BIOSTATISTIKA



L07

Mění klima původu tloušťku listů křídlatky?

Více prediktorů: aditivní model a korelace prediktorů

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

Ilustrace vytvořená pomocí AI. Fotografie zahrad: Cao et al. (2025), CC BY 4.0.

Mění klima původu tloušťku listů křídlatky?

Která klimatická stopa nese vlastní informaci?

- Průměrná teplota?
- Nejteplejší měsíc?
- Původ populace?

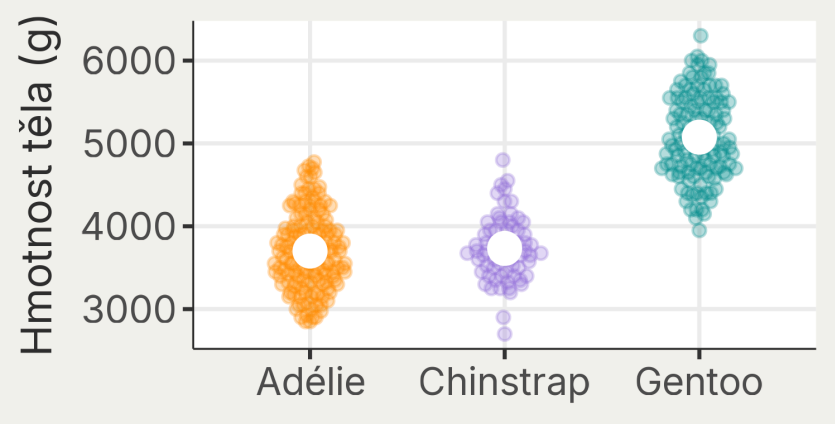


Ilustrace vytvořená pomocí AI.

Co si pamatujete z minulé lekce?

Odpovězte samostatně na tři krátké otázky. Potom společně projdeme správné odpovědi.

V modelu hmotnost_tela_g ~ druh je referenční skupinou Adélie. Co znamená intercept?

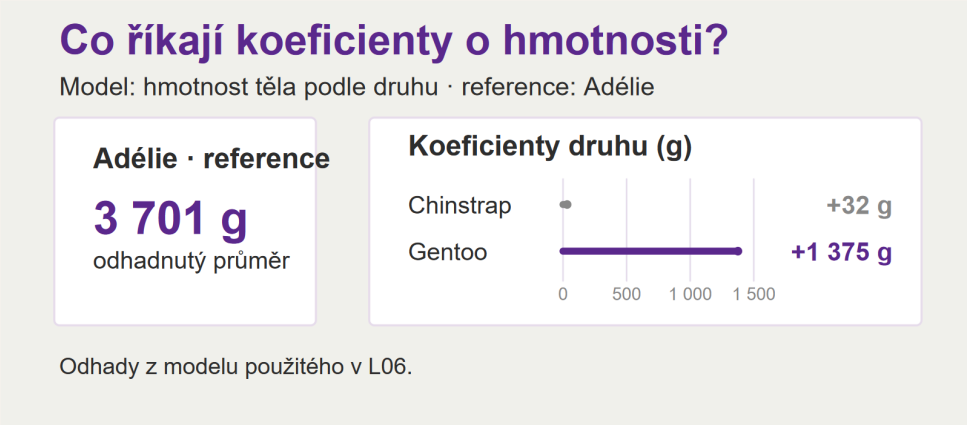


- Průměrnou hmotnost všech tří druhů dohromady.
- Rozdíl mezi Gentoo a Adéliemi.
- Odhadnutou průměrnou hmotnost Adélií.
- Hodnotu celkového F-testu.

SPRÁVNÁ ODPOVĚĎ

: Odhadnutou průměrnou hmotnost Adélií.
Při referenci Adélie představuje intercept její odhadnutý průměr; v datech L06 je přibližně 3 701 g.

Koeficient druhGentoo ve stejném modelu je přibližně +1 375 g. Jak ho číst?



Gentoo mají podle modelu průměr přibližně 1 375 g.

Odhadnutý průměr Gentoo je přibližně o 1 375 g vyšší než průměr Adélií.

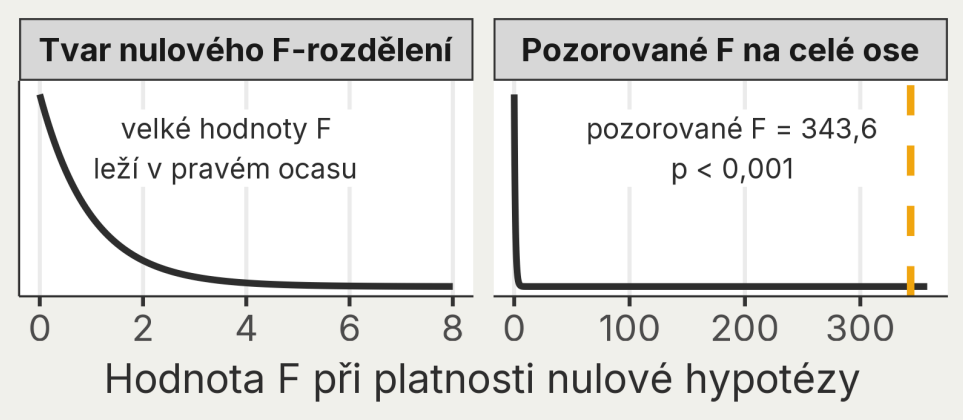
Každý jednotlivý tučňák Gentoo váží přesně o 1 375 g více než Adélie.

Gentoo jsou o 1 375 g těžší než Chinstrap.

SPRÁVNÁ ODPOVĚĎ

: Odhadnutý průměr Gentoo je přibližně o 1 375 g vyšší než průměr Adélií.
Koeficient druhu je rozdíl jeho odhadnutého průměru vůči referenci Adélie, nikoli samostatný průměr ani rozdíl každé dvojice jedinců.

Celkový F-test modelu se třemi druhy dal velmi malou p-hodnotu. Co z něj plyne přímo?



- Všechny tři dvojice druhů se navzájem liší.
- Víme, která konkrétní dvojice se liší, bez dalšího porovnání.
- Druh tučňáka je prokázanou příčinou rozdílu hmotnosti.
- Alespoň jeden druhový průměr se liší od některého dalšího.

SPRÁVNÁ ODPOVĚĎ

: Alespoň jeden druhový průměr se liší od některého dalšího.
Které dvojice se liší, určí až párová porovnání. Samotné srovnání druhů neprokazuje příčinu.

Výsledky učení

Po absolvování této přednášky budete schopni:

- fitovat aditivní model s více prediktory a vyložit každý koeficient „při stejných hodnotách ostatních prediktorů“;
- vysvětlit, proč se odhad efektu po přidání prediktoru může změnit, a odlišit celkový a podmíněný vztah;
- rozpoznat silně korelované prediktory a popsat, jak zvyšují nejistotu a komplikují výklad;
- porovnat skupiny a vytvářet predikce jen pro kombinace hodnot podložené daty;
- odlišit test koeficientu v `summary()` od sekvenčního testu v `anova()`.

Jedna teplota, jeden sklon

Co už umíme s jedním prediktorem?

Číselný prediktor

$$\text{odezva} = \beta_0 + \beta_1 x + \varepsilon$$

- sklon v biologických jednotkách
- interval spolehlivosti
- residua

Kategoriální prediktor

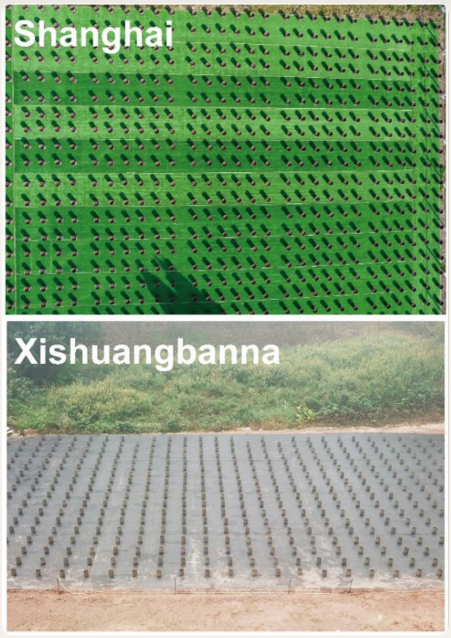
- referenční skupina
- rozdíl mezi skupinami
- koeficientový a modelový test

> Novinka: jedna odezva, několik prediktorů současně — a s nimi jiná otázka.

Křídlatky z různých částí světa

🌍 128 populací z různých míst → 🪴 stejná zahrada v Šanghaji → 🌿 průměrná tloušťka listu





Křídlatka: Jasper Shide / Wikimedia Commons, CC0. Zahrady: obr. 1b z Cao et al. (2025), CC BY 4.0.

Co je jednotka analýzy?

Klima se liší mezi zdrojovými populacemi. Co je tedy jednou nezávislou informací?

A

Jeden změřený list nebo jedna rostlina

B

Průměr jedné zdrojové populace

Jeden bod představuje jednu zdrojovou populaci

128 populačních průměrů (řádků)

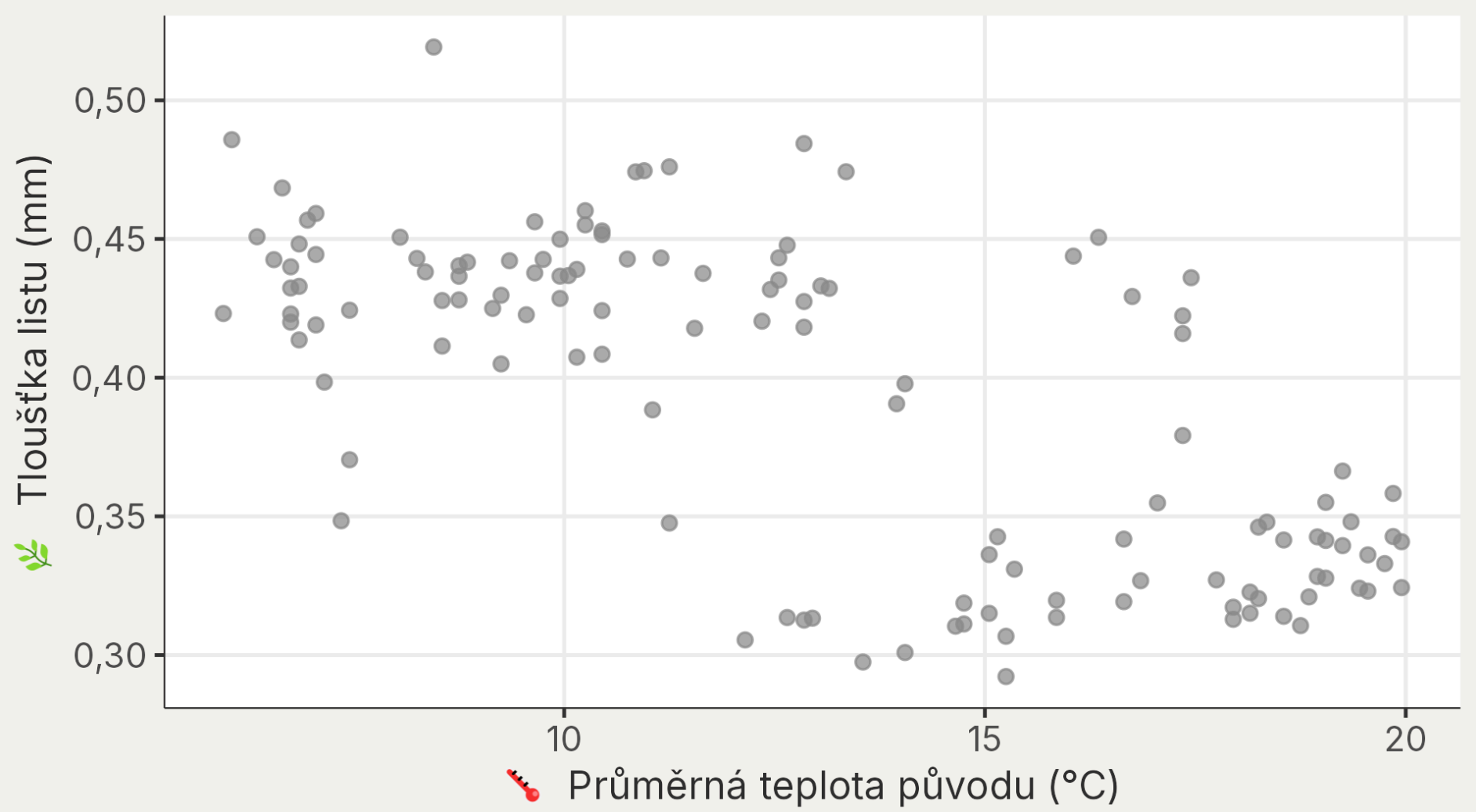
Jednotlivé rostliny poskytly měření, ale klima původu je společné celé populaci.

