

Co se změní, když přidáme prediktor?

Vícenásobný lineární model na klimatických datech křídlatky

11. října 2026

Obsah

1 Úvod	1
2 Výsledky učení	2
3 Každý řádek je jedna zdrojová populace	2
4 Populace z teplejších oblastí mají v průměru tenčí listy	10
5 Jedna teplota, jeden sklon	11
5.1 Sedí přímka k datům?	13
6 Také nejteplejší měsíc souvisí s tloušťkou listu	14
7 Jeden model, dvě teploty	17
8 Klimatické proměnné nenesou nezávislé informace	19
9 Proč se odhad roční teploty změnil?	22
9.1 Změnil druhý prediktor strukturu residuí?	24
10 Původ populací mění příběh vztahu	25
10.1 Sedí model s teplotou a původem k datům?	29
11 Při stejném původu sklon teploty téměř mizí	31
12 Jak porovnat skupiny při stejné teplotě?	33
13 Jak přesně známe modelové průměry?	36
14 Záleží u testu na pořadí prediktorů?	39
14.1 Co se změní v tabulce koeficientů?	40
14.2 Sekvenční analýza rozptylu klade otázky zleva doprava	41
14.3 Výběr prediktorů patří do dalších lekcí	45
15 Jak postupovat u modelu s více prediktory	45
16 Co tedy můžeme říct o křídlatce?	45
17 Co bude následovat dál	46
18 Shrnutí	46
18.1 Závěrečná otázka	46
Slovníček pojmů	47

1 Úvod

V předchozích lekcích jsme pracovali s lineárním modelem, který měl jediný prediktor. Číselný prediktor, například zeměpisná šířka, nám dal sklon: o kolik se v průměru liší odezva při rozdílu prediktoru o jednu jednotku. Kategoriální prediktor, například druh tučňáka, nám dal rozdíly mezi

skupinami vůči referenční skupině. Biologické vlastnosti se ale v přírodě obvykle nemění jedna po druhé. Teplé místo bývá zároveň místem s horkým létem, jiným úhrnem srážek a někdy i jinou historií populací.

Budeme sledovat křídlatku japonskou (*Reynoutria japonica*), jednu z nejznámějších invazních rostlin Evropy i Severní Ameriky. Výzkumníci odebrali oddenky populací z původního areálu ve východní Asii i ze zavlečeného areálu a pěstovali je vedle sebe ve **společné zahradě**. Všechny rostliny tak rostly ve stejném aktuálním prostředí a rozdíly mezi nimi mohou odrážet hlavně to, odkud populace pocházejí.

Hlavní otázka této lekce

Mění klima původu tloušťku listů křídlatky?

Co čekáte ještě před prvním grafem?

Představte si, že populace z teplejších míst mají v průměru tenčí listy. Znamená to už, že teplo listy ztenčuje? Co dalšího se může s teplotou místa původu měnit zároveň?

Možná odpověď

Samotná souvislost ještě nevysvětluje proč. S teplotou se obvykle mění i další části klimatu, například nejteplejší měsíc nebo srážky. Teplá a chladná místa se navíc mohou lišit tím, zda jde o původní, nebo zavlečené populace s odlišnou historií. Všechny tyto vlastnosti se mohou měnit společně. V této lekci uvidíme, jak takové souběžné vlastnosti zapojit do jednoho modelu a jak se tím změní otázka, na kterou model odpovídá.

2 Výsledky učení

Po prostudování této lekce budete schopni:

- fitovat aditivní lineární 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 podmínný 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()`.

3 Každý řádek je jedna zdrojová populace

Data pocházejí ze studie Cao et al. (2025). Autoři pěstovali populace křídlatky z Číny, Japonska, Evropy a Severní Ameriky ve dvou společných zahradách v Číně; my používáme zahradu v

Šanghaji. Tloušťku listů jednotlivých rostlin jsme zprůměrovali za každou populaci. Klimatické údaje o místě původu každé populace jsme doplnili z otevřeného souboru CHELSA-BIOCLIM 2.1, který shrnuje dlouhodobé klima let 1981–2010.

Soubor [kridlatka_populace.csv](#) je součástí materiálů. Pro následující kód jej uložte do složky `data` ve svém projektu v R. Načteme ho a zkontrolujeme názvy a typy sloupců:

```
# Načteme tabulku s jedním řádkem pro každou zdrojovou populaci
data_kridlatka <-
  read.csv(file = "data/kridlatka_populace.csv")

# Zkontrolujeme názvy a typy sloupců
str(object = data_kridlatka)
```

```
'data.frame':  128 obs. of  14 variables:
 $ populace      : chr  "CN1" "CN10" "CN11" "CN12" ...
 $ oblast        : chr  "CN" "CN" "CN" "CN" ...
 $ puvod         : chr  "Puvodni" "Puvodni" "Puvodni" "Puvodni" ...
 $ zemepisna_sirka : num  29.3 26.8 26.6 26.3 26 ...
 $ zemepisna_delka : num  118 116 116 116 116 ...
 $ tloustka_listu_mm : num  0.346 0.343 0.323 0.328 0.323 ...
 $ teplota_rocni_C : num  18.3 18.9 18.1 19.1 19.6 ...
 $ teplota_sezonnost : num  818 771 706 712 720 ...
 $ teplota_max_C   : num  32.9 33.1 31.3 32.1 33 ...
 $ teplota_min_C   : num  1.65 2.45 2.45 3.05 3.35 ...
 $ srazky_rocni_mm : num  2073 1874 2484 1906 1945 ...
 $ srazky_nejvlhci_mesic_mm: num  382 325 421 323 317 ...
 $ srazky_nejsussi_mesic_mm: num  52.8 46.8 60.3 45.6 47.3 43.9 37.4 38.3
 41.7 42.1 ...
 $ srazky_sezonnost : num  56.5 58.1 58.4 59.6 59.2 56.6 55 54.6
 54.9 56.5 ...
```

Sloupec `puvod` obsahuje text `Puvodni` nebo `Zavleceeny`. Převédeme ho na faktor a pořadím úrovní určíme, že referenční skupinou budou původní populace. Potom zkontrolujeme rozsahy sloupců, které použijeme v modelech, a chybějící hodnoty:

```
# Původ převedeme na faktor; první úroveň bude referenční skupinou
data_kridlatka$puvod <-
  factor(
    x = data_kridlatka$puvod,
    levels = c("Puvodni", "Zavleceeny")
  )
```

```
# Počet populací v původním a zavlečeném areálu
table(data_kridlatka$puvod)
```

```
Puvodni  Zavleceny
      55      73
```

```
# Vybereme sloupce, které použijeme v hlavních modelech
vec_sloupce_modelu <-
  c("tloustka_listu_mm", "teplota_rocni_C", "teplota_max_C", "puvod")

# Minimum, kvartily, průměr a maximum vybraných sloupců
summary(object = data_kridlatka[, vec_sloupce_modelu])
```

```
tloustka_listu_mm  teplota_rocni_C  teplota_max_C      puvod
Min.   :0.2922    Min.   : 5.95    Min.   :19.25    Puvodni   :55
1st Qu.:0.3353    1st Qu.: 9.25    1st Qu.:23.52    Zavleceny :73
Median :0.4180    Median :12.60    Median :27.90
Mean   :0.3928    Mean   :12.89    Mean   :27.17
3rd Qu.:0.4401    3rd Qu.:17.12    3rd Qu.:30.77
Max.   :0.5192    Max.   :19.95    Max.   :33.95
```

```
# Počet chybějících hodnot v každém sloupci
colSums(x = is.na(x = data_kridlatka))
```

```
      populace      oblast      puvod
      0          0          0
zemepisna_sirka  zemepisna_delka  tloustka_listu_mm
      0          0          0
teplota_rocni_C  teplota_sezonnost  teplota_max_C
      0          0          0
teplota_min_C    srazky_rocni_mm  srazky_nejvlhci_mesic_mm
      0          0          0
srazky_nejsussi_mesic_mm  srazky_sezonnost
      0          0
```

Tabulka obsahuje 128 populací: 55 z původního a 73 ze zavlečeného areálu. Žádný sloupec nemá chybějící hodnoty. Průměrná tloušťka listu se pohybuje od 0,29 do 0,52 mm a průměrná roční teplota místa původu od 5,95 do 19,95 °C. Takové hodnoty odpovídají listům bylin i klimatu

mírného a subtropického pásu, takže nic nenaznačuje chybu jednotek. Rozsahy teplot zároveň vymezují podmínky, pro které jsou pozdější predikce podloženy daty.

Význam jednotlivých sloupců je:

Sloupec	Co obsahuje	Role v analýze
populace	identifikátor zdrojové populace	jednotka pozorování
oblast	země nebo oblast původu	popis
puvod	původní nebo zavlečený areál	potenciální kategoriální prediktor
zemepisna_sirka, zemepisna_delka	souřadnice místa původu	popis
tloustka_listu_mm	průměrná tloušťka listu v mm	odezva
teplota_rocni_C	průměrná roční teplota v °C	potenciální spojitý klimatický prediktor
teplota_sezonnost	proměnlivost teplot během roku	potenciální spojitý klimatický prediktor
teplota_max_C	maximum teploty nejteplejšího měsíce v °C	potenciální spojitý klimatický prediktor
teplota_min_C	minimum teploty nejchladnějšího měsíce v °C	potenciální spojitý klimatický prediktor
srazky_rocni_mm	roční úhrn srážek v mm	potenciální spojitý klimatický prediktor
srazky_nejvlhci_mesic_mm	srážky nejvlhčího měsíce v mm	potenciální spojitý klimatický prediktor
srazky_nejsussi_mesic_mm	srážky nejsuššího měsíce v mm	potenciální spojitý klimatický prediktor
srazky_sezonnost	proměnlivost srážek během roku	potenciální spojitý klimatický prediktor

Co je jednotkou analýzy?

Rostliny ze stejné populace mají stejné klima původu. Proč by bylo chybou analyzovat všechny rostliny jako nezávislá pozorování klimatu?

Možná odpověď

Klimatický prediktor se mění jen mezi populacemi, nikoli mezi rostlinami jedné populace. Několik rostlin z jednoho místa nepřináší několik nezávislých informací o klimatu, ale jedenkrát totéž klima. Kdybychom je počítali jako samostatná pozorování, nadhodnotili bychom množství informace a model by působil přesněji, než ve skutečnosti je. Proto je jedním řádkem a jednou jednotkou analýzy zdrojová populace.

Doplňující: stažení a příprava dat

Pro běžnou práci používáme hotovou lokální tabulku, takže není potřeba internetové připojení. Volitelný postup níže umožňuje celý vznik tabulky zopakovat: stáhne původní znaky z repozitáře Dryad, ověří soubor, získá klimatická data CHELSA a spojí oba zdroje. Jednotlivé rostliny shrneme na úroveň populace, protože klima se mezi rostlinami stejné populace nemění. Pokud automatické stažení z Dryad selže, stáhněte soubor herb-trait.csv ze [stránky datasetu](#) ručně do složky data-raw; kontrolní součet v kódu ověří správnou verzi.

```
# Balíčky stačí nainstalovat jednou:
# install.packages(pkgs = c("digest", "here", "terra"))

# 1. Připravíme adresáře pro stažená a výsledná data
soubor_znaky <- here::here("data-raw", "herb-trait.csv")
soubor_vyuka <- here::here("data", "kridlatka_populace.csv")

dir.create(path = dirname(path = soubor_znaky), recursive = TRUE,
showWarnings = FALSE)
dir.create(path = dirname(path = soubor_vyuka), recursive = TRUE,
showWarnings = FALSE)

# 2. Stáhneme původní tabulku znaků přímo z Dryad
url_znaky <- "https://datadryad.org/downloads/file_stream/3893167"

if (
  !file.exists(soubor_znaky)) {
  download.file(
    url = url_znaky,
    destfile = soubor_znaky,
    mode = "wb"
  )
}

# Kontrolní součet odhalí neúplné nebo později změněné stažení
sha256_ocekavany <- paste0(
```

```

    "c6a5afae067597870156fc1fe4b8e688",
    "3784d5eb40f967a2ca95dde6d4fc74a8"
  )
sha256_stazeny <- digest::digest(
  object = soubor_znaky,
  algo = "sha256",
  file = TRUE
)

stopifnot(identical(x = sha256_stazeny, y = sha256_ocekavany))

# 3. Načteme tabulku a dáme sloupcům krátké pracovní názvy
data_znaky <- read.csv(
  file = soubor_znaky,
  check.names = FALSE
)

names(data_znaky) <- c(
  "garden", "range", "population", "individual", "latitude", "longitude",
  "block", "root_weight_g", "source", "leaf_thickness_mm",
  "leaf_toughness_n", "leaf_cn", "alkaloids_mg_g", "flavonoids_mg_g",
  "lignin_mg_g", "herb_beetle_pct", "herb_caterpillar_pct"
)

# 4. Vybereme šanghajskou zahradu a vytvoříme průměr za populaci
data_znaky_shanghai <- data_znaky[data_znaky$garden == "SH", ]

data_populace <- aggregate(
  x = leaf_thickness_mm ~ population + range + source + latitude + longitude,
  data = data_znaky_shanghai,
  FUN = mean
)

# 5. Otevřeme osm vrstev CHELSA-BIOCLIM 2.1 (1981–2010)
vec_vybrane_bioclim <- c(
  "bio1", "bio4", "bio5", "bio6",
  "bio12", "bio13", "bio14", "bio15"
)
zaklad_klima <- paste0(
  "https://os.zhdk.cloud.switch.ch/chelsav2/GLOBAL/",
  "climatologies/1981-2010/bio/"
)

```

```

Sys.setenv(
  GDAL_DISABLE_READDIR_ON_OPEN = "YES",
  CPL_VSIL_CURL_ALLOWED_EXTENSIONS = ".tif"
)
vec_adresy_vrstev <- paste0(
  "/vsicurl/", zaklad_klima, "CHELSA_", vec_vybrane_bioclim,
  "_1981-2010_V.2.1.tif"
)

# 6. Z vrstev vyzvedneme klima v souřadnicích každé populace
body_populaci <- terra::vect(
  x = data_populace,
  geom = c("longitude", "latitude"),
  crs = "EPSG:4326"
)
list_klima_populaci <- lapply(
  X = vec_adresy_vrstev,
  FUN = function(adresa) {
    vrstva <- terra::rast(x = adresa)
    terra::extract(x = vrstva, y = body_populaci, ID = FALSE)[[1]]
  }
)
data_klima_populaci <- as.data.frame(list_klima_populaci)
names(data_klima_populaci) <- vec_vybrane_bioclim

# 7. Spojíme znaky s osmi vybranými BIOCLIM proměnnými
data_vyuka <- cbind(data_populace, data_klima_populaci)

# 8. Použijeme stejné české názvy a kódování původu jako v lekci
names(data_vyuka) <- c(
  "populace", "oblast", "puvod", "zemepisna_sirka", "zemepisna_delka",
  "tloustka_listu_mm", "teplota_rocni_C", "teplota_sezonnost",
  "teplota_max_C", "teplota_min_C", "srazky_rocni_mm",
  "srazky_nejvlhci_mesic_mm", "srazky_nejsussi_mesic_mm",
  "srazky_sezonnost"
)

data_vyuka$puvod <- ifelse(
  test = data_vyuka$puvod == "Native",
  yes = "Puvodni",
  no = "Zavleceeny"
)

```

```

data_vyuka <- data_vyuka[order(data_vyuka$populace), ]

# 9. Před uložením ověříme očekávaný počet a úplnost dat
stopifnot(
  nrow(x = data_vyuka) == 128,
  !anyNA(x = data_vyuka),
  abs(x = cor(x = data_vyuka$teplota_rocni_C, y = data_vyuka$teplota_max_C)
- 0.855) < 0.002
)

write.csv(
  x = data_vyuka,
  file = soubor_vyuka,
  row.names = FALSE,
  na = ""
)

```

Původní tabulka z Dryad má přibližně 121 kB. Osm vrstev CHELSA-BIOCLIM čteme po malých úsecích přímo ze vzdálených souborů, takže postup vyžaduje internetové připojení a může několik minut běžet. Vrstva má rozlišení 30 úhlových sekund (přibližně 1 km). Studie Cao et al. použila jiný klimatický zdroj; naše výukové hodnoty proto nejsou kopií klimatické tabulky studie.

Zdroj znaků: Cao et al. (2025), Dryad, CC0, DOI [10.5061/dryad.qbzh18r4](https://doi.org/10.5061/dryad.qbzh18r4). Klimatické vrstvy: [CHELSA-BIOCLIM 2.1](#), období 1981–2010, CC0; Brun et al. (2022), DOI [10.16904/envidat.332](https://doi.org/10.16904/envidat.332).

4 Populace z teplejších oblastí mají v průměru tenčí listy



Obrázek 1: Průměrná tloušťka listu a průměrná roční teplota místa původu. Jeden bod je jedna populace pěstovaná v Šanghaji.

Co dokážeme vyčíst z grafu?

Popište směr a těsnost vztahu. Jak těsně body sledují klesající směr a vidíte neobvyklá pozorování?

Možná odpověď

Body převážně klesají: populace z teplejších míst mají v průměru tenčí listy. Vztah je zřetelný, ale body neleží na přímce; při stejné teplotě se tloušťka populací liší. Žádný bod nestojí úplně mimo ostatní, i když několik populací leží výrazněji nad nebo pod hlavním mrakem.

Z předchozích lekcí známe korelaci, která směr a těsnost takového vztahu shrne jedním číslem:

```
# Pearsonova korelace roční teploty a tloušťky listu
cor(
  x = data_kridlatka$teplota_rocni_C,
```

```
y = data_kridlatka$tloustka_listu_mm  
)
```

```
[1] -0.6931225
```

Záporná hodnota $-0,69$ potvrzuje klesající a poměrně těsný vztah. Korelace však nemá jednotky; neřekne nám, o kolik milimetrů se tloušťka listu liší při rozdílu teploty $1\text{ }^{\circ}\text{C}$. K tomu použijeme známý lineární model.

5 Jedna teplota, jeden sklon

Model s průměrnou roční teplotou jako jediným prediktorem fitujeme stejně jako v předchozích lekcích:

```
# Fitujeme přímku s jedinou klimatickou proměnnou  
mod_rocni_teplota <-  
  lm(  
    formula = tloustka_listu_mm ~ teplota_rocni_C,  
    data = data_kridlatka  
  )
```

```
# Zobrazíme celý přehled modelu  
summary(object = mod_rocni_teplota)
```

```
Call:  
lm(formula = tloustka_listu_mm ~ teplota_rocni_C, data = data_kridlatka)
```

```
Residuals:  
      Min       1Q   Median       3Q      Max   
-0.095017 -0.026242  0.003225  0.025530  0.091216
```

```
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)      
(Intercept)   0.5105454  0.0115134   44.34  <2e-16 ***  
teplota_rocni_C -0.0091332  0.0008462  -10.79  <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.04161 on 126 degrees of freedom  
Multiple R-squared:  0.4804,    Adjusted R-squared:  0.4763   
F-statistic: 116.5 on 1 and 126 DF,  p-value: < 2.2e-16
```

Co už umíme přečíst z výstupu?

Najděte ve výstupu řádek `teplota_rocni_C` a odpovězte:

1. Jaké znaménko má odhad koeficientu?
2. V jakých jednotkách koeficient interpretujeme?
3. Jakou změnu tloušťky model očekává při rozdílu 5 °C?

Možná odpověď

Koeficient je záporný, $-0,00913$ mm na 1 °C. Dvě populace, jejichž místa původu se liší průměrnou roční teplotou o 1 °C, se podle modelu liší v tloušťce listu v průměru o 0,00913 mm; teplejší populace má tenčí list. Výpočet pro rozdíl 5 °C ukazuje následující odstavec.

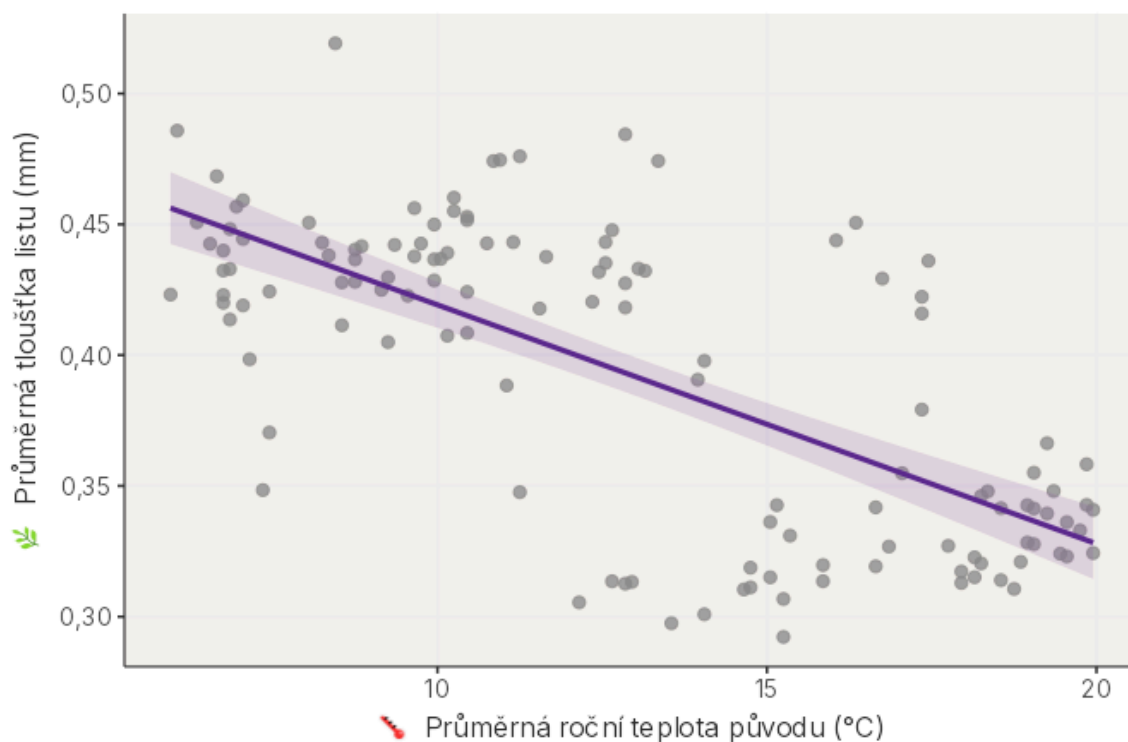
Odhadnutý rozdíl tloušťky dvou populací získáme vynásobením sklonu rozdílem jejich teplot:

$$\text{odhadnutý rozdíl tloušťky} = \text{sklon roční teploty} \times \text{rozdíl roční teploty} .$$

Pro rozdíl 5 °C to je

$$-0,00913 \times 5 = -0,046 \text{ mm}.$$

Populace z místa teplejšího o 5 °C má tedy podle modelu v průměru list tenčí asi o 0,046 mm. Přímka v grafu ukazuje odhadnutou tloušťku pro jednotlivé roční teploty a fialový pás nejistotu průměrného trendu:



Obrázek 2: Pozorované populace a trend odhadnutý modelem s průměrnou roční teplotou.

Doplňující: údaj o popsané variabilitě

Ve výstupu `summary()` je také $R^2 = 0,480$. Tato hodnota říká, že přímka s roční teplotou popisuje přibližně 48 % variability populačních průměrů tloušťky. R^2 samo nerozhoduje, který z několika modelů máme upřednostnit; jeho využití při porovnávání modelů patří do pozdější lekce.

5.1 Sedí přímka k datům?

Po interpretaci koeficientu zkontrolujeme residua: rozdíly mezi pozorovanou tloušťkou a hodnotou na přímce. Vykreslíme je proti hodnotám odhadnutým modelem:

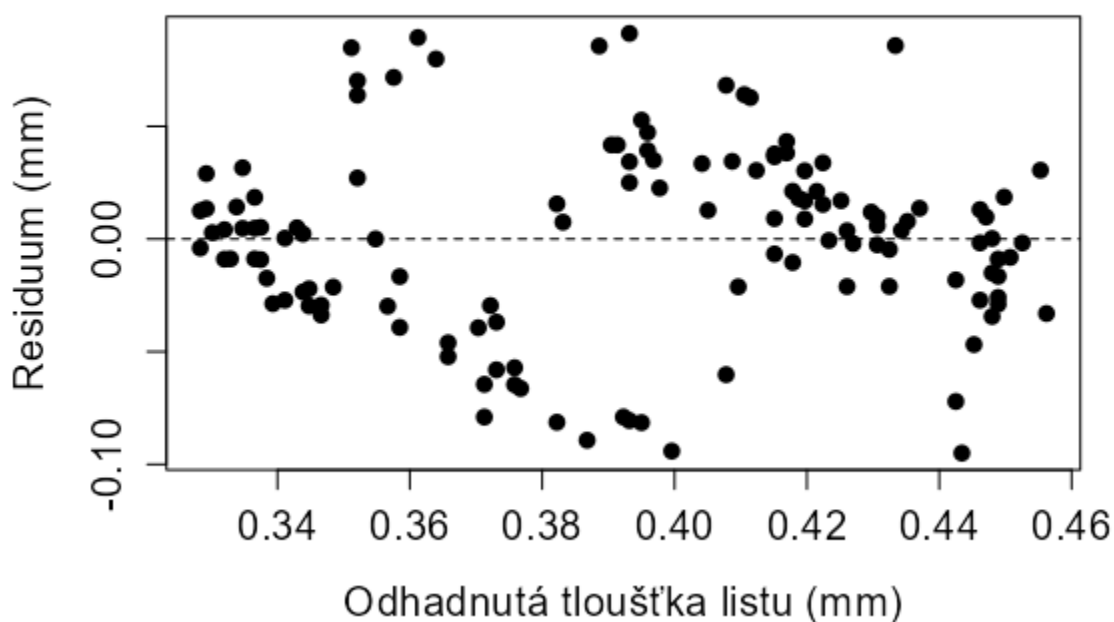
```
# Z modelu získáme residua a odhadnuté hodnoty
vec_residua_rocni <- resid(object = mod_rocni_teplota)
vec_odhad_rocni <- fitted(object = mod_rocni_teplota)

# Residua proti odhadnutým hodnotám s přerušovanou čarou v nule
plot(
  x = vec_odhad_rocni,
  y = vec_residua_rocni,
  xlab = "Odhadnutá tloušťka listu (mm)",
```

```

ylab = "Residuum (mm)",
pch = 16,
cex = 1.2,
cex.axis = 1.3,
cex.lab = 1.3
)
abline(h = 0, lty = 2)

```



Body by měly být rozptýlené kolem vodorovné nuly bez nápadného oblouku nebo trychtýře. Oblouk ani trychtýř zde nevidíme, takže přímka je pro hrubý popis vztahu přijatelná. Body ale nejsou rozmístěné úplně náhodně: při nízkých až středních odhadnutých hodnotách leží spíš pod nulou, potom spíš nad ní. Přímka tedy něco nezachytila. Residua jsou navíc poměrně velká: roční teplota zdaleka nevysvětluje všechny rozdíly mezi populacemi. Co v residuích zůstalo, ukáže část o původu populací.

6 Také nejteplejší měsíc souvisí s tloušťkou listu

V tabulce máme také maximum teploty nejteplejšího měsíce. Stejným postupem ověříme, jak souvisí s tloušťkou listu samo o sobě:

```
# Fitujeme přímku s maximem nejteplejšího měsíce
mod_max_teplota <-
  lm(
    formula = tloustka_listu_mm ~ teplota_max_C,
    data = data_kridlatka
  )

# Zobrazíme celý přehled modelu
summary(object = mod_max_teplota)
```

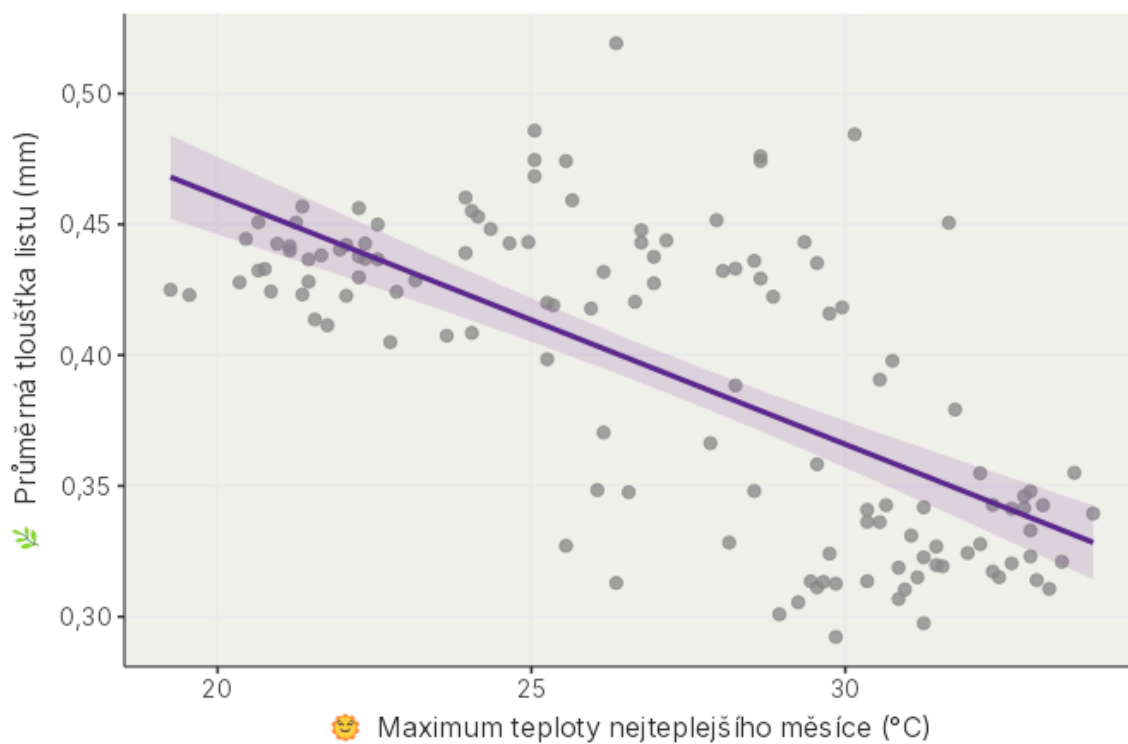
```
Call:
lm(formula = tloustka_listu_mm ~ teplota_max_C, data = data_kridlatka)

Residuals:
      Min       1Q   Median       3Q      Max
-0.087708 -0.026667 -0.007747  0.029279  0.119924

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6509898   0.0247701   26.28  <2e-16 ***
teplota_max_C -0.0095029   0.0009013  -10.54  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04207 on 126 degrees of freedom
Multiple R-squared:  0.4687,    Adjusted R-squared:  0.4645
F-statistic: 111.2 on 1 and 126 DF,  p-value: < 2.2e-16
```

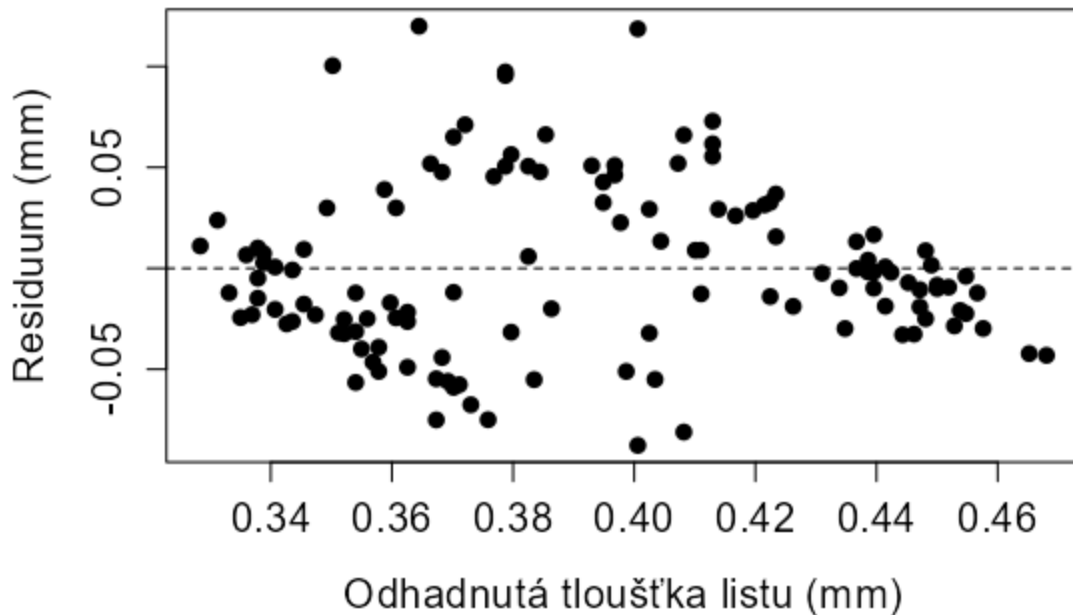
Model odhaduje rozdíl $-0,00950$ mm při maximu nejteplejšího měsíce vyšším o 1 °C. Také tento sklon popisuje celkový vztah k jedné části teplotního režimu, bez ohledu na ostatní klimatické vlastnosti.



Obrázek 3: Pozorované populace a trend odhadnutý modelem s maximem teploty nejteplejšího měsíce.

Také u tohoto modelu zkontrolujeme residua stejným grafem:

```
# Residua proti odhadnutým hodnotám modelu s maximem
plot(
  x = fitted(object = mod_max_teplota),
  y = resid(object = mod_max_teplota),
  xlab = "Odhadnutá tloušťka listu (mm)",
  ylab = "Residuum (mm)",
  pch = 16,
  cex = 1.2,
  cex.axis = 1.3,
  cex.lab = 1.3
)
abline(h = 0, lty = 2)
```



Residua jsou rozptýlená kolem nuly bez výrazného oblouku nebo trychtýře. Oba jednoduché modely tedy ukazují klesající klimatický trend. Každý z nich ale používá jen jednu část teplotního režimu a kolem přímky zůstává značná variabilita.

7 Jeden model, dvě teploty

Chceme odhadovat tloušťku listu podle průměrné roční i maximální teploty současně. Do formule přidáme druhý prediktor znaménkem +:

```
# Jeden model se dvěma teplotními prediktory
mod_dve_teploty <-
  lm(
    formula = tloustka_listu_mm ~ teplota_rocni_C + teplota_max_C,
    data = data_kridlatka
  )

# Zobrazíme celý přehled modelu
summary(object = mod_dve_teploty)
```

```
Call:
lm(formula = tloustka_listu_mm ~ teplota_rocni_C + teplota_max_C,
    data = data_kridlatka)

Residuals:
    Min       1Q   Median       3Q      Max
-0.081384 -0.024798 -0.003083  0.025821  0.105537

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.589936   0.030071  19.618 < 2e-16 ***
teplota_rocni_C -0.005279   0.001585  -3.330  0.00114 **
teplota_max_C   -0.004751   0.001670  -2.845  0.00519 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04048 on 125 degrees of freedom
Multiple R-squared:  0.512, Adjusted R-squared:  0.5042
F-statistic: 65.58 on 2 and 125 DF, p-value: < 2.2e-16
```

Znaménko + ve formuli říká, že model odhaduje příspěvek **obou prediktorů současně**. Neznamená, že se dvě naměřené teploty prostě sečtou. Koeficient roční teploty popisuje její vztah k tloušťce **při stejném maximu nejteplejšího měsíce**; koeficient maxima popisuje jeho vztah **při stejné roční teplotě**. Takovému modelu, ve kterém se příspěvky prediktorů jen sčítají, říkáme **aditivní model**.

Které tři parametry model odhadl?

Ve výstupu `summary()` najděte tři parametry modelu. Jak byste každý z nich popsali jednou větou a v jakých jednotkách?

Možná odpověď

Řádek **(Intercept)** je intercept, 0,590 mm: odhadnutá tloušťka při nulové roční i nulové maximální teplotě. Taková kombinace v datech není, takže intercept slouží hlavně jako výchozí bod rovnice. Koeficient **teplota_rocni_C**, -0,00528 mm na 1 °C, porovnává populace se stejným maximem nejteplejšího měsíce, které se liší roční teplotou. Koeficient **teplota_max_C**, -0,00475 mm na 1 °C, porovnává populace se stejnou roční teplotou, které se liší maximem nejteplejšího měsíce.

Slovně model skládá odhadnutou hodnotu ze tří částí:

odhadnutá tloušťka listu = intercept
 + sklon roční teploty × roční teplota
 + sklon maxima × maximum nejteplejšího měsíce .

Konkrétní dosazení ukážeme na jedné skutečné populaci:

```
# Vybereme populaci CN35 a její dvě teploty
data_priklad <-
  data_kridlatka[
    data_kridlatka$populace == "CN35",
    c("populace", "teplota_rocni_C", "teplota_max_C")
  ]

data_priklad
```

	populace	teplota_rocni_C	teplota_max_C
29	CN35	15.25	29.85

Populace CN35 pochází z místa s průměrnou roční teplotou 15,25 °C a maximem nejteplejšího měsíce 29,85 °C. Dosazení odhadnutých parametrů a těchto teplot je

$$\begin{aligned}\widehat{\text{tloušťka}} &= 0,5899 + (-0,00528) \times 15,25 \\ &\quad + (-0,00475) \times 29,85 \\ &= 0,5899 - 0,0805 - 0,1418 \\ &= 0,368 \text{ mm.}\end{aligned}$$

Stejnou strukturu zapisujeme pro dva prediktory obecně jako

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

a pro p prediktorů jako

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ji} + \varepsilon_i.$$

Y_i je tloušťka listu populace i ; β_0 je intercept; β_j je koeficient prediktoru j ; X_{ji} je hodnota prediktoru j u populace i ; p je počet prediktorů; j označuje konkrétní prediktor a ε_i odchylku populace od modelu. Každý koeficient má **podmíněný význam**: β_1 porovnává populace lišící se v X_1 při stejných hodnotách ostatních prediktorů.

8 Klimatické proměnné nenesou nezávislé informace

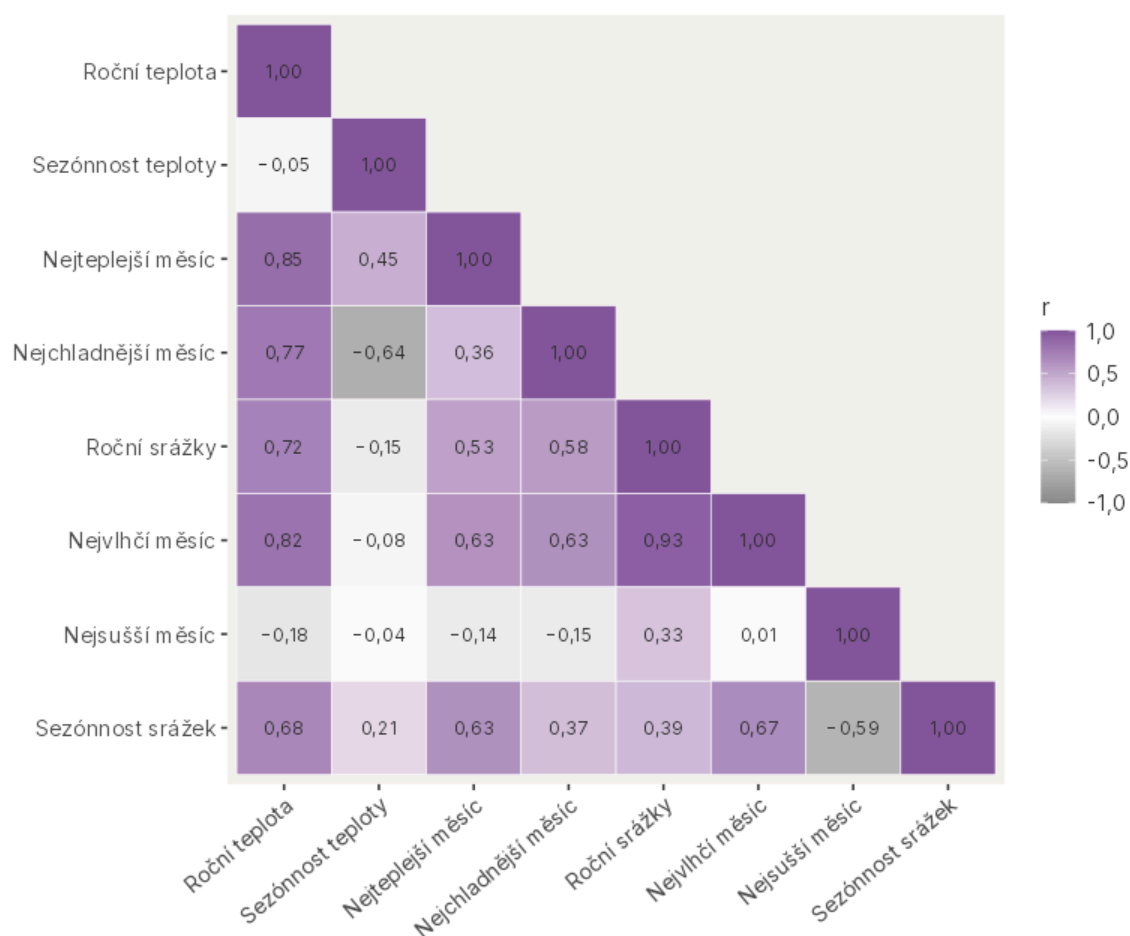
Osm klimatických proměnných v tabulce popisuje různé části stejného klimatu. Proto se vyplatí podívat se na vztahy mezi samotnými prediktory.

Které dvojice budou silně souviset?

Než se podíváte na následující obrázek, vyberte dvě klimatické proměnné, u nichž očekáváte silnou kladnou korelaci. Který biologický nebo fyzikální vztah váš odhad podporuje?

Možná odpověď

Silně kladně by měly souviset například průměrná roční teplota a maximum nejteplejšího měsíce: obě popisují, jak teplé místo je. Podobně spolu budou souviset roční srážky a srážky nejvlhčího měsíce, protože nejvlhčí měsíc tvoří velkou část ročního úhrnu.



Obrázek 4: Korelační matice osmi klimatických proměnných. Čísla jsou Pearsonovy korelační koeficienty; fialová značí kladnou, šedá zápornou korelaci.

Doplňující: jak korelační matici spočítat

Matici spočítáme funkcí `cor()`, které předáme tabulku jen s číselnými klimatickými sloupci. Výsledkem je čtvercová tabulka: v řádku i sloupci jsou stejné proměnné a každá buňka obsahuje korelaci jedné dvojice. Graf výše zobrazuje stejná čísla barevně a každou dvojici jen jednou.

```
# Vybereme osm číselných klimatických sloupců
vec_klima <-
  c(
    "teplota_rocni_C", "teplota_sezonnost", "teplota_max_C",
    "teplota_min_C", "srazky_rocni_mm", "srazky_nejvlhci_mesic_mm",
    "srazky_nejsussi_mesic_mm", "srazky_sezonnost"
  )

# Korelace všech dvojic, zaokrouhlené na dvě desetinná místa
round(
  x = cor(x = data_kridlatka[, vec_klima]),
  digits = 2
)
```

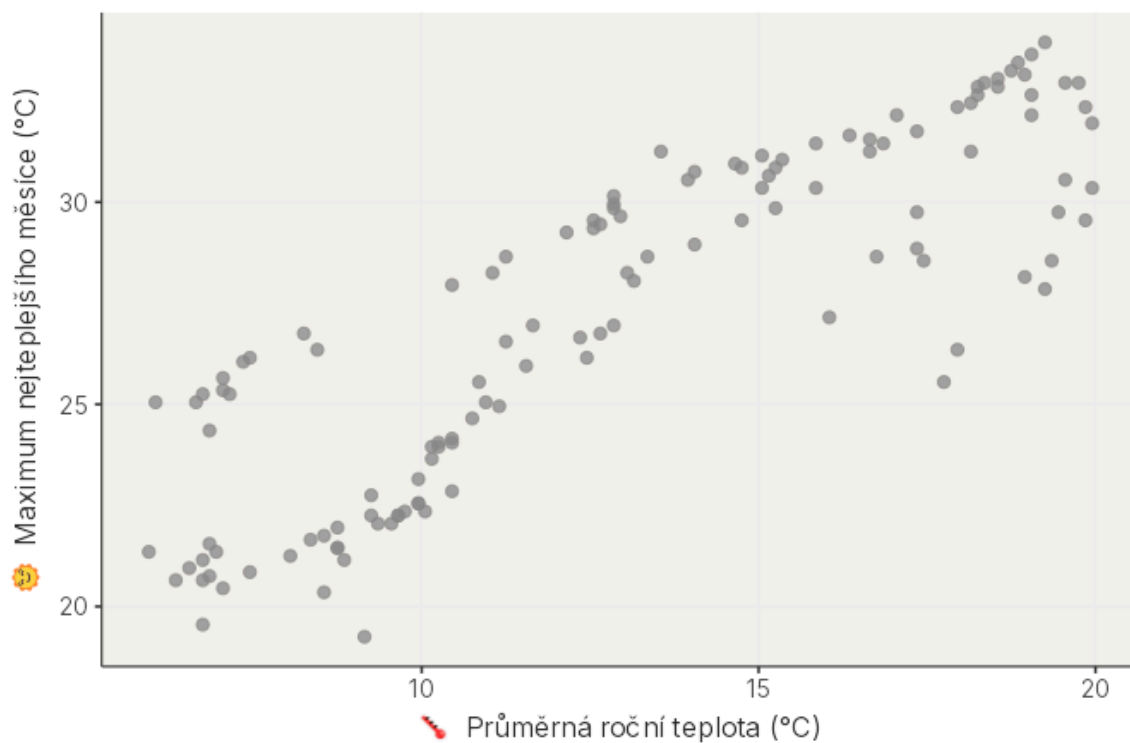
	teplota_rocni_C	teplota_sezonnost	teplota_max_C
teplota_rocni_C	1.00	-0.05	0.85
teplota_sezonnost	-0.05	1.00	0.45
teplota_max_C	0.85	0.45	1.00
teplota_min_C	0.77	-0.64	0.36
srazky_rocni_mm	0.72	-0.15	0.53
srazky_nejvlhci_mesic_mm	0.82	-0.08	0.63
srazky_nejsussi_mesic_mm	-0.18	-0.04	-0.14
srazky_sezonnost	0.68	0.21	0.63

	teplota_min_C	srazky_rocni_mm	srazky_nejvlhci_mesic_mm
teplota_rocni_C	0.77	0.72	0.82
teplota_sezonnost	-0.64	-0.15	-0.08
teplota_max_C	0.36	0.53	0.63
teplota_min_C	1.00	0.58	0.63
srazky_rocni_mm	0.58	1.00	0.93
srazky_nejvlhci_mesic_mm	0.63	0.93	1.00
srazky_nejsussi_mesic_mm	-0.15	0.33	0.01
srazky_sezonnost	0.37	0.39	0.67

	srazky_nejsussi_mesic_mm	srazky_sezonnost
teplota_rocni_C	-0.18	0.68
teplota_sezonnost	-0.04	0.21
teplota_max_C	-0.14	0.63
teplota_min_C	-0.15	0.37

srazky_rocni_mm	0.33	0.39
srazky_nejvlhci_mesic_mm	0.01	0.67
srazky_nejsussi_mesic_mm	1.00	-0.59
srazky_sezonnost	-0.59	1.00

Průměrná roční teplota a maximum nejteplejšího měsíce koreluje 0,855. Obě proměnné proto nesou z velké části stejnou informaci o tom, jak teplé je místo původu:



Obrázek 5: Průměrná roční teplota a maximum nejteplejšího měsíce místa původu. Jeden bod je jedna populace.

Když prediktory v modelu spolu silně souvisejí, mluvíme o korelaci prediktorů. Co to znamená pro model, uvidíme v následující části.

9 Proč se odhad roční teploty změnil?

Porovnáme koeficient roční teploty v modelu, kde je jediným prediktorem, a v modelu se dvěma teplotami. Oba modely používají stejných 128 populací, takže rozdíl nevzniká změnou analyzovaných řádků.

```
# Celá tabulka koeficientů modelu s jednou teplotou
coef(object = summary(object = mod_rocni_teplota))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.510545418	0.01151336	44.34376	9.711812e-79
teplota_rocni_C	-0.009133159	0.00084616	-10.79365	1.237814e-19

```
# Celá tabulka koeficientů modelu se dvěma teplotami
coef(object = summary(object = mod_dve_teploty))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.589935700	0.030071033	19.618073	5.666443e-40
teplota_rocni_C	-0.005278927	0.001585346	-3.329828	1.142457e-03
teplota_max_C	-0.004750850	0.001669974	-2.844864	5.193218e-03

V každé tabulce najdeme řádek `teplota_rocni_C`. Sloupec `Estimate` obsahuje odhad koeficientu a sloupec `Std. Error` jeho standardní chybu.

Odhad se změnil z $-0,00913$ na $-0,00528$ mm na 1°C . V prvním modelu koeficient roční teploty zahrnuje veškerou informaci, kterou tato proměnná nese o celkovém teplotním režimu, včetně té, kterou sdílí s maximem nejteplejšího měsíce. Po přidání maxima se ptáme na užší otázku: jak se liší tloušťka listu mezi populacemi se **stejným maximem**, ale odlišnou průměrnou roční teplotou.

Standardní chyba současně vzroste z $0,00085$ na $0,00159$ mm na 1°C . Proč přesnost klesla, napoví graf dvou teplot z předchozí části.

Jakou otázku zodpovídá koeficient roční teploty?

Koeficient roční teploty v modelu se dvěma teplotami porovnává populace se stejným maximem nejteplejšího měsíce, které se liší průměrnou roční teplotou. Najděte v předchozím bodovém grafu dvou teplot populace, které takové srovnání skutečně umožňují. Je jich hodně, nebo málo?

Možná odpověď

Hledáme body, které leží na téměř stejné výšce (stejně maximum), ale daleko od sebe ve vodorovném směru (různá roční teplota). Protože body tvoří úzký rostoucí mrak, je takových dvojic málo a leží hlavně na okrajích mraku.

Na podmíněnou otázku proto odpovídá jen malá část dat. Model o ní ví méně než o celkovém vztahu, a samostatné příspěvky obou teplot se tak oddělují s větší nejistotou, přestože model lze bez problému fitovat.

Doplňující: korelované prediktory mohou vytvářet nestabilní koeficienty

Když dva prediktory nesou téměř stejnou informaci, model má málo pozorování, která by je dokázala od sebe oddělit. Malá změna dat nebo přidání několika populací pak může výrazně změnit jejich jednotlivé koeficienty, znaménka i standardní chyby. Proto je