

L09

Jak poznáme, že je model dobrý?



Porovnání kandidátních modelů: R^2 , adjustované R^2 a AIC

cuni-natur-biostatistics.github.io/materialy.html

Ilustrace vytvořená pomocí AI.

Co vlastně na stromě měříme?



Ilustrace vytvořená pomocí AI.

U každého jehličnanu známe

- **druh stromu** — tři severoamerické jehličnany;
- **průměr kmene** ve výšce 1,3 m (cm);
- **výšku stromu** (m).

Mají stejně silné stromy různých druhů stejnou výšku?

Co si pamatujete z minulé lekce?

Odpovězte samostatně na tři krátké otázky. Společné odpovědi projdeme po hlasování.

Jaký je podle modelu s interakcí rozdíl zadní šířky samec – samice při délce krunýře 30 mm?

Koeficienty modelu s interakcí: modří krabi

zadni_sirka_mm ~ delka_krunyre_mm * pohlavi · referenční skupina: samice

Koeficient	Odhad (mm, mm/mm)
(Intercept)	0,7077
delka_krunyre_mm	0,4067
pohlavisamec	1,9731
delka_krunyre_mm:pohlavisamec	-0,1245

+1,97 mm

-0,12 mm

+5,71 mm

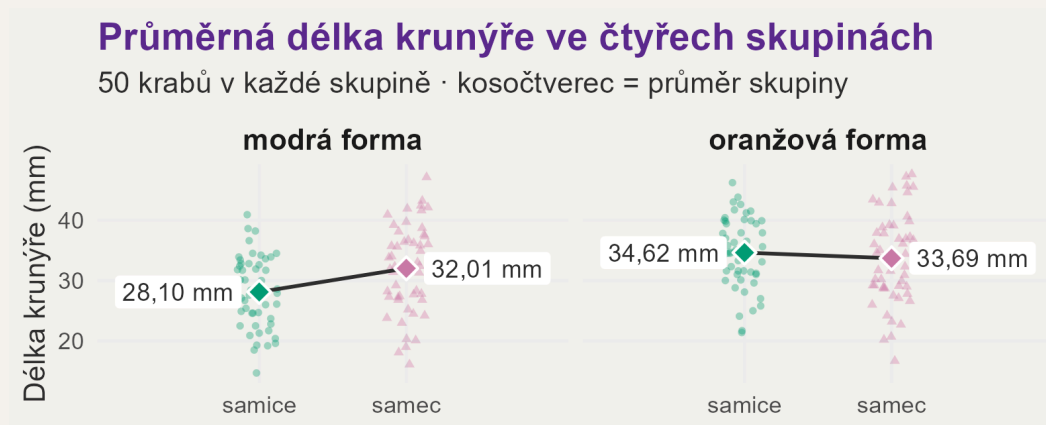
-1,76 mm

SPRÁVNÁ ODPOVĚĎ

: -1,76 mm

Koeficient pohlavisamec platí jen při délce 0 mm. Při 30 mm k němu přičteme $30 \times$ interakční koeficient: $1,973 + 30 \times (-0,1245) = -1,76$ mm. Samec má tedy při 30 mm užší zadní část krunýře.

Jakou hodnotu má v modelu forma * pohlaví koeficient formaoranžová:pohlavísamec?



-4,84 mm

-0,93 mm

+3,91 mm

+6,52 mm

SPRÁVNÁ ODPOVĚĎ

: -4,84 mm

Interakce je rozdíl dvou rozdílů: $(33,69 - 34,62) - (32,01 - 28,10) = -0,93 - 3,91 = -4,84$ mm. O tolik je rozdíl samec – samice u oranžové formy menší než u modré.

Co platí o přírůstku objemu na 1 cm průměru u stromu vysokého 35 m?

Koeficienty modelu objemu dřeva

$\text{objem_m3} \sim \text{prumer_cm} * \text{vyska_m} \cdot 31$ stromů
změřené průměry 21,1–52,3 cm, výšky 19,2–26,5 m

Koeficient	Odhad (m^3 , m^3/cm , m^3/m)
(Intercept)	1,96508
prumer_cm	-0,06528
vyska_m	-0,12050
prumer_cm:vyska_m	0,00493

Je $+0,107 \text{ m}^3$ na 1 cm; vzorec platí pro každou výšku.

Je $-0,065 \text{ m}^3$ na 1 cm; to je koeficient prumer_cm.

Vyjde $+0,107 \text{ m}^3$ na 1 cm; 35 m je ale mimo data.

Nelze ho spočítat; u dvou číselných prediktorů sklon nemá smysl.

SPRÁVNÁ ODPOVĚĎ

: Vyjde $+0,107 \text{ m}^3$ na 1 cm; 35 m je ale mimo data.

Vzorec $-0,065 + 35 \times 0,00493$ dá $+0,107 \text{ m}^3$ na 1 cm, ale změřené stromy byly vysoké jen 19–27 m. Mimo tento rozsah data tvar vztahu nepodporují, takže číslo je extrapolace, odhad mimo pozorované výšky. Uvnitř dat je sklon např. $+0,038$ (21 m) a $+0,058$ (25 m) m^3 na 1 cm.

Výsledky učení

Po absolvování této přednášky budete schopni:

- vypočítat R^2 z celkového součtu čtverců a součtu čtverců residuí a vysvětlit, co popisuje;
- vysvětlit, proč nejvyšší R^2 automaticky neznamena nejlepší model;
- rozlišit, jak R^2 , $\text{adjustované } R^2$ a AIC zacházejí se shodou a složitostí, a používat AIC jen k relativnímu porovnání;
- sestavit malou sadu biologicky zdůvodněných kandidátních modelů se stejnou odezvou a stejnými pozorováními;
- obhájit volbu modelu pomocí biologické otázky, složitosti, diagnostiky a interpretovatelnosti a pojmenovat její omezení.

Jedna přímka pro všechny
stromy?

Jeden řádek tabulky představuje jeden strom

```
1 data_stromy <-  
2   read.csv(file = "data/allometrie_stromu.csv")  
3  
4 data_stromy$druh_stromu <-  
5   factor(x = data_stromy$druh_stromu)  
6  
7 vec_promenne <-  
8   c("druh_stromu", "prumer_kmene_cm", "vyska_stromu_m")  
9  
10 head(  
11   x = data_stromy[, vec_promenne],  
12   n = 4  
13 )
```

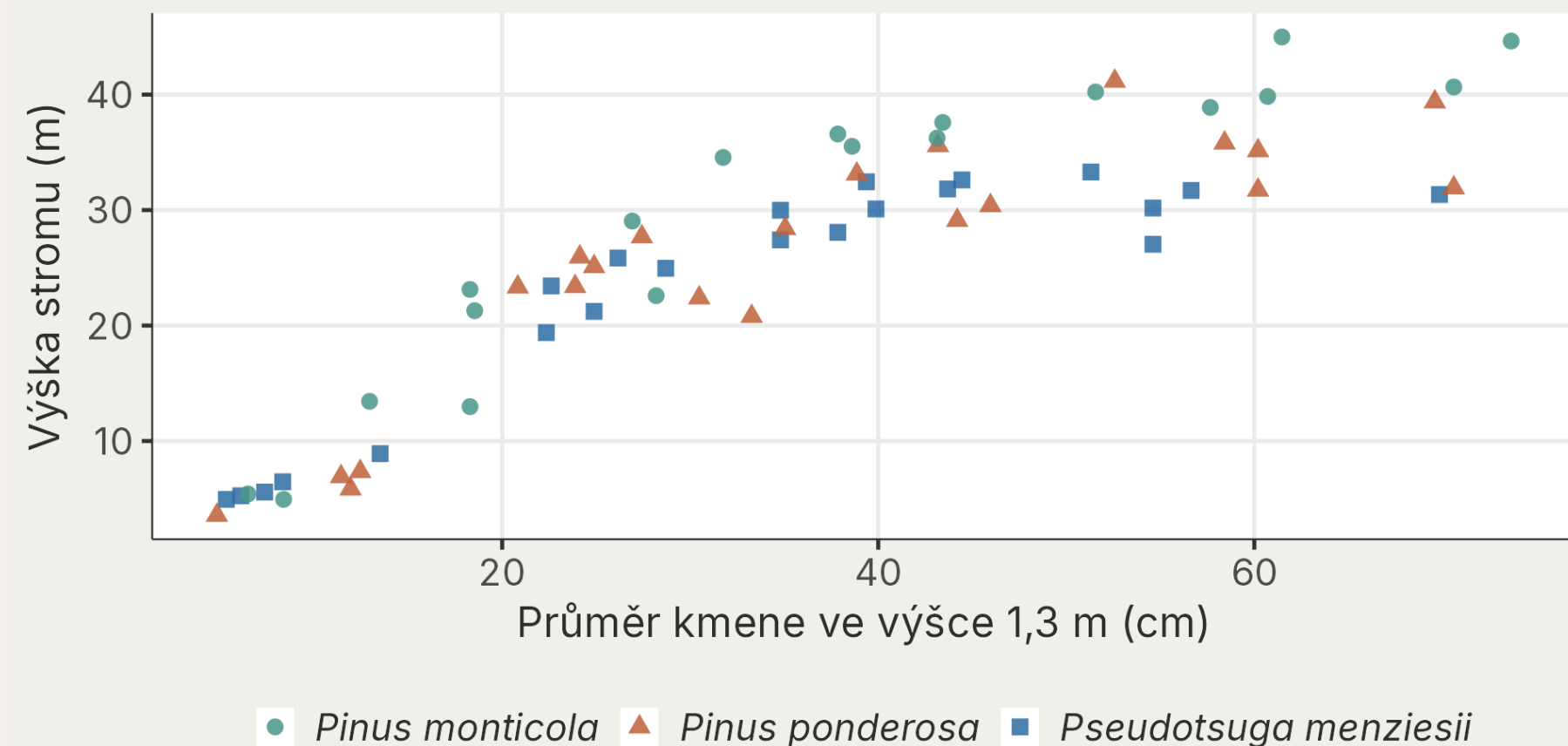
	druh_stromu	prumer_kmene_cm	vyska_stromu_m
1	Pseudotsuga menziesii	54.61	27.04
2	Pseudotsuga menziesii	34.80	27.42
3	Pseudotsuga menziesii	24.89	21.23
4	Pseudotsuga menziesii	28.70	24.96

63 stromů

- *Pinus monticola*: 19 stromů
- *Pinus ponderosa*: 22 stromů
- *Pseudotsuga menziesii*: 22 stromů
- chybějící hodnoty: 0
- průměr kmene: 4,8–73,7 cm
- výška: 3,6–45,0 m

Jiné stromy než minule: tři druhy jehličnanů a **výška je tentokrát odezva.**

Jak se výška mění s průměrem kmene?

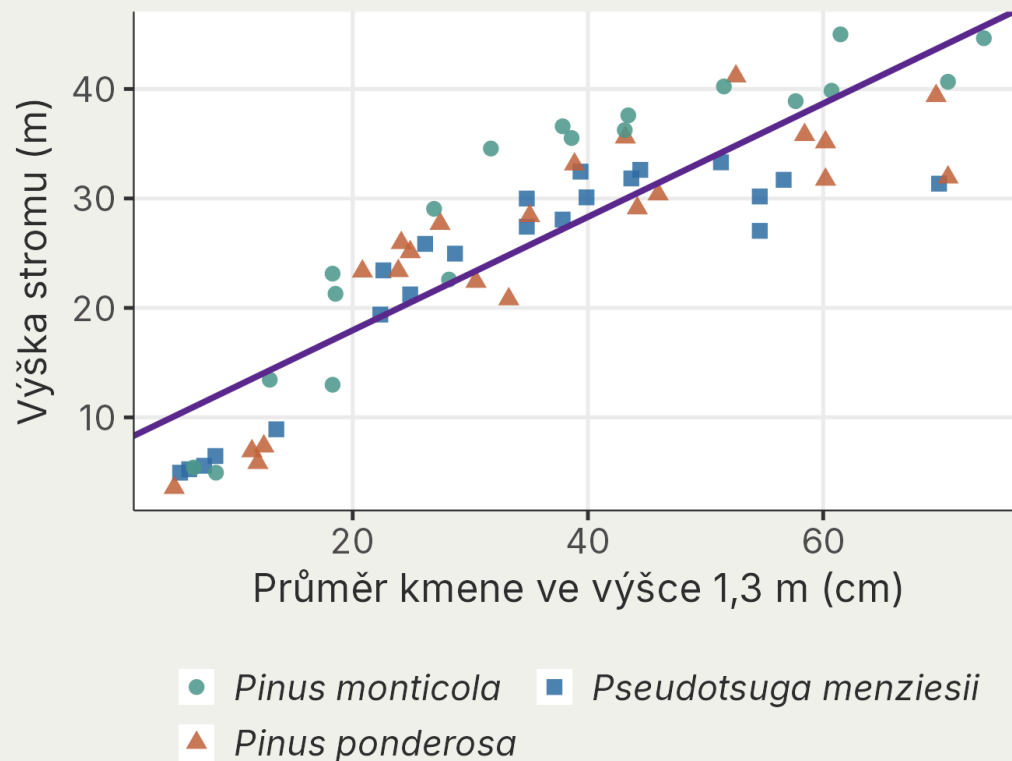


Ve dvojici: Jaký vzorec v datech převládá? Co by mohlo vadit jedné přímce pro všechny stromy?

Jedna společná přímka pro všechny druhy

```
1 mod_spolecny <-  
2   lm(  
3     formula = vyska_stromu_m ~  
4       prumer_kmene_cm,  
5     data = data_stromy  
6   )  
7  
8 coef(object = mod_spolecny)
```

```
(Intercept) prumer_kmene_cm  
7.5967386    0.5179077
```



Sklon: o **0,52 m vyšší** odhadovaná výška na každý centimetr průměru —
stejně pro všechny druhy.

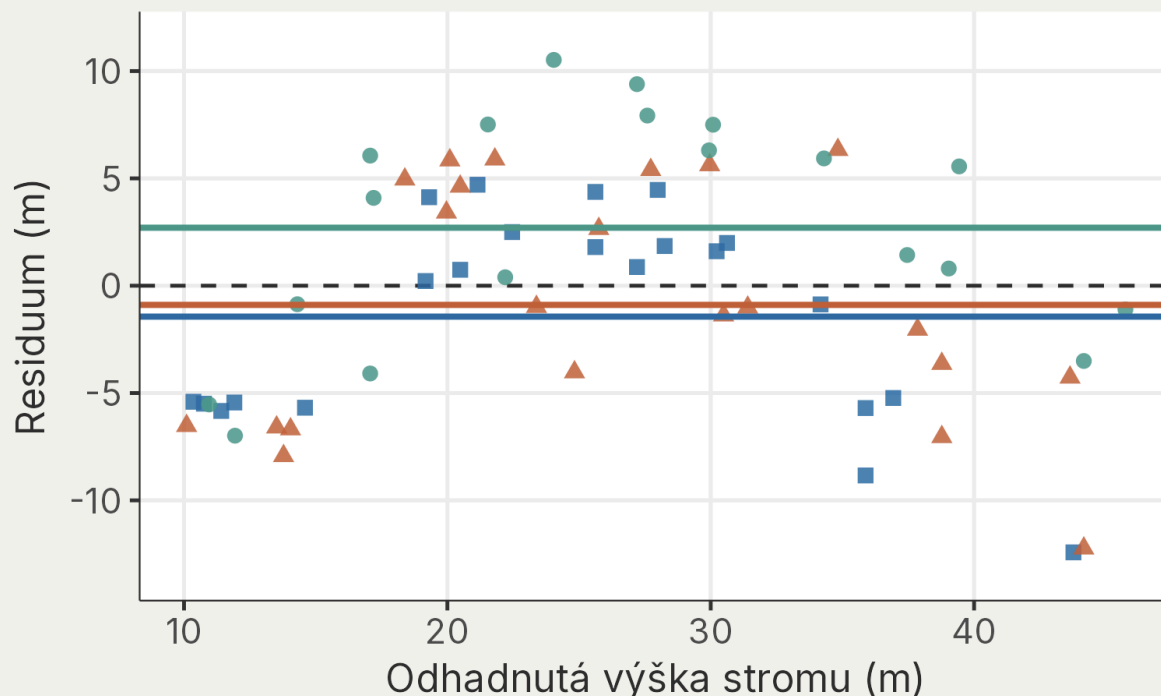
Co už umíme přechíst ze `summary()` ?

```
1 summary(object = mod_spolecny)

2 Call:
3 lm(formula = vyska_stromu_m ~ prumer_kmene_cm, data = data_stromy)
4
5 Residuals:
6      Min       1Q   Median       3Q      Max
7 -12.4226  -5.3191   0.7425   4.6637  10.5197
8
9 Coefficients:
10              Estimate Std. Error t value Pr(>|t|)
11 (Intercept)    7.59674    1.47310   5.157 2.88e-06 ***
12 prumer_kmene_cm 0.51791    0.03647  14.201 < 2e-16 ***
13 ---
14 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
15
16 Residual standard error: 5.548 on 61 degrees of freedom
17 Multiple R-squared:  0.7678,    Adjusted R-squared:  0.764
18 F-statistic: 201.7 on 1 and 61 DF,  p-value: < 2.2e-16
```

Známe: volání modelu · residua · koeficienty s nejistotou

Zůstává v residuích rozdíl mezi druhy?



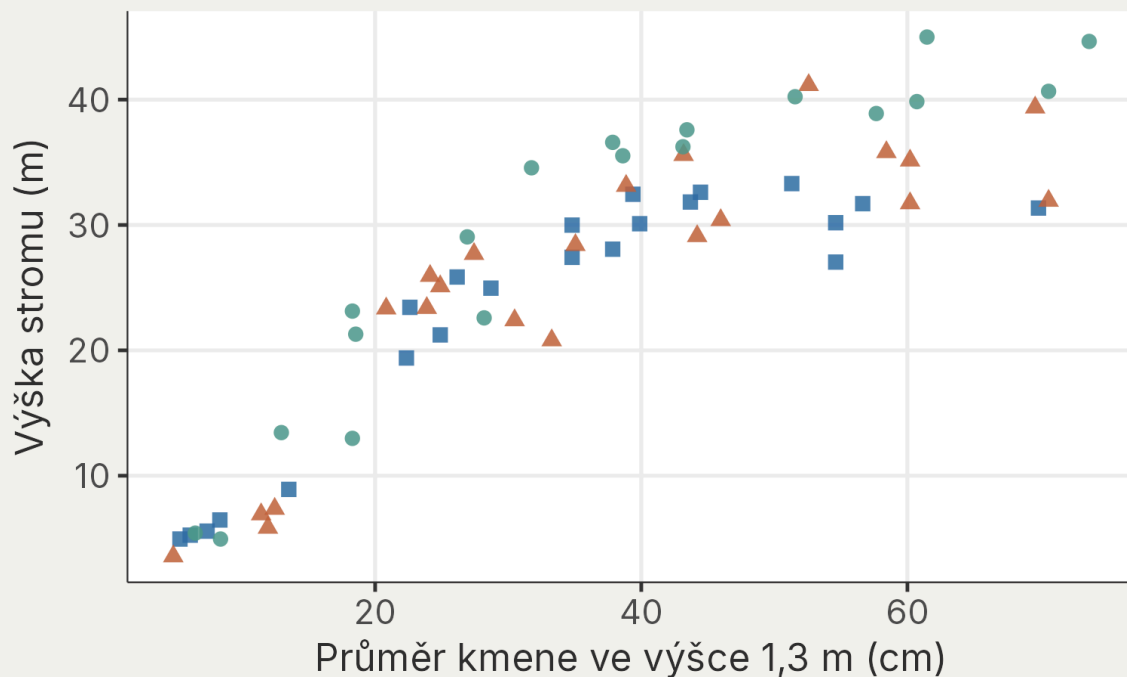
Leží residua některého druhu převážně nad nulou, nebo pod ní?

Průměrné residuum *P. monticola*: **+2,7 m**

Společná přímka přehlíží rozdíly mezi druhy.

Tři představy o růstu stromů

Jaké jiné přímky by mohly stromy popsat?



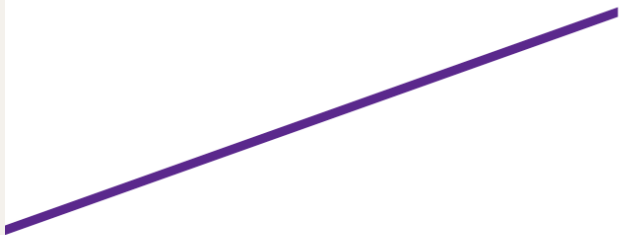
● *Pinus monticola* ■ *Pseudotsuga menziesii*
▲ *Pinus ponderosa*

Ve dvojici: Načrtněte dvě jiné možnosti než jednu společnou přímku. Co by každá z nich tvrdila o stromech?

Pomůcka: modely z minulé lekce — $+$ a $*$.

Tři modely vyjadřují tři biologické představy

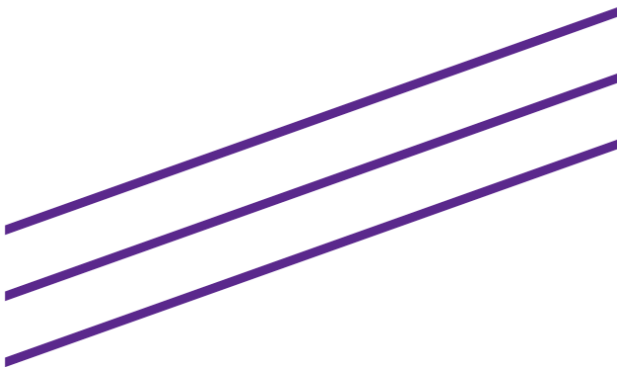
Společná přímka



$\text{vyska} \sim \text{prumer}$

Všechny druhy sdílejí
**stejný intercept i
sklon**.

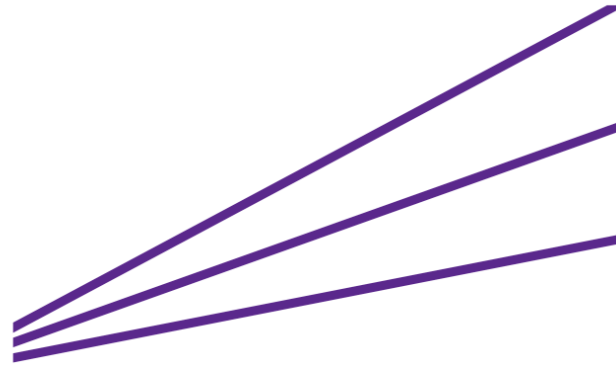
Posun mezi druhy



$\text{vyska} \sim \text{prumer} +$
druh

Při stejném průměru
se druhy mohou lišit
výškou, sklon sdílejí.

Různé sklony



$\text{vyska} \sim \text{prumer} *$
druh

Vztah průměru a výšky
se může **mezi druhy
lišit**.

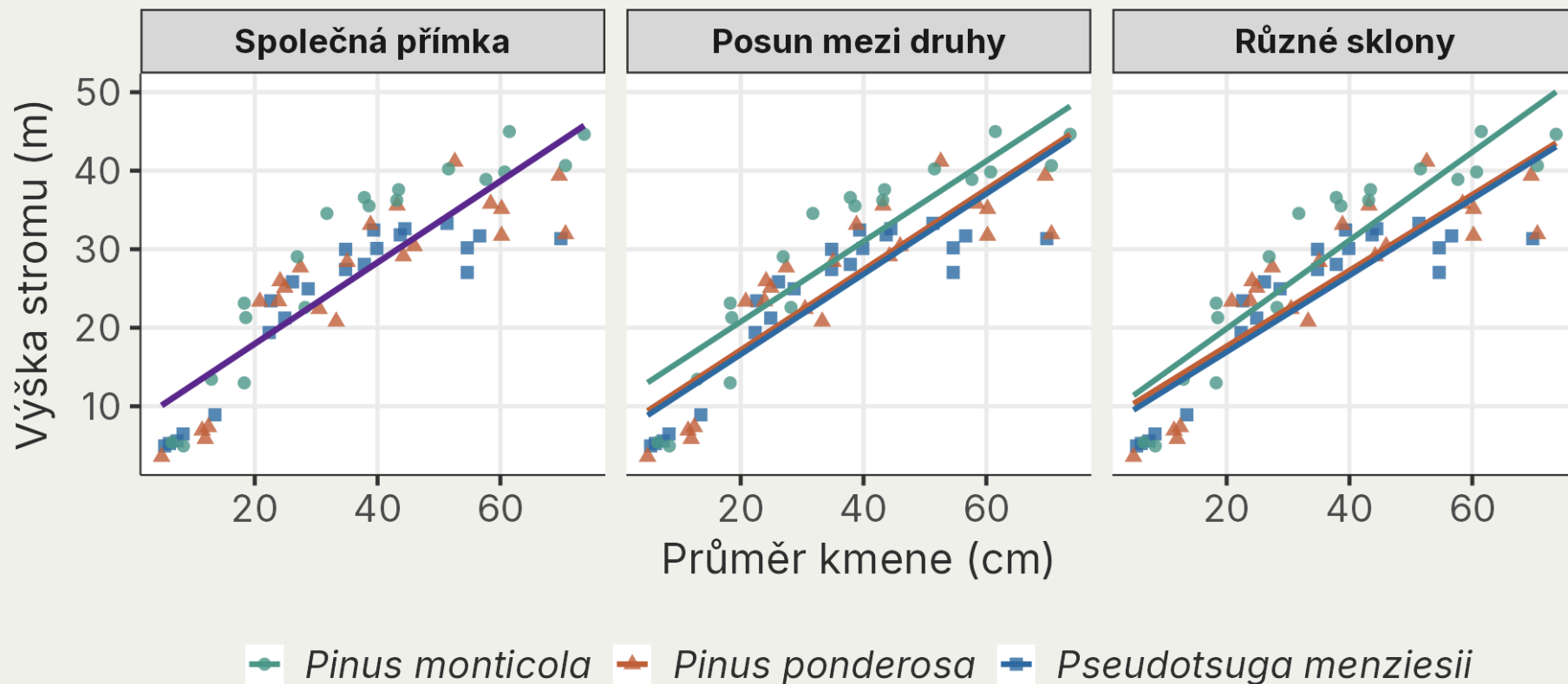
Tuto sadu kandidátů určíme **předem** — z biologické otázky, ne z výsledků.

Všechny tři modely odhadneme ze stejných 63 stromů

```
1 # Všechny tři kandidáty zapíšeme vedle sebe
2 mod_spolecny <-
3   lm(
4     formula = vyska_stromu_m ~ prumer_kmene_cm,
5     data = data_stromy
6   )
7
8 mod_aditivni <-
9   lm(
10    formula = vyska_stromu_m ~ prumer_kmene_cm + druh_stromu,
11    data = data_stromy
12  )
13
14 mod_interakce <-
15   lm(
16     formula = vyska_stromu_m ~ prumer_kmene_cm * druh_stromu,
17     data = data_stromy
18   )
```

Stejná odezva `vyska_stromu_m` · stejných **63** stromů v každém modelu

Který model prochází nejbližše pozorovaným stromům?

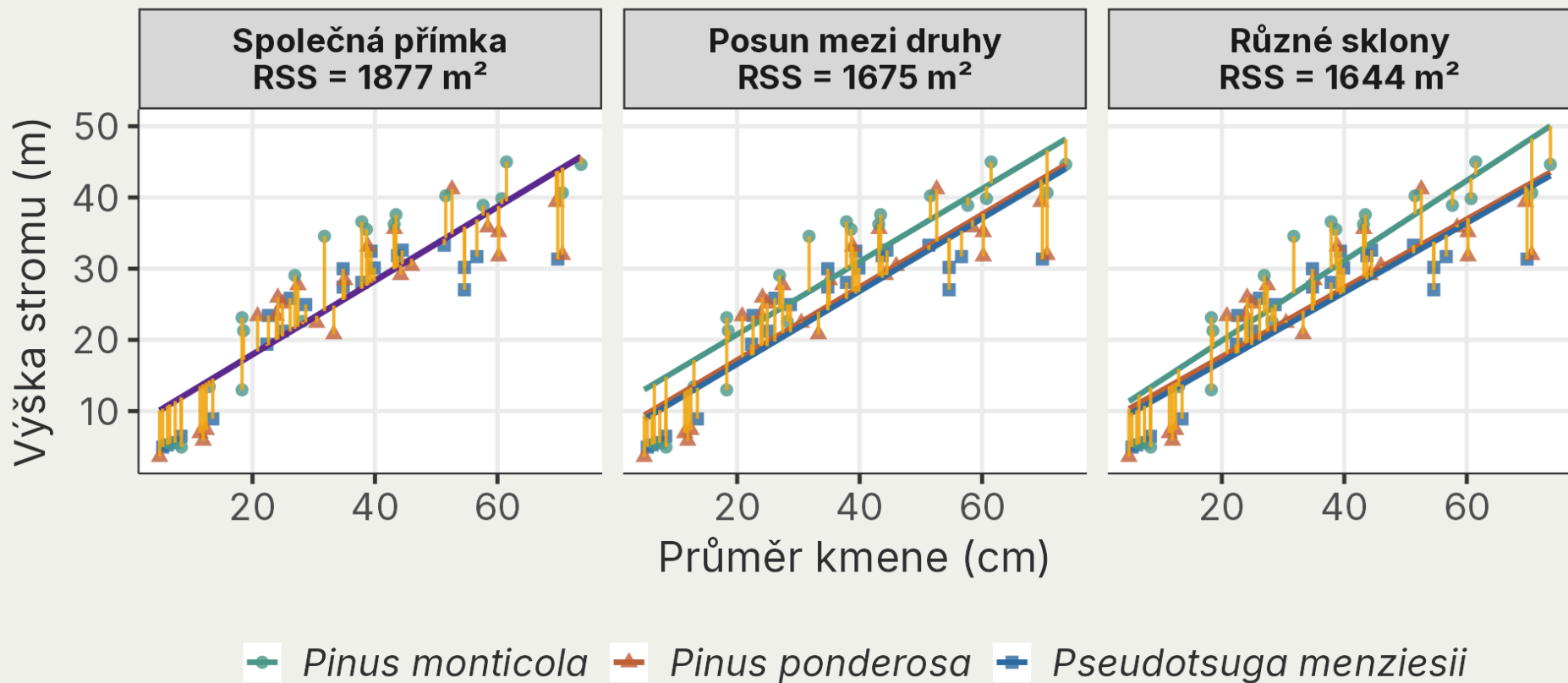


A · Společná přímka

B · Posun mezi druhy

C · Různé sklony

Nejblíže procházejí přímky s různými sklony



Je proto nejlepší?

Podle čeho poznáme lepší model?

Podle čeho byste mezi modely vybírali?

- jak těsně model odpovídá pozorovaným datům;
- kolik koeficientů potřebuje;
- zda v residuích nezůstává systematický vzorec;
- zda odpovídá biologické otázce;
- zda jeho výsledek dokážeme smysluplně interpretovat.

Žádné jediné číslo neodpoví na všechny tyto otázky.

Jak změřit shodu modelu s daty?

Co už víme o součtu čtverců residuí?

HLEDÁNÍ PŘÍMKY

Kandidátní přímky jsme porovnávali **součtem čtverců residuí**.

ZKRATKA RSS

Stejnou veličinu jsme označili **RSS** (*residual sum of squares*).

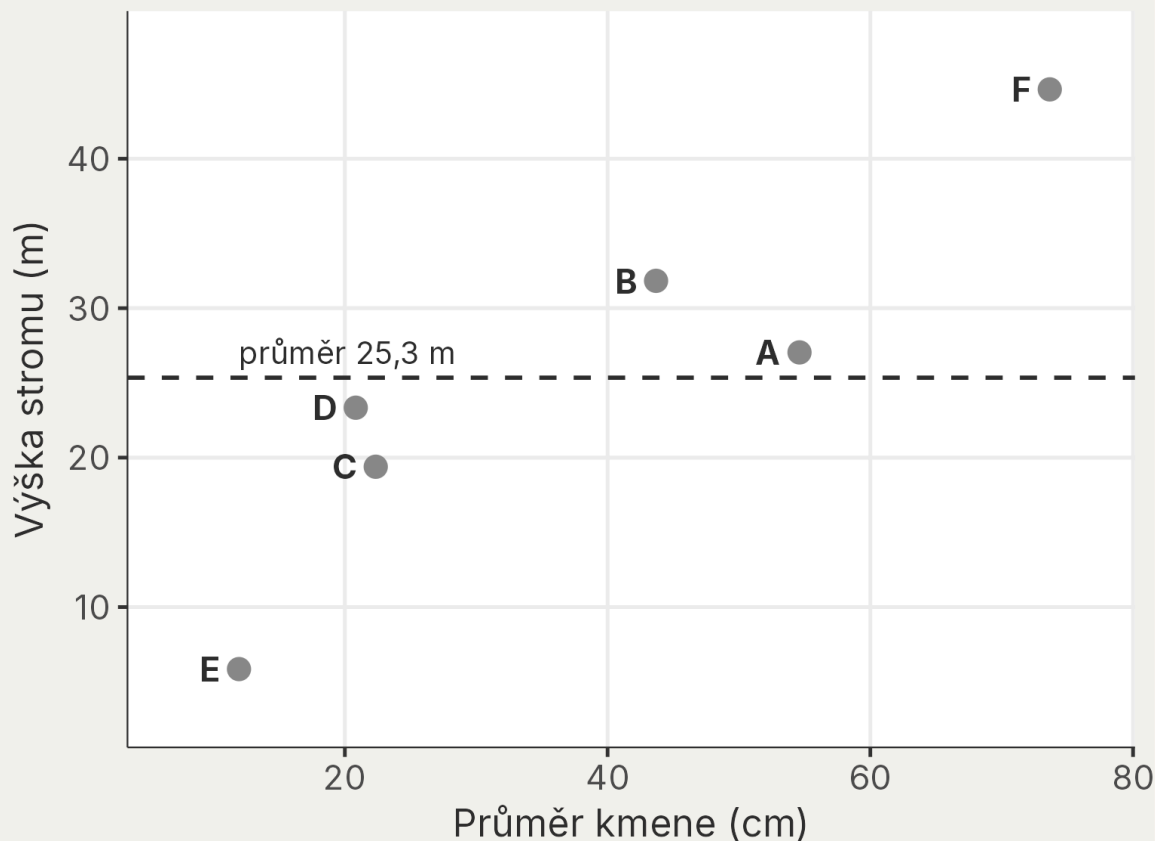
SKUPINOVÉ PRŮMĚRY

Skupinové průměry RSS zmenšily oproti jednomu společnému průměru.

Co se stane s RSS, když přímka prochází blíže pozorovaným bodům?

Kratší residua → menší čtverce → **menší RSS** = těsnější shoda s daty.

Je přímka lepší než předpověď průměrem?

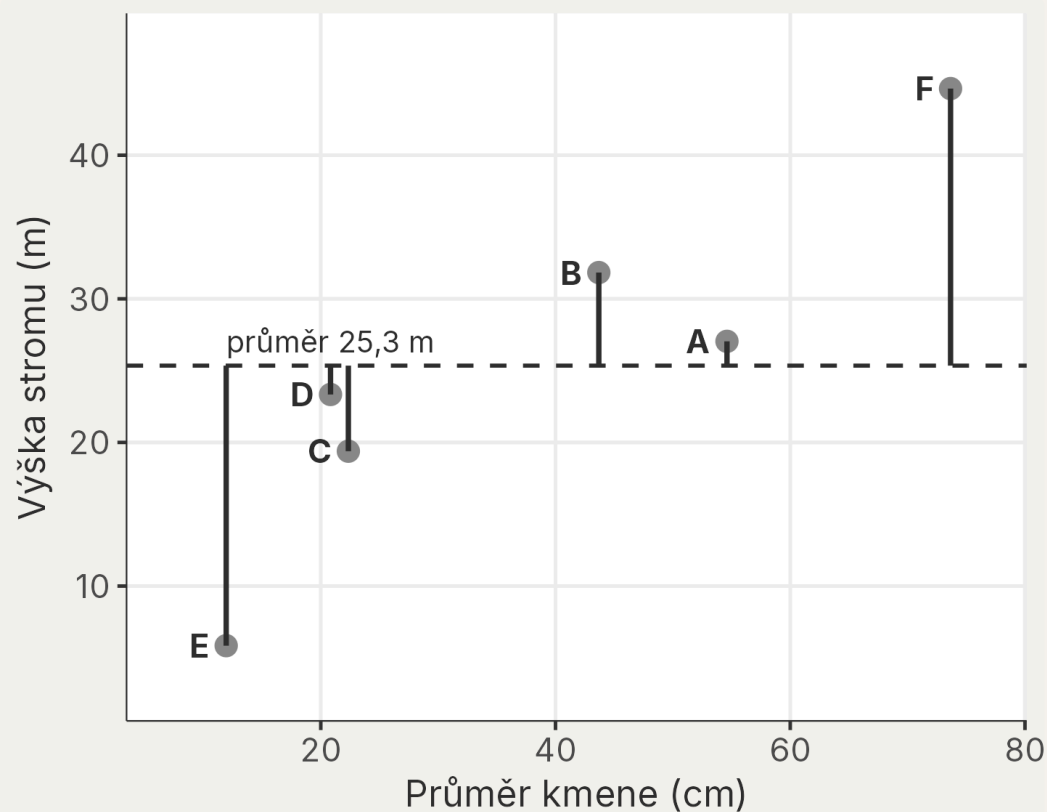


Šest skutečných stromů z tabulky.

Předpověď bez prediktoru: každému stromu **průměrnou výšku**.

Prošla by přímka podle průměru kmene body blíže než průměr?

Jak velké jsou rozdíly výšek kolem průměru?

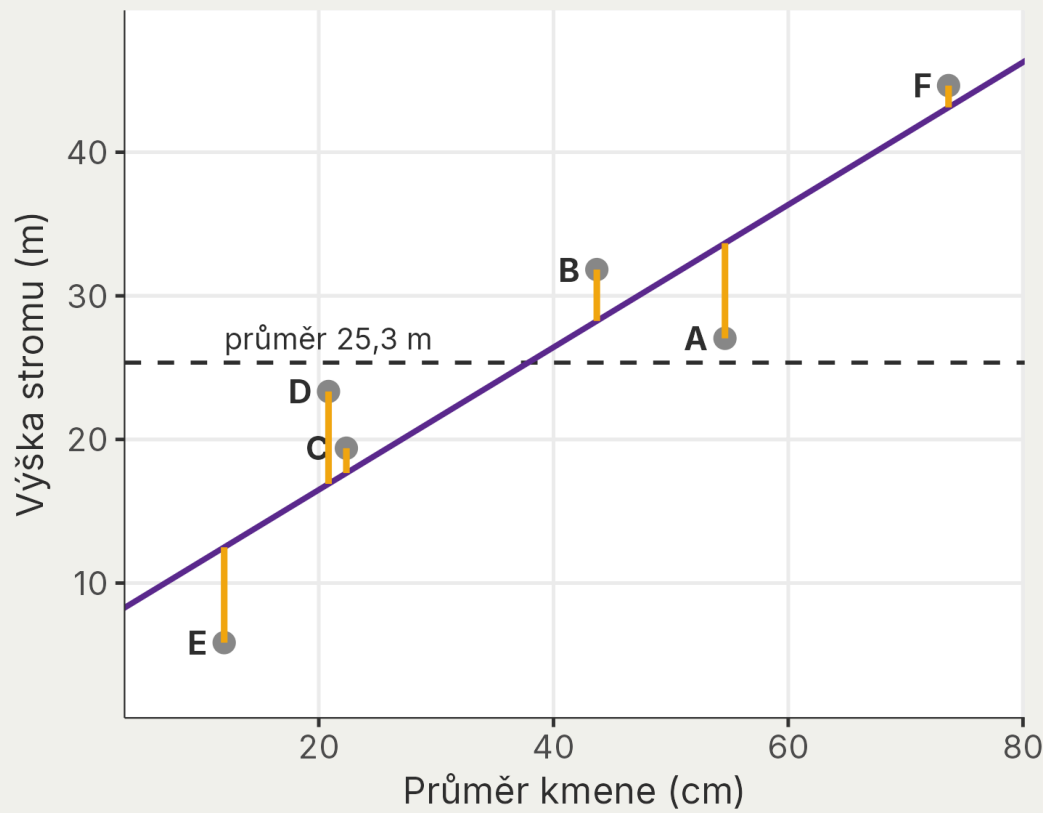


Strom	Odchylka	Čtverec
A	1,7	2,9
B	6,5	42,0
C	-6,0	35,5
D	-2,0	4,0
E	-19,5	380,2
F	19,3	372,2

Součet čtverců: **836,8 m²**

Tomuto součtu říkáme **celkový součet čtverců** (*total sum of squares*).

Kolik rozdílů zůstane po odhadnutí přímky?



Strom	Residuum	Čtverec
A	-6,6	43,9
B	3,6	12,8
C	1,7	3,0
D	6,4	41,4
E	-6,6	44,1
F	1,5	2,3

RSS: 147,6 m² · celkový: 836,8 m²

Který strom leží nejvýše nad přímkou? A který nejvíce přispívá k RSS?
Strom D; k RSS nejvíce E (rozhoduje velikost, ne znaménko).

Jakou část rozdílů přímka zachytila?

$$\text{zachycený podíl} = 1 - \frac{\text{součet čtverců residuí}}{\text{celkový součet čtverců}}$$

Jaká část rozdílů ve výšce **zůstala** a jakou část přímka **zachytila**? Dosad'te RSS 147,6 m² a celkový součet čtverců 836,8 m².

$$\text{zachycený podíl} = 1 - \frac{147,6}{836,8}$$

$$\text{zachycený podíl} = 0,824$$

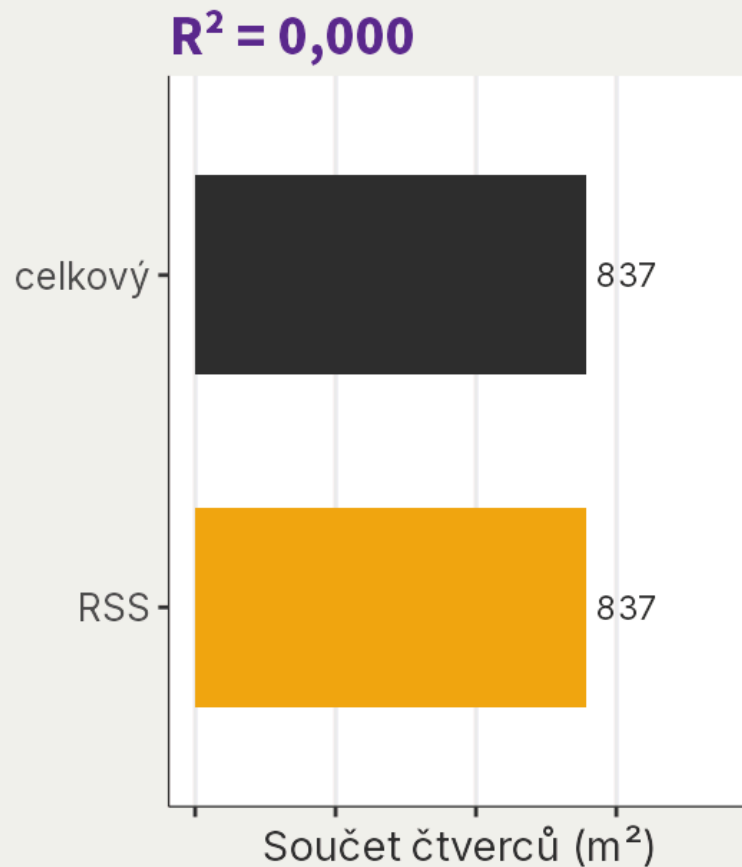
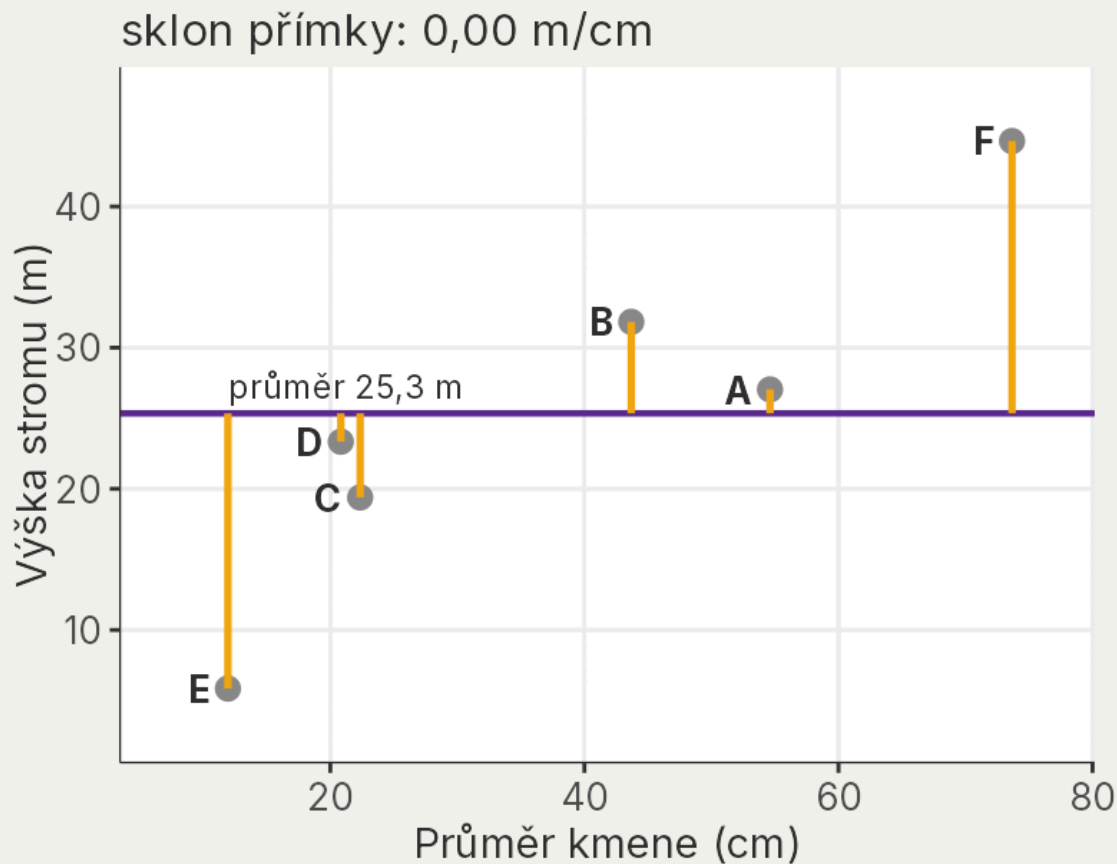
Zachycený podíl se jmenuje R^2

KOEFICIENT DETERMINACE R^2 (*COEFFICIENT OF DETERMINATION*)

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{total}}}$$

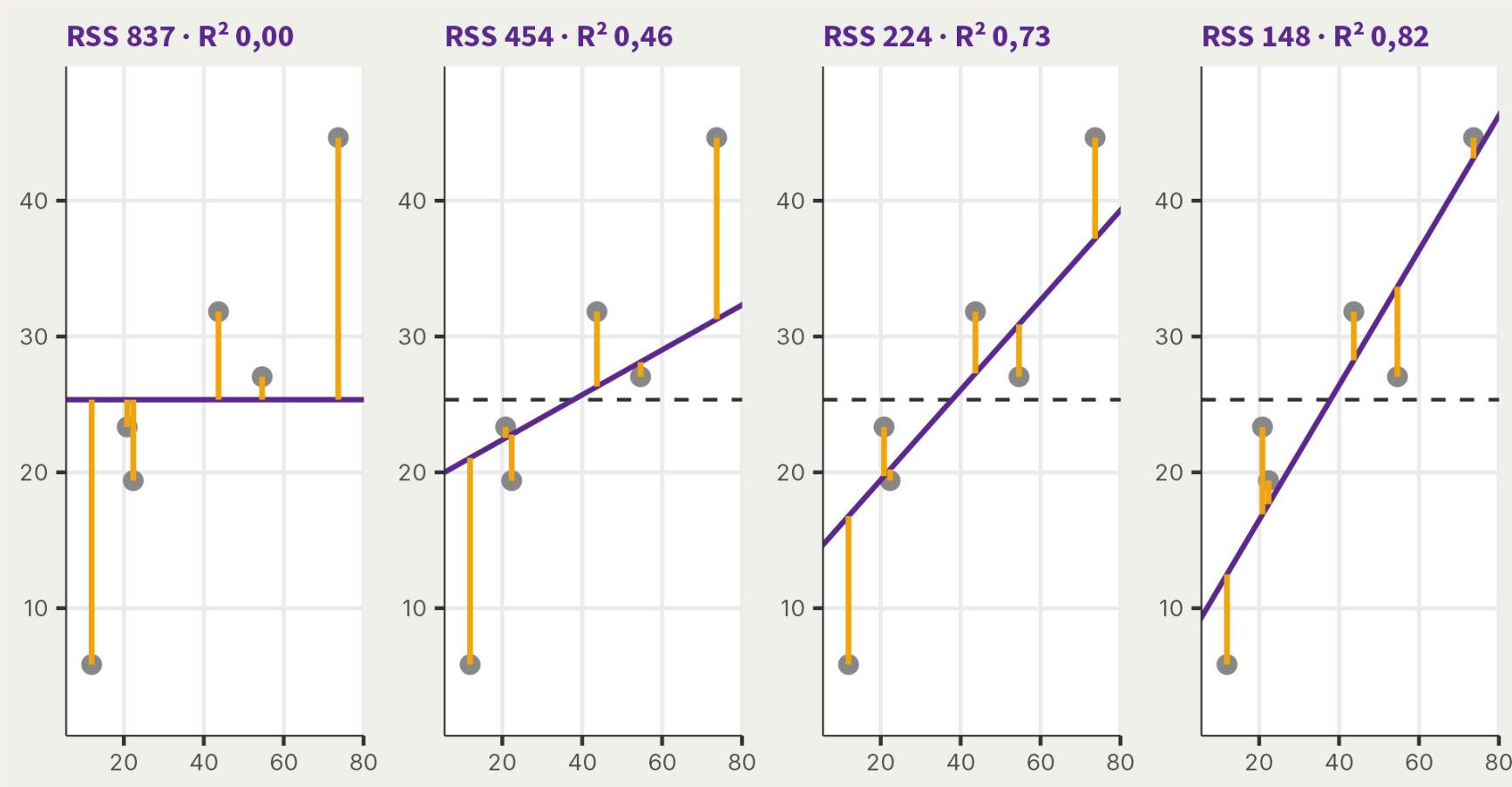
- SS_{res} — součet čtverců residuí (RSS) od odhadnuté přímky;
- SS_{total} — celkový součet čtverců odchylek od průměru;
- u šesti stromů: $R^2 = 0,824$.

Jak se mění RSS a R^2 , když přímku natáčíme?



Sledujte oranžové úsečky a sloupec RSS: kdy je R^2 nulové a kdy největší?

Stejných šest stromů, čtyři natočení přímky



Vodorovná přímka = předpověď **průměrem**: $R^2 = 0$. Odhadnutá přímka: **nejmenší RSS**, a proto **největší R^2** .

Dává `summary()` stejné R^2 pro všech 63 stromů?

$$R^2 = 1 - \frac{1877,4}{8084,6} = 0,768$$

```
1 summary(object = mod_spolecny)$r.squared
```

```
[1] 0.7677804
```

Multiple R-squared ve výpisu je **tentýž výpočet** — nic nového.

Společná přímka má $R^2 = 0,768$. Co to znamená?

A

Model správně předpověděl 76,8 % stromů.

B

Model zachytil 76,8 % variability výšky kolem průměru v těchto datech.

C

S pravděpodobností 0,768 je model pravdivý.

Přesvědčte souseda, proč jste vybrali svoji možnost.

Obhajitelná je **B**: R^2 popisuje podíl zachycené variability v datech, ze kterých byl model odhadnut.

Co R^2 popisuje a na co nestačí?

R^2 POPISUJE

- podíl **variability odezvy** zachycený modelem;
- ve **stejných datech**, na kterých byl model odhadnut.

R^2 NEŘÍKÁ

- zda model splňuje předpoklady;
- zda je vztah příčinný;
- jak dobře model poslouží pro nové stromy;
- **zda je to nejlepší z našich kandidátů?**

Pomáhá každý koeficient navíc?

Co udělá R^2 po přidání druhu a interakce?

prumer $\rightarrow$ prumer + druh $\rightarrow$ prumer * druh

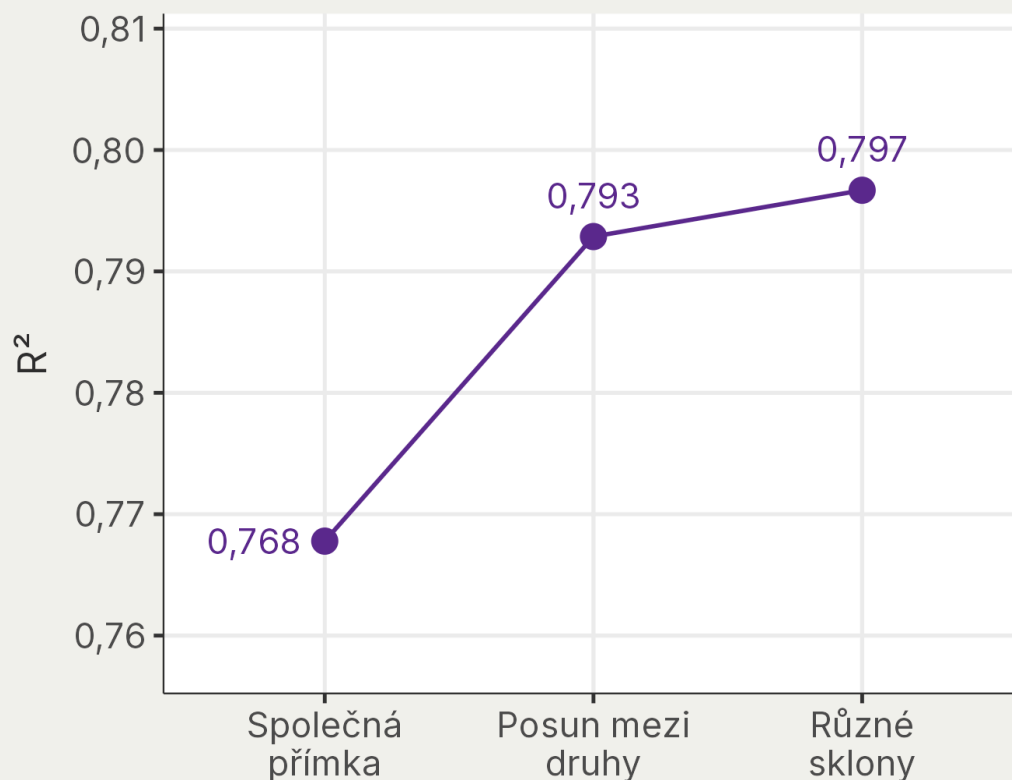
Co se stane s R^2 , když přidáme druh a interakci?

A · R^2 poroste

B · R^2 může klesnout

C · R^2 zůstane stejné

R^2 po přidání členů neklesne



Model	k	RSS
Společná přímka	2	1877
Posun mezi druhy	4	1675
Různé sklony	6	1644

Složitější model umí **zopakovat jednodušší**, a tak R^2 nejhůř zůstane stejné; klesnout nemůže.

Vyšší R^2 není automatické vítězství. I téměř zbytečný prediktor R^2 nesníží.

Které koeficienty jednotlivé modely odhadují?

Který seznam patří společné přímce, který posunu mezi druhy a který různým sklonům?

A· POSUN MEZI DRUHY · K = 4

```
(Intercept)
prumer_kmene_cm
druh_stromuPinus ponderosa
druh_stromuPseudotsuga menziesii
```

B· SPOLEČNÁ PŘÍMKA · K = 2

```
(Intercept)
prumer_kmene_cm
```

C· RŮZNÉ SKLONY · K = 6

```
(Intercept)
prumer_kmene_cm
druh_stromuPinus ponderosa
druh_stromuPseudotsuga menziesii
prumer_kmene_cm:druh_stromuPinus ponderosa
prumer_kmene_cm:druh_stromuPseudotsuga menziesii
```

Jak R^2 opravit za počet koeficientů?

adjustované $R^2 = 1 - \text{nezachycený podíl} \times \text{úprava za počet stromů a koeficientů}$

Posun mezi druhy: $R^2 = 0,793$, $n = 63$ stromů, $k = 4$ koeficienty

$$R_{\text{adj}}^2 = 1 - (1 - 0,793) \times \frac{63 - 1}{63 - 4} = 0,782$$

Adjustované R^2 (*adjusted R-squared*):

$$R_{\text{adj}}^2 = 1 - (1 - R^2) \times \frac{n - 1}{n - k}$$

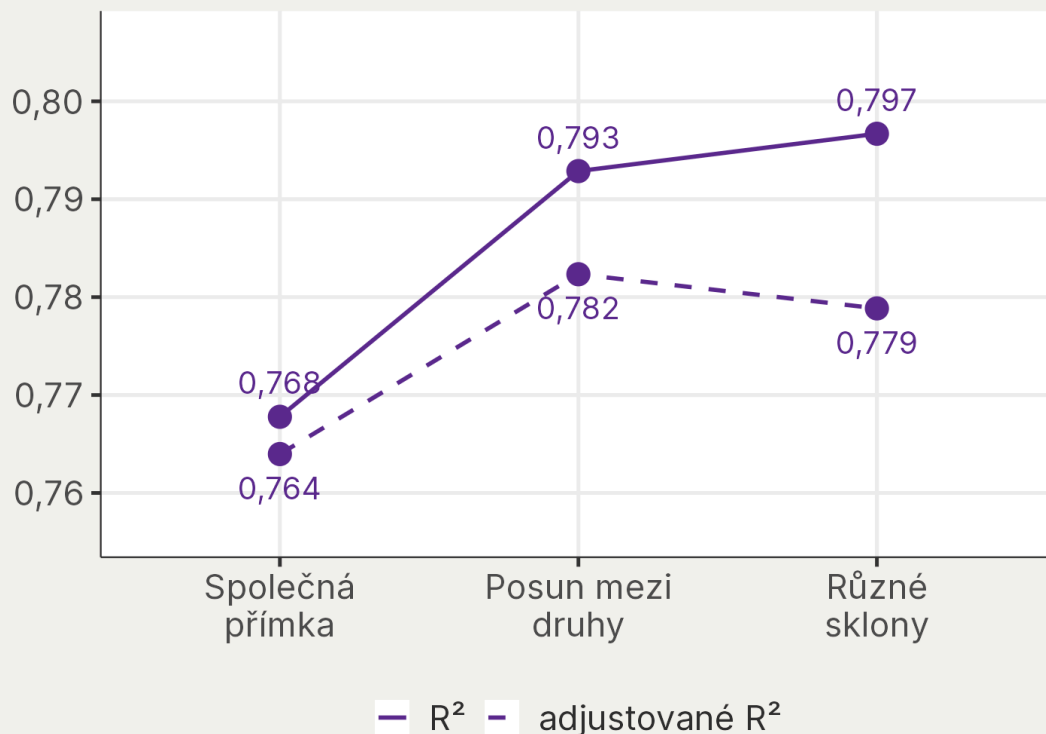
Co udělá adjustované R^2 po přidání interakce?

Posun mezi druhy $\rightarrow$ Různé sklony: R^2 0,793 $\rightarrow$ 0,797 $\cdot k$ 4 $\rightarrow$ 6 $\cdot n = 63$

A $\cdot$ adjustované R^2 vzroste

B $\cdot$ adjustované R^2 klesne

Adjustované R^2 po přidání interakce klesne



Dva sklony navíc zlepšily R^2 jen o **0,004**.

Zlomek $\frac{n-1}{n-k}$ vzrostl víc, než se zmenšil nezachycený podíl → **úprava zvítězila**.

Co už o třech modelech víme?

Jak poznáme, že je model dobrý?

UŽ UMÍME ŘÍCT

- R^2 měří shodu s daty a vždy upřednostní nejpružnější model.
- Adjustované R^2 upřednostní posun mezi druhy.

ZATÍM NEVÍME

- Kdy se složitost obecně vyplatí?
- Vypadají residua vybraného modelu v pořádku?

Kdy se složitost vyplatí?

Vyplatí se model šitý na míru?



Ilustrace vytvořená pomocí AI.

Ve dvojici: Posun mezi druhy a různé sklony popisují stromy téměř stejně dobře (R^2 0,793 a 0,797). První potřebuje **4** koeficienty, druhý **6**. Který byste upřednostnili? Co byste chtěli vědět, než se rozhodnete?

Princip úspornosti (*parsimony*)

Složitější model upřednostníme jen tehdy, když **dostatečně zlepší shodu** s daty.

AIC váží shodu i složitost

AKAIKEHO INFORMAČNÍ KRITÉRIUM (*AKAIKE INFORMATION CRITERION*, AIC)

- lepší shoda AIC **snižuje**;
- každá odhadovaná veličina navíc AIC **zvyšuje**;
- preferujeme **nižší AIC**.

```
1 AIC(object = mod_spolecny)
```

```
[1] 398.64
```

AIC **není podíl** a nemá stejnou škálu jako R^2 .

Model má AIC 398,64. Je to dobrý model?

A · Ano, AIC je nízké.

B · Ne, AIC je vysoké.

C · Z jedné hodnoty to nepoznáme.

AIC je **relativní**. Porovnáváme jen modely se **stejnou odezvou**, **stejnými pozorováními** a stejným předpokladem o rozdělení residuí.

O kolik je AIC každého kandidáta vyšší než nejnižší?

$$\Delta\text{AIC} = \text{AIC kandidáta} - \text{nejnižší AIC v sadě}$$

Společná přímka: $398,64 - 395,44 = 3,20$

Model	AIC	ΔAIC
Společná přímka	398,64	3,20
Posun mezi druhy	395,44	0,00
Různé sklony	398,27	2,83

Posun mezi druhy má $\Delta\text{AIC} = 0$ · univerzální práh **nepoužíváme**

Ukazují R^2 , adjustované R^2 a AIC na stejný model?

Model	k	$R^2 \uparrow$	adj. $R^2 \uparrow$	AIC $\downarrow$
Společná přímka	2	0,768	0,764	398,64
Posun mezi druhy	4	0,793	0,782	395,44
Různé sklony	6	0,797	0,779	398,27

Ve dvojici: Proč R^2 upřednostní jiný model než adjustované R^2 a AIC?

R^2 hodnotí **jen shodu**. Adjustované R^2 a AIC **penalizují složitost** — dva sklony navíc se nevyplatily. Nejsou to tři hlasy o pravdě.

Nejlépe podporovaný kandidát nemusí být dobrý model



Ilustrace vytvořená pomocí AI.

Porovnání řadí jen modely, které jsme do sady zařadili.

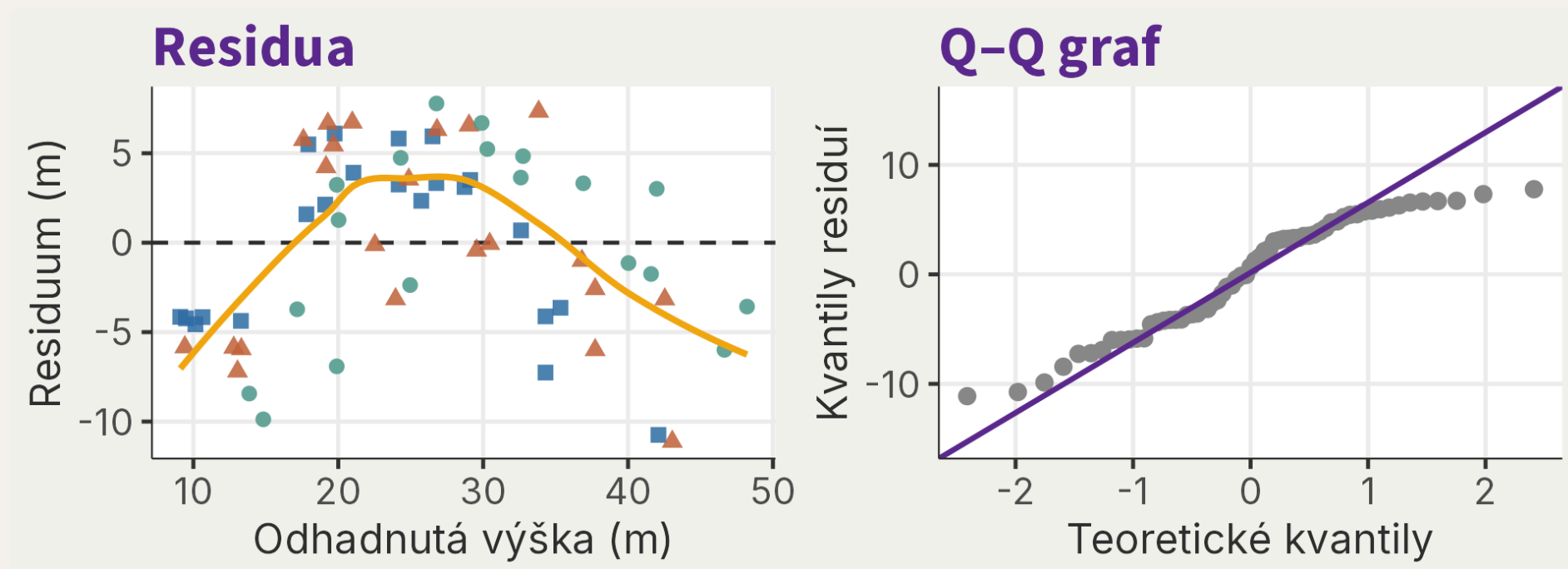
Neříká, že je vítěz obecně dobrý.

Model, který jsme vůbec neposuzovali, může stromy popsat lépe.

Proto vybraný model ještě zkontrolujeme.

Obstojí vybraný model?

Zůstává v residuích modelu s posunem mezi druhy nějaký vzorec?

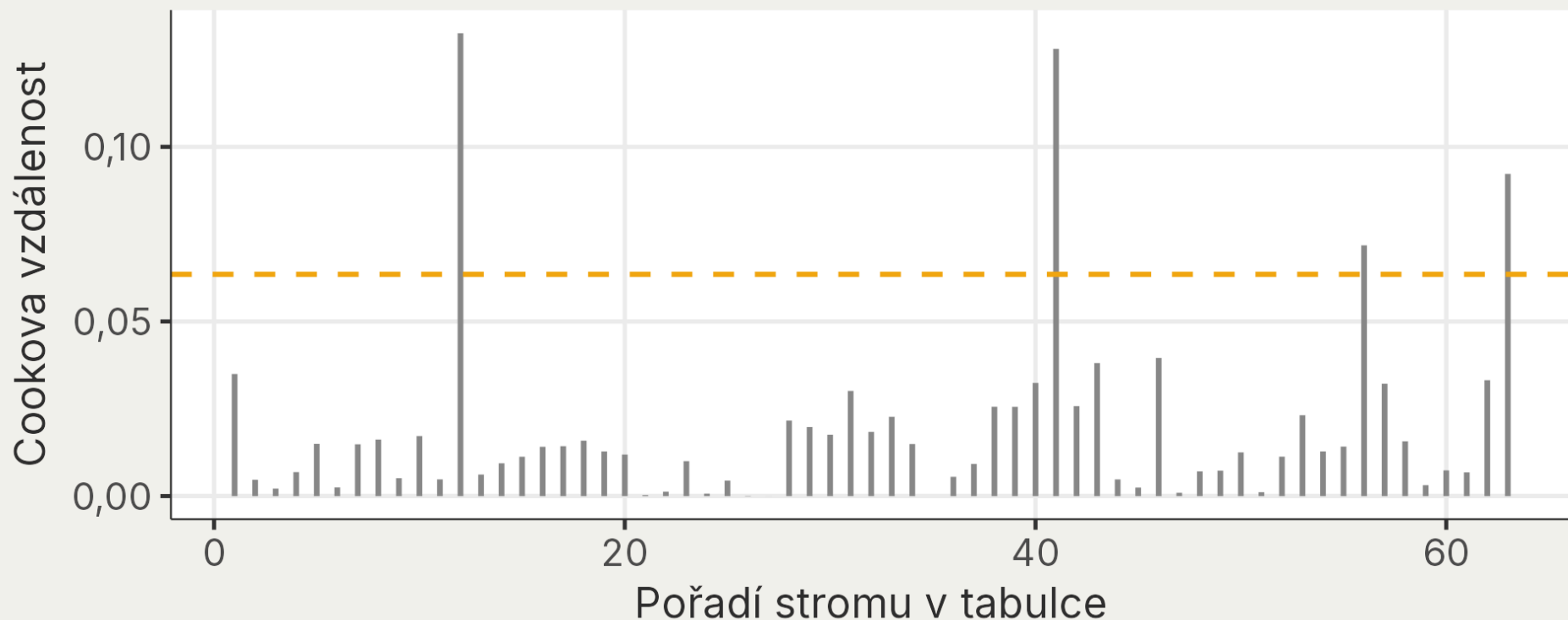


Ve dvojici: Leží residua v celém rozsahu odhadnutých výšek náhodně kolem nuly?

Oranžová křivka jen vyhlazuje průběh bodů, aby byl vzor lépe vidět: residua tvoří **oblouk**

• Q-Q graf **není přímý**.

Které stromy ovlivňují model nejvíc?



Cookova vzdálenost: o kolik by se odhadnuté hodnoty změnily, kdyby strom chyběl. Nad orientační čarou $4/n$: **4** stromy.

Důvod **stromy prověřit** — ne je automaticky odstranit.

Obhajte model dvěma větami

40 S SAMI · VE DVOJICI POROVNEJTE

„Ze tří kandidátů vybíráme model ____, protože ____.“

„Omezením tohoto závěru je ____.“

Opřete se aspoň o **dva různé druhy důkazů**: shodu, složitost, diagnostiku, biologický smysl.

Který model lze pro tyto stromy nejlépe obhájit?

POSUN MEZI DRUHY

- zachytí rozdíly mezi druhy, které společná přímka přehlíží;
- nejvyšší adjustované R^2 (0,782) a nejnižší AIC.

OMEZENÍ

- vybírali jsme jen ze tří kandidátů;
- residua tvoří oblouk: přímka nevystihuje tvar vztahu;
- 63 stromů, Q-Q graf není přímý;
- pozorovací data: nejde o příčinný efekt druhu.

Dobrý model nepoznáme podle jediného čísla. Obhajoba propojuje otázku, shodu, složitost, diagnostiku a interpretovatelnost.

Jak porovnávat modely bez automatického hledání?

1. Položte biologickou otázku.
2. Prohlédněte a zobrazte data.
3. Určete malou sadu **zdůvodněných kandidátů**.
4. Použijte **stejnou odezvu a stejná pozorování**.
5. Porovnejte přímky, residua, R^2 , adjustované R^2 a AIC.
6. Zkontrolujte předpoklady a vlivná pozorování.
7. Obhajte závěr a **pojmenujte omezení**.

Automatické přidávání a odebírání členů vytvoří mnoho modelů, ale **neurčí, které dávají biologický smysl**.

Co si odnést

- R^2 = podíl zachycené variability; počítá se z celkového součtu čtverců a RSS.
- Po přidání členu R^2 **neklesne**, a proto samo model nevybírá.
- Adjustované R^2 a AIC **penalizují složitost**, každé jinak; AIC je **relativní**.
- Kandidáty určujeme **předem**: stejná odezva, stejná pozorování.
- Obhajoba spojuje **několik druhů důkazů** včetně diagnostiky a omezení.

Co když žádná přímka nestačí?



Všechny tři kandidáty jsme kreslili jako přímky — a residua přesto tvoří oblouk.

Ilustrace vytvořená pomocí AI.

Jak popsat vztah, který není přímkový?