55 populací z **původního areálu** (Čína, Japonsko) a **73** ze **zavlečeného areálu** (Evropa, Severní Amerika).

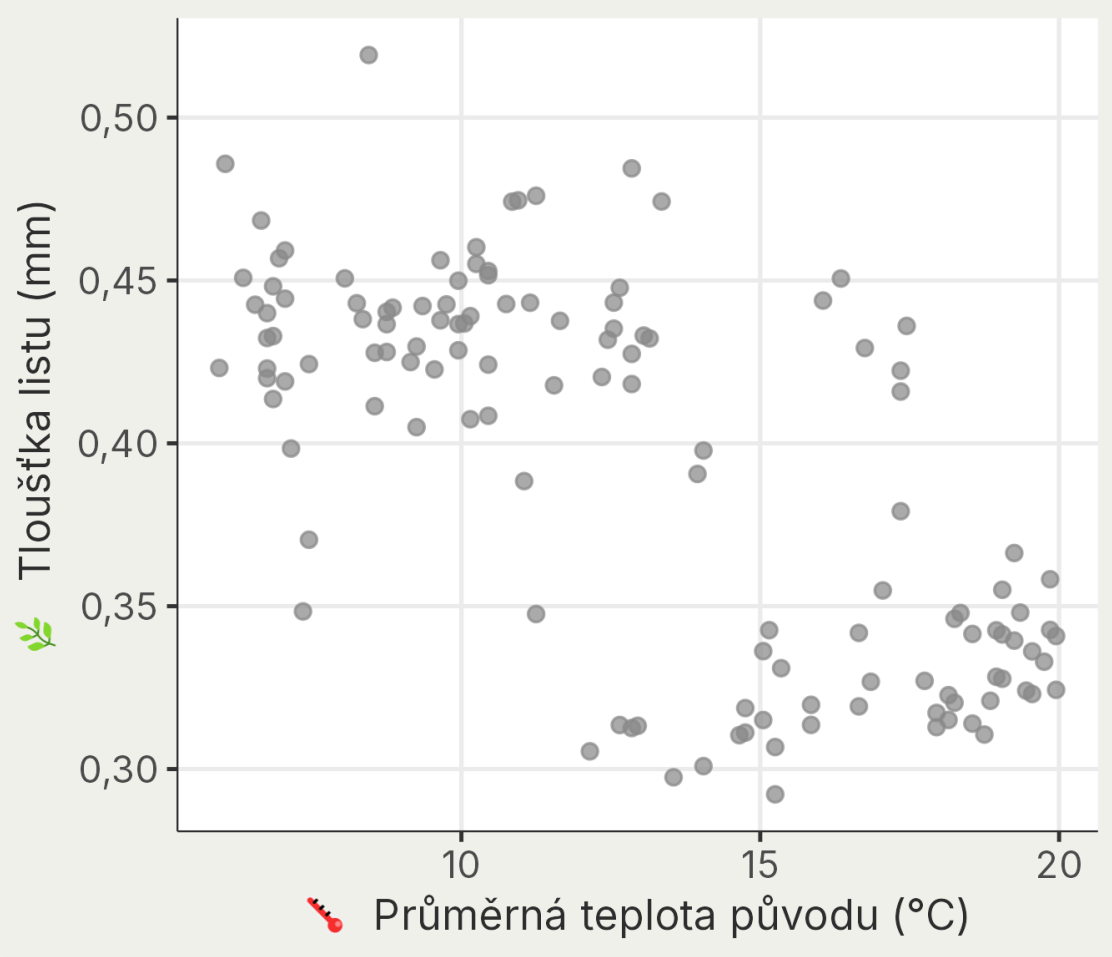
Úroveň	Role
rostlina	měření listů
populace	řádek analýzy
zahrada	společné aktuální prostředí

Klima původu: CHELSA-BIOCLIM 2.1, 1981–2010, Brun et al. (2022), CC0.

Tloušťka listu a teplota místa původu



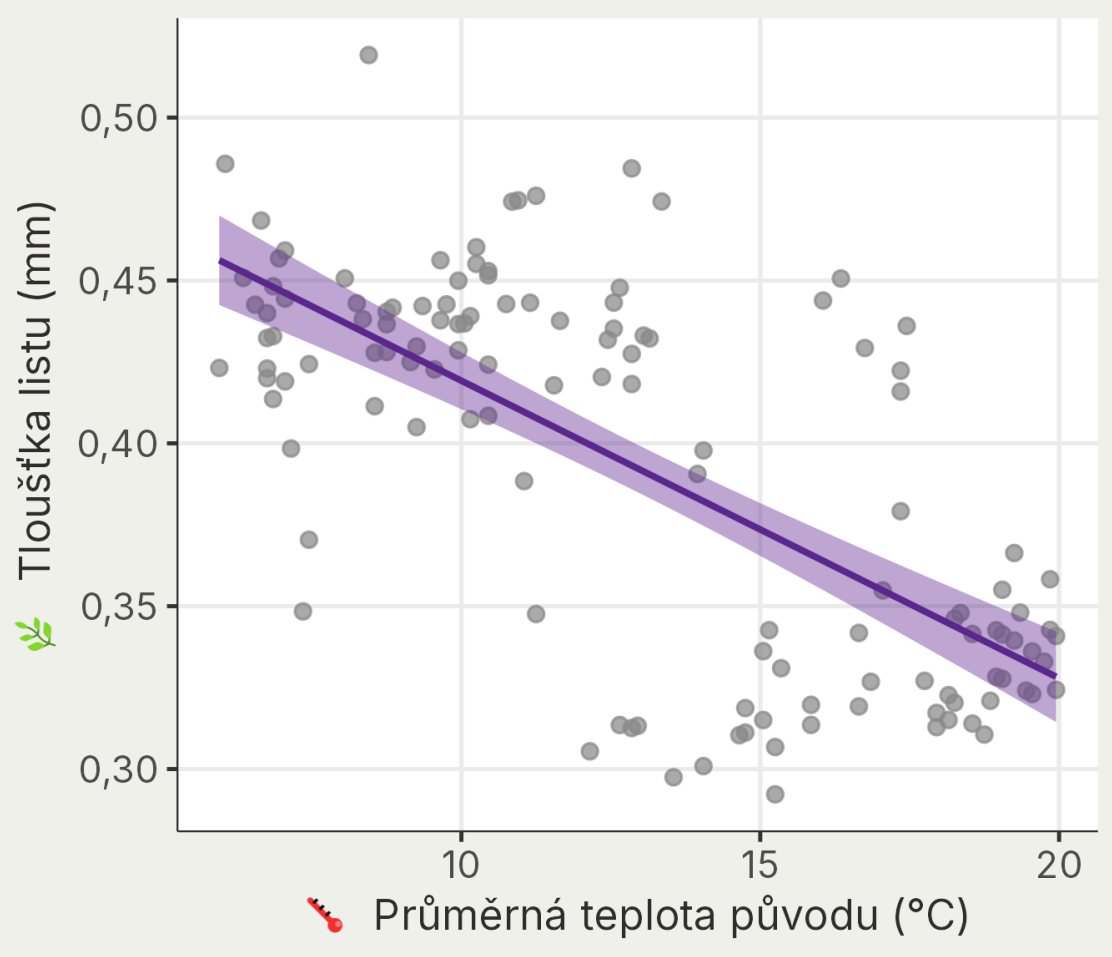
Co dokážeme vyčíst z grafu?



Ve dvojici popište:

1. směr vztahu;
2. jeho těsnost;
3. neobvyklé body;
4. jedno tvrzení, které graf **zatím nepodporuje**.

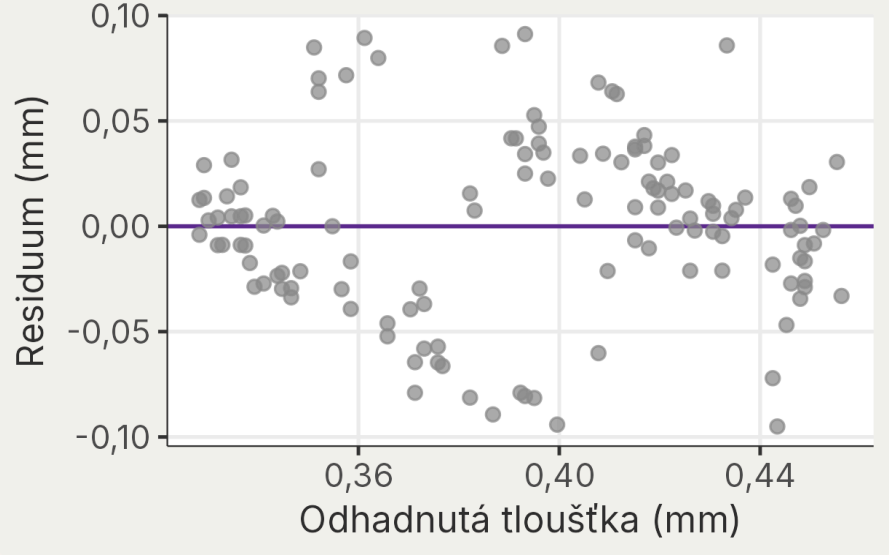
Průměrná teplota má záporný sklon

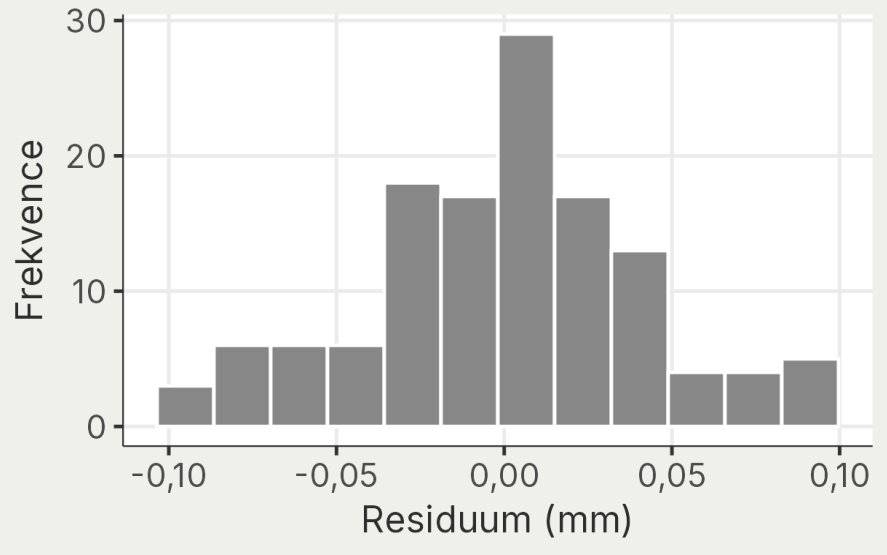


−0,00913 mm na 1 °C

O 5 °C teplejší původ odpovídá podle modelu přibližně o 0,046 mm tenčímu listu.

Odpovídá přímka struktuře reziduí?

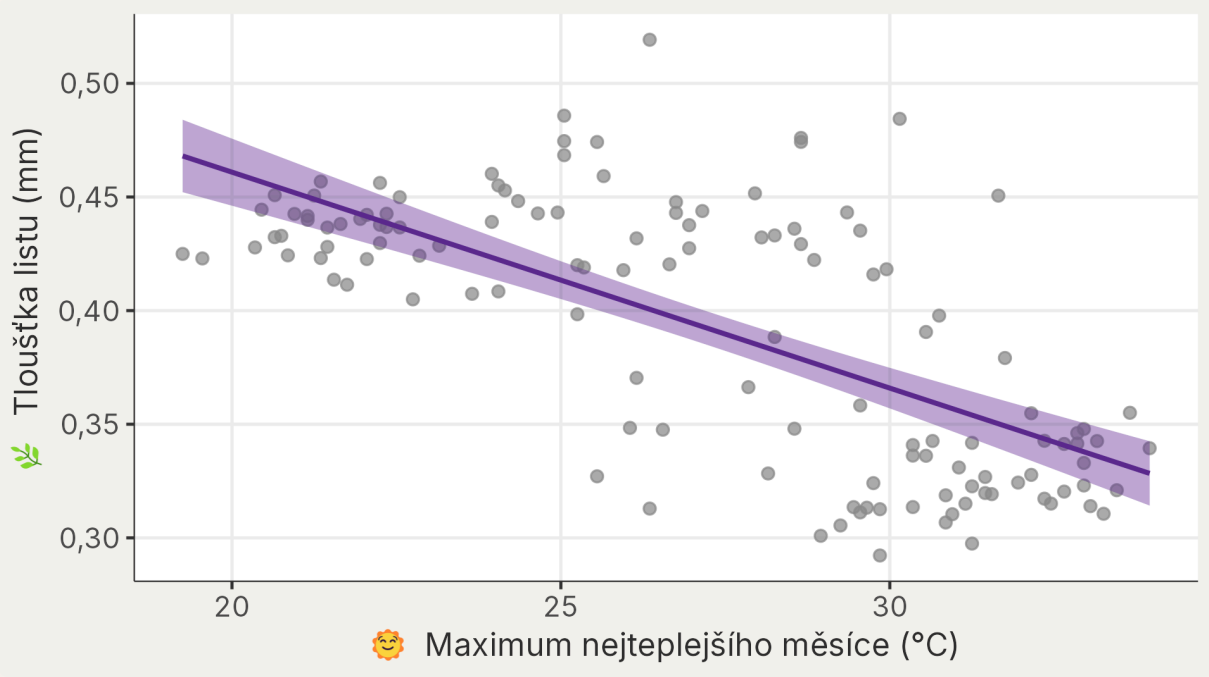




> Hledáme náhodný mrak kolem nuly — bez oblouku, vln, trychtýře nebo samostatných pásů.

Při nízkých až středních odhadech leží residua spíš pod nulou, potom nad ní: přímka něco nezachytila. Co by to mohlo být?

Také nejteplejší měsíc souvisí s tloušťkou listu



Stejná odezva

🌿 Tloušťka listu

Jiná osa x

🌞 Maximum nejteplejšího měsíce

−0,00950 mm na 1 °C maxima

Dva prediktory

Co ukáže model s oběma teplotami?

průměrná teplota nebo **maximum**?

Oba samostatné vztahy jsou záporné.

> Nese každá teplota vlastní informaci, nebo obě totéž?

Co znamená znaménko `+` ve formuli?

```
tloustka_listu_mm ~ teplota_rocni_C +
teplota_max_C
```

A) Oba prediktory vstupují do jednoho modelu.

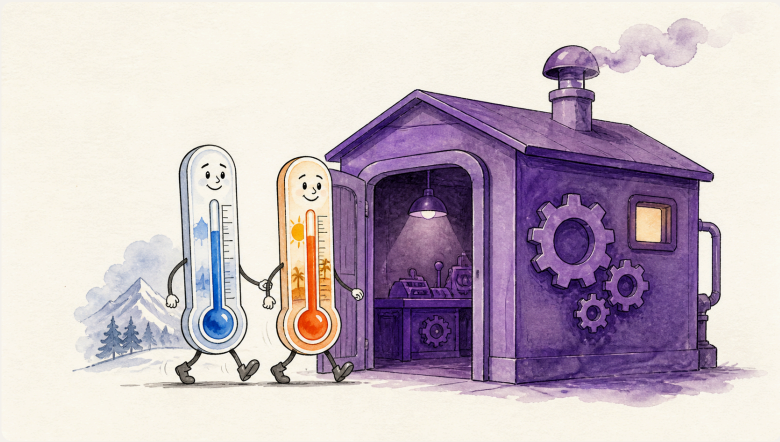
B) Sečteme naměřené teploty.

C) Oba koeficienty musí být kladné.

Znaménko určuje složení modelu, ne sčítání jednotek ani směr efektu.

Jeden model může použít obě teploty

```r
mod_dve_teploty <- lm(
  formula = tloustka_listu_mm ~
    teplota_rocni_C + teplota_max_C,
  data = data_kridlatka
)
```



Ilustrace vytvořená pomocí AI.

odhad tloušťky = intercept + příspěvek průměrné teploty + příspěvek maxima

Každý sklon platí při stejné druhé teplotě

PRŮMĚRNÁ TEPLOTA

Při stejné maximální teplotě:
−0,00528 mm / °C

MAXIMÁLNÍ TEPLOTA

Při stejné průměrné teplotě:
−0,00475 mm / °C

„Při stejné hodnotě ostatních prediktorů“ je součást významu koeficientu.

Jak model odhadne jednu populaci?

Skutečná populace **CN35**: průměrná teplota 15,25 °C; maximum 29,85 °C.

Parametr	Odhad
intercept (mm)	0,5899
průměrná teplota (mm / °C)	−0,00528
maximum (mm / °C)	−0,00475

$$0{,}5899$$

$$+ (-0{,}00528) \times 15{,}25$$

$$+ (-0{,}00475) \times 29{,}85$$

$$= \mathbf{0{,}368}\ \mathrm{mm}$$

Model skládá odhad ze tří příspěvků

$$\widehat{\text{tloušťka}} = \text{intercept} + \text{příspěvek průměrné teploty} + \text{příspěvek maxima}$$

$\hat{\beta}_0$
intercept modelu

$\hat{\beta}_1 x_1$
první příspěvek

$\hat{\beta}_2 x_2$
druhý příspěvek

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Korelace prediktorů

Které klimatické dvojice budou silně souviset?

Vyberte dvě dvojice. Proč by spolu měly souviset?

- 🌡️ průměrná teplota ↔ 🌞 maximum
- 🌧️ roční srážky ↔ ☔ nejvlhčí měsíc
- 🌡️ průměrná teplota ↔ ❄️ minimum
- 🌡️ teplota ↔ 🌧️ srážky

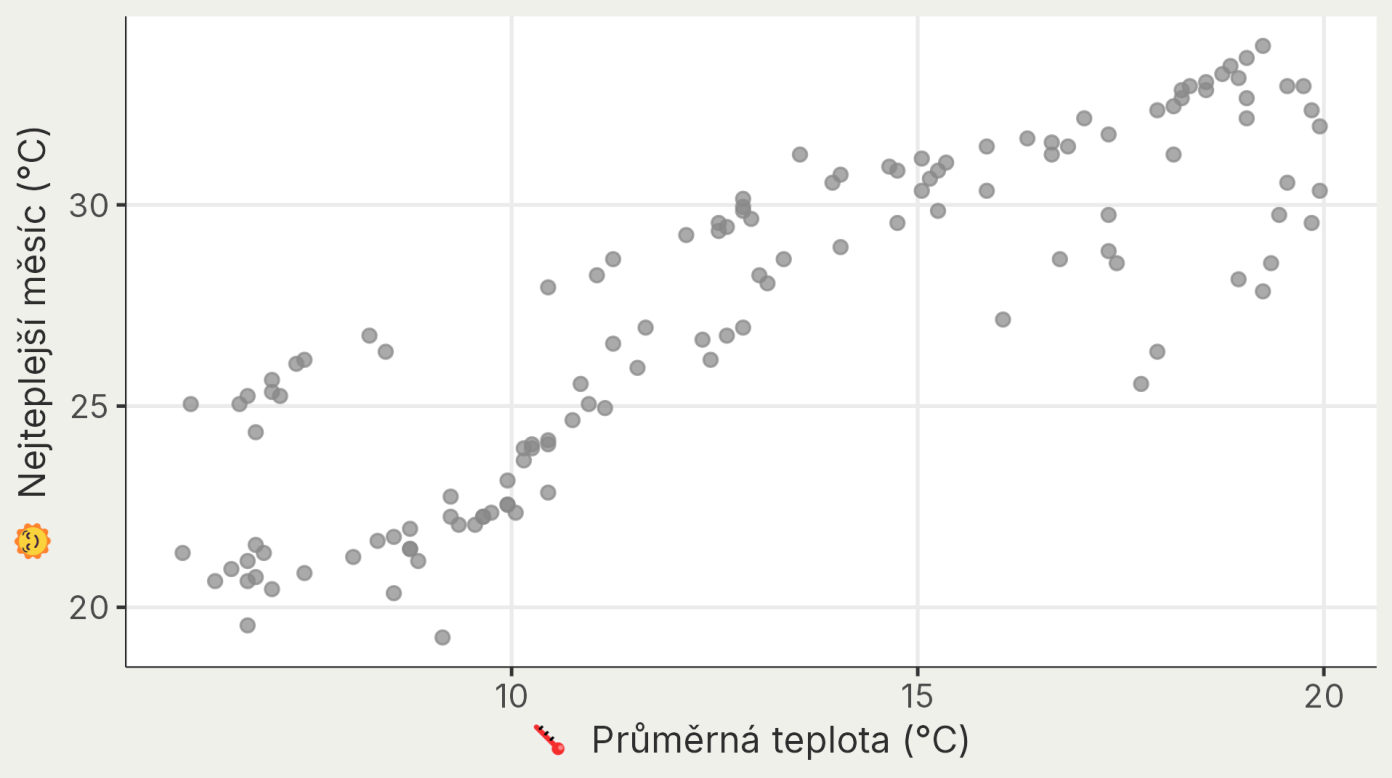
Osm klimatických proměnných spolu souvisí



Čtěte: **pozice buňky → znaménko → velikost → biologický smysl**.

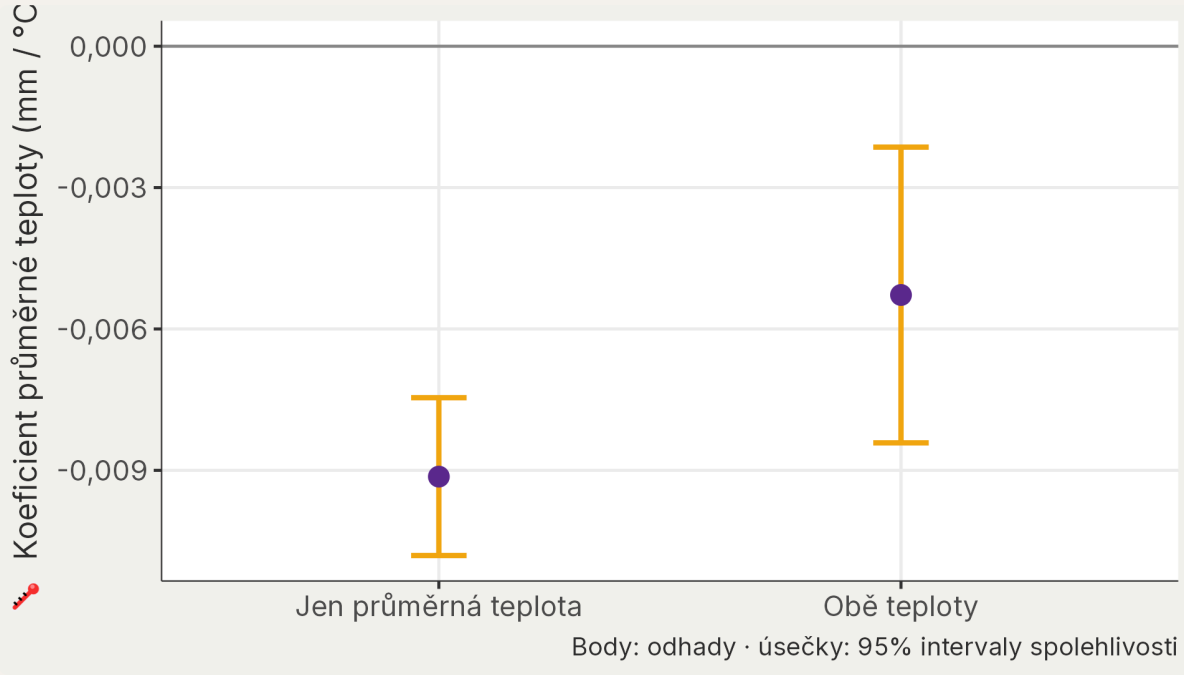
Nejsilnější z vašich dvojic: roční srážky ↔ nejvlhčí měsíc (r = 0,93) a průměrná ↔ maximální teplota (r = 0,85).

Průměrná a maximální teplota se mění společně (r = 0,855)



Hledejte body se skoro stejnou maximální teplotou, ale rozdílnou průměrnou teplotou.

Překrývající se informace zvětšuje nejistotu



Jen průměrná teplota

0,00085
standardní chyba

Obě teploty

0,00159 standardní chyba

Nejistota se zvětšila.

Přidáním prediktoru se změnila otázka



Ilustrace vytvořená pomocí AI.

Jednoprediktorová otázka

Jaký je celkový vztah k průměrné teplotě?

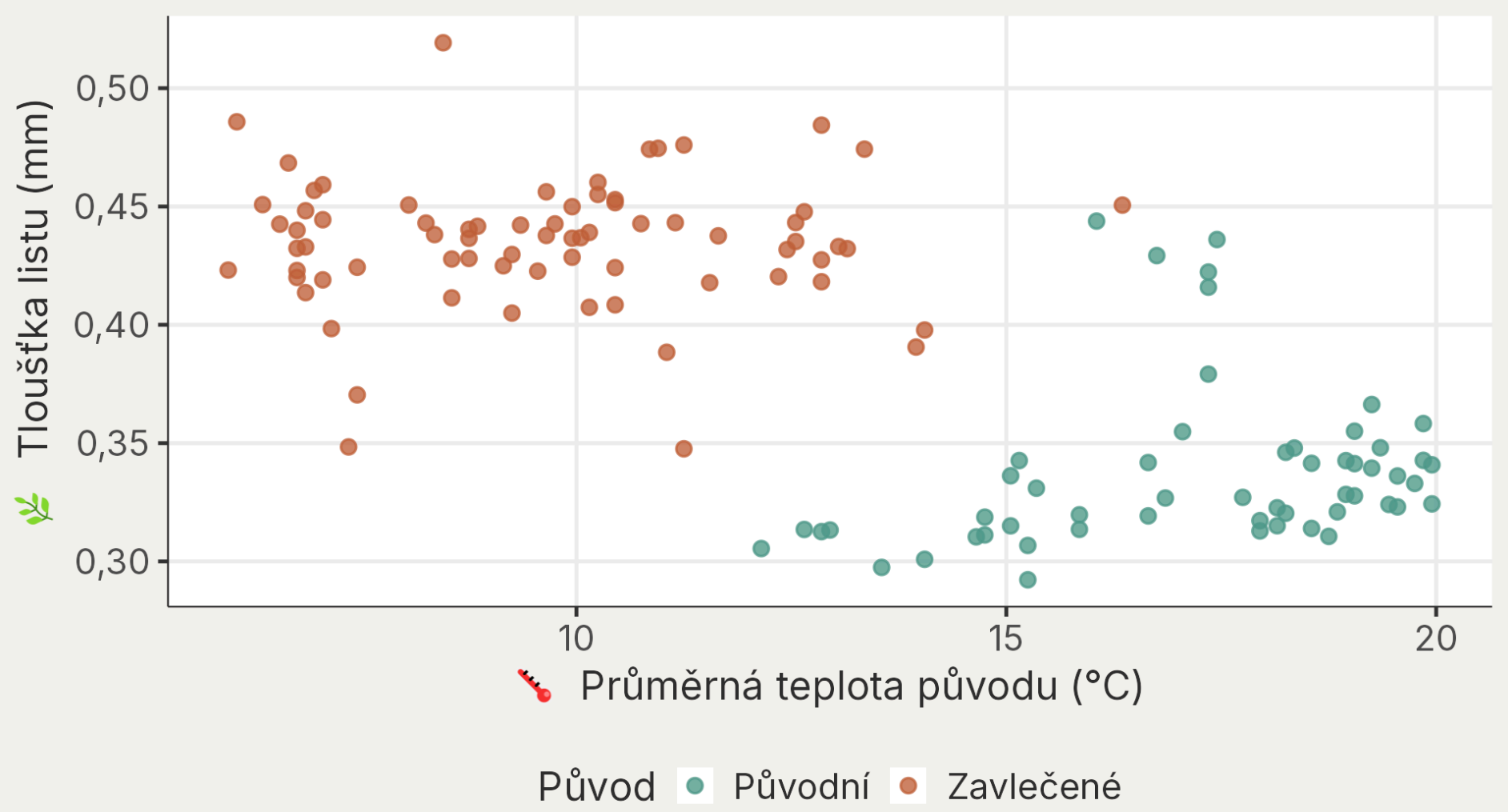
Podmíněná otázka

Jaký je vztah k průměrné teplotě při stejné maximální teplotě?

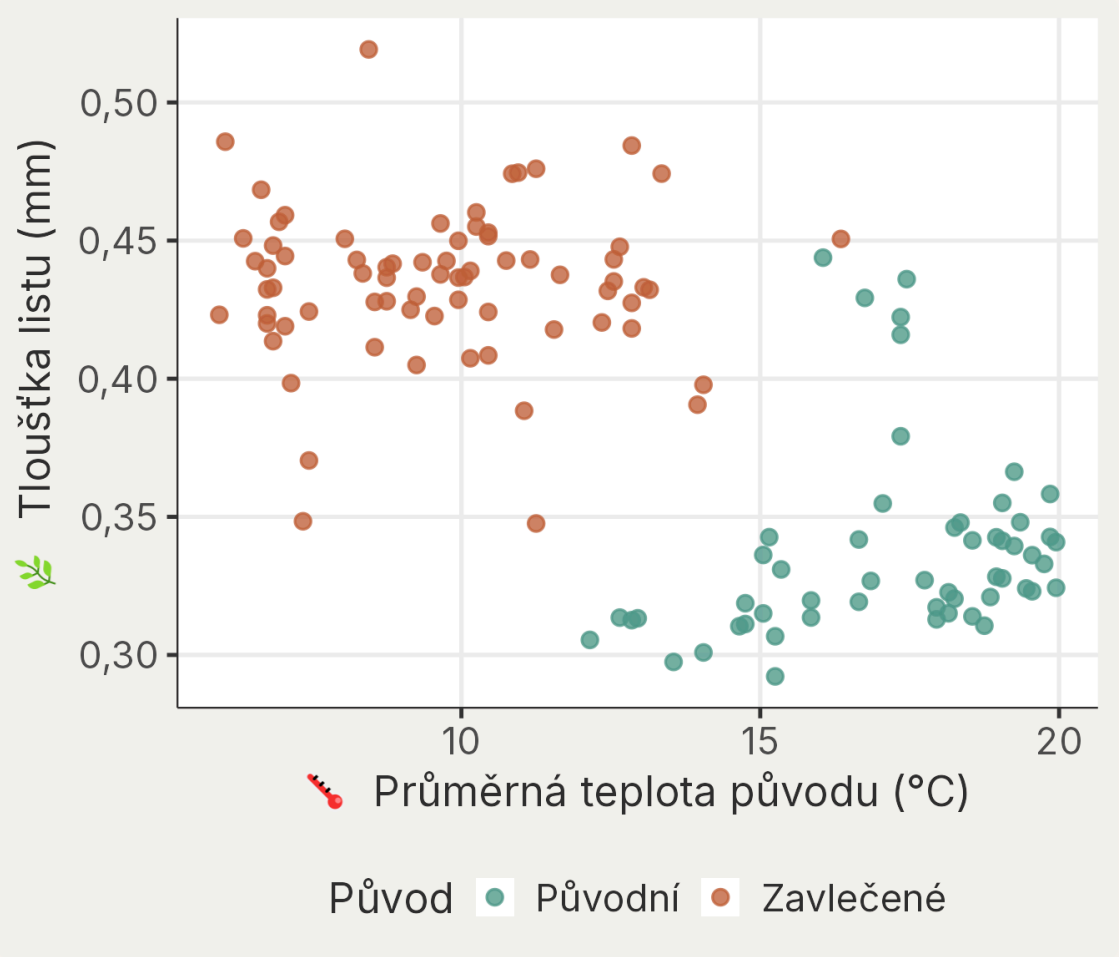
> Ani jeden koeficient není „jediný pravý“. Odpovídají na **jiné otázky**.

Původ populací

Původ populací mění obraz vztahu



Co porovnává celkový sklon?



Která část klesajícího vztahu vzniká...

- rozdíly **uvnitř** skupin?
- rozdílnými rozsahy **mezi** skupinami?

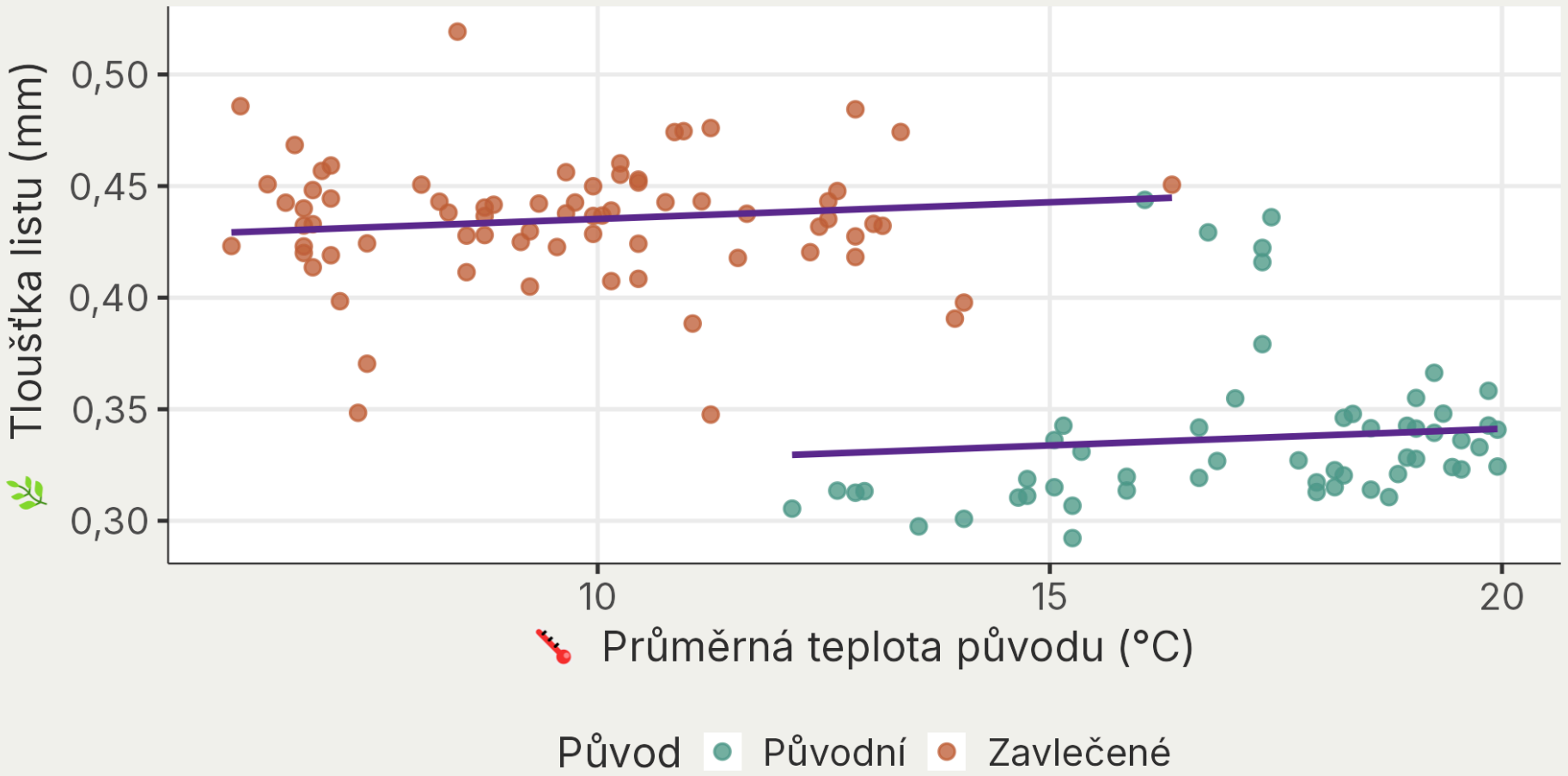
Jeden model spojí teplotu i původ

```r
mod_puvod <- lm(
  formula = tloustka_listu_mm ~
    puvod + teplota_rocni_C,
  data = data_kridlatka
)

coef(object = mod_puvod)
```

Koeficienty z coef(), doplněné o biologickou roli		
Parametr	Odhad	Biologická role
intercept (mm)	0,31143	průměr reference
zavlečené – původní (mm)	0,10895	rozdíl skupin
průměrná teplota (mm / °C)	0,00149	společný sklon

Aditivní model: rovnoběžné přímky



Stejný sklon pro obě skupiny je **znak aditivního modelu**.

Jak čteme společný sklon a posun skupiny?

> Referenční skupina modelu: **původní** populace.

+0,00149 mm / °C

Doplňte:

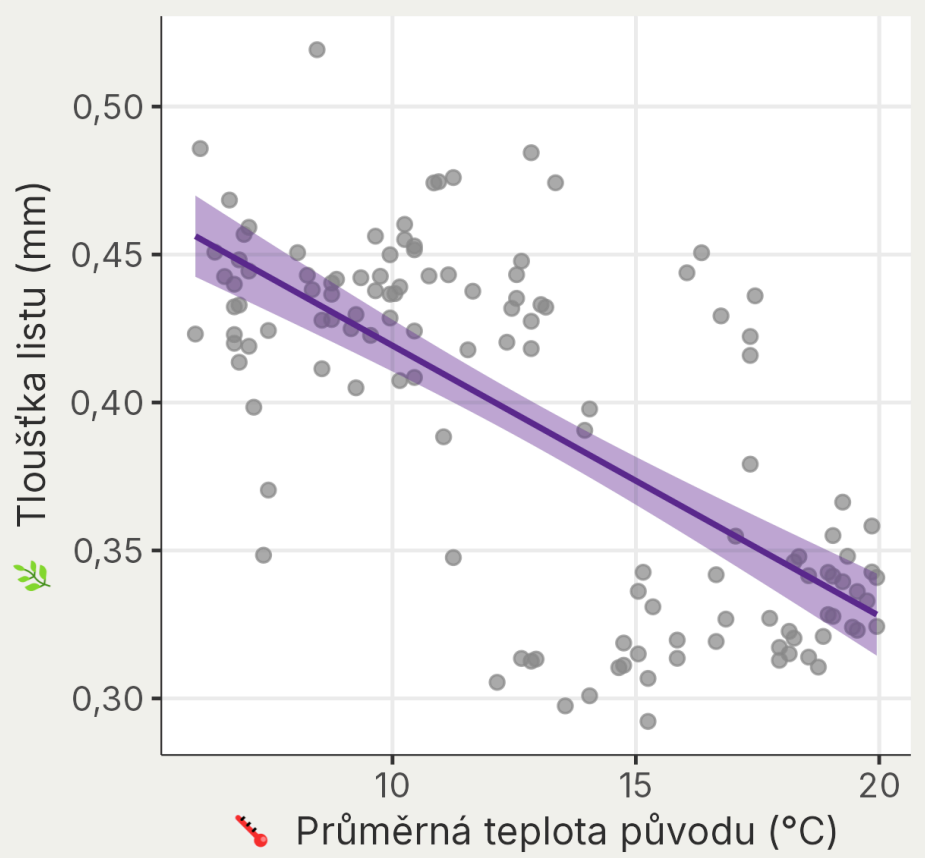
„Při stejném _____ se odhadovaná tloušťka…“

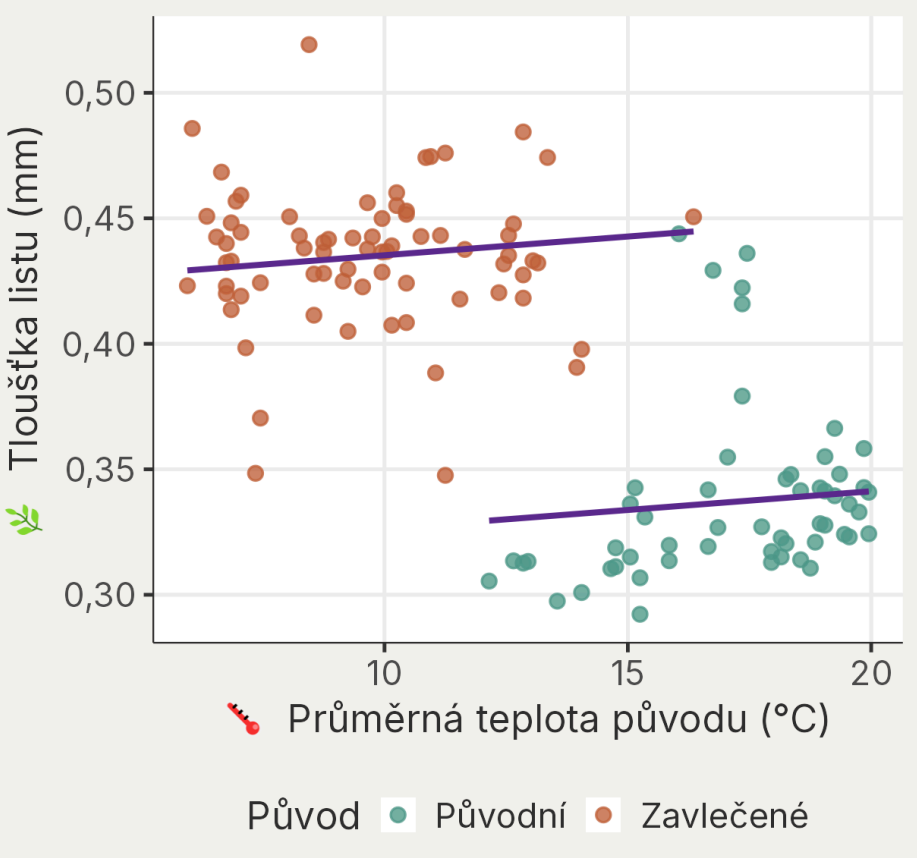
+0,10895 mm

Doplňte:

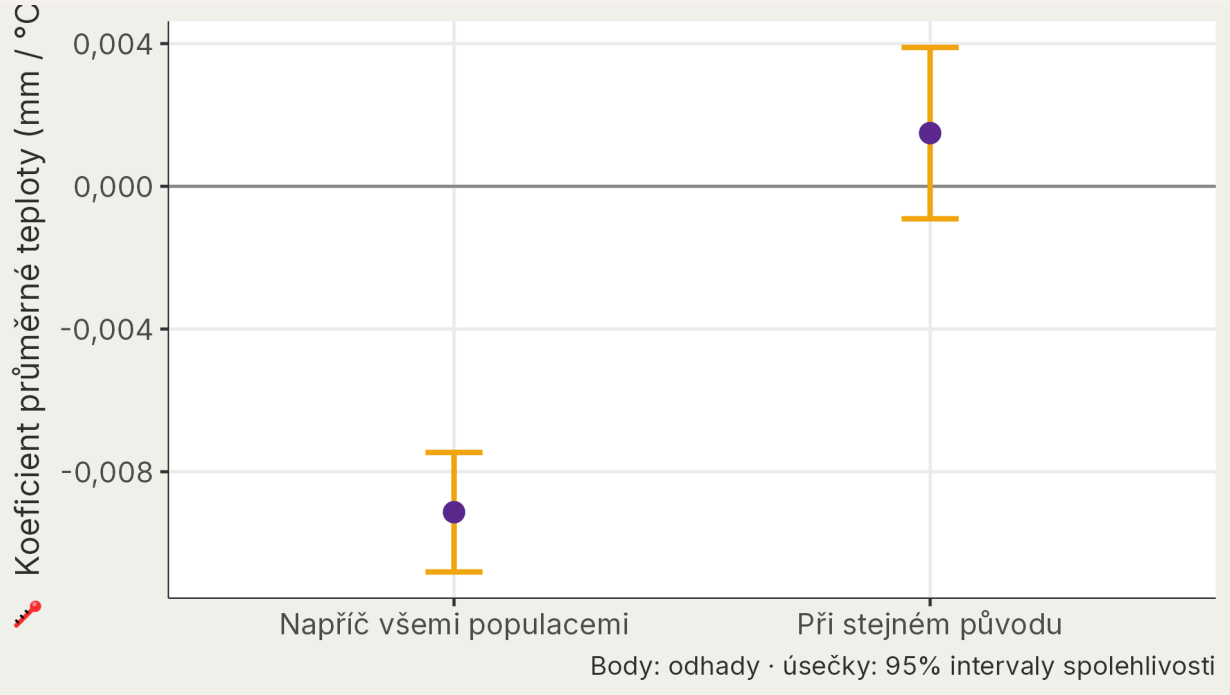
„Při stejné _____ mají zavlečené populace…“

Při zohlednění původu sklon téměř mizí





Při zohlednění původu sklon téměř mizí



Napříč populacemi:

−0,00913 mm/°C

Při stejném původu:

+0,00149 mm/°C

Původ souvisí s **teplotou** i s **tloušťkou listu**. Celkový sklon proto mísí rozdíl mezi areály s případným vztahem k teplotě: **konfundování**.

Co už víme o klimatické otázce?

Mění klima původu tloušťku listů křídlatky?

VÍME

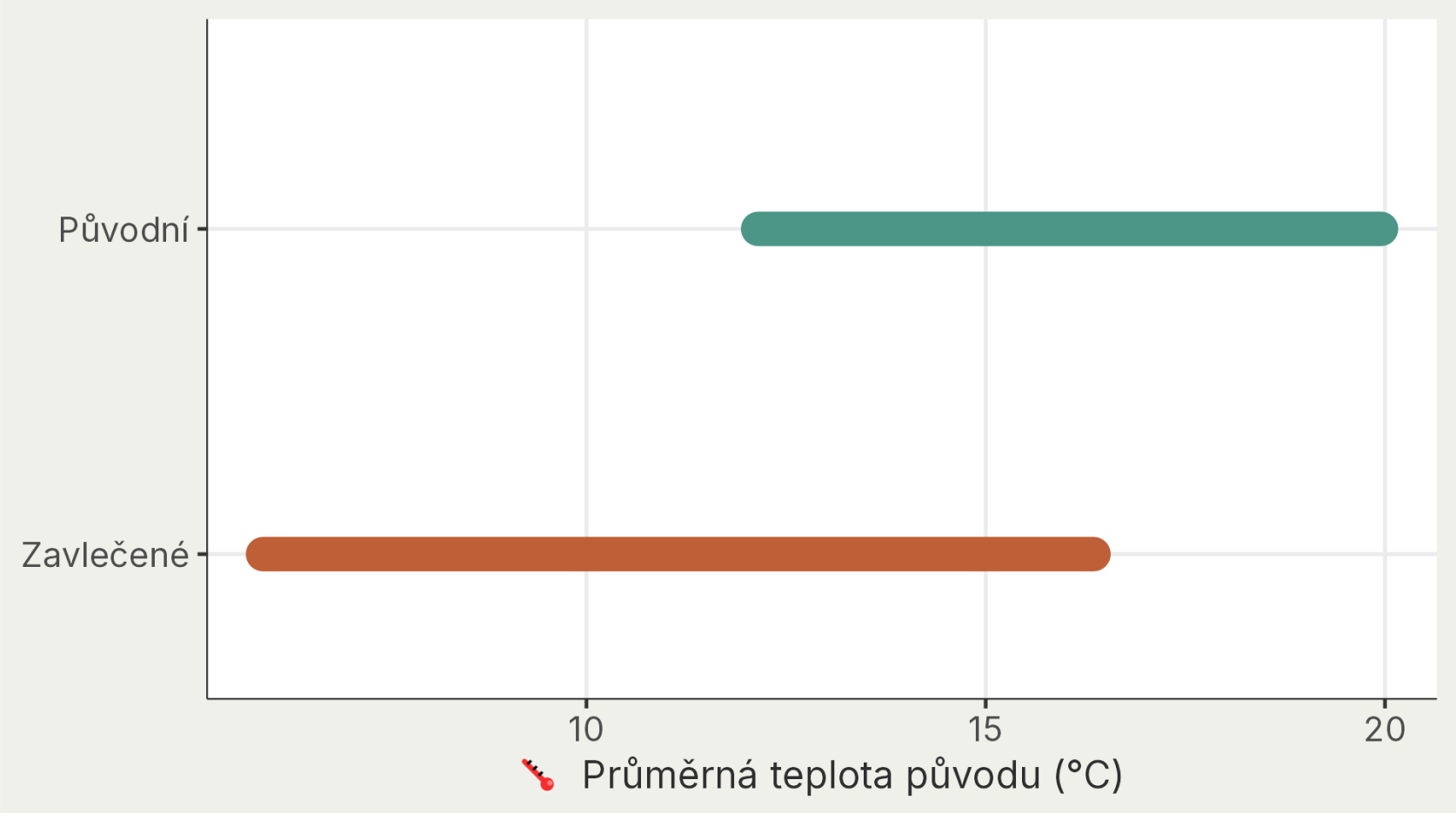
Původ a klimatické proměnné nesou překrývající se informaci. Podmíněné asociace se liší od celkového sklonu.

JEŠTĚ NEVÍME

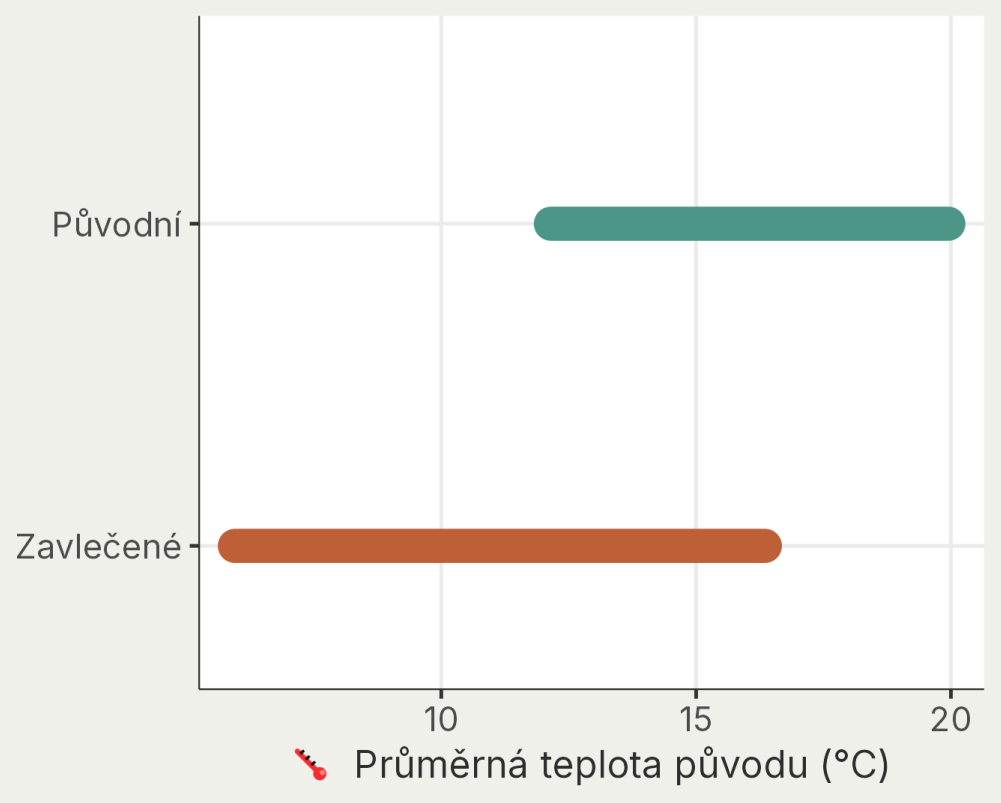
Jak se skupiny liší při stejné, daty podložené teplotě?

Predikce jen v podložené oblasti

Pro obě skupiny potřebujeme společnou teplotu



Která teplota leží v překryvu obou skupin?



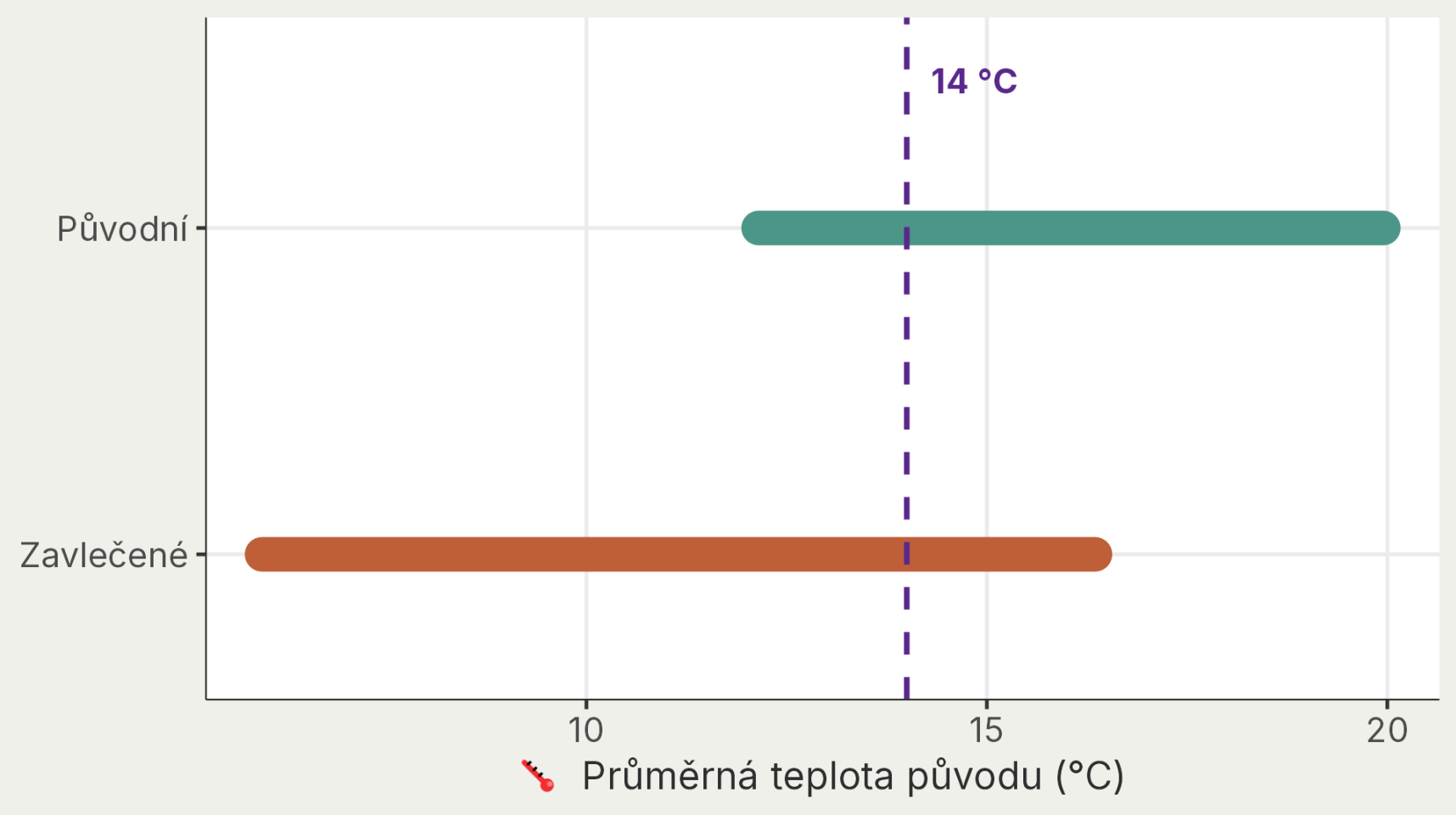
Vyberte podložené srovnání:

A) 21 °C

B) 1 °C

C) 14 °C

14 °C leží uvnitř překryvu obou skupin



Obě skupiny mají poblíž pozorované populace.

Dvě modelové populace se liší pouze původem

```r
teplota_srovnani_C <- 14

data_scenare <- data.frame(
  puvod = c("Původní", "Zavlečené"),
  teplota_rocni_C = rep(
    x = teplota_srovnani_C,
    times = 2
  )
)
```

puvod	teplota_rocni_C
Původní	14
Zavlečené	14

> Teplotu držíme **stejnou**. Měníme **pouze původ**.

Odhad při 14 °C

Odhad = intercept 0,31143 + posun zavlečených 0,10895 + sklon 0,00149 × teplota

Původní

$$0{,}31143 + 0 + 0{,}00149 \times 14 = \mathbf{0{,}332}\ \text{mm}$$

Zavlečené

$$0{,}31143 + 0{,}10895 + 0{,}00149 \times 14 = \mathbf{0{,}441}\ \text{mm}$$

```
predict(
  object = mod_puvod,
  newdata = data_scenare
)
```

```
        1         2 
0.3323001 0.4412474 
```

Bodový odhad ještě neukazuje přesnost

```
1  coef(object = mod_puvod)
```

```
   (Intercept)  puvodZavlečené teplota_rocni_C
   0.311426537     0.108947301     0.001490971
```

Jak přesně známe oba modelové průměry při 14 °C?

Potřebujeme intervaly spolehlivosti!

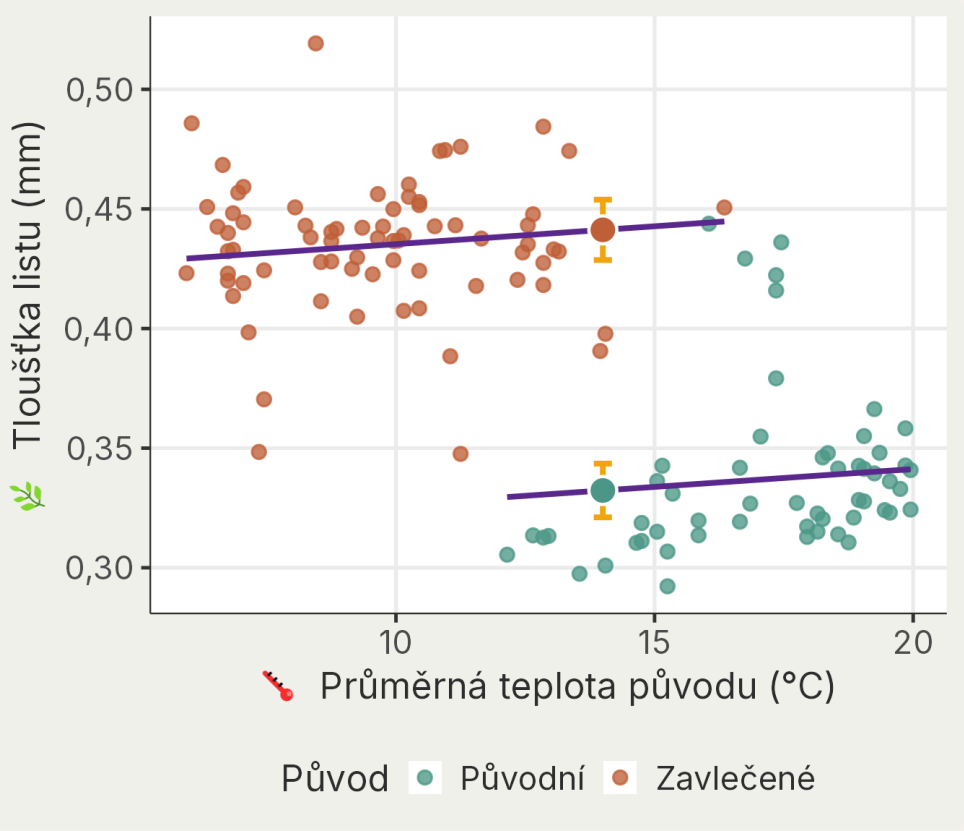
Modelové průměry doplníme o intervaly spolehlivosti

```
emmeans::emmeans(
  object = mod_puvod,
  specs = ~ puvod,
  at = list(teplota_rocni_C = teplota_srovnani_C)
)
```

Původ	Průměr (mm)	95% interval
Původní	0,332	0,321 – 0,344
Zavlečené	0,441	0,429 – 0,454

> Intervaly obou průměrů se nepřekrývají: rozdíl mezi areály je zřetelný.

Co zjistíme při stejné teplotě?



Napište jednu větu:

- 14 °C jako srovnávací podmínku;
- rozdíl přibližně 0,109 mm;
- intervaly: původní 0,321–0,344 mm, zavlečené 0,429–0,454 mm;
- slova „souvisí“ nebo „asociace“, nikoli „způsobuje“.

Model spočítá i kombinaci mimo pozorovaný mrak



Ilustrace vytvořená pomocí AI.

Numericky platné

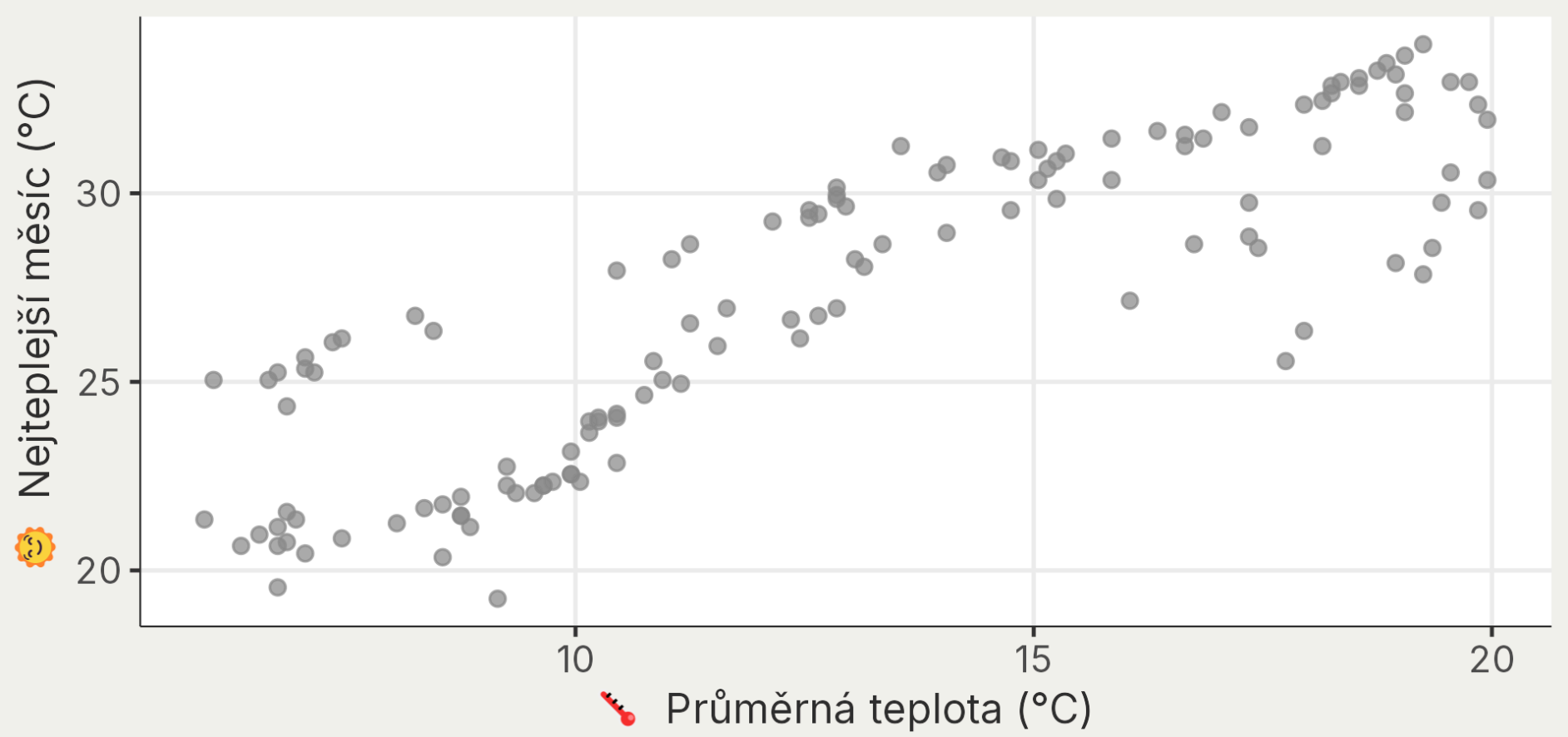
`predict()` vrátí číslo.

Empiricky podložené

V datech existují blízké kombinace obou prediktorů.

> Dva jednotlivě možné údaje nemusí tvořit společně realistický scénář.

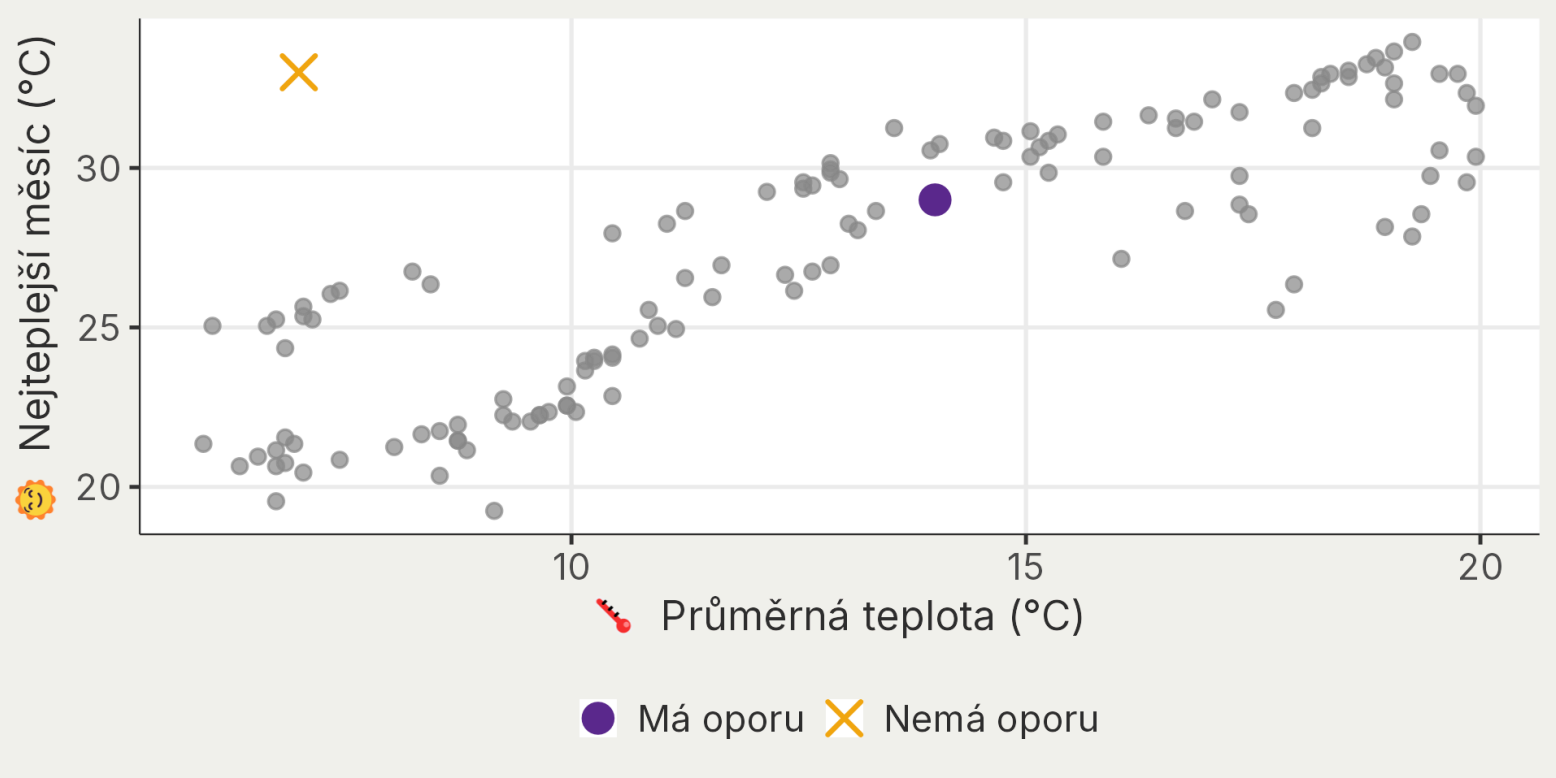
Který scénář má oporu v datech?



A) průměrná 14 °C, maximum 29 °C

B) průměrná 7 °C, maximum 33 °C

Jedna kombinace leží v mraku, druhá mimo něj



A má oporu: poblíž leží pozorované kombinace. **B oporu nemá:** oba údaje se v datech vyskytují, ale ne společně.

Pořadí prediktorů

Změní pořadí prediktorů výsledek modelu?

```r
mod_dve_teploty <- lm(
  formula = tloustka_listu_mm ~ teplota_rocni_C + teplota_max_C,
  data = data_kridlatka
)
mod_dve_teploty_opacne <- lm(
  formula = tloustka_listu_mm ~ teplota_max_C + teplota_rocni_C,
  data = data_kridlatka
)
```

Co se může změnit?	Ano / ne?
odhadnuté hodnoty	?
koeficienty a jejich testy	?
výsledek `anova()`	?

Koeficienty se po změně pořadí nezmění

Koeficient	průměrná + maximum	maximum + průměrná
průměrná teplota	−0,00528	−0,00528
maximální teplota	−0,00475	−0,00475

- stejné odhadnuté hodnoty (`all.equal()`: TRUE);
- stejná residua (`all.equal()`: TRUE);
- stejné koeficienty a standardní chyby.

Podmíněná otázka modelu se pořadím zápisu nemění.

Sdílenou informaci si vezme první prediktor ve formuli



Ilustrace vytvořená pomocí AI.

Sekvenční test se ptá zleva doprava.

První prediktor dostane i sdílenou část.

Druhý prediktor dostane pouze informaci navíc.

Sekvenční test čte formuli zleva doprava

Pořadí	Termín	Součet čtverců	F	p
průměrná → max	průměrná	0,20170	123,06	$< 2\times10^{-16}$
	maximum po průměrné	0,01326	8,09	0,005193
max → průměrná	maximum	0,19679	120,07	$< 2\times10^{-16}$
	průměrná po maximu	0,01817	11,09	0,001142

Sdílená informace se započítá prvnímu termínu. Proto se řádky po prohození změní.

Pořadí mění sekvenční test, nikoli fit modelu

NEMĚNÍ SE

koeficienty, residua, testy koeficientů

MĚNÍ SE

sekvenční součty čtverců, F a p v `anova()`

> Přehazování pořadí není strategie výběru prediktorů ani cesta k „hezčí“ p-hodnotě.

Závěr

Jak postupovat u modelu s více prediktory

1. Otázka
Biologická otázka a jednotka analýzy

2. Data
Prediktory prohlédněte i společně

3. Model
Koeficienty při stejných ostatních prediktorech

4. Predikce
Jen v podložené oblasti

5. Kontrola
Residua · nejistota · typ testu

6. Výklad
Zpět k biologii

Společná zahrada sjednocuje prostředí, ne historii populací



Ilustrace vytvořená pomocí AI.

V zahradě stejné: aktuální podmínky pěstování

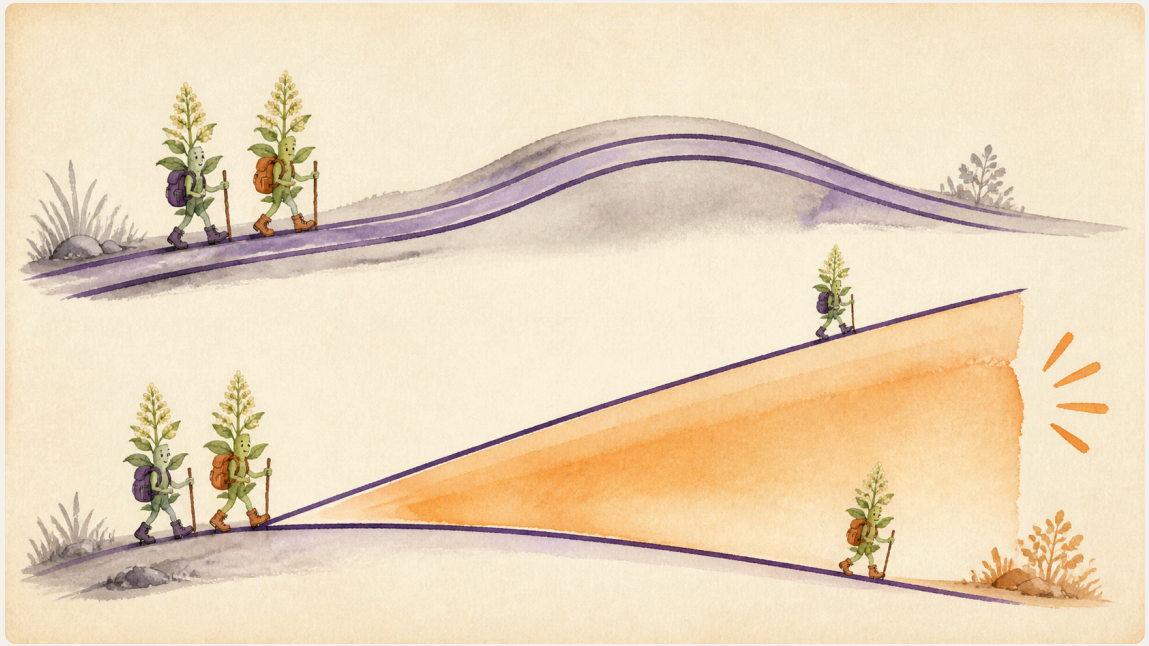
Mezi populacemi různé: klima místa původu, geografická a evoluční historie

Proto mluvíme o **asociaci**, ne o důkazu, že klima rozdíly **způsobilo**.

Shrnutí

- Koeficient popisuje rozdíl **při stejných hodnotách ostatních prediktorů**.
- Přidání prediktoru mění **otázku, nejen číslo**; celkový vztah může mísit více cest.
- **Korelované prediktory** sdílejí informaci a zvětšují nejistotu odhadů.
- Skupiny porovnáváme a predikujeme jen pro **kombinace hodnot podložené daty**.
- Pořadí mění **sekvenční test** v `anova()`, ne koeficienty v `summary()`.

Kdy by rovnoběžné přímky nestačily?



Ilustrace vytvořená pomocí AI.

Jaký biologický vzorec by vyžadoval různé sklony pro obě skupiny?

To je otázka pro další lekci.