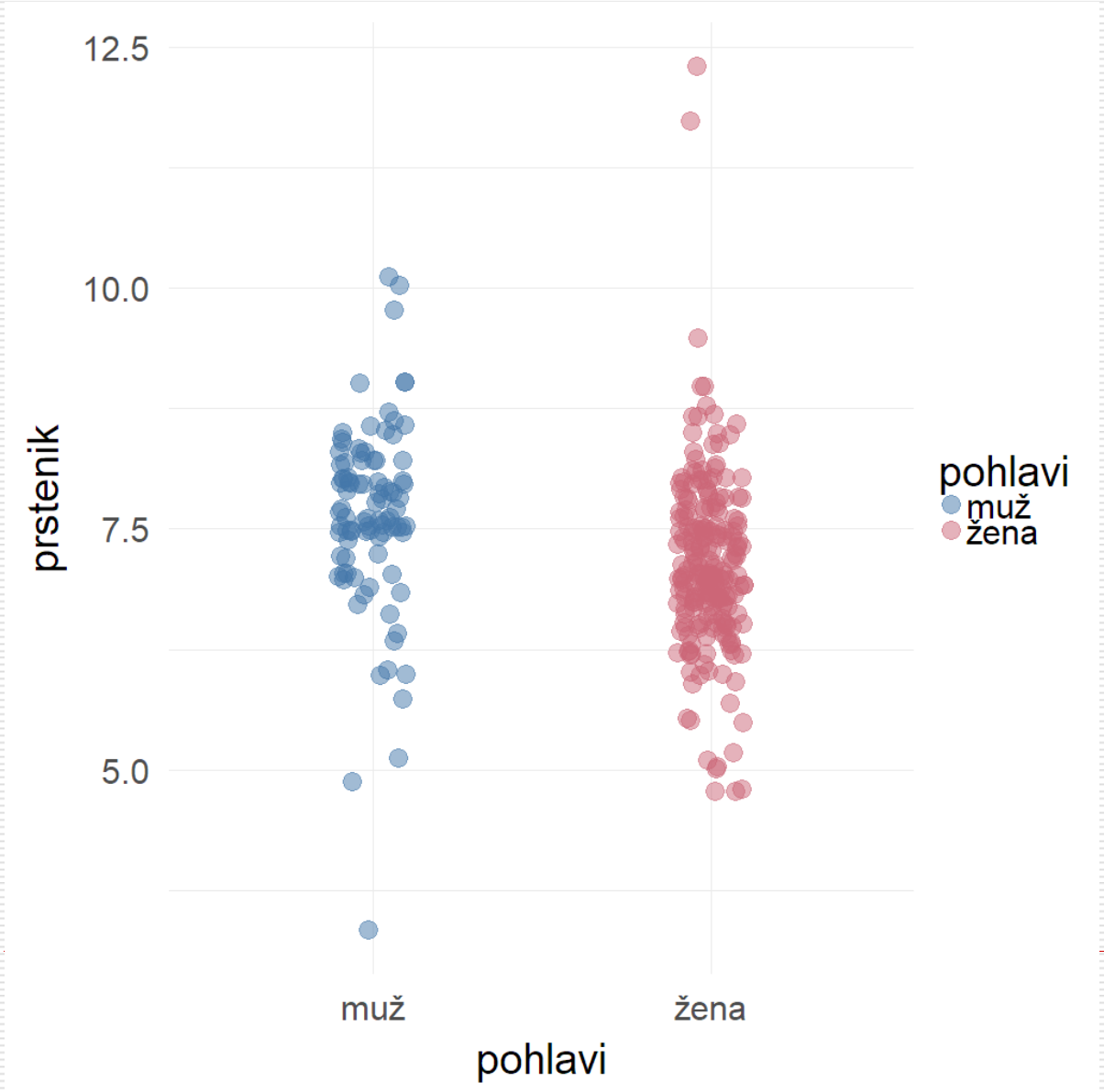
Predikce délky prsteníku z pohlaví



Výchozí znalost:
data o 312 lidech.

Predikce:
Jak dlouhý prsteník udělat muži, který si přišel nechat vyrobit jeho náhradu po úraze?

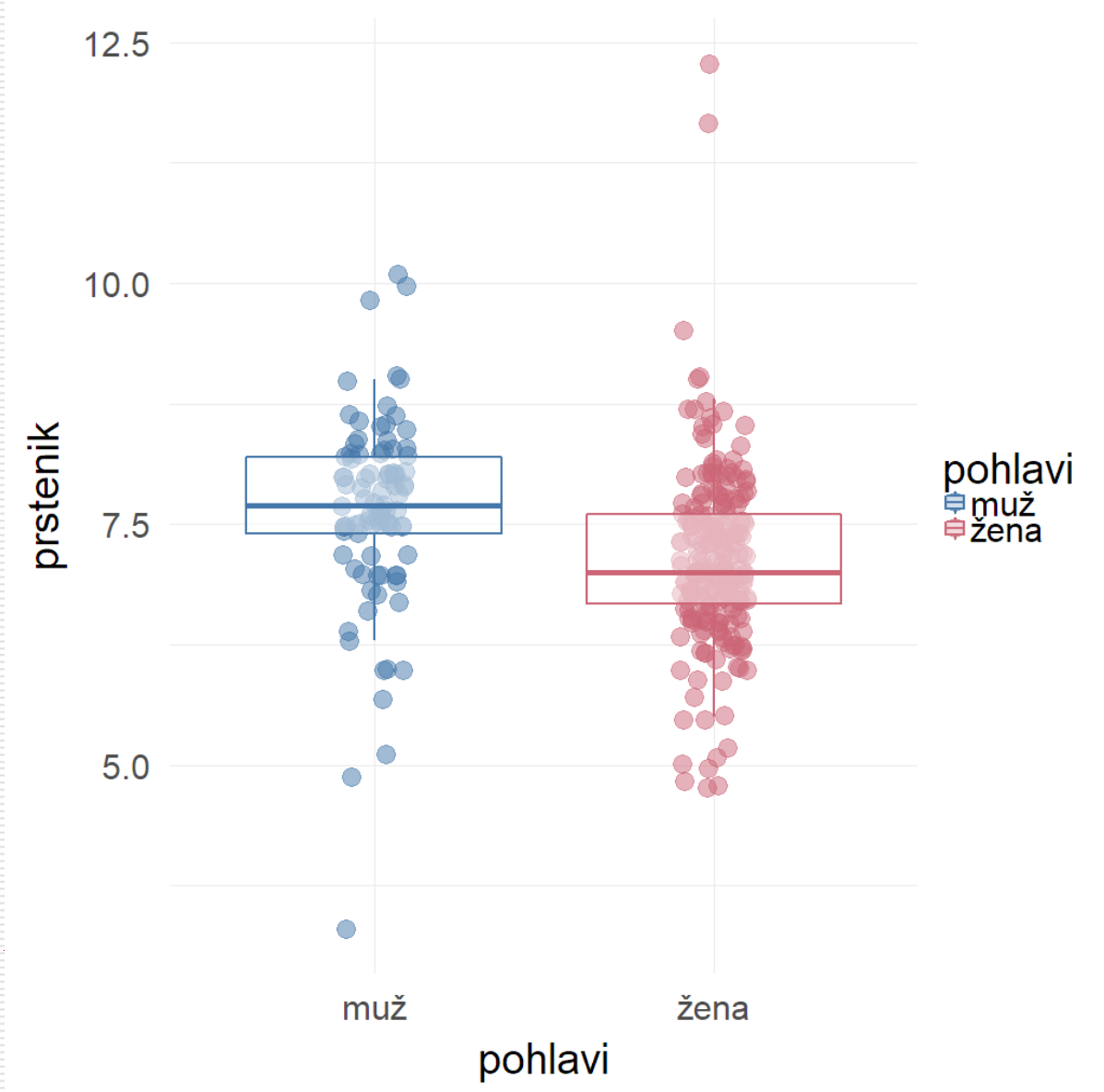
Predikce délky prsteníku z pohlaví



Výchozí znalost:
data o 312 lidech.

Predikce:
Jak dlouhý prsteník udělat muži, který si přišel nechat vyrobit jeho náhradu po úraze?

Predikce délky prsteníku z pohlaví

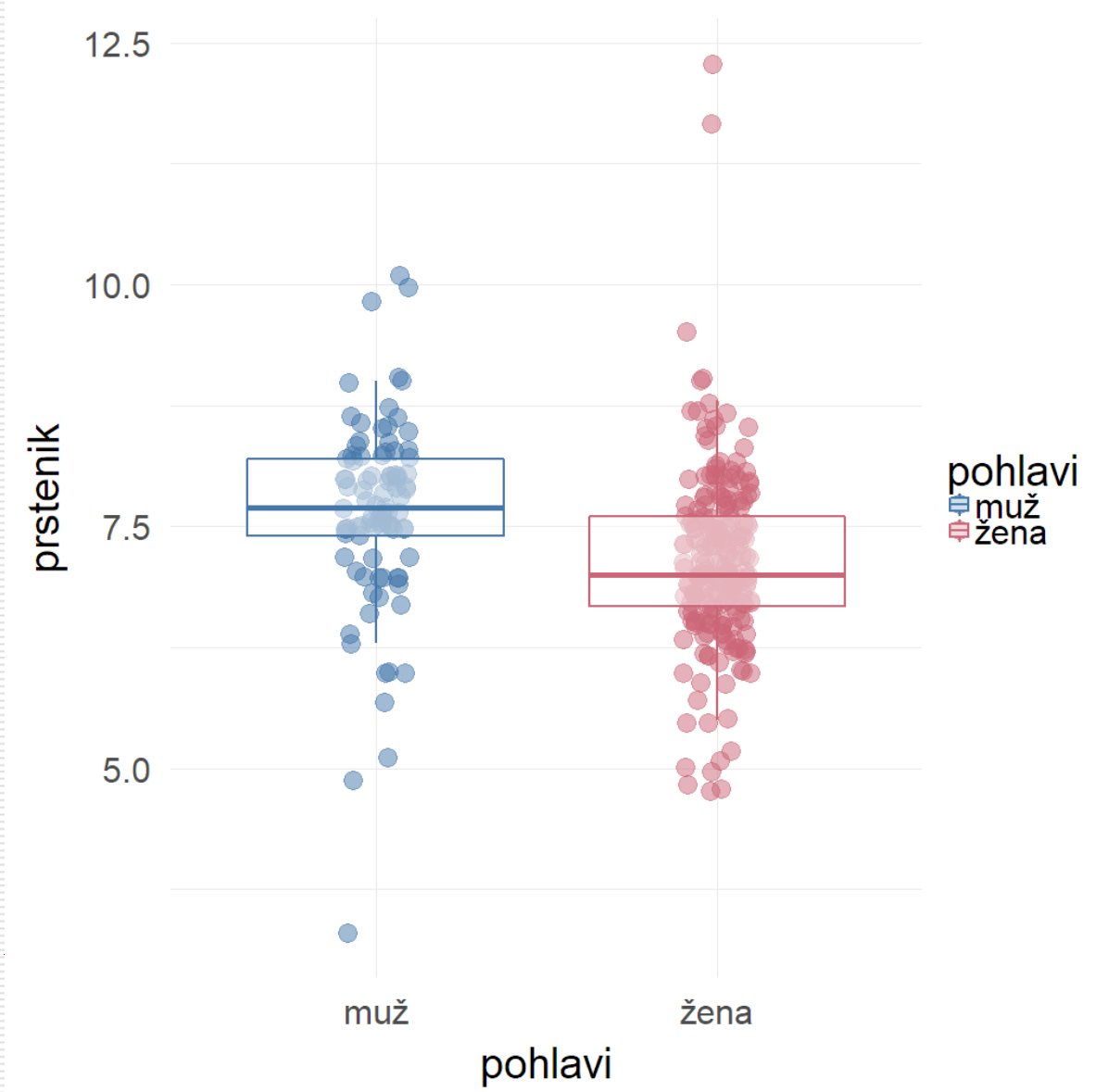


Výchozí znalost:
data o 312 lidech.

Predikce:
Jak dlouhý prsteník udělat muži, který si přišel nechat vyrobit jeho náhradu po úraze?

$M(Y|X=\text{muž})=7{,}64$
$M(Y|X=\text{žena})=7{,}13$
$SD(Y|X=\text{muž})=0{,}99$
$SD(Y|X=\text{žena})=0{,}93$

Predikce délky prsteníku z pohlaví



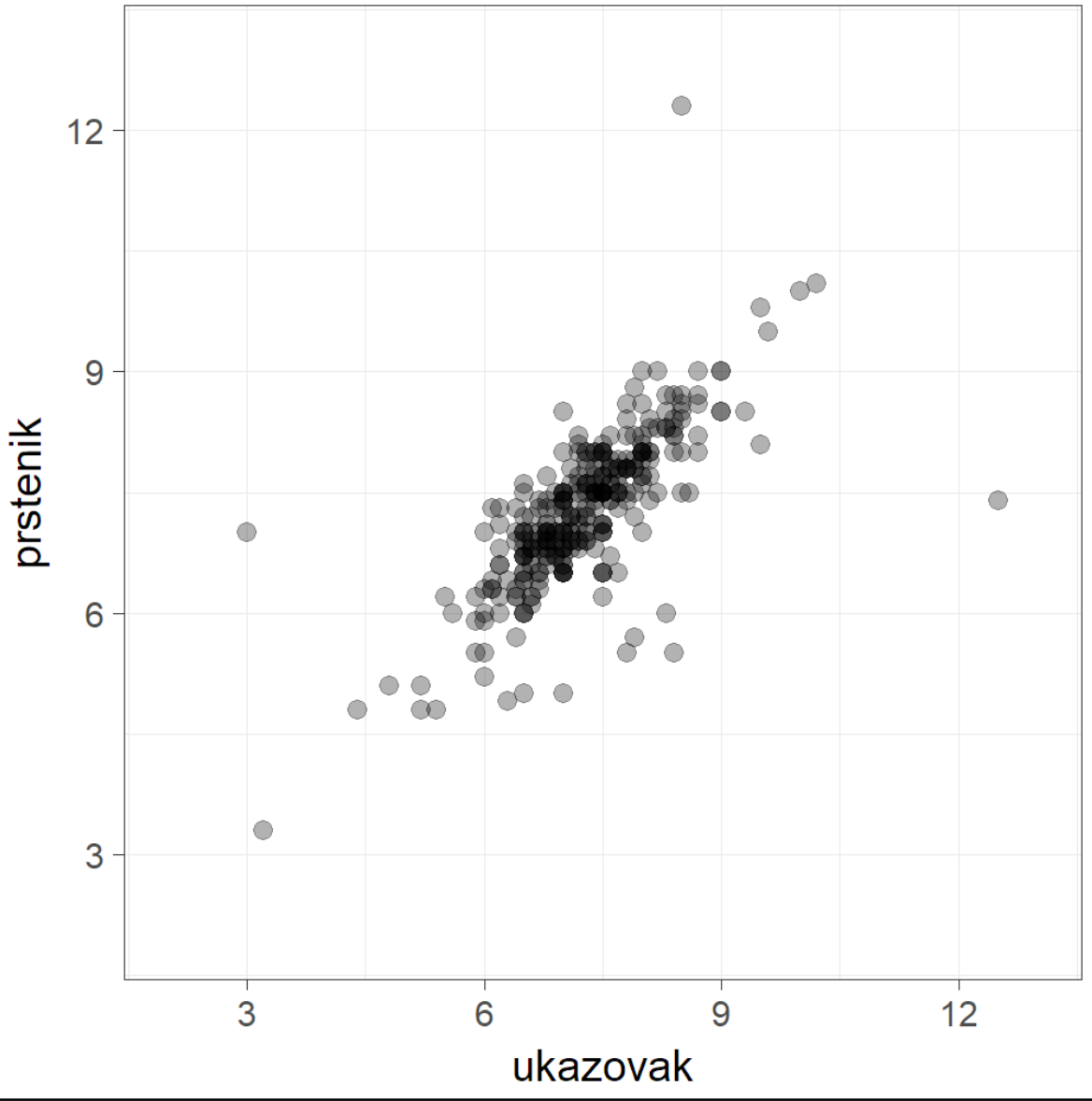
Výchozí znalost:
data o 312 lidech.

Predikce:
Jak dlouhý prsteník udělat muži, který si přišel nechat vyrobit jeho náhradu po úraze?

Zapsáno jako funkce:

$$f(X) = \begin{cases} 7{,}6 \ , X = muž, \\ 7{,}1 \ , X = žena \end{cases}$$

Predikce délky prsteníku z ukazováku

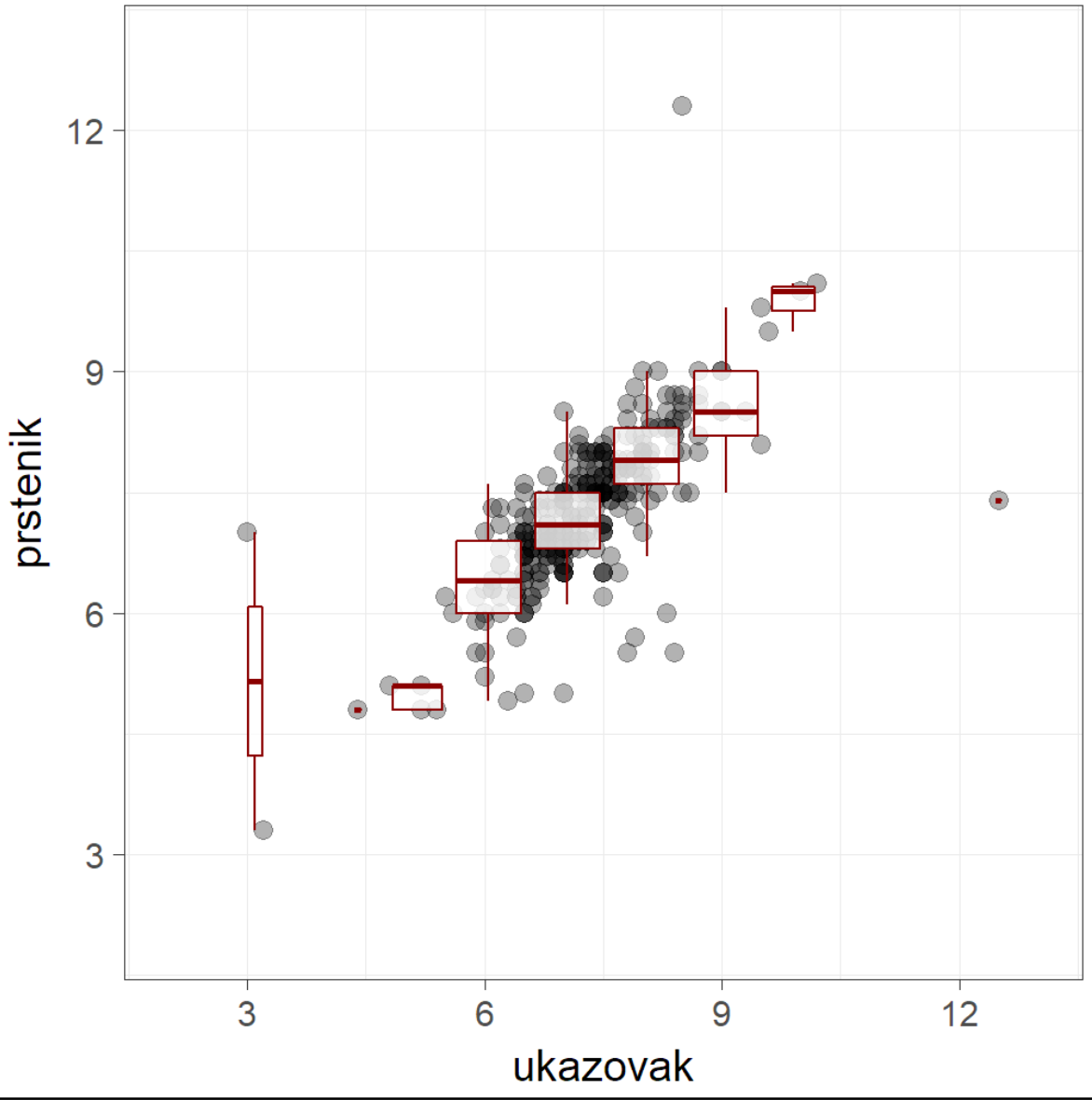


…

Predikce:

Jak dlouhý prsteník udělat člověku, který si přišel nechat vyrobit jeho náhradu po úraze? Ukazovák má 7cm dlouhý.

Predikce délky prsteníku z ukazováku



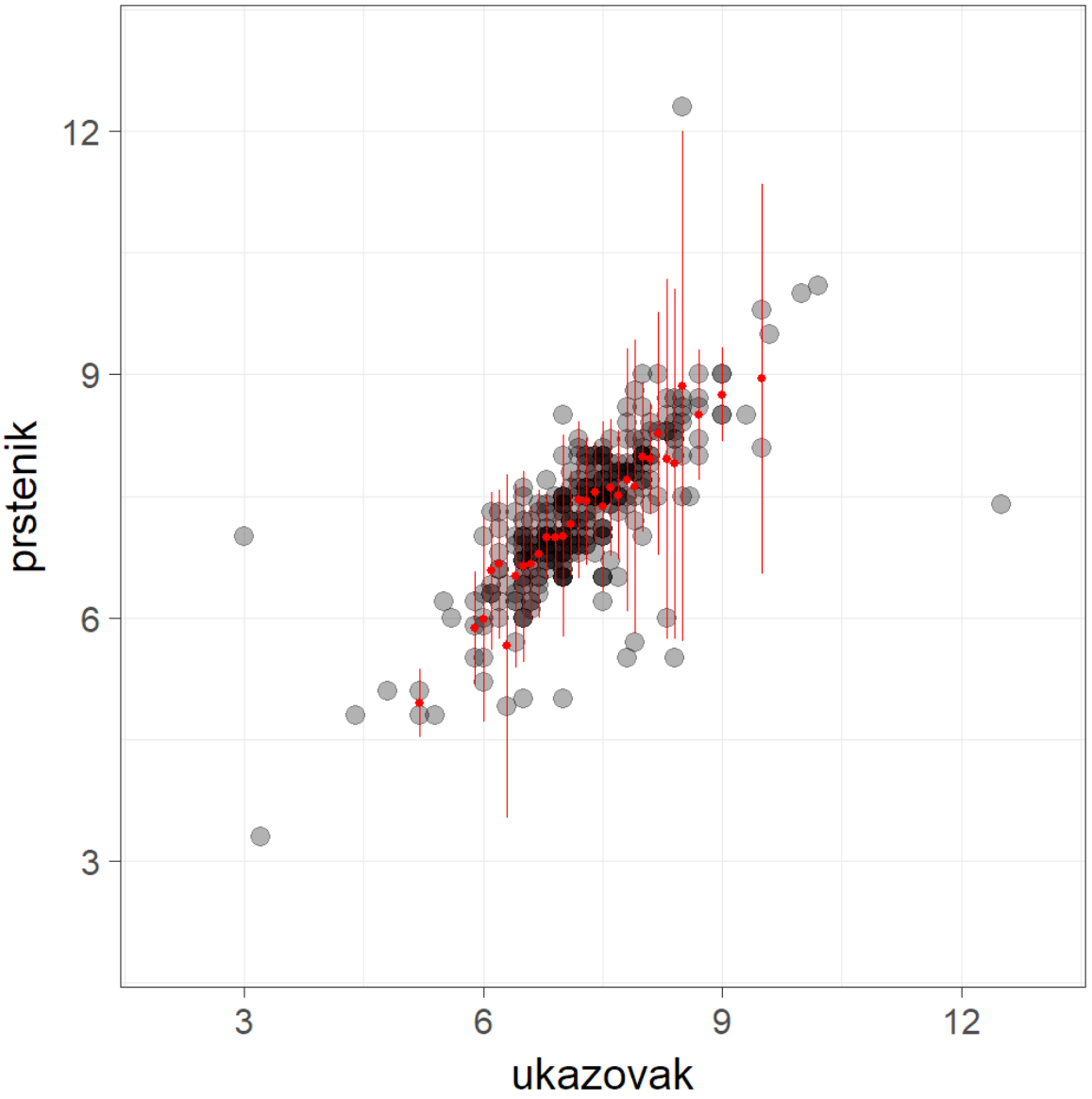
...

Predikce:

Jak dlouhý prsteník udělat člověku, který si přišel nechat vyrobit jeho náhradu po úraze? Ukazovák má 7cm dlouhý.

$$f(X) = \begin{cases} 5{,}1 \ , x \in (4{,}5; 5{,}5] \\ 6{,}3 \ , x \in (5{,}5; 6{,}5] \\ 7{,}1 \ , x \in (6{,}5; 7{,}5] \\ \dots \end{cases}$$

Predikce délky prsteníku z ukazováku

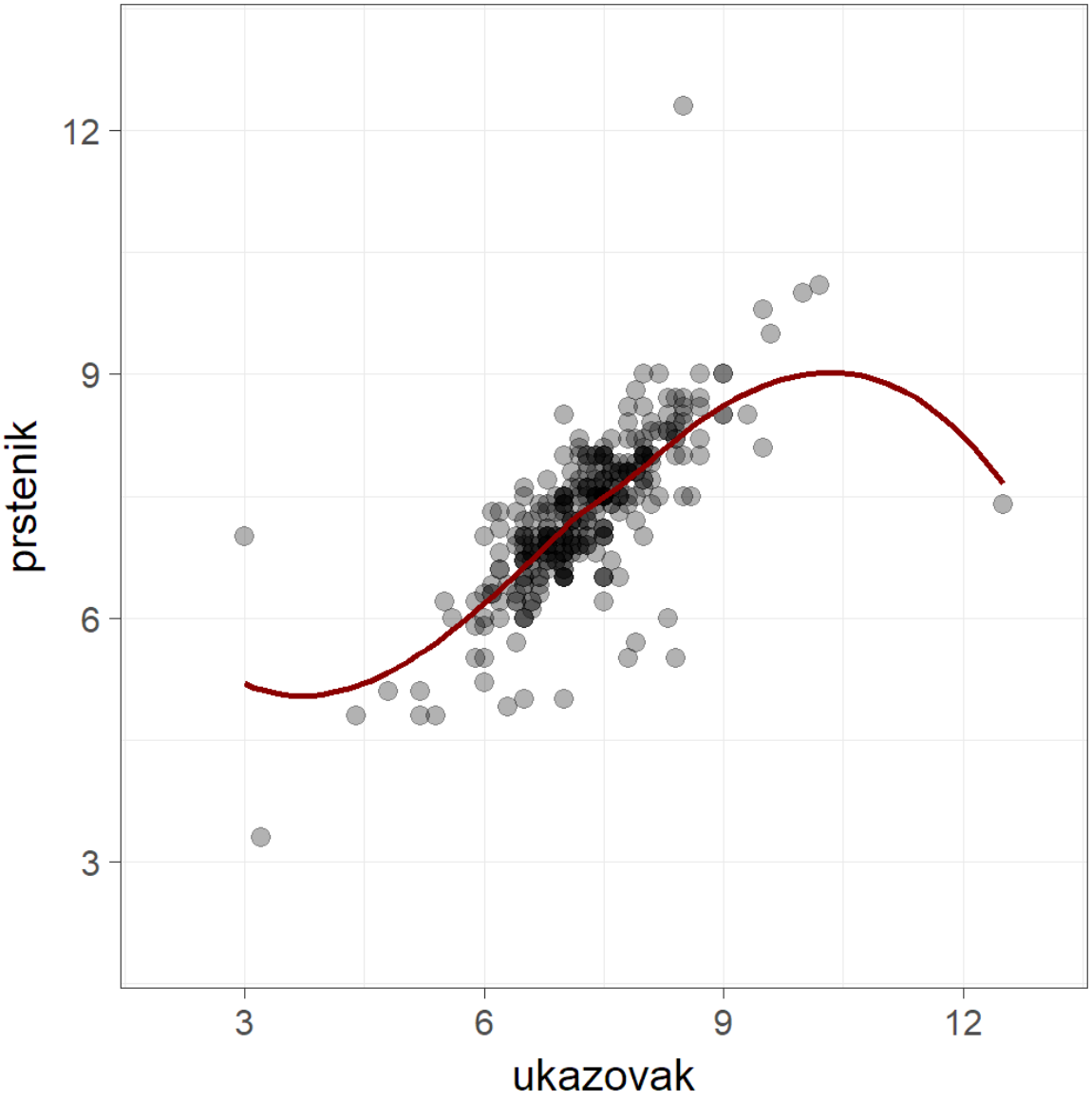


…

Predikce:

Jak dlouhý prsteník udělat člověku, který si přišel nechat vyrobit jeho náhradu po úraze? Ukazovák má 7cm dlouhý.

Predikce délky prsteníku z ukazováku



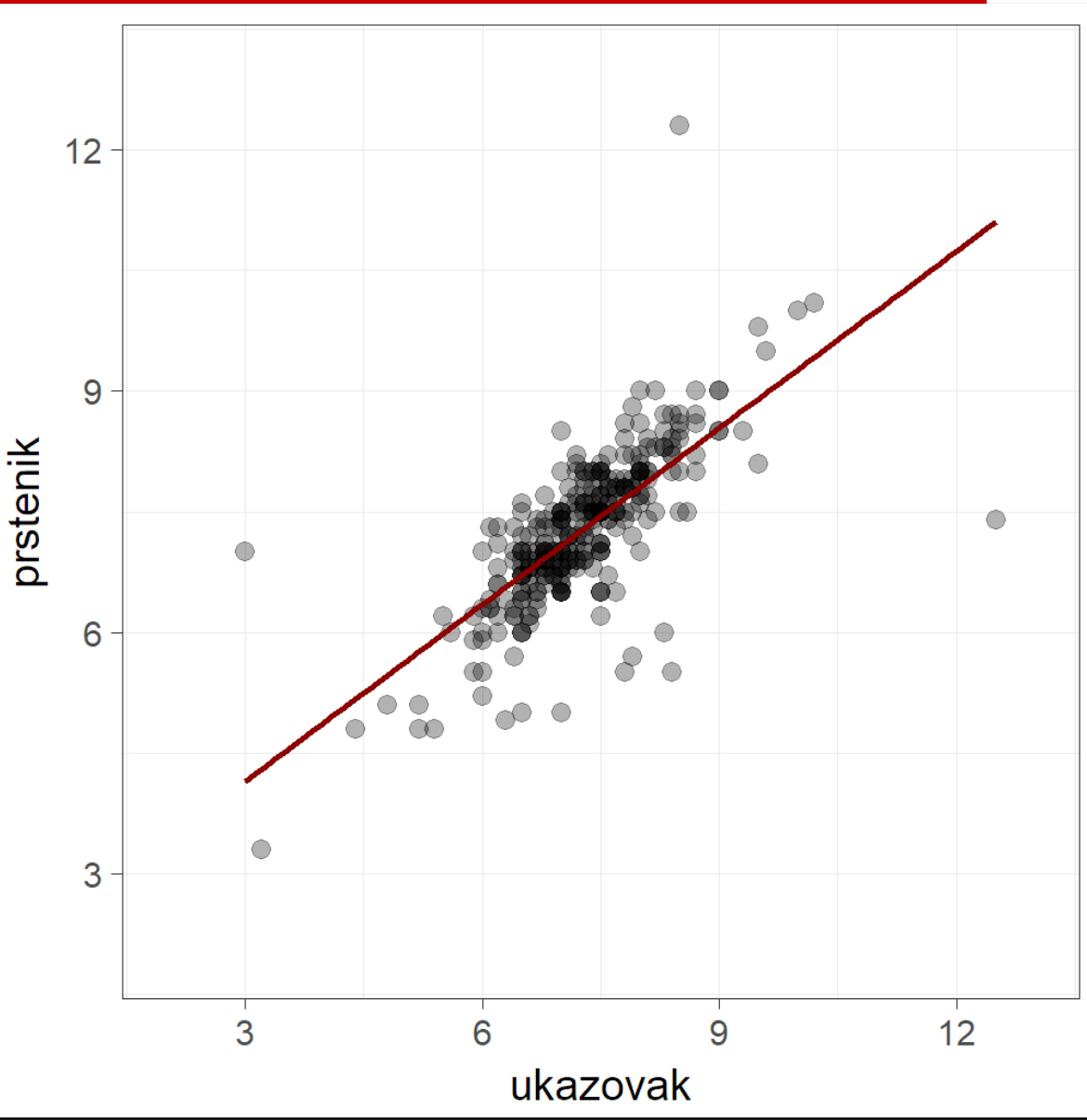
…

Predikce:

Jak dlouhý prsteník udělat člověku, který si přišel nechat vyrobit jeho náhradu po úraze? Ukazovák má 7cm dlouhý.

$f(X)$ = výška křivky v bodě x

Predikce délky prsteníku z ukazováku



...

Predikce:

Jak dlouhý prsteník udělat člověku, který si přišel nechat vyrobit jeho náhradu po úraze? Ukazovák má 7cm dlouhý.

$f(X)=0{,}6X+2{,}7$

1. Stanovení modelu

Model je funkce

jak ze známé hodnoty X vypočítat tu neznámou Y

$$\boldsymbol{Y = f(X)}$$

Funkce

- stanovené výčtem
- trigonometrické, exponenciální a logaritmické …
- polynomické:
  - lineární: $Y = bX + a$ (rovná čára … Pearsonova r)
  - kvadratické: $Y = cX^2 + bX + a$ (jedna zatáčka)…

1. Stanovení modelu

1. Tuto funkci (po racionální úvaze) volíme, **specifikujeme**…
   - Např. lineární funkce $Y = bX + a$
2. …a stanovujeme (**odhadujeme**) její parametry.
   - Např. jaké jsou vhodné hodnoty parametrů *a* a *b* …… *a=2 b=3*
3. Funkci pak použijeme k predikci - výsledkem je **predikovaná hodnota** *Y'*
   - Např. má-li člověk hodnotu *X*=5, pak mu predikujeme 5*3+2=17
4. Známe-li skutečné hodnoty *Y*, můžeme je porovnat s predikovanými hodnotami
   - $Y_i = Y_i' + e_i = f(X_i) + e_i$, kde $e_i = Y_i - Y_i'$ a *i* je číslo účastníka
   - e_i je reziduální hodnota (reziduum),
   - *Y* je závislá proměnná, *X* je prediktor, nezávislá proměnná
   - e_i představuje všechny ostatní zdroje variability vyjma *X*

Tradičně modelu, jeho stanovení a použití říkáme **regrese** (regrese *Y* na *X*)

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

1. Stanovení modelu

Základní otázka je

Jak specifikovat model a stanovit jeho parametry tak, aby byla predikce co nejpřesnější?

Volba a specifikace modelu

Mnoho možností

- Funkce s několika málo parametry
- Výčty – mnoho parametrů

Můžeme volit sami, nebo to nechat na chytrých algoritmech – strojové učení, neuronové sítě

Zpočátku volba jasná – **lineární model(regrese)**

Lineární regrese I. – odhad přímou úměrou

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

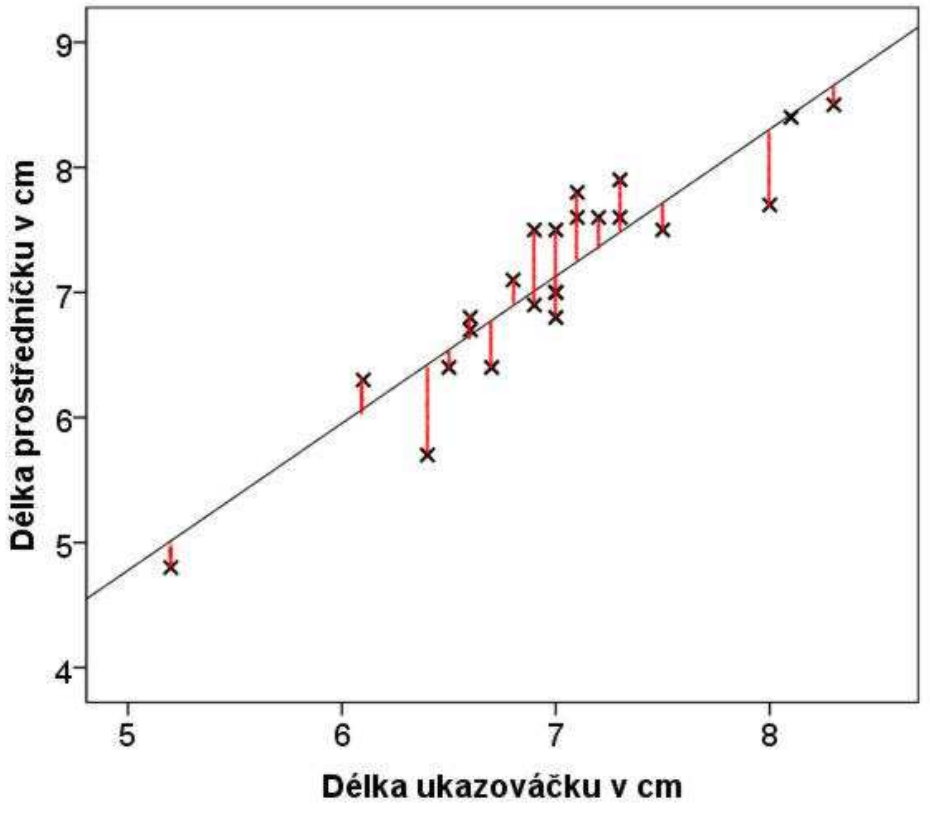
$$Y' = a + bX$$

b ... směrnice

a ... průsečík

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

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



Nejlepší přímka?

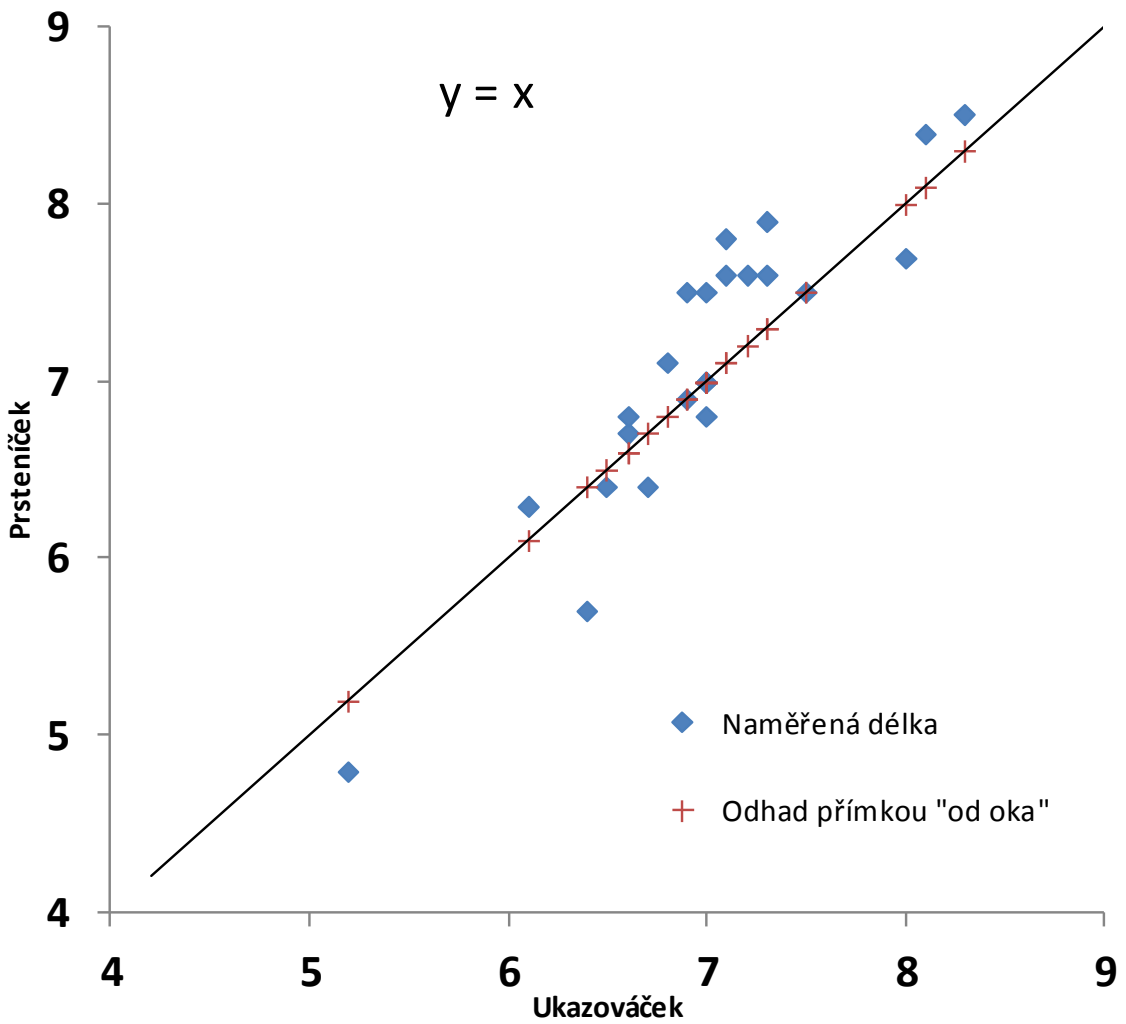
AJ: slope, intercept, least squares (estimation), regression coefficents (a,b)

Stanovení parametrů modelu (přímky)

Nastavme si parametry regresní přímky ručně

průsečík	0
směrnice	1

X	Y	
UKA	PRS	Odhad přímkou od oka
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

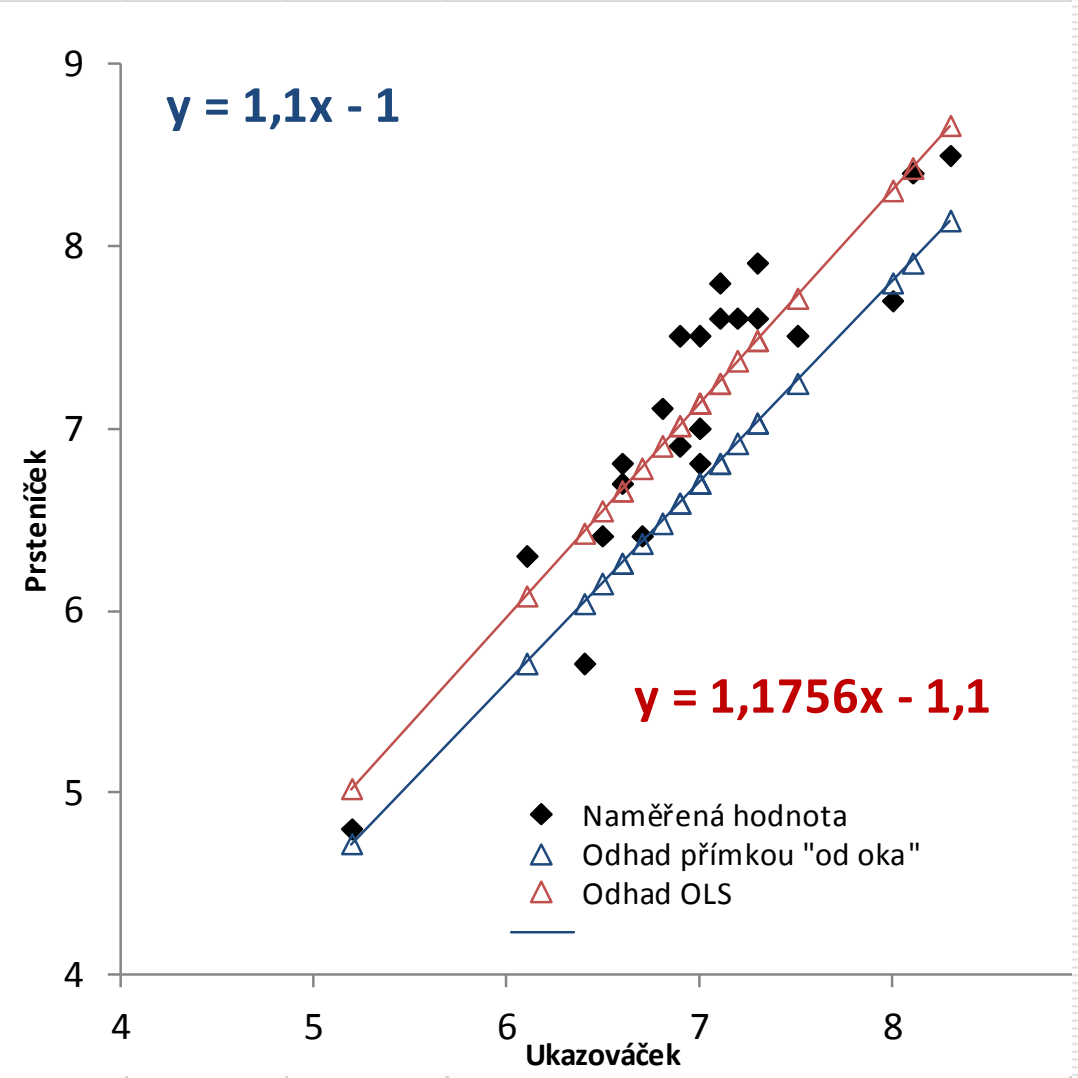


Jak stanovit „nejlepší přímku“?

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

Nastavme si parametry regresní přímky ručně

						průsečík	-1	**-1,10** **OLS**
						směrnice	1,1	**1,18**
	X	*Y*						
	UKA	PRS	Odhad přímkou "od oka"	Odhad OLS	e."od oka"	e.OLS		
	6,5	6,4	6,2	6,5	-0,25	0,14		
	7,0	7,0	6,7	7,1	-0,30	0,13		
	7,5	7,5	7,3	7,7	-0,25	0,22		
	5,2	4,8	4,7	5,0	-0,08	0,21		
	6,6	6,7	6,3	6,7	-0,44	-0,04		
	6,6	6,8	6,3	6,7	-0,54	-0,14		
	7,0	7,0	6,7	7,1	-0,30	0,13		
	7,1	7,8	6,8	7,2	-0,99	-0,55		
	6,8	7,1	6,5	6,9	-0,62	-0,21		
	7,0	6,8	6,7	7,1	-0,10	0,33		
	6,9	6,9	6,6	7,0	-0,31	0,11		
	6,7	6,4	6,4	6,8	-0,03	0,38		
	6,4	5,7	6,0	6,4	0,34	0,72		
	7,1	7,6	6,8	7,2	-0,79	-0,35		
	6,9	7,5	6,6	7,0	-0,91	-0,49		
	7,2	7,6	6,9	7,4	-0,68	-0,24		
	6,1	6,3	5,7	6,1	-0,59	-0,23		
	7,3	7,6	7,0	7,5	-0,57	-0,12		
	7,0	7,5	6,7	7,1	-0,80	-0,37		
	8,3	8,5	8,1	8,7	-0,37	0,16		
	7,3	7,9	7,0	7,5	-0,87	-0,42		
	8,0	7,7	7,8	8,3	0,10	0,60		
	8,1	8,4	7,9	8,4	-0,49	0,02		
				SUMA:	-9,84	0,00		
				SUMA.Č	6,76	2,49		



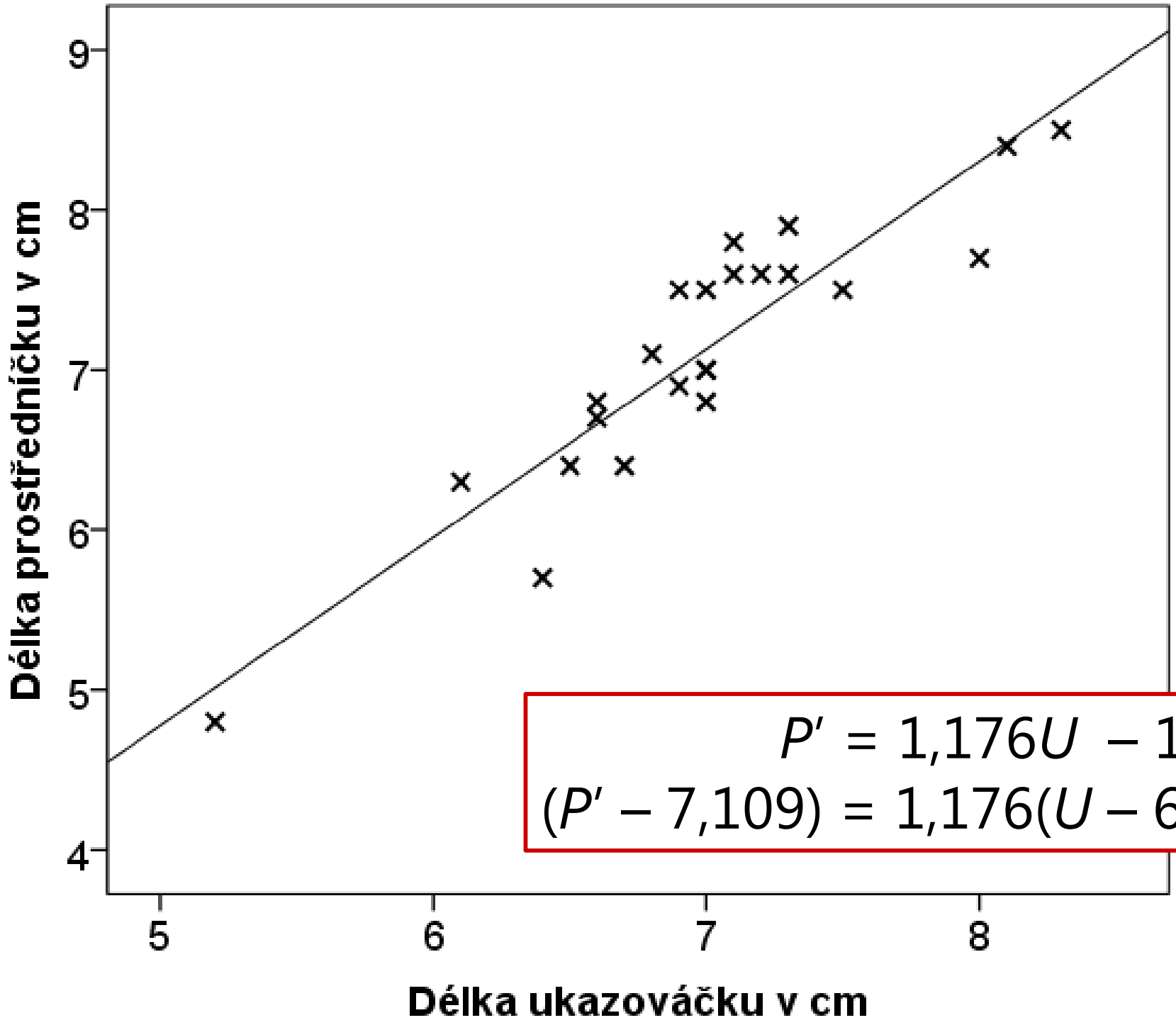
Řešení metodou nejmenších čtverců

$Y' = a + bX$:odhad metodou **nejmenších čtverců**

$$b = r_{YX}\frac{s_Y}{s_X} \qquad a = m_Y - bm_X$$

- Jsou-li X a Y vyjádřeny v z-skórech, pak $b = r_{YX}$
- Přímka prochází m_X a m_Y
- Průměr Y a Y' je stejný
- Součet reziduí je nulový, součet reziduí umocněných na druhou nejmenší možný

AJ: (ordinary) least squares estimation



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

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

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

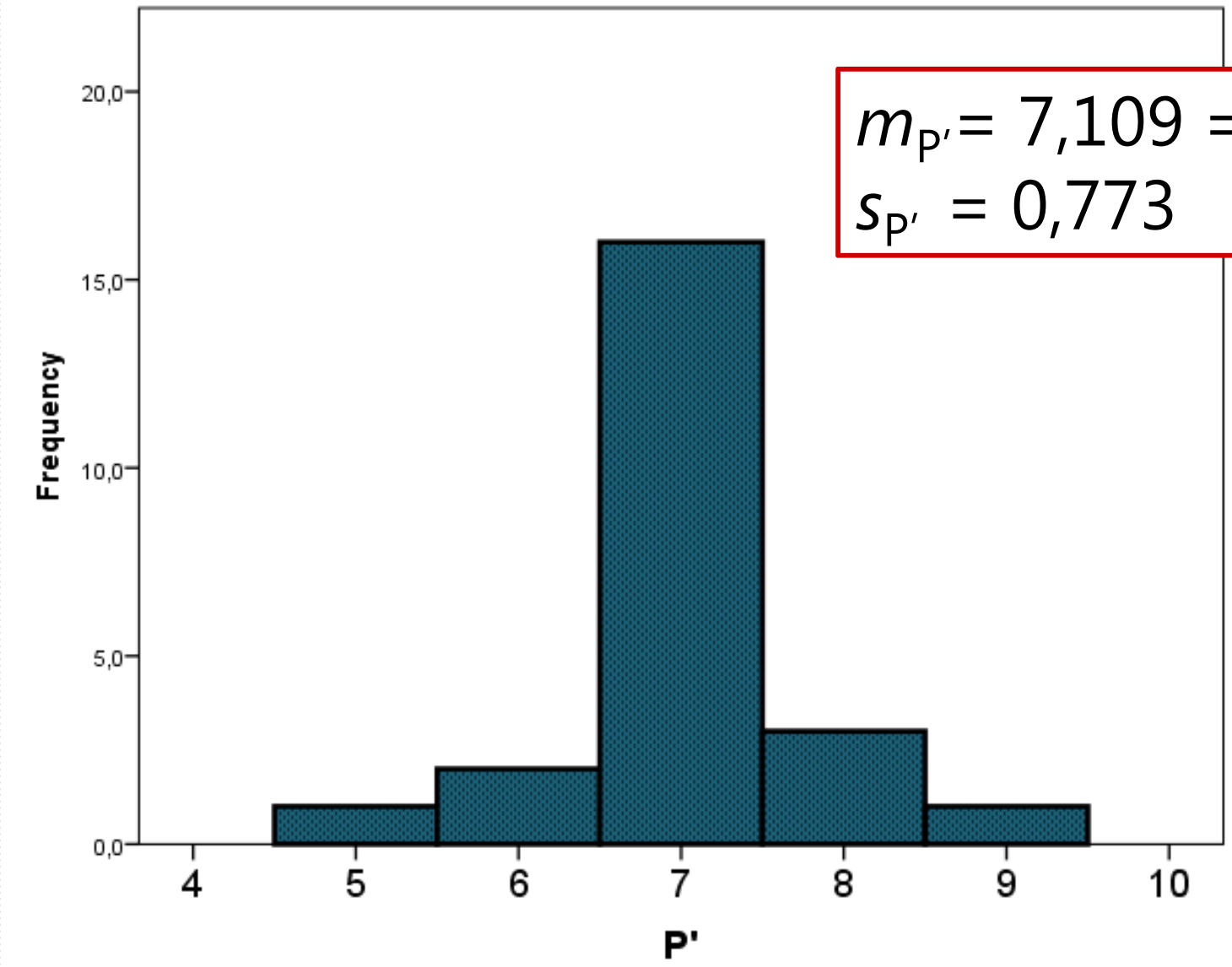
$$P' = 1{,}176U - 1{,}100$$
$$(P' - 7{,}109) = 1{,}176(U - 6{,}983)$$

Použití modelu

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



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

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

S jakou přesností predikujeme?

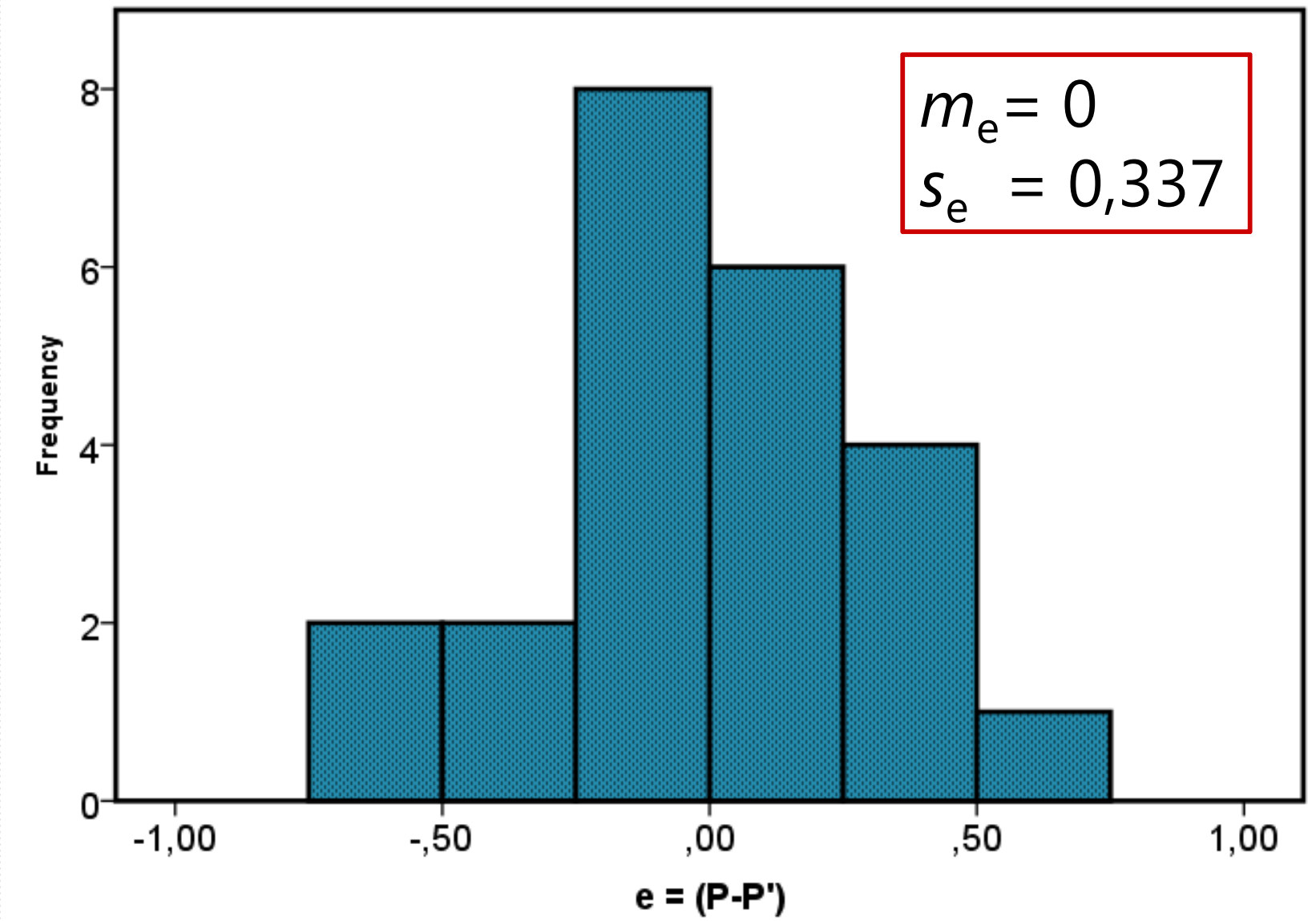
Model nejmenších čtverců říká, že „nepřesnost“ je nejmenší možná. Jaká ale je?

Lineární regrese II. – úspěšnost predikce

- Jak *dobré* jsou takto predikované hodnoty?
- Dobré ≈ přesné ≈ s co nejmenšími rezidui
  - Kritériem úspěšnosti je suma čtverců reziduí
- Jak velká jsou rezidua?

U	P	P'	e = (P-P')	e^2
6,5	6,4	6,54	-0,14	0,020
7	7	7,13	-0,13	0,017
7,5	7,5	7,72	-0,22	0,047
5,2	4,8	5,01	-0,21	0,045
6,6	6,7	6,66	0,04	0,002
6,6	6,8	6,66	0,14	0,020
7	7	7,13	-0,13	0,017

Rozložení reziduí



Přesnost predikce

- s_e vyjadřuje míru chyby při individuální predikci způsobenou nedokonalou těsností lineárního vztahu
  - vzhledem k (předpokládanému) 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

$m_P = 7{,}109$

$s_P = 0{,}843$

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

$s_{P'} = 0{,}773$ + $s_e = 0{,}337$ = s_P

Rozložení predikovaných hodnot a reziduí

$m_P = 7{,}109$

$s^2_P = 0{,}711$

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

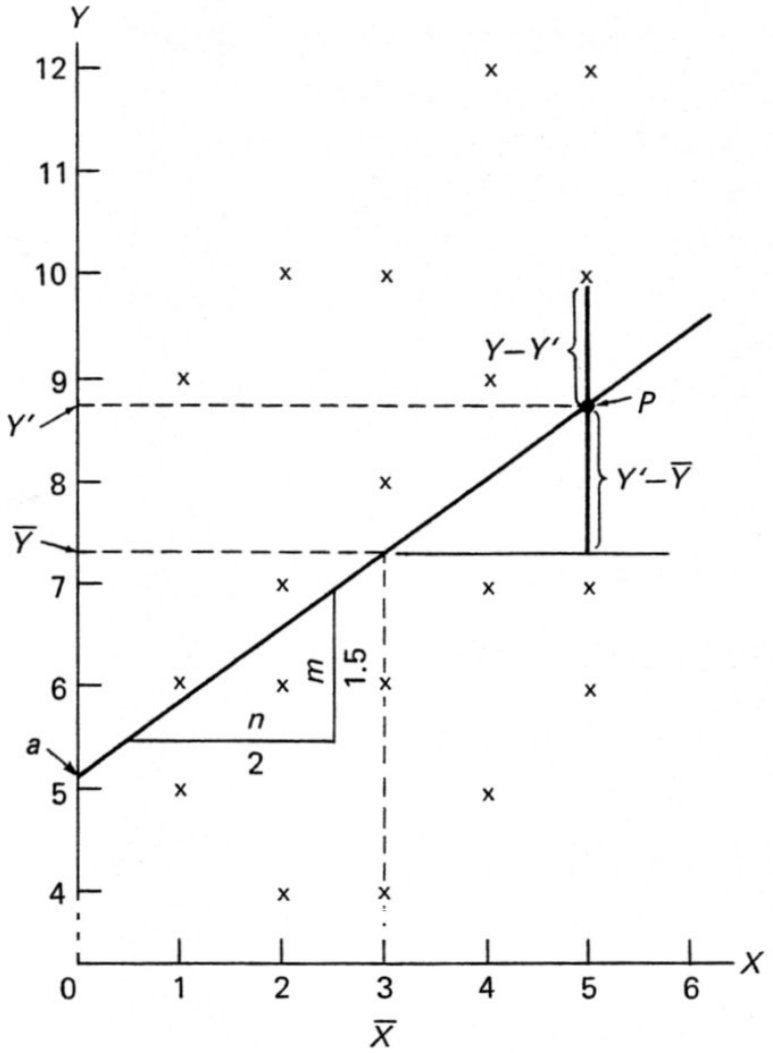
$s^2_{P'} = 0{,}598$ $\quad + \quad$ $s^2_e = 0{,}113$ $\quad = \quad$ s^2_P

Lineární regrese II. – úspěšnost predikce

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

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

- $s_Y^2 = s_{reg}^2 + s_{res}^2$ $(ss_Y = ss_{res} + ss_{reg})$
- $R^2 = s_{reg}^2 / s_y^2$ … $s_{res}^2 = s_Y^2(1-R^2)$
- 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)

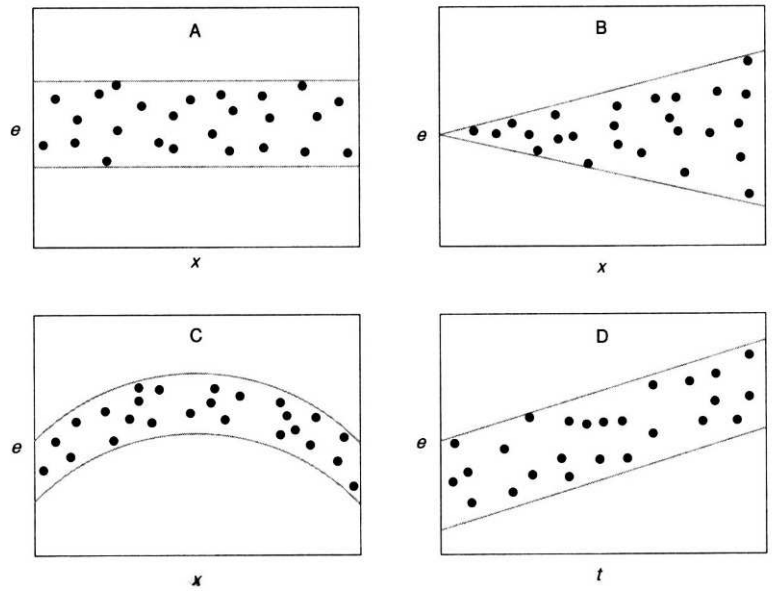
Pozn. Zde uvedené vzorce jsou pro s^2_{res}. Pro populační parametr,tj. nejlepší odhad z výběrových dat σ^2_{res} počítáme $ss_{res}/(n-2)$.

Chyby při volbě modelu

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,

Dvě základní otázky predikce

1. Jakou hodnotu predikovat?
   - Stanovení modelu
     - výběr z mnoha „šablon“ – **lineární regrese**
     - stanovení parametrů modelu – **výpočet hodnot**
   - Použití modelu k predikci – **dosazení do rovnice**
2. S jakou přesností predikujeme?
   - Chyby ve volbě modelu – **linearita, homoskedascita**
   - Chyby ve stanovení par. – **outlieři, výběrová chyba**
   - Chyby implikované modelem – **chyba odhadu s_{res}**

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 a kauzalita

K predikci stačí korelace.

Kauzální vztah mezi prediktorem a závislou není třeba.

Predikce *Y* pro nového jedince

- Dosazením do regresní rovnice získáme odhad *Y'*
- Jak přesný?
  - Rezidua 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