Neparametrické odhady Z-skóre

Z. Hlávka

KPMS MFF UK

Robust 2014

Z-skóre

Definice podle wikipedie:

Z-skóre je matematická transformace souboru číselných údajů tak, aby výsledná čísla po transformaci měla průměr 0 a směrodatnou odchylku 1.

Standard score (z-score) is the (signed) number of standard deviations an observation is above the mean:

$$Z = \frac{X - \mu}{\sigma}$$

The use of "Z" is because the normal distribution is also known as the "Z distribution".

Z-skóre

Definice podle wikipedie:

Z-skóre je matematická transformace souboru číselných údajů tak, aby výsledná čísla po transformaci měla průměr 0 a směrodatnou odchylku 1.

Standard score (z-score) is the (signed) number of standard deviations an observation is above the mean:

$$Z = \frac{X - \mu}{\sigma}$$

The use of "Z" is because the normal distribution is also known as the "Z distribution".

Cíl: odvodit věková Z-skóre výšky výskoku (X):

$$Z_i = \frac{X_i - \mu(\text{věk}_i)}{\sigma(\text{věk}_i).}$$

Popis dat (800 dětí od 6 do 19 let)

Základní popisné statistiky:

	průměr(+−s.odch.)	p-hodnota
Věk	11.83(+−3.453)	$< 1e{-}04^{***}$
Výška	149.021(+−18.047)	$< 1e{-}04^{***}$
Hmotnost	43.228(+−15.51)	$< 1e{-}04^{***}$
BMI	18.903(+−3.217)	$< 1e{-}04^{***}$

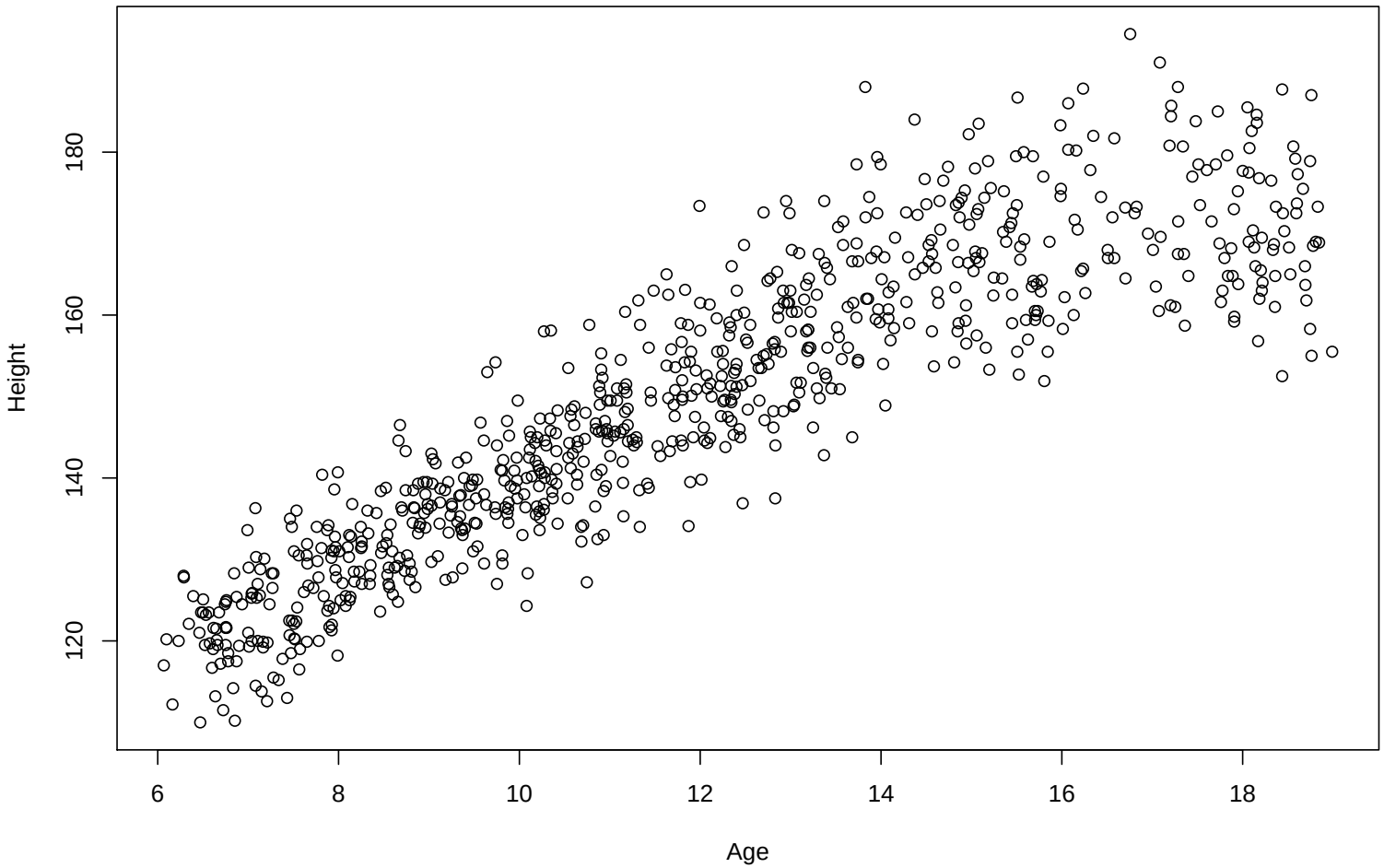
Popis dat (800 dětí od 6 do 19 let)

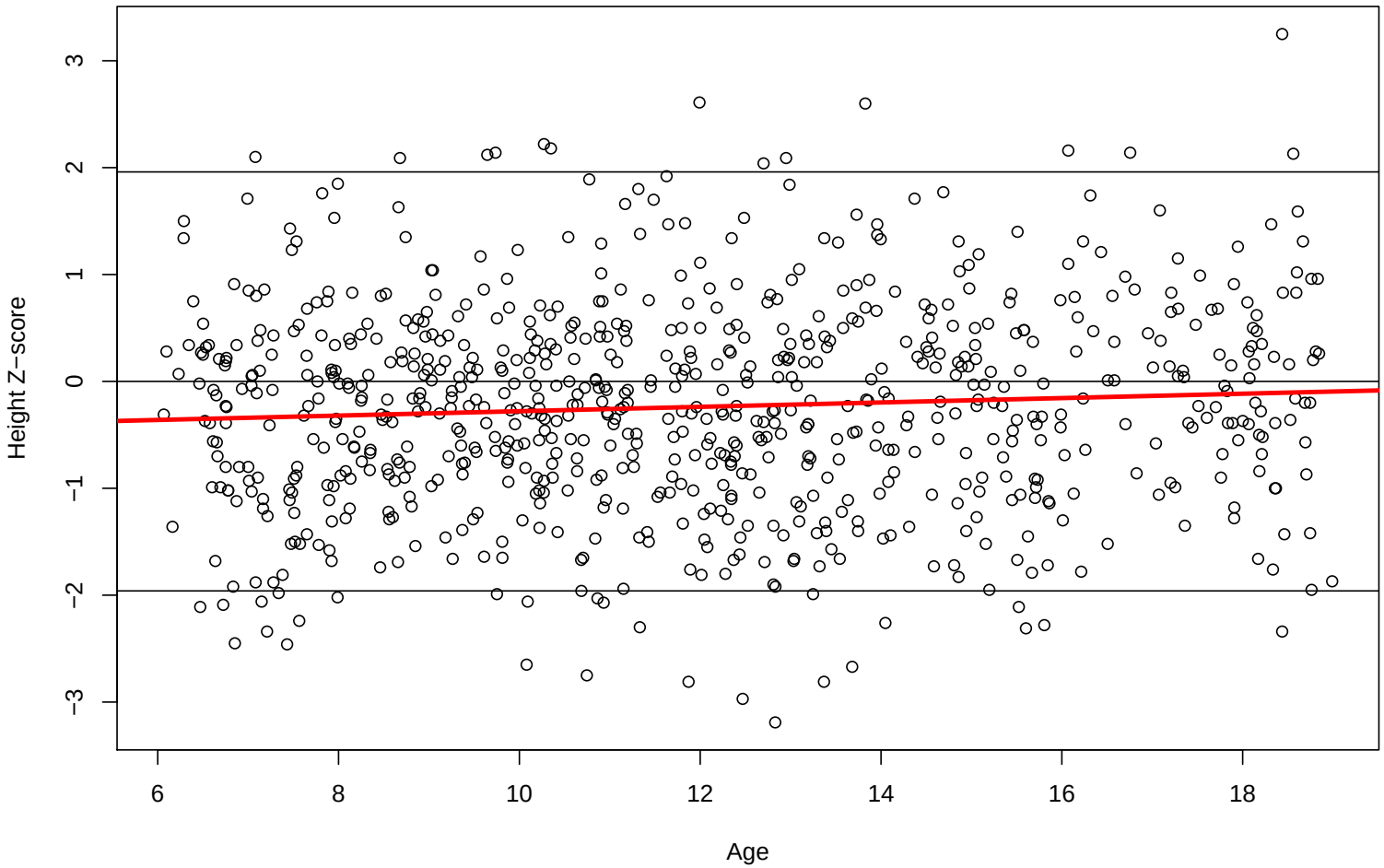
Základní popisné statistiky:

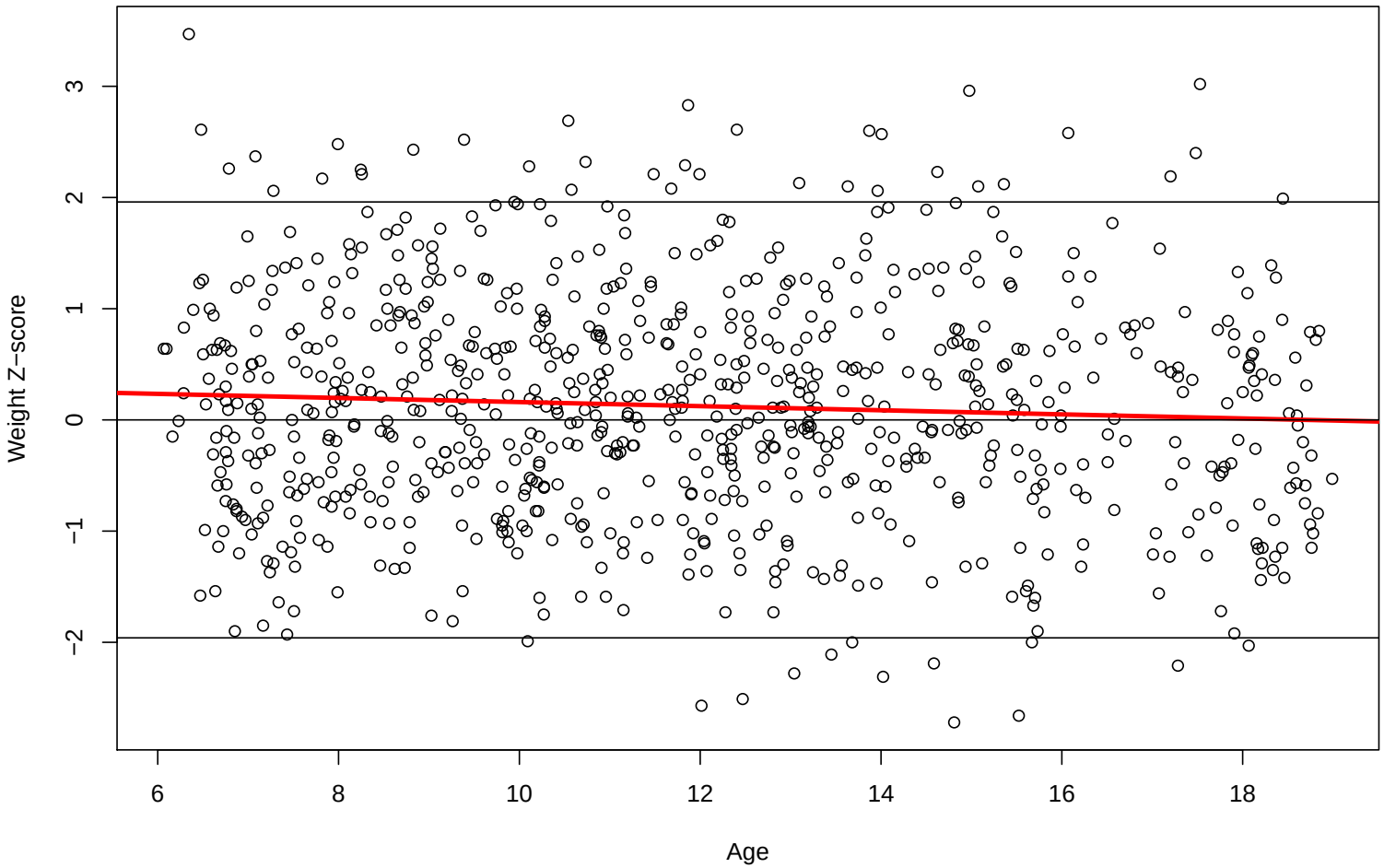
	průměr(+−s.odch.)	p-hodnota
Věk	11.83(+−3.453)	$< 1e{-}04$***
Výška	149.021(+−18.047)	$< 1e{-}04$***
Hmotnost	43.228(+−15.51)	$< 1e{-}04$***
BMI	18.903(+−3.217)	$< 1e{-}04$***

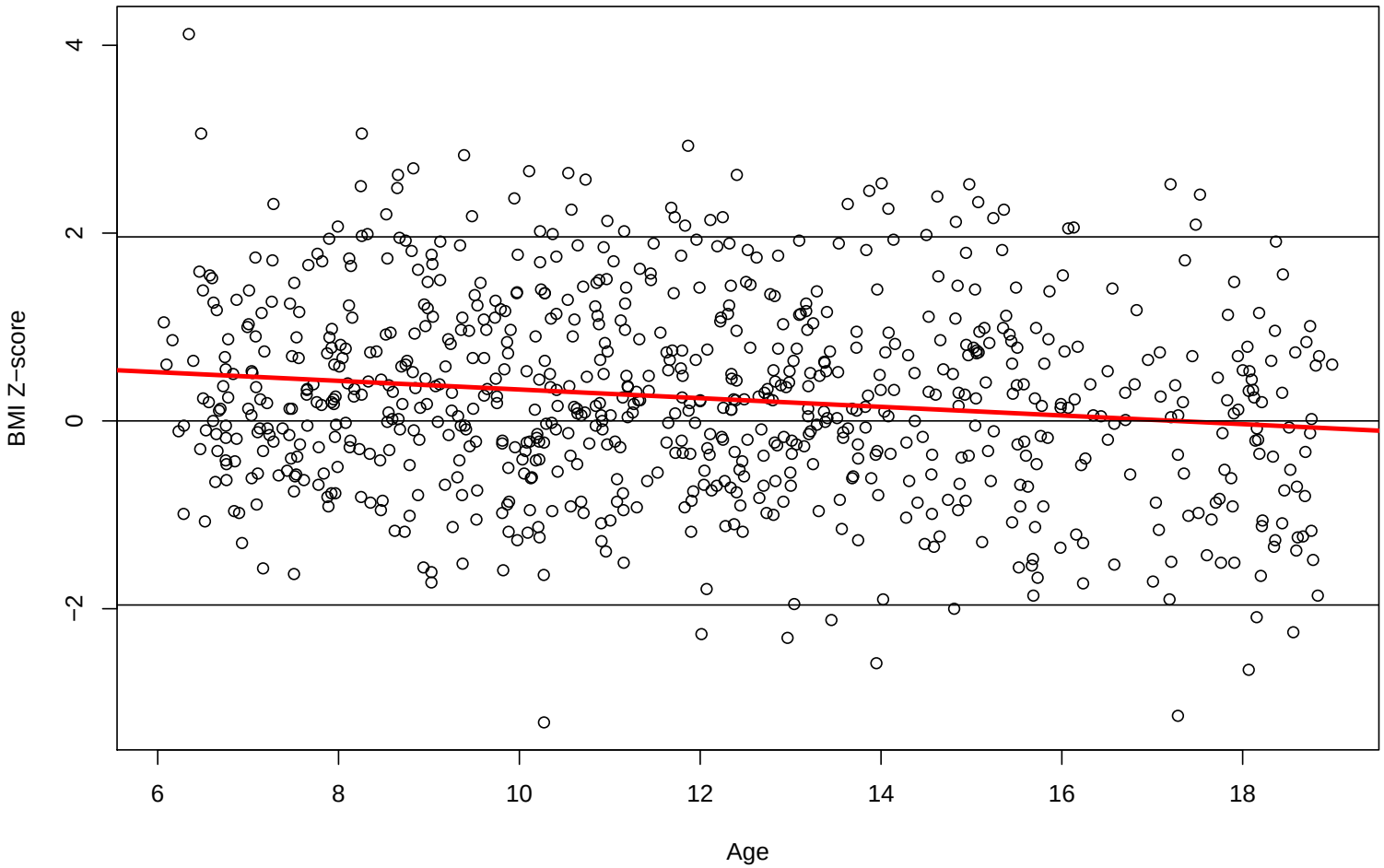
Tabulka věkových Z-skóre:

	průměr(+−s.odch.)	p-hodnota
Z-skóre výšky	−0.24(+−0.969)	$< 1e{-}04$***
Z-skóre hmotnosti	0.127(+−1.027)	0.000532***
Z-skóre BMI	0.247(+−1.043)	$< 1e{-}04$***


Height
120
140
160
180
6
8
10
12
14
16
18
Age


Height Z−score
3
2
1
0
−1
−2
−3
6
8
10
12
14
16
18
Age


Weight Z−score
3
2
1
0
−1
−2
6
8
10
12
14
16
18
Age


BMI Z−score
4
2
0
−2
6
8
10
12
14
16
18
Age

Jak se Z-skóre obvykle počítá...

Cole (1990) The LMS method for constructing normalized growth standards. *Eur J Clin Nutr* **44**: 45–60.

L: Box-Coxova transformace původních dat (kvůli nesymetrii),
M: střední hodnota,
S: variační koeficient (měřítko).

Po zvolení L, např. 0 (logaritmus) nebo 0.5 (odmocnina), spočítáme Z-skóre jako:

$$Z = \frac{(X/M)^L - 1}{LS}, \quad \text{pro } L \neq 0,$$

$$Z = \frac{\log(X/M)}{S}, \quad \text{pro } L = 0.$$

Vysvětlení vzorečků:

Např. pro $L \neq 0$:

$$Z = \frac{(X/M)^L - 1}{LS} = \frac{X^L - M^L}{LSM^L} = \frac{Y - \mu_Y}{\sigma_Y},$$

kde
$\mu_Y = M^L$ je střední hodnota $Y = X^L$
a
$\sigma_Y = LSM^L$ je směrodatná odchylka $Y = X^L$.

Obráceně: $M = \mu_Y^{1/L}$ a $S = \sigma_Y \mu_Y^{1/L}/L$.

V praxi se M a S obvykle odhadují jako polynomy (funkce věku).
Cole (1990) ale mluví i o neparametrických odhadech (spliny).

Měření výšky výskoku

Leonardo Mechanograph GRFP (Ground Reaction Force Plate)

Měření výšky výskoku

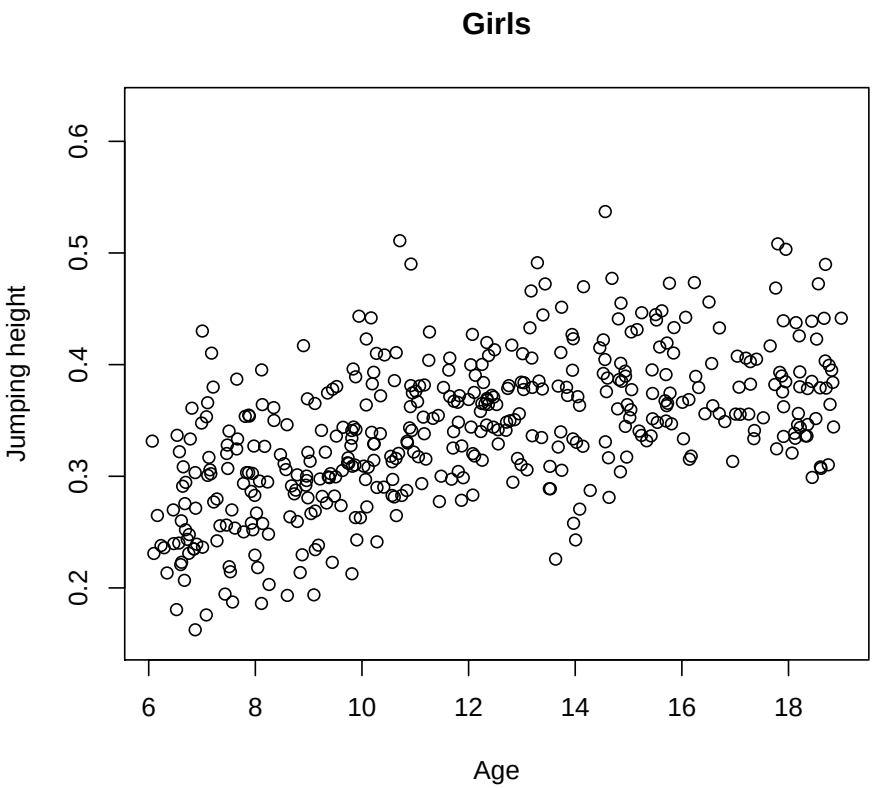
Celkem 796 dětí (432 dívek, 364 chlapců).

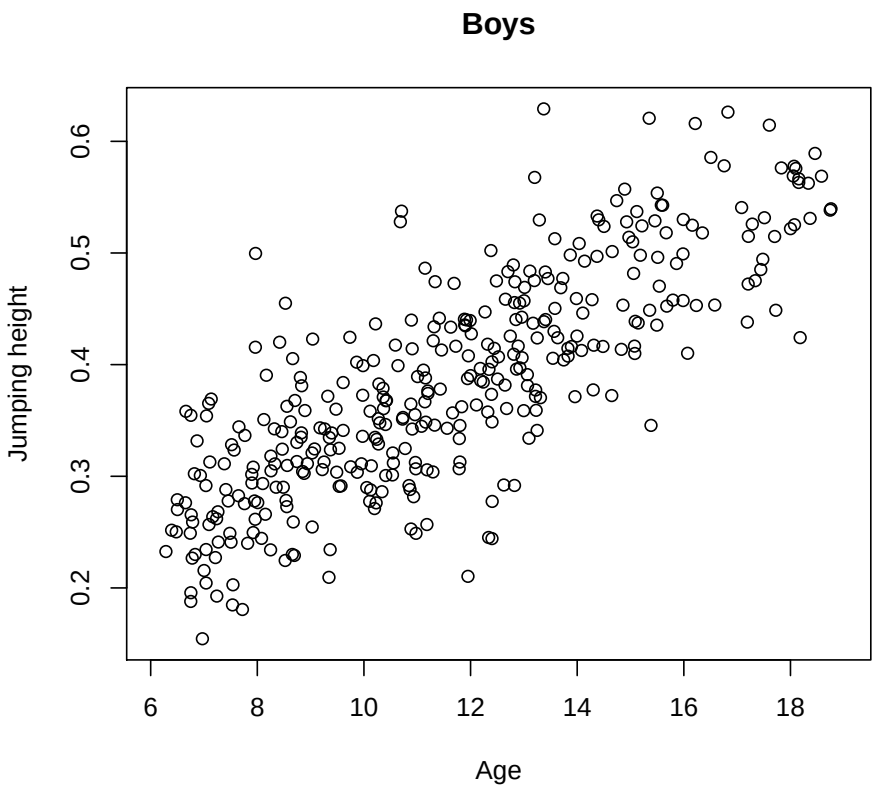
Single two-legged jump (výskok snožmo)
–aims to achieve *maximum jump height*.

Multiple one-legged hopping (poskakování na jedné noze)
–aims to achieve maximum voluntary forefoot ground reaction force during landing. One possible application of this test is to evaluate the maximal force to which the tibia is exposed, and thus it might serve to evaluate the muscle-bone unit.

Data nasbírali kolegové z 2.LF a PřF: Šumník, Z., Matysková, J., Hlávka, Z., Durdilová, L., Souček, O., & Zemková, D. (2013). Reference data for jumping mechanography in healthy children and adolescents aged 6-18 years. *Journal of musculoskeletal & neuronal interactions*, 13(3), 259-273.

Výška výskoku

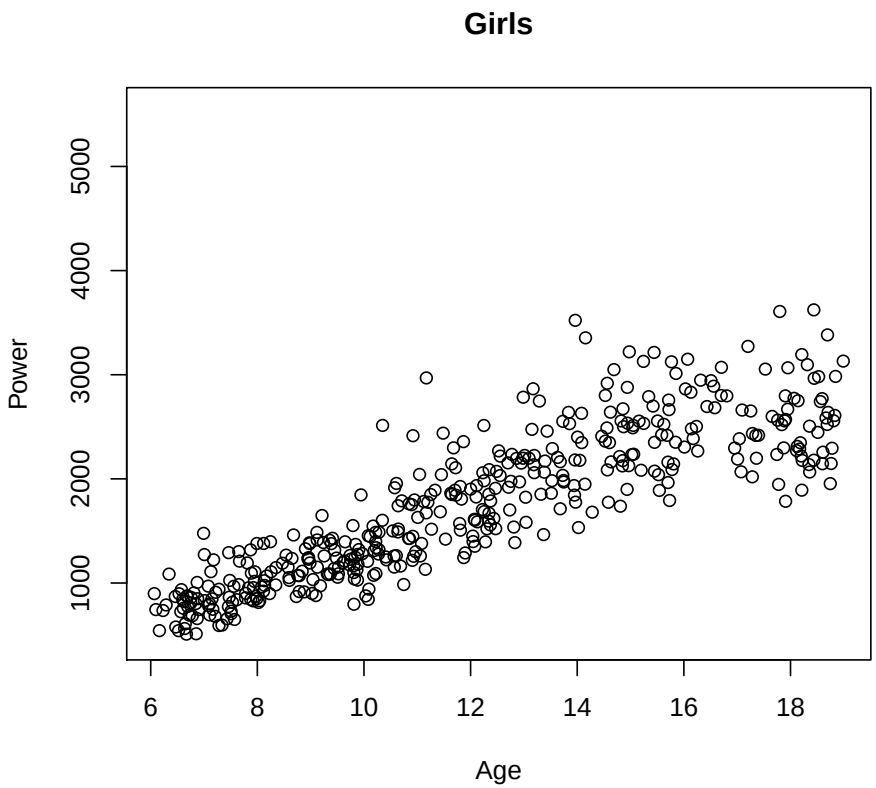


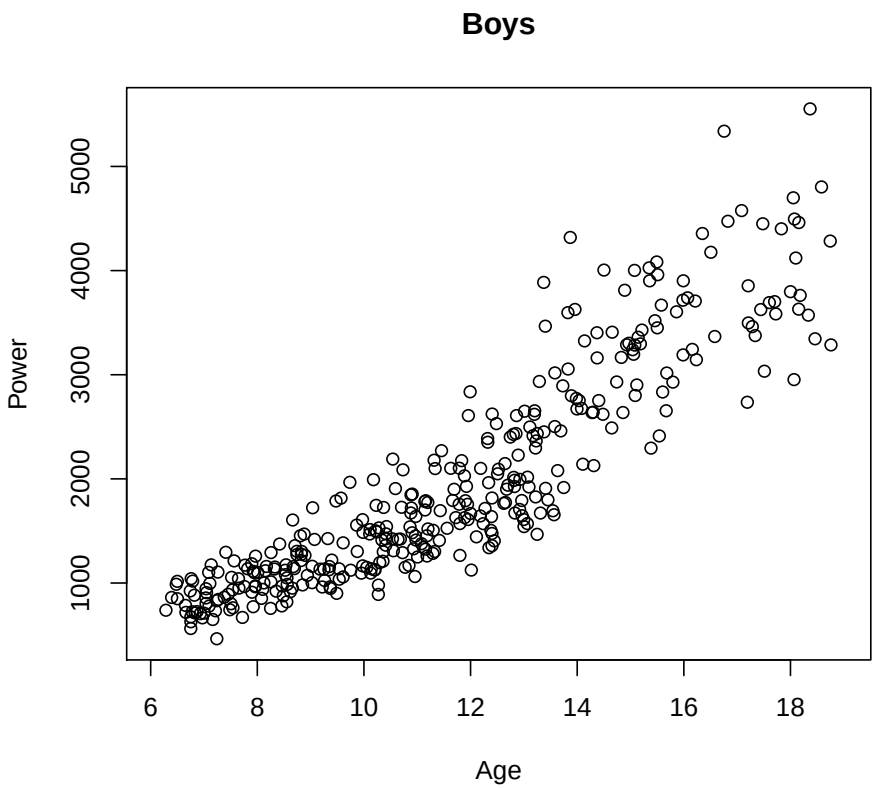


Průměrná výška výskoku podle pohlaví a věku:

	mean.f	n.f	mean.m	n.m	pval	
6	0.26	33	0.26	19	0.99221	
7	0.30	43	0.28	38	0.33882	
8	0.29	33	0.32	38	0.05454	.
9	0.31	42	0.33	29	0.07565	.
10	0.34	42	0.34	45	0.88786	
11	0.35	30	0.38	37	0.00736	**
12	0.36	41	0.40	40	0.00201	**
13	0.37	32	0.44	36	5.4e-05	***
14	0.37	31	0.47	20	5.1e-07	***
15	0.39	29	0.49	26	4.4e-09	***
16	0.38	17	0.53	9	0.00021	***
17	0.39	25	0.51	13	2.3e-07	***
18	0.38	34	0.55	14	5.3e-13	***

Další proměnné: výkon při výskoku, síla při dopadu





Odhad Z-skóre

Postup:
zvolíme L a pak vypočítáme:

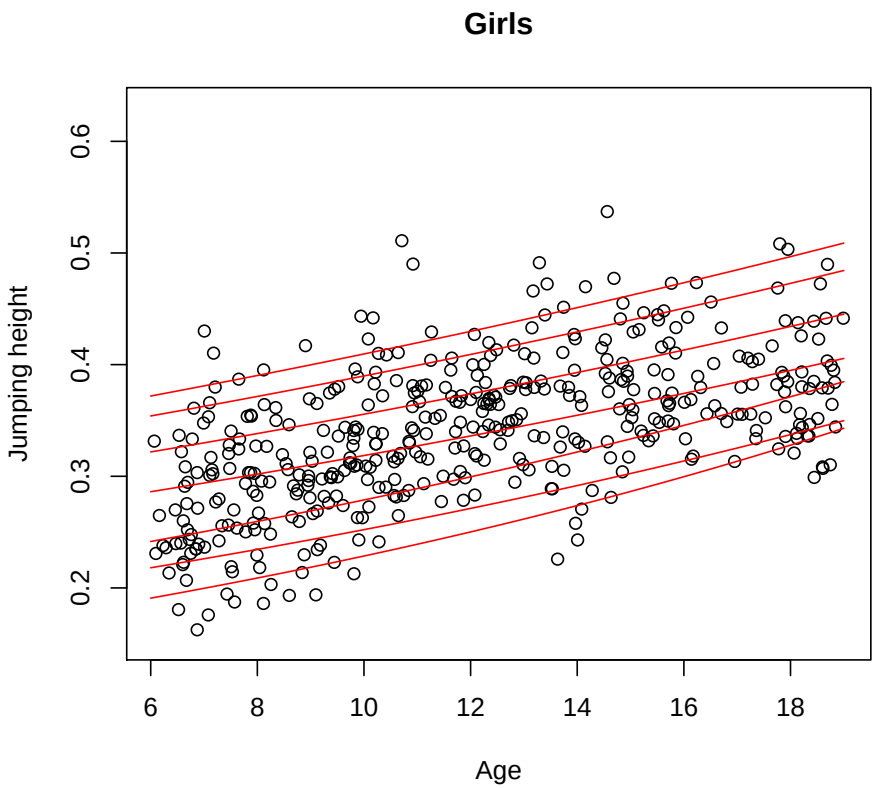
$$Z_i = \frac{(X/M)^L - 1}{LS} = \frac{Y_i - \mu_Y(\text{věk}_i)}{\sigma_Y(\text{věk}_i)}.$$

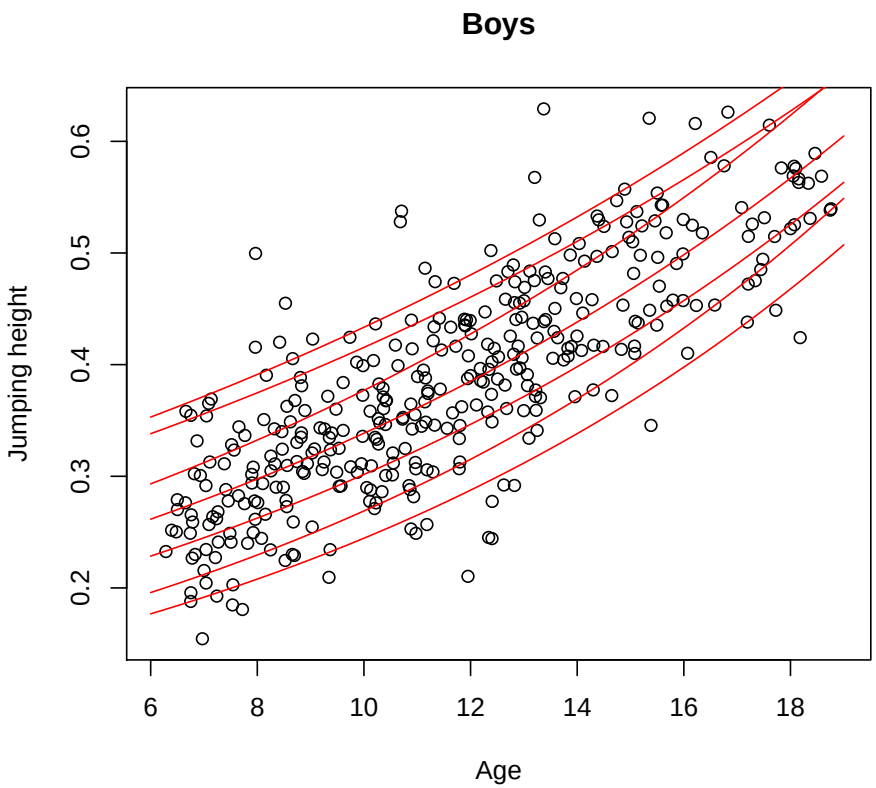
Z hodnot Y_i a věk$_i$ tedy potřebujeme odhadnout $\mu_Y(\text{věk})$ a $\sigma^2_Y(\text{věk})$, tj. podmíněnou střední hodnotu a rozptyl Y (jako funkci věku).

Možnosti:

1. polynomická regrese (postupně $\widehat{\mu}_Y(\text{věk})$ a $\widehat{\sigma}^2_Y(\text{věk})$),
2. neparametrická regrese,
3. kvantilová regrese,
4. další postupy (většinou odvozené od kvantilové regrese).

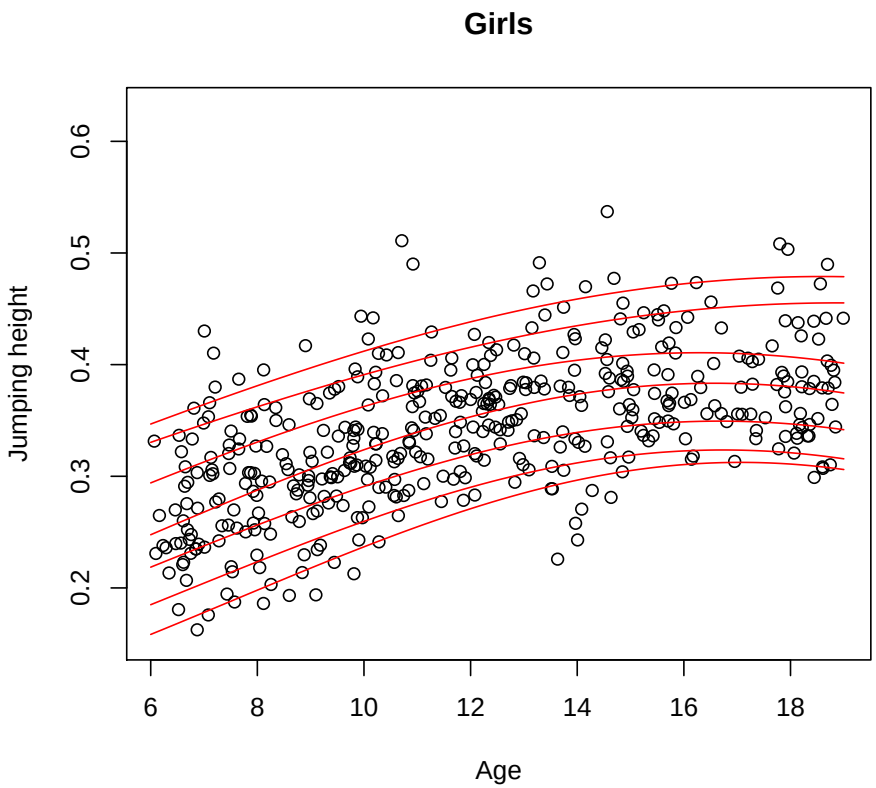
Výška výskoku: lineární kvantilová regrese ($L = 0$)

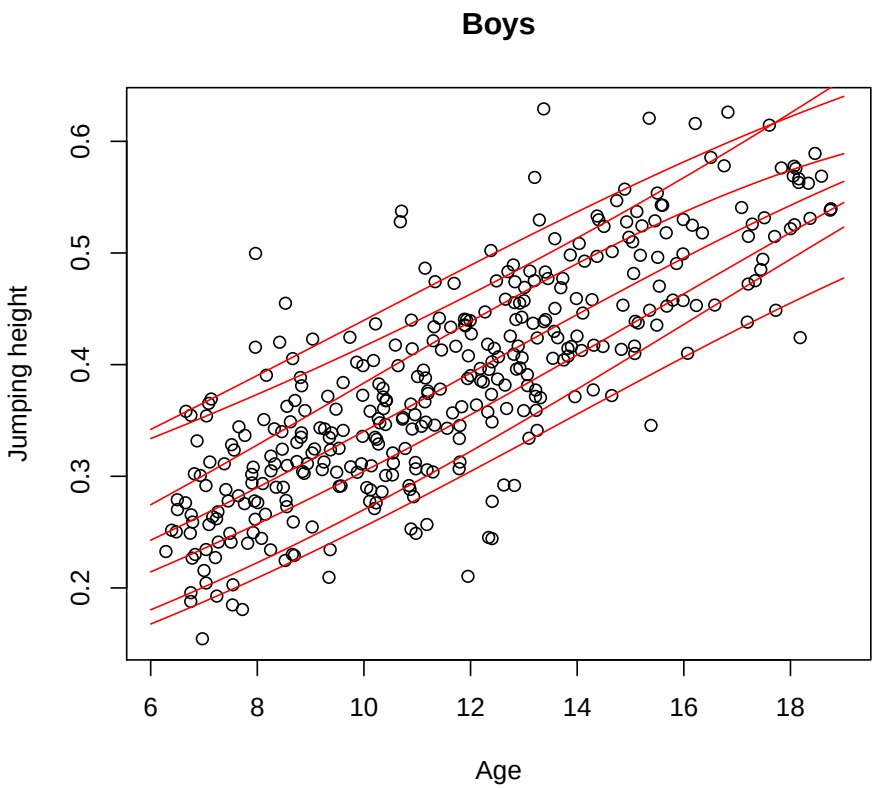




Nemusí se zvlášť odhadovat měřítko, ale výsledek nevypadá příliš hezky.

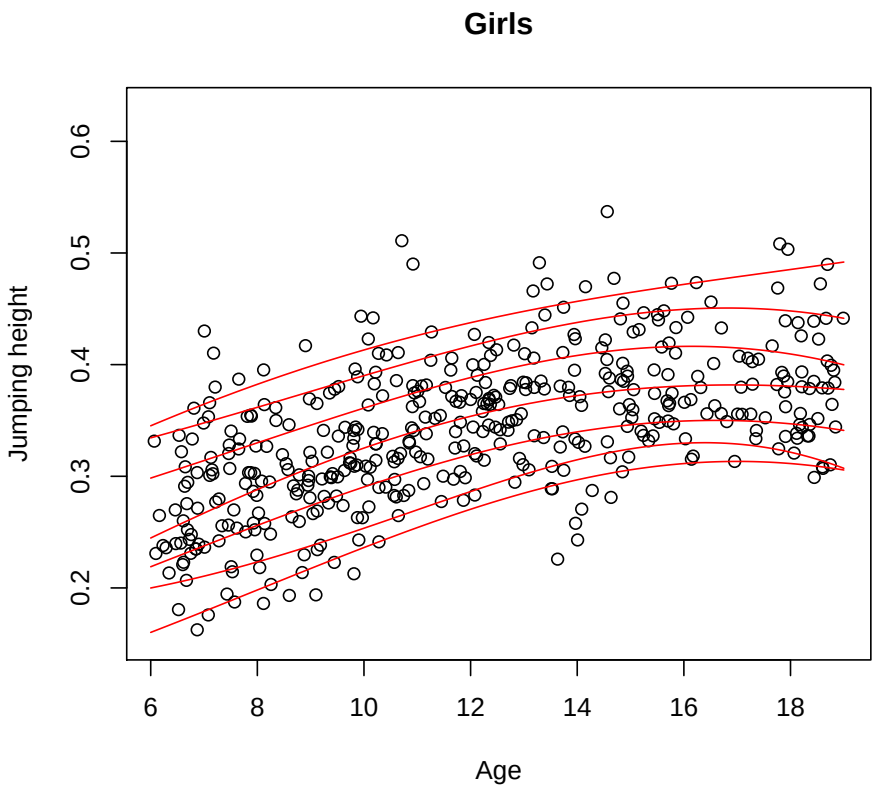
Výška výskoku: kvadratická kvantilová regrese

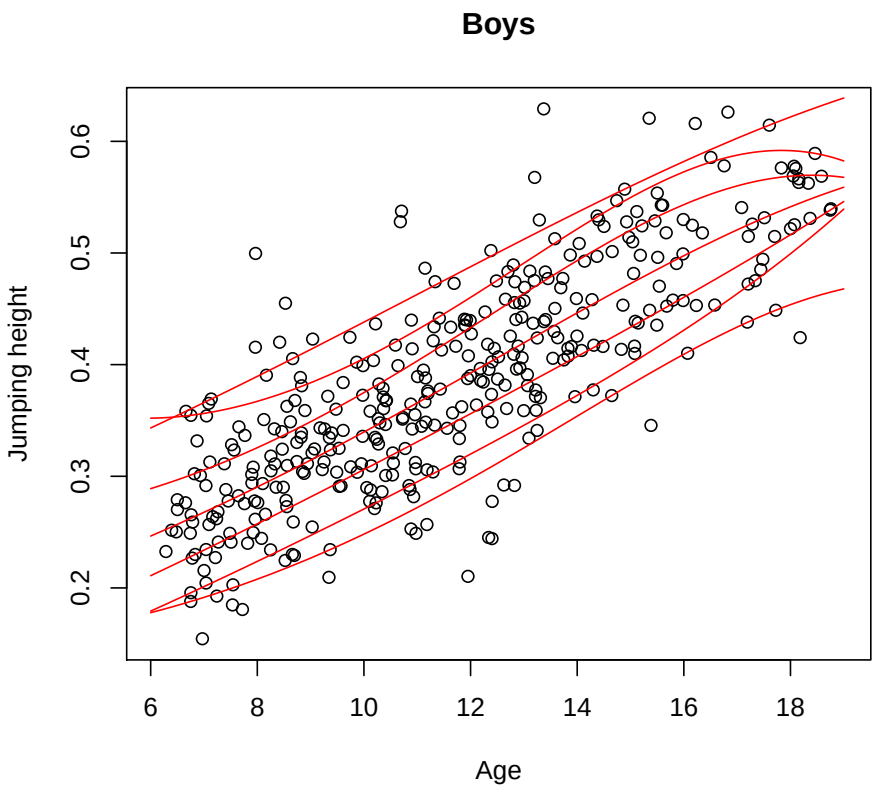




Odhadnuté kvantilové křivky se někdy i protínají.

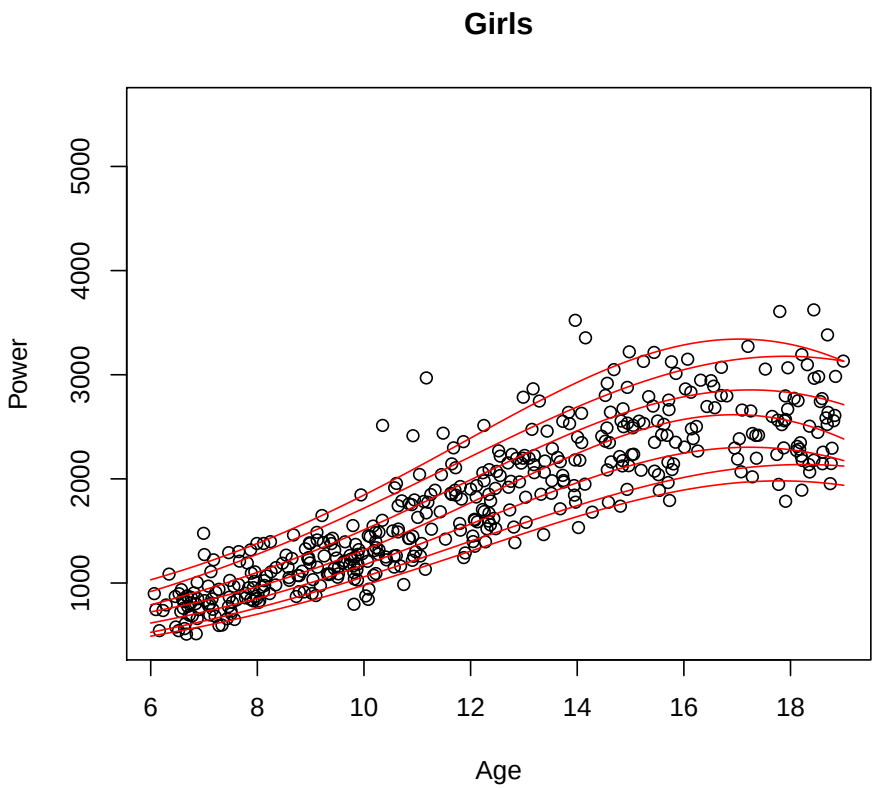
Výška výskoku: kubická kvantilová regrese

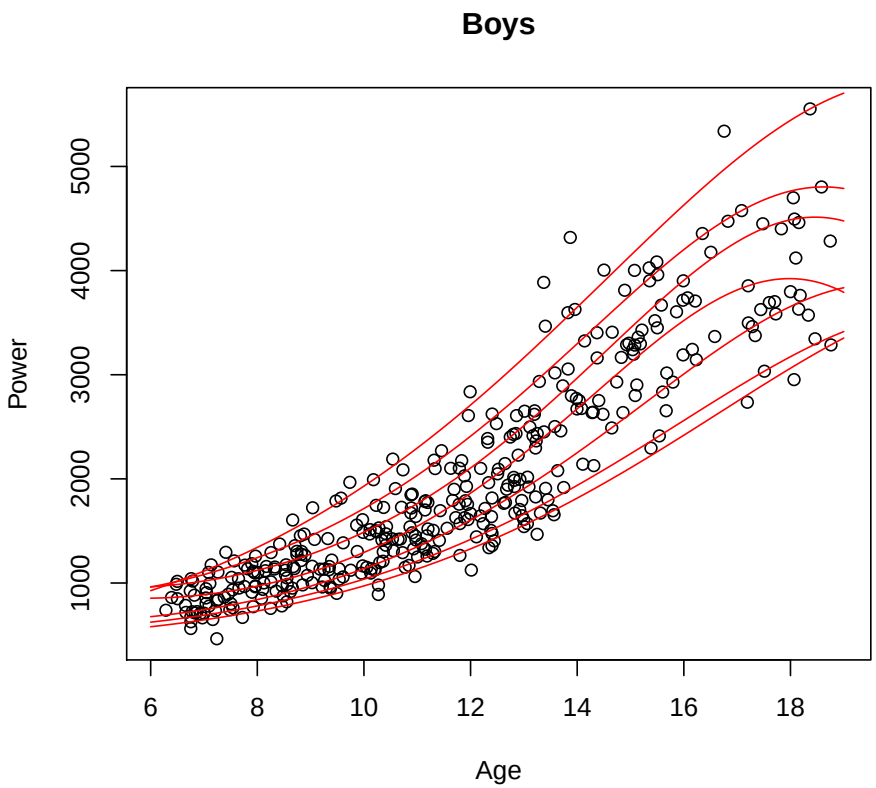




Ani s polynomy vyšších řádů není výsledek použitelný.

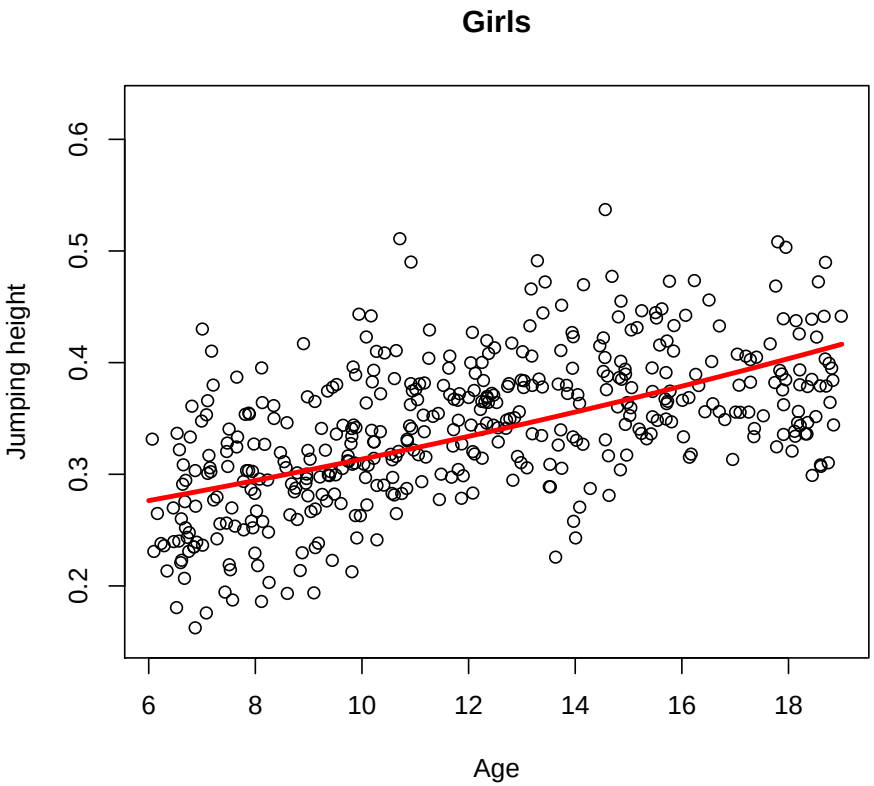
Výkon při výskoku: kubická kvantilová regrese

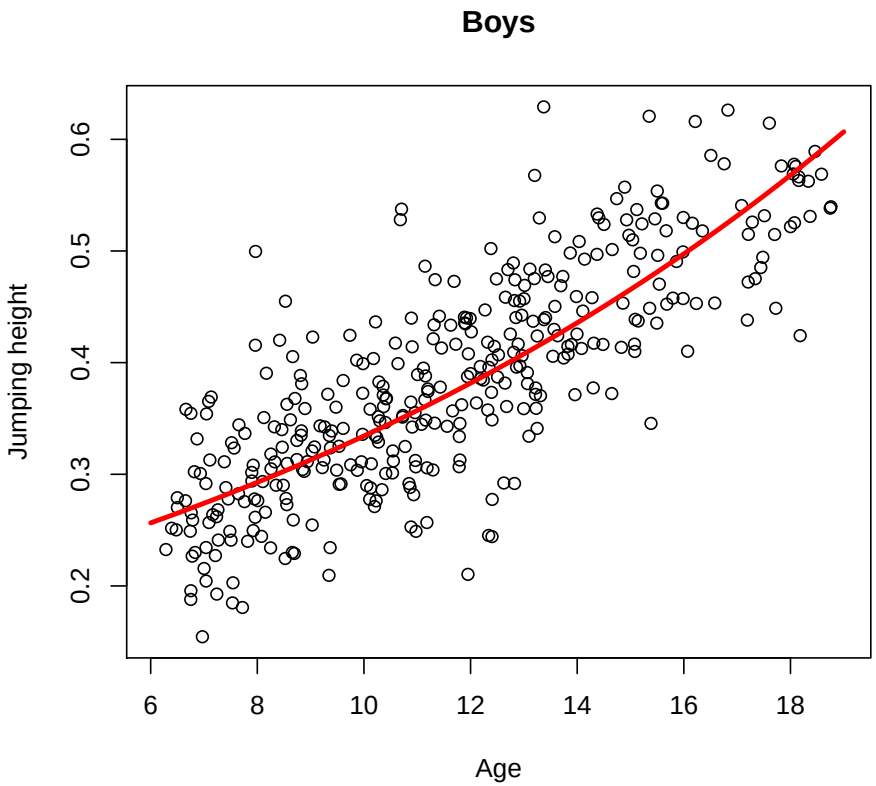




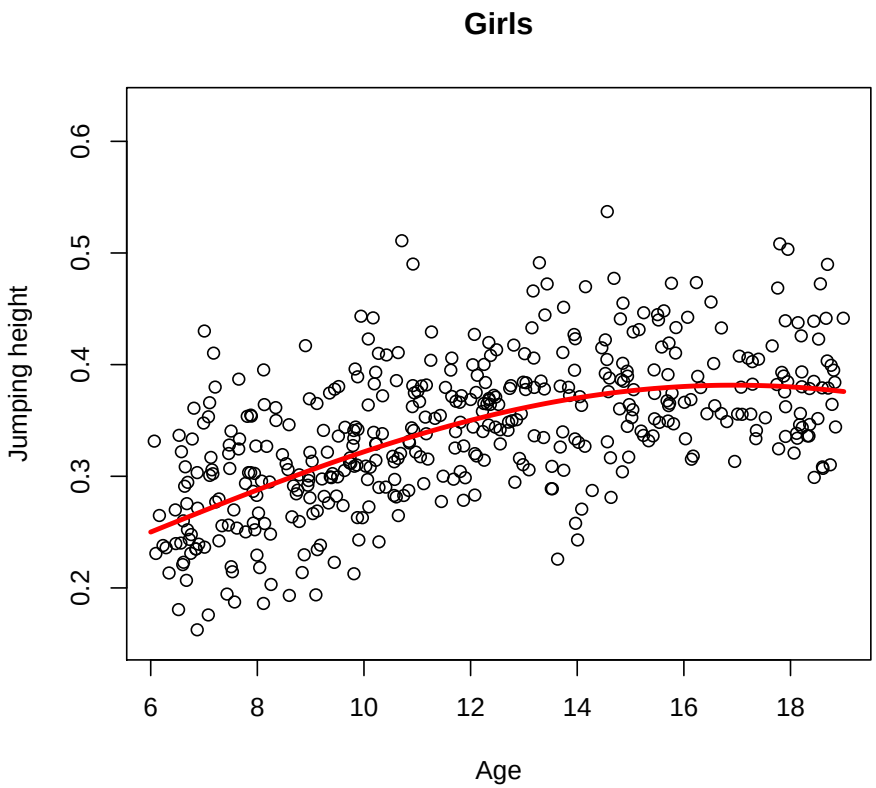
U dalších proměnných je to podobné.

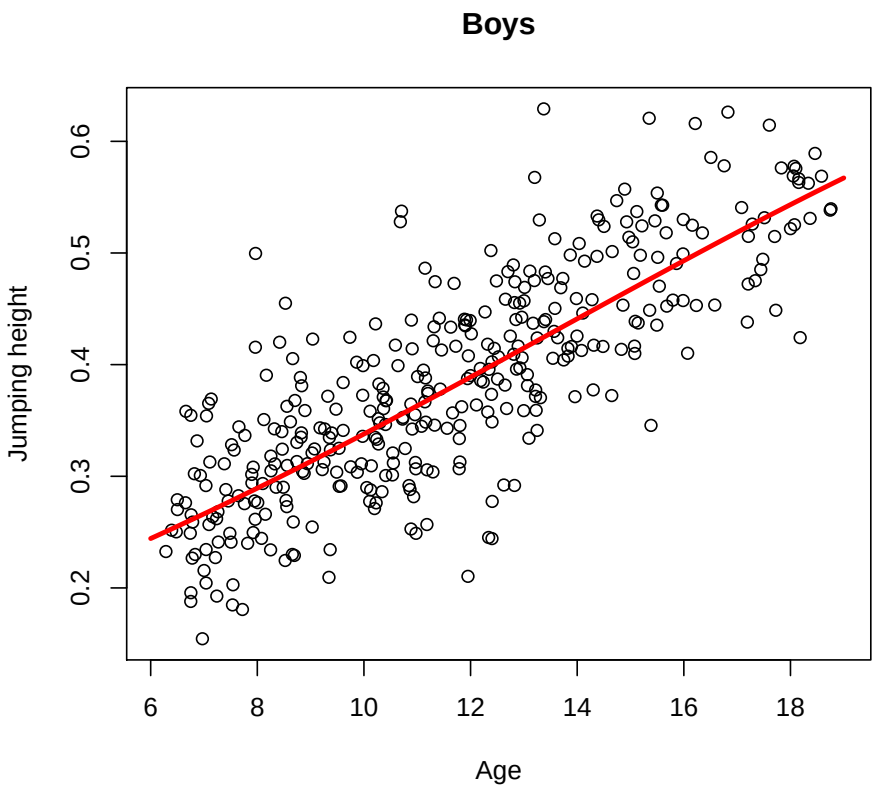
Výška výskoku: lineární regrese (odhad střední hodnoty)





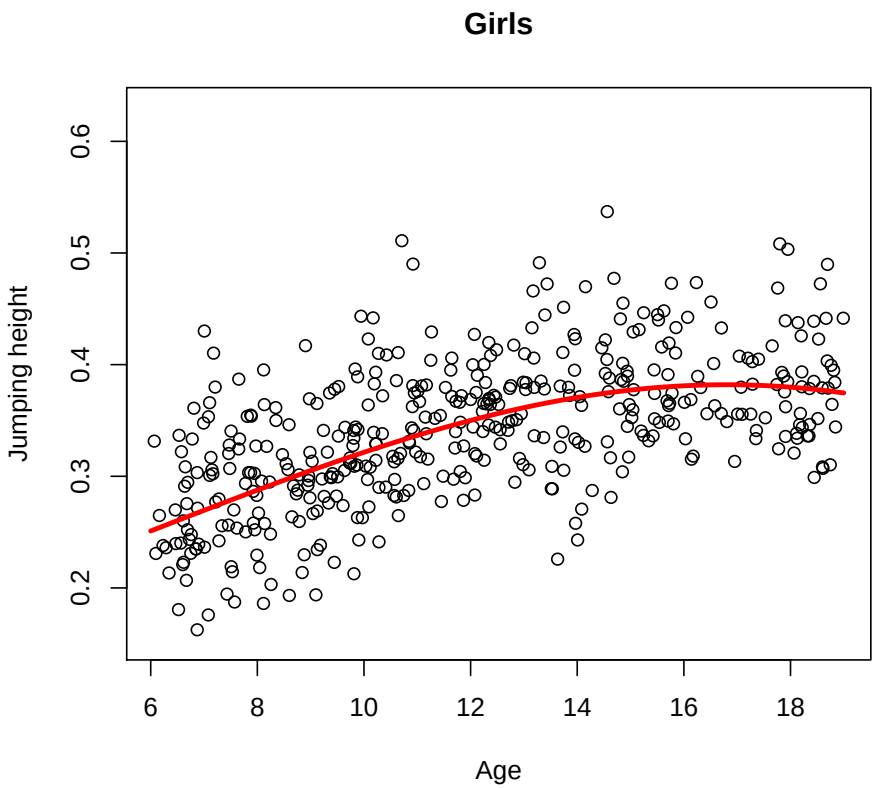
Výška výskoku: kvadratická regrese

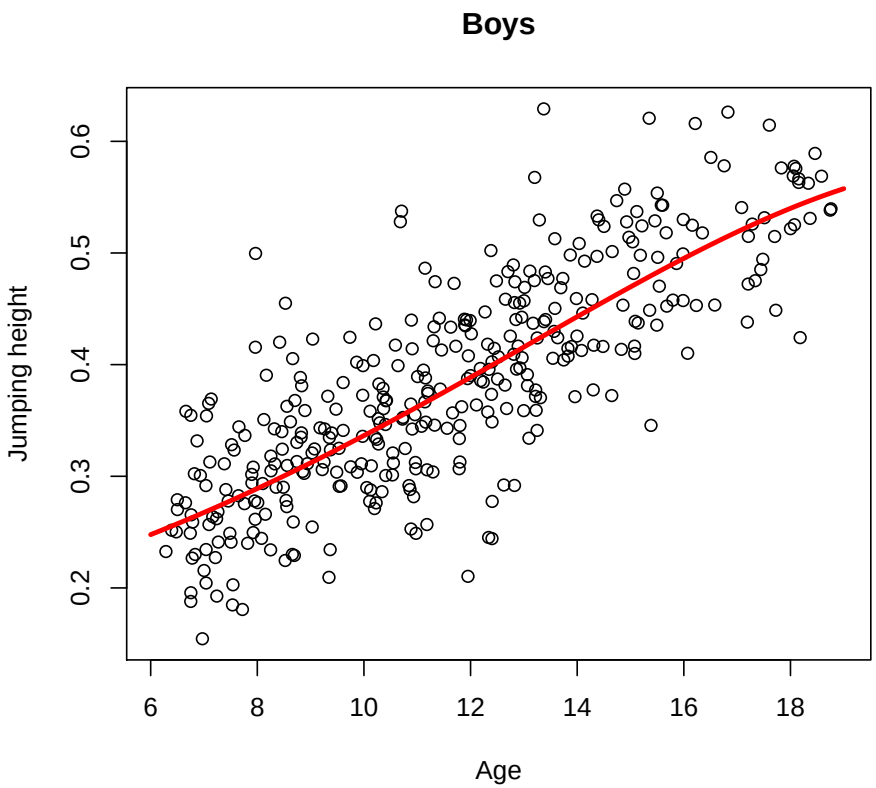




Neklesá pro nejstarší dívky? Je to dobrý odhad pro nejstarší chlapce?

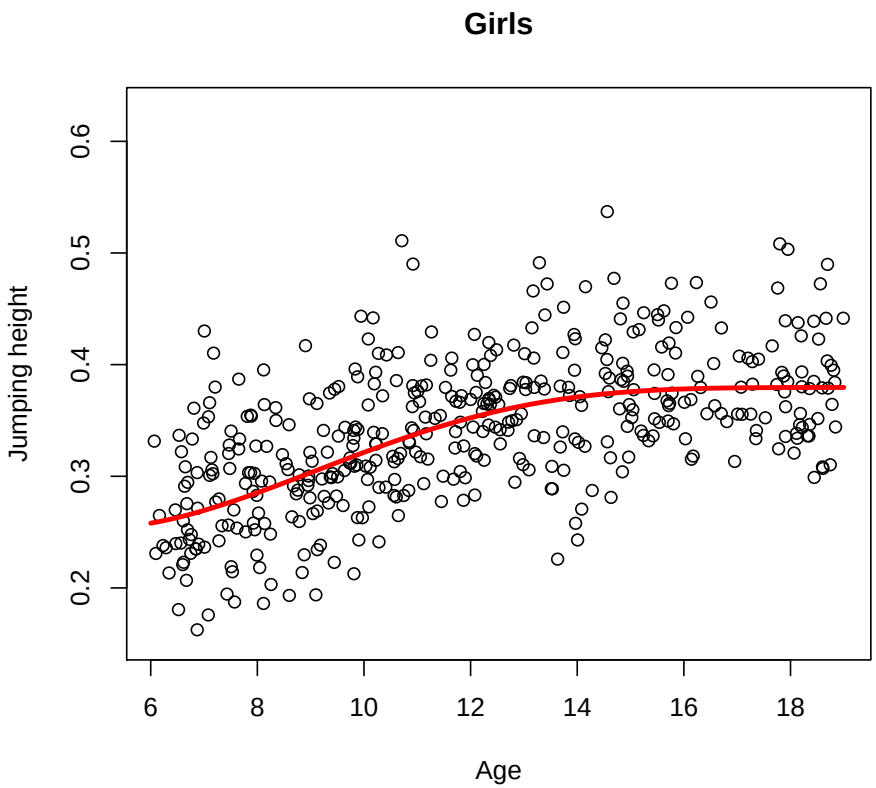
Výška výskoku: kubická regrese

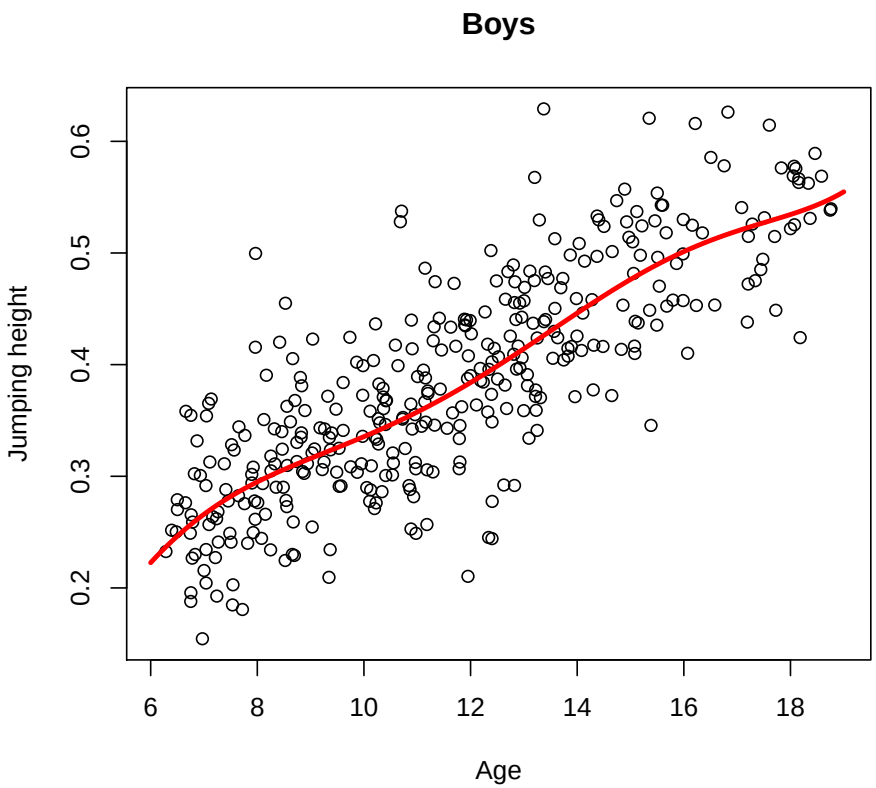




Polynomy vyššího stupně nic neřeší.

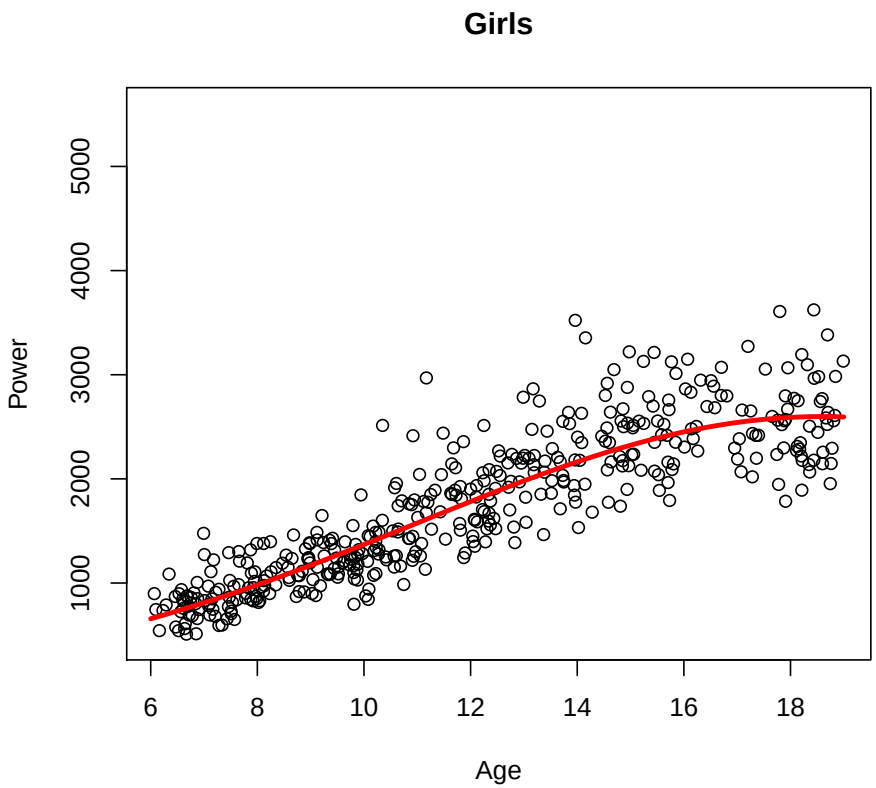
Výška výskoku: polynom pátého stupně

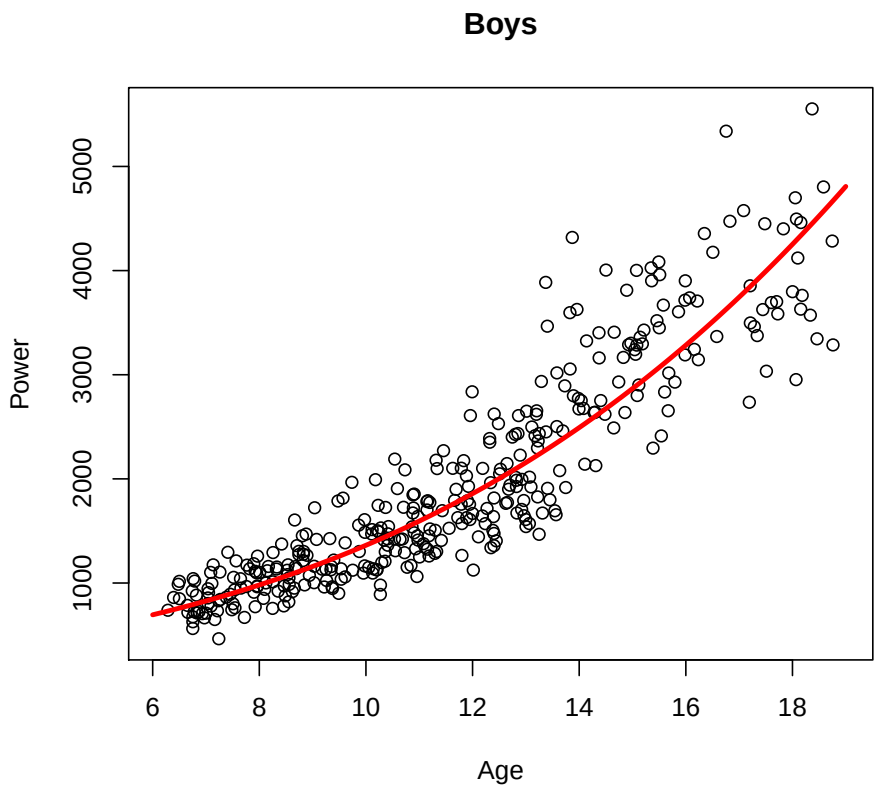




Polynomy vyššího stupně nic neřeší. Jak to funguje pro jiné proměnné?

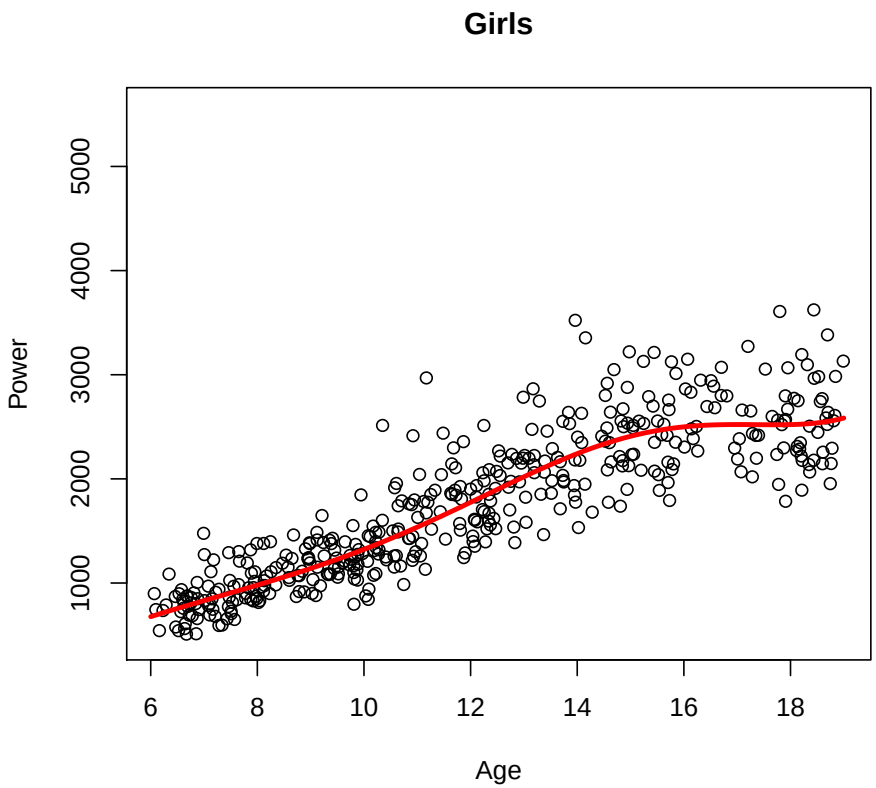
Výkon při výskoku: kvadratická regrese

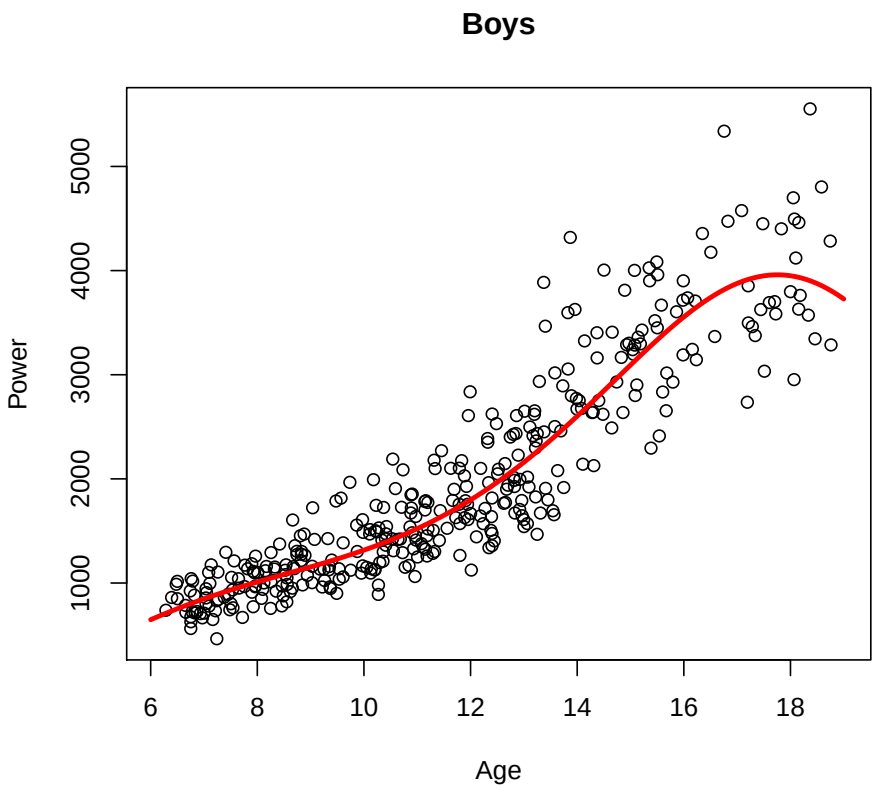




Neklesá pro nejstarší dívky? Je to dobrý odhad pro nejstarší chlapce?

Výkon při výskoku: polynom pátého stupně





Polynomy vyššího stupně nic neřeší. Pomůže neparametrická regrese?

Požadavky na rozumný postup:

- jednoduchost (praktická analýza pro lékařský časopis),
- automatizace (shodný postup pro všechny proměnné),
- flexibilita (závislost na věku nebo váze není lineární ani polynomická),
- výsledek musí být rozumný (např. monotonicita).

Parametrické modely splňují pouze první požadavek.

Můžeme použít neparametrickou regresi?

Požadavky na rozumný postup:

- jednoduchost (praktická analýza pro lékařský časopis),
- automatizace (shodný postup pro všechny proměnné),
- flexibilita (závislost na věku nebo váze není lineární ani polynomická),
- výsledek musí být rozumný (např. monotonicita).

Parametrické modely splňují pouze první požadavek.

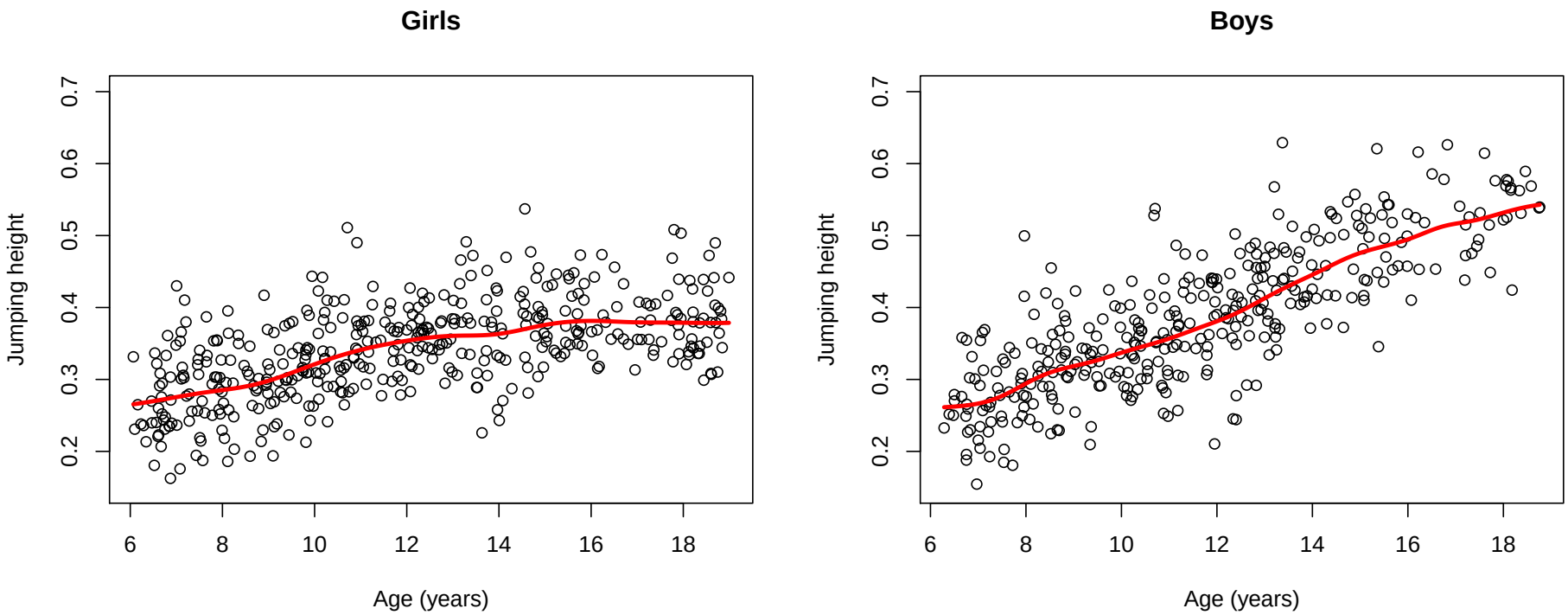
Můžeme použít neparametrickou regresi?

Např. při použití R funkce `sm.regression(x, y, h, method, ...)` narazíme na obvyklé „praktické problémy“:

volba bandwidth: nedokumentovaný parametr `method` (CV?),

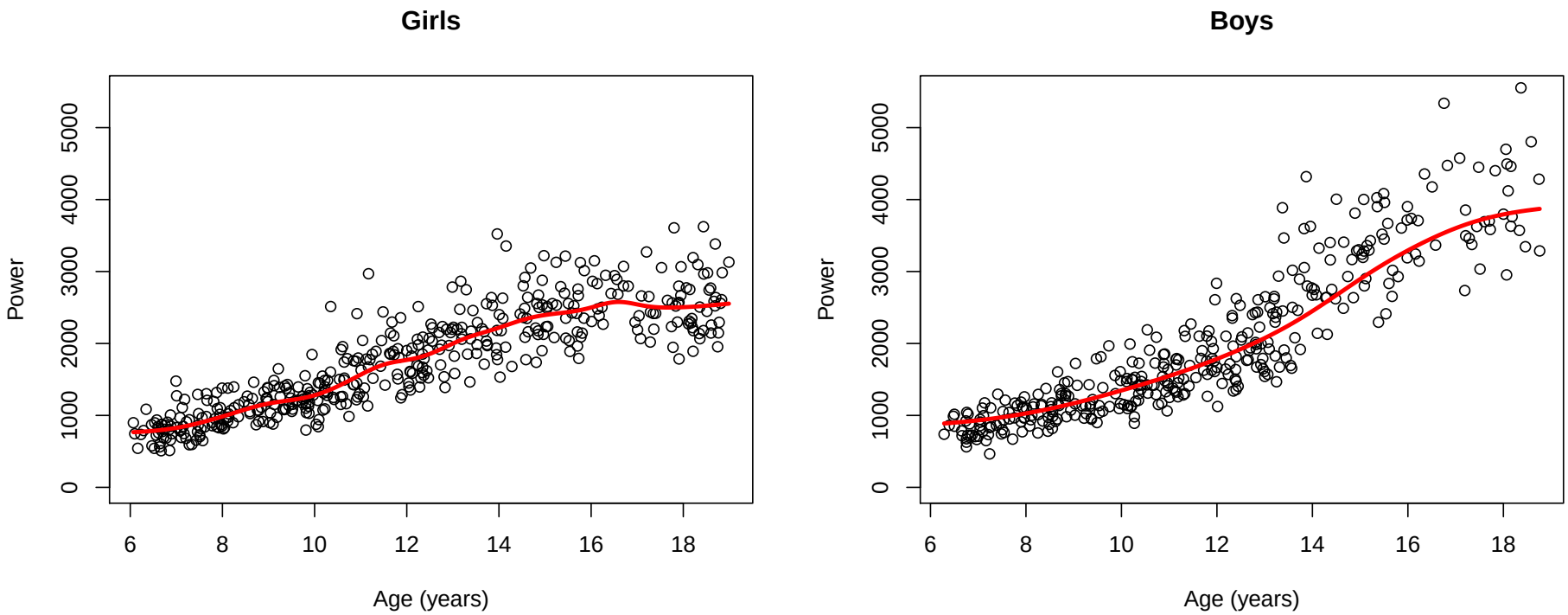
okrajové efekty: lokálně polynomické odhady?

Výška výskoku: lokálně konstantní $\widehat{m}_0(x)$



Bandwidth: automaticky podle CV (křížové ověřování).
Jak to vypadá pro výkon?

Výkon při výskoku: lokálně konstantní $\widehat{m}_0(x)$



Bandwidth: automaticky podle CV (křížové ověřování).
Metoda CV zde tedy není vhodná. Jiné možnosti?

Volba bandwidth: AICC

Parametr `method` funkce `sm.regression()` sice není zdokumentovaný, ale v helpu lze dohledat možnost nastavení AICC podle článku:

Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. JRSSB 60(2), 271–293.

Z článku: Classical bandwidth selectors (particularly GCV and the AIC) have to some extent fallen into disuse (particularly in application to local polynomial and kernel estimators) because of two unfavourable properties: the selectors lead to highly variable choices of smoothing parameter, and they have a noticeable tendency towards undersmoothing...

Volba bandwidth: AICC

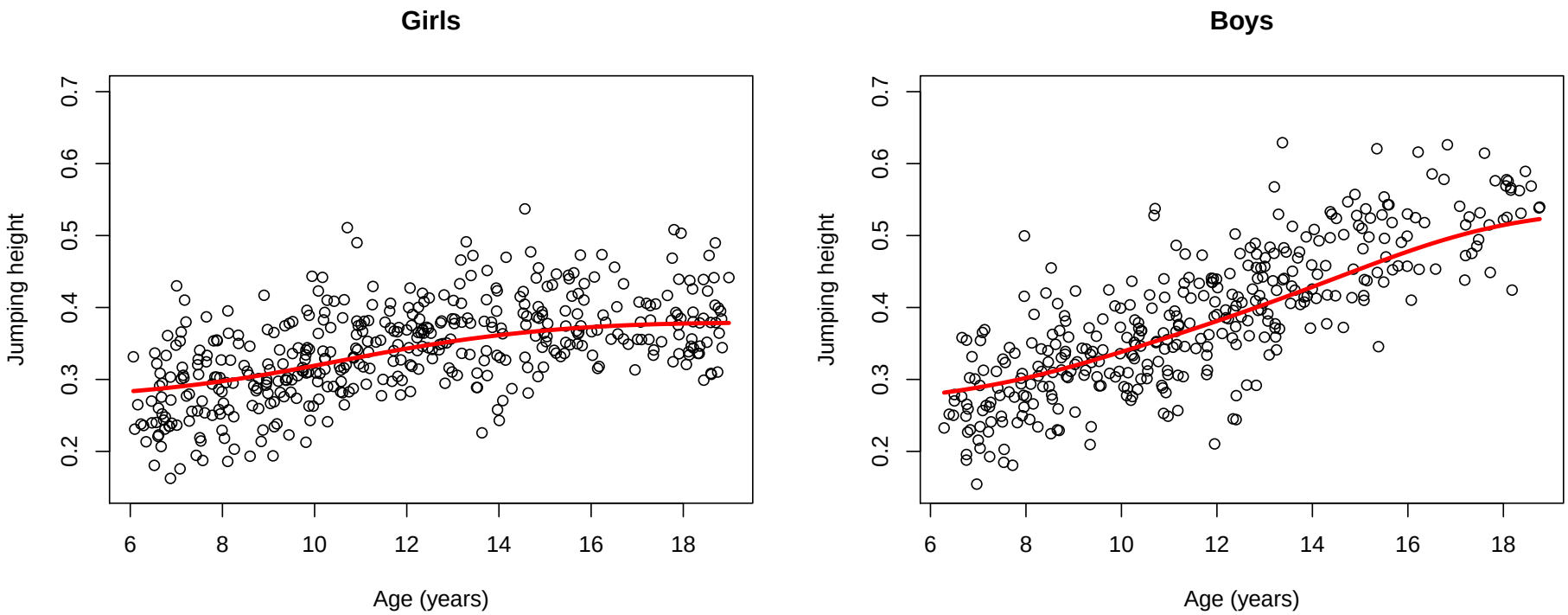
Pokud můžeme zapsat neparametrický regresní model ve tvaru $\hat{m}(\text{věk}) = HY$, pak

$$\text{AICC} = \log(\hat{\sigma}^2) + 1 + \frac{2\{\text{tr}(H) + 1\}}{n - \text{tr}(H) + 2}$$

je odvozeno jako aproximace odhadu střední hodnoty Kullback-Leiblerovy informace (porovnání modelu a skutečnosti).

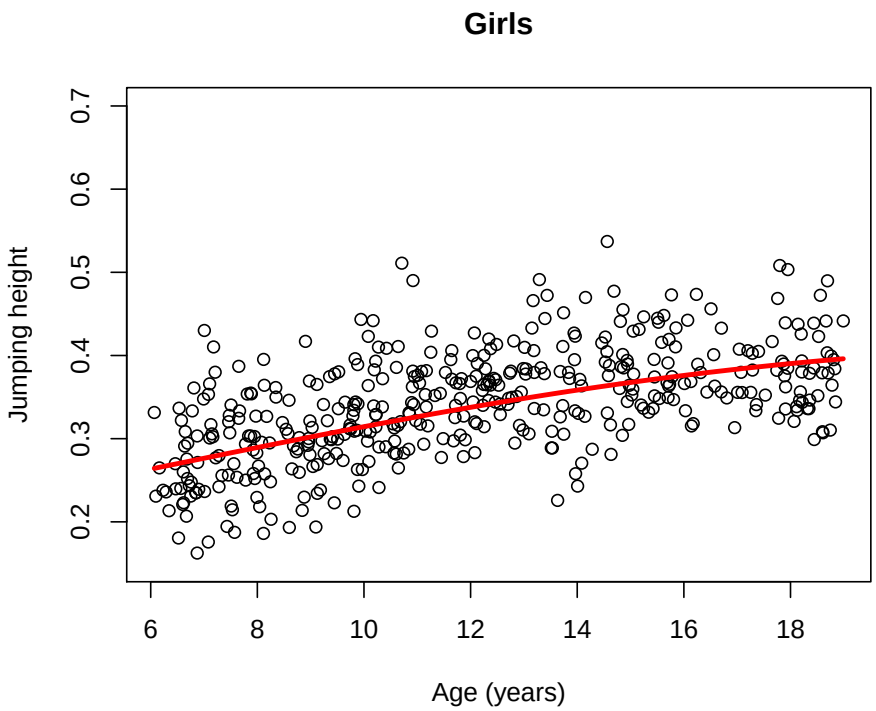
Z abstraktu článku Hurvich et al (1998): . . . The use of AICC avoids the large variability and tendency to undersmooth (compared with the actual minimizer of average squared error) seen when other 'classical' approaches (such as generalized cross-validation (GCV) or the AIC) are used to choose the smoothing parameter.

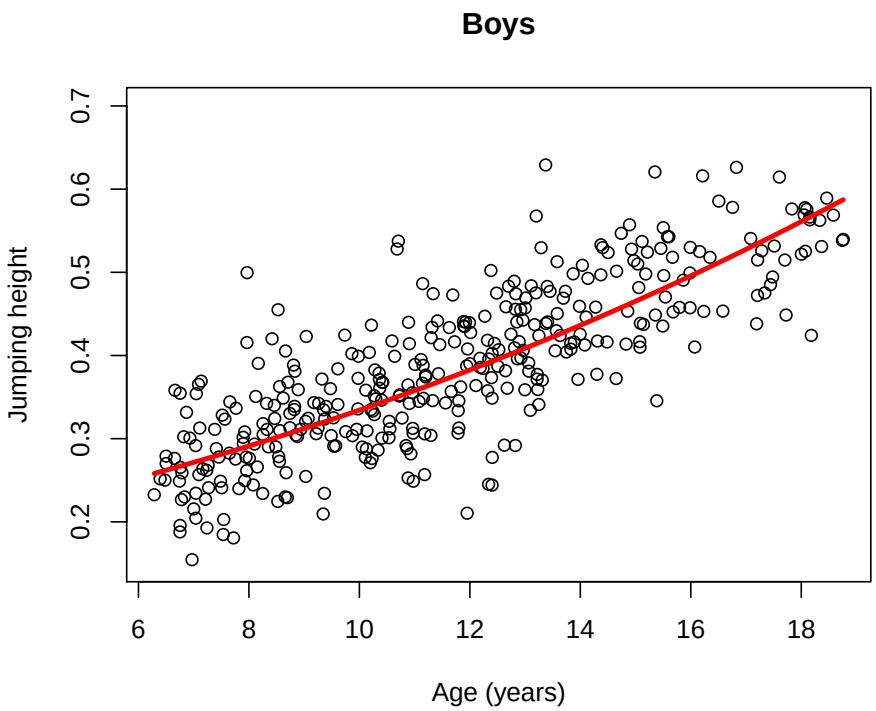
Výška výskoku: lokálně konstantní $\widehat{m}_0(x)$, AICC



Bandwidth: automaticky podle AICC (větší než při použití CV).
Odhad nevypadá dobře pro nejmladší děti (boundary effect).

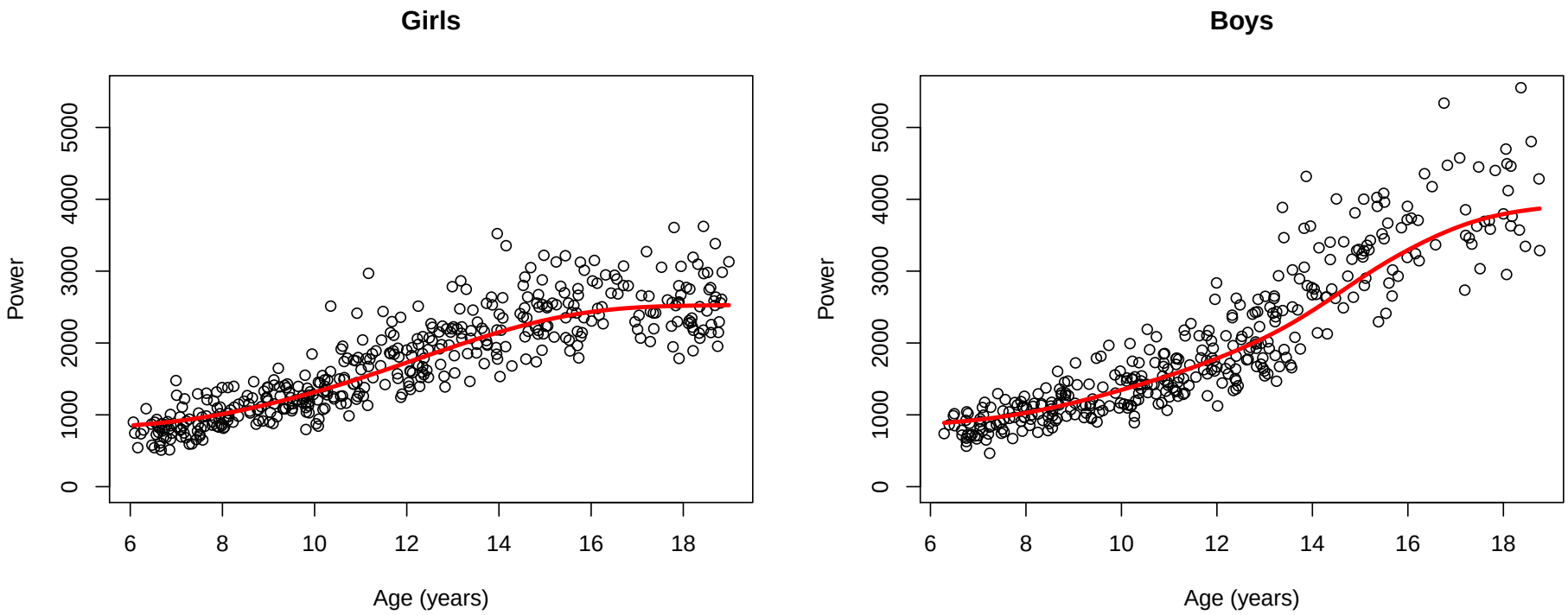
Výška výskoku: lokálně lineární $\widehat{m}_1(x)$





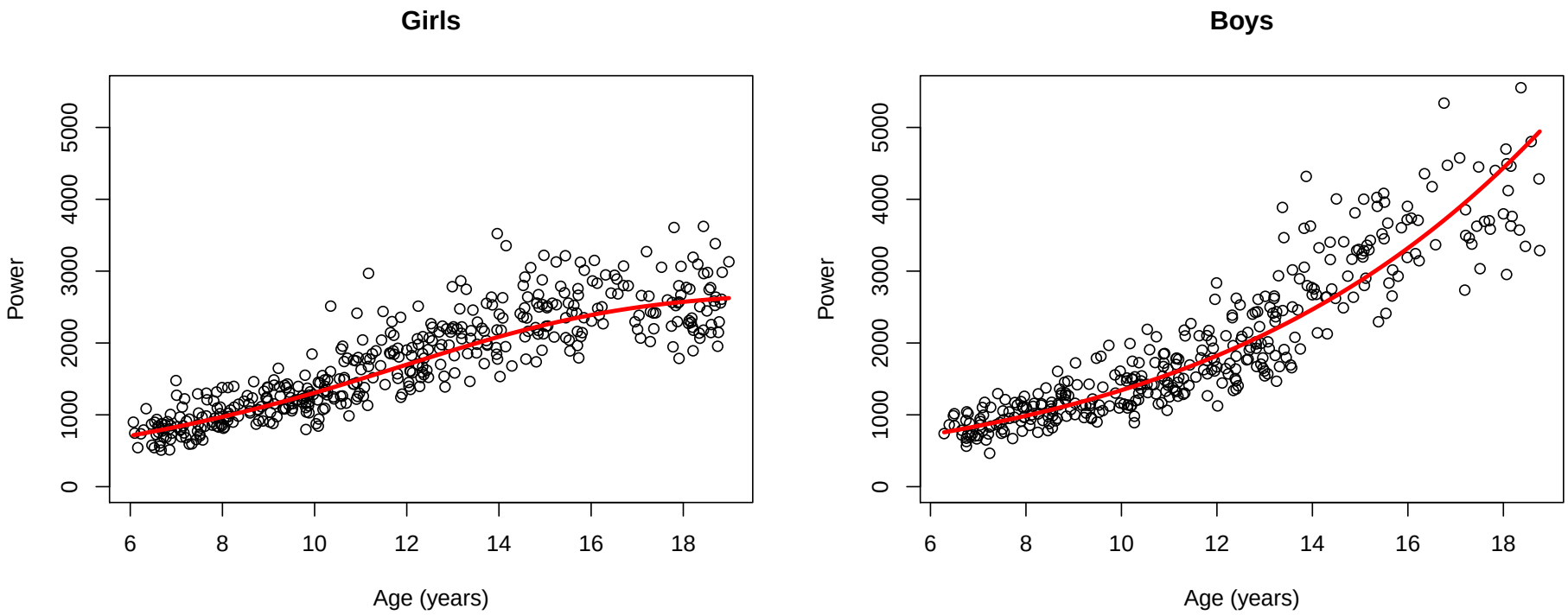
Odhad nevypadá dobře pro nejstarší chlapce (boundary effect).
Je to stejné pro výkon?

Výkon při výskoku: lokálně konstantní $\widehat{m}_0(x)$, AICC



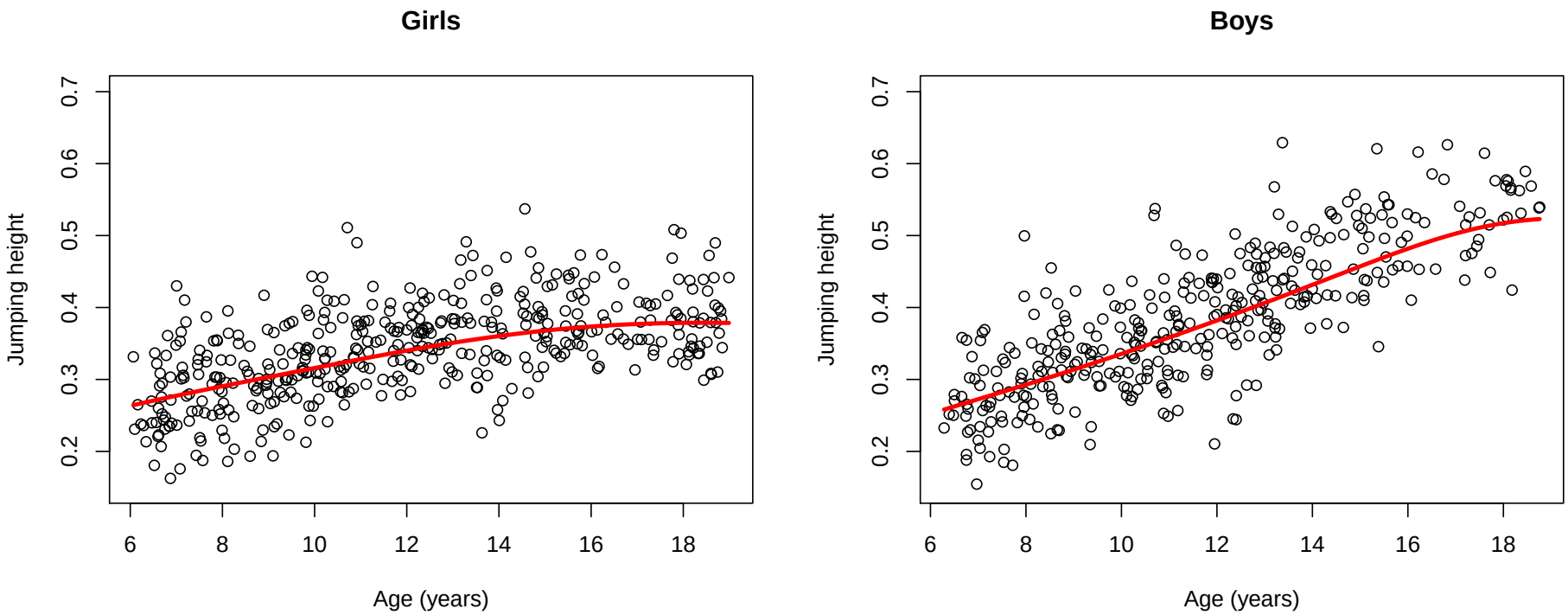
Ani zde to nevypadá dobře pro ty nejmladší.

Výkon při výskoku: lokálně lineární $\widehat{m}_1(x)$



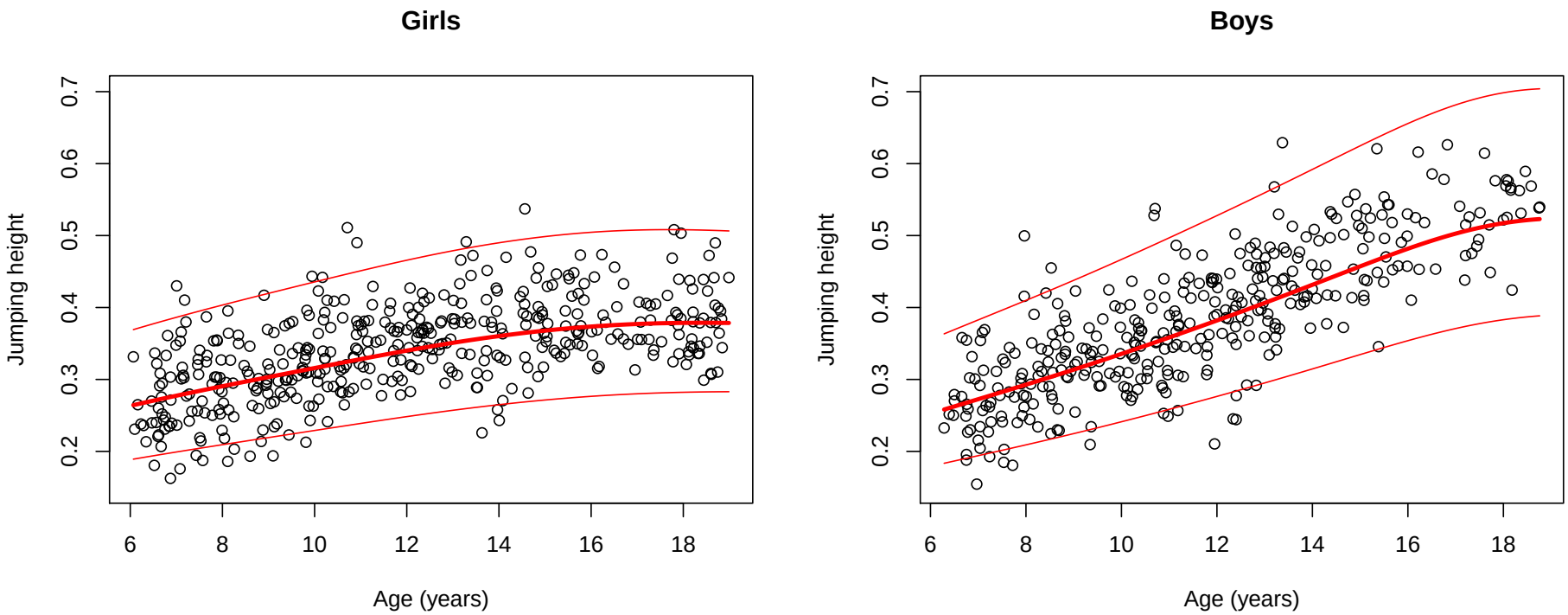
Ani zde odhad nevypadá vůbec dobře pro nejstarší chlapce.
Můžeme ze dvou nepoužitelných odhadů vyrobit jeden použitelný?

Výška výskoku: kombinovaný $(1 - x)\hat{m}_1(x) + x\hat{m}_0(x)$



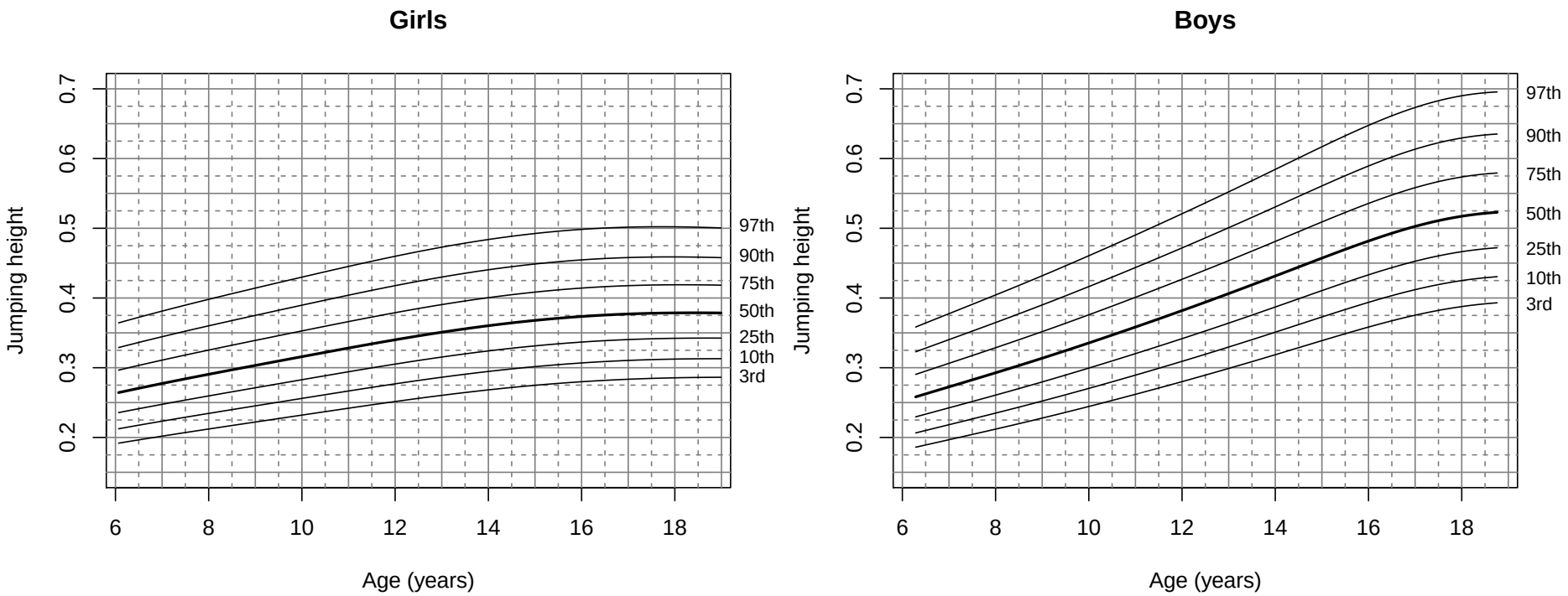
Kombinovaný odhad eliminuje vychýlení lokálně konstantního odhadu pro nejmladší a zároveň vychýlení lokálně lineárního odhadu pro nejstarší děti.

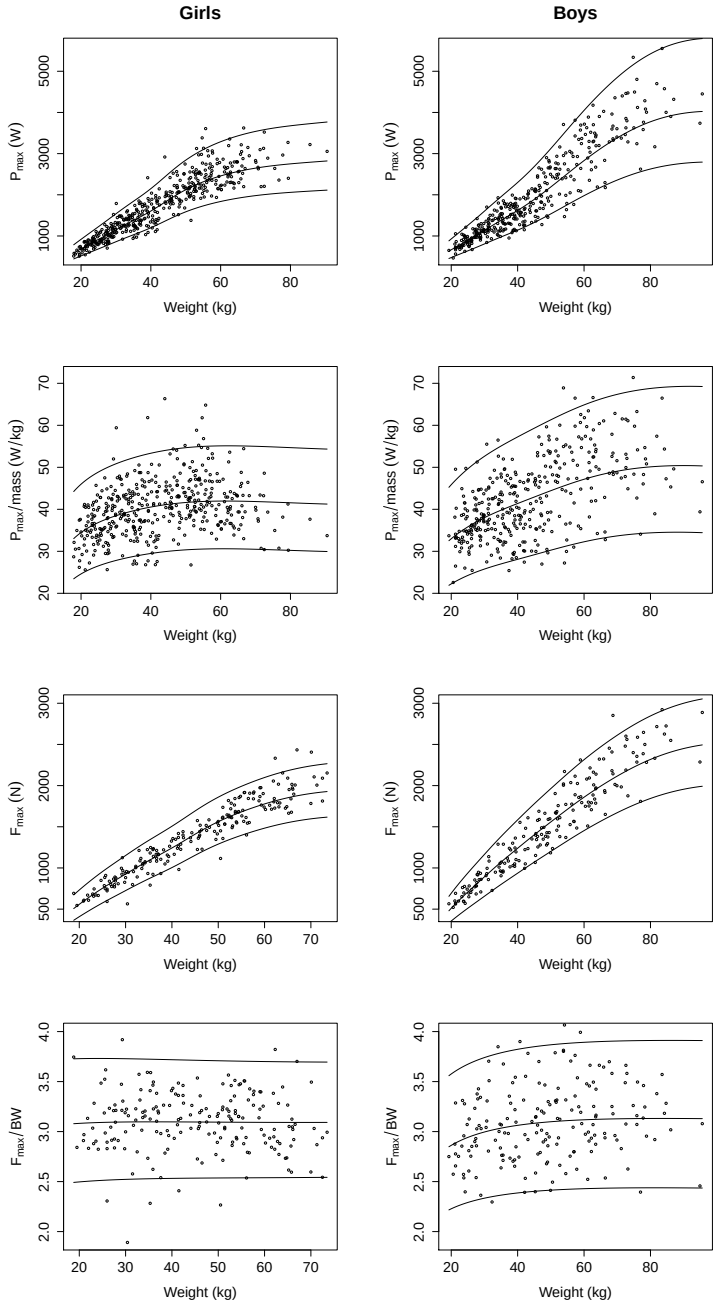
Výška výskoku: s odhadem měřítka

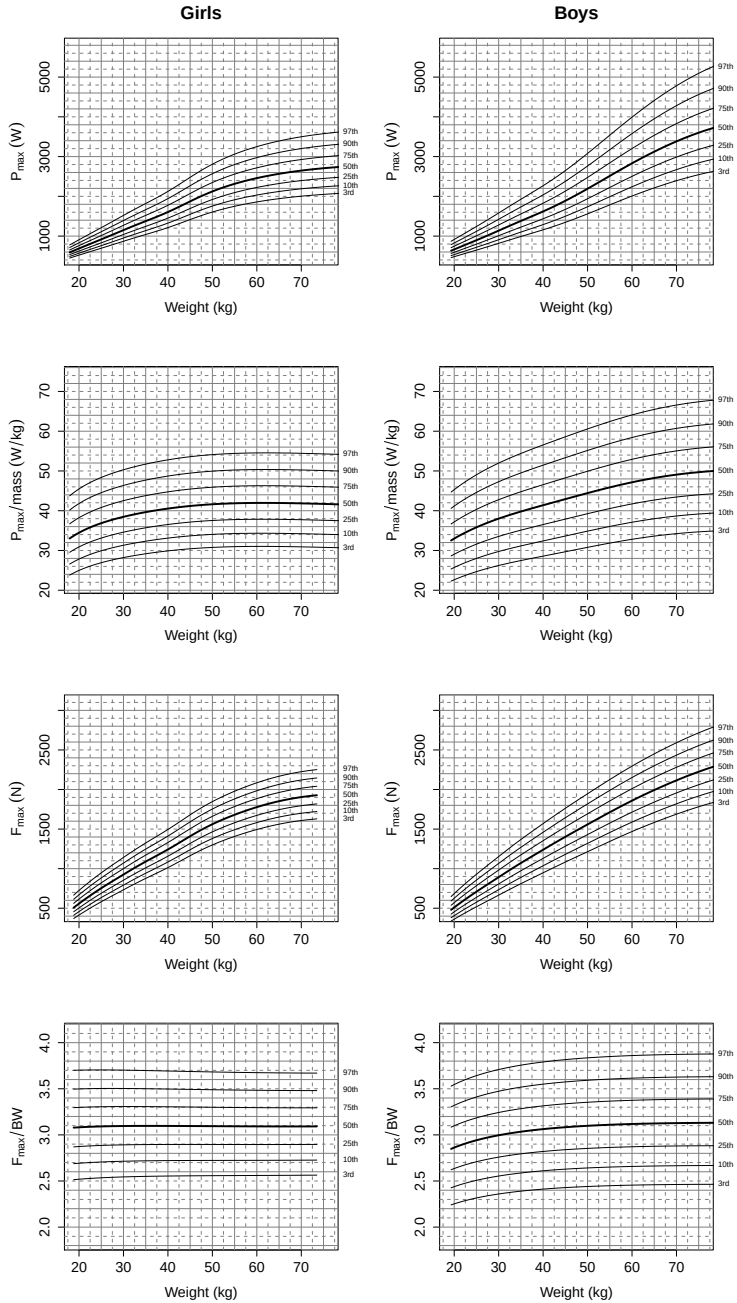


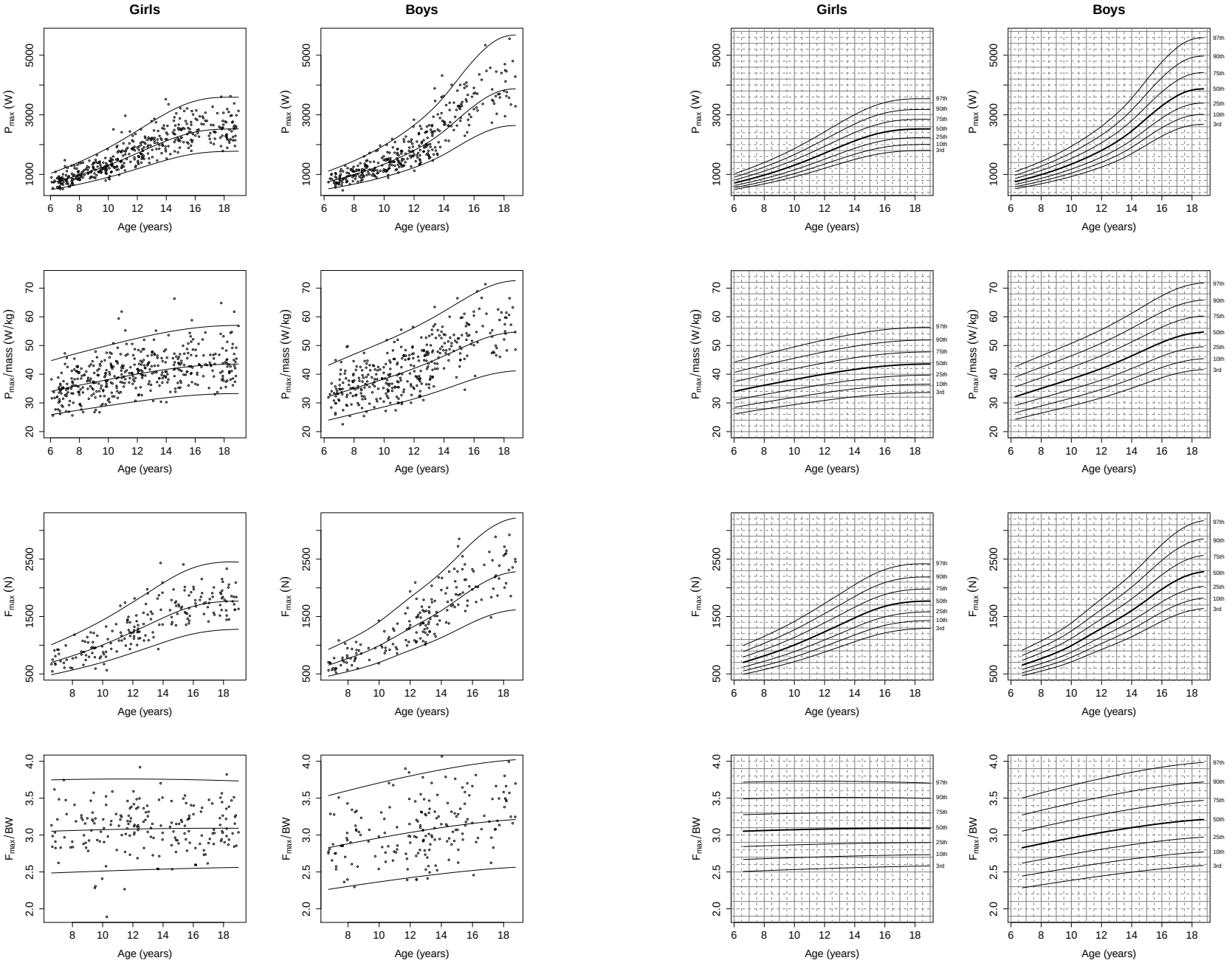
Odhad měřítka spočítáme jako lokálně konstantní jádrový odhad ze čtverců reziduí.

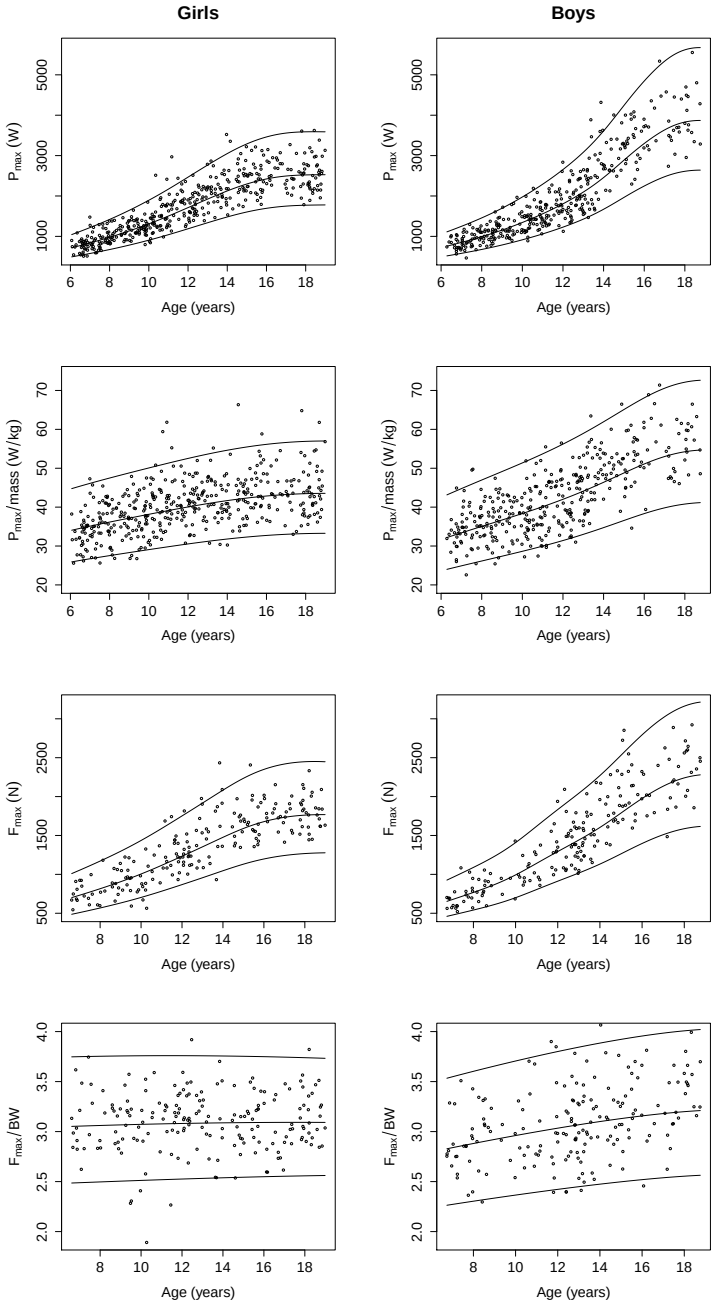
Výška výskoku

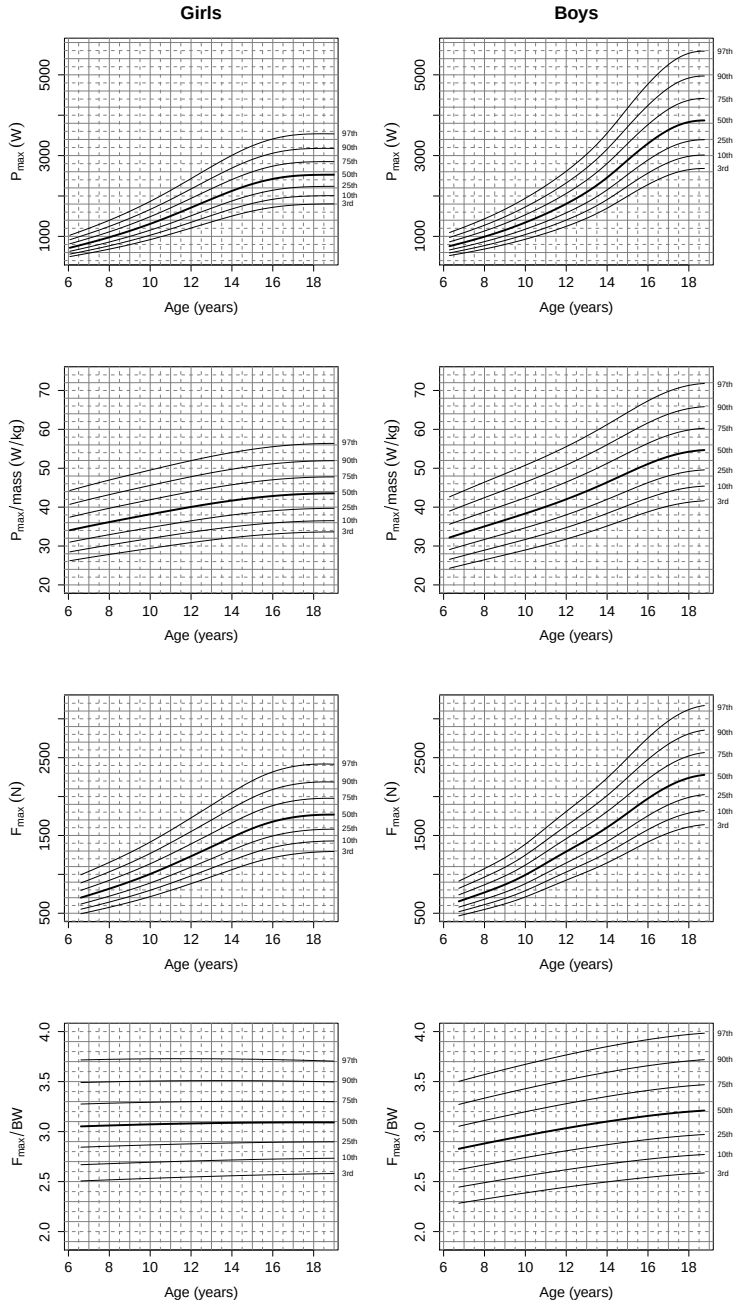



Girls
Boys
P_{max} (W)
P_{max}/mass (W/kg)
F_{max} (N)
F_{max}/BW
Weight (kg)


Girls
Boys
P_{max} (W)
P_{max}/mass (W/kg)
F_{max} (N)
F_{max}/BW
Weight (kg)
97th
90th
75th
50th
25th
10th
3rd


Girls
Boys
Girls
Boys
P_{max} (W)
P_{max}/mass (W/kg)
F_{max} (N)
F_{max}/BW
Age (years)
97th
90th
75th
50th
25th
10th
3rd


Girls
Boys
P_{max} (W)
P_{max}/mass (W/kg)
F_{max} (N)
F_{max}/BW
Age (years)


Girls
Boys
P_{max} (W)
P_{max}/mass (W/kg)
F_{max} (N)
F_{max}/BW
Age (years)
97th
90th
75th
50th
25th
10th
3rd

Porovnání parametrických a neparametrických přístupů:

Kvantilová regrese řeší vše v jednom kroku, ale výsledek opticky nevypadá přijatelně.

Porovnání parametrických a neparametrických přístupů:

Kvantilová regrese řeší vše v jednom kroku, ale výsledek opticky nevypadá přijatelně.

Lineární model není pro podobná data dostatečně flexibilní.

Porovnání parametrických a neparametrických přístupů:

Kvantilová regrese řeší vše v jednom kroku, ale výsledek opticky nevypadá přijatelně.

Lineární model není pro podobná data dostatečně flexibilní.

Pomocí lineární kombinace lokálně konstantního a lokálně lineárního jádrového odhadu (s automatickou volbou bandwidth) lze eliminovat vychýlení pro nejmladší a nejstarší děti.

Hlavní výhoda: tento postup velice dobře funguje i pro váhová Z-skóre a pro další proměnné.

Další (robustní) výzkum: ověřování předpokladů

Data pocházejí z pediatrické kliniky a odhadnuté křivky končí asi v 19 letech.

Další (robustní) výzkum: ověřování předpokladů

Data pocházejí z pediatrické kliniky a odhadnuté křivky končí asi v 19 letech.

Je odhad pro dospělé (od 19 let) skutečně „lokálně konstantní“?

Další (robustní) výzkum: ověřování předpokladů

Data pocházejí z pediatrické kliniky a odhadnuté křivky končí asi v 19 letech.

Je odhad pro dospělé (od 19 let) skutečně „lokálně konstantní“?

Na posteru lze doplnit údaje i pro účastníky Robustu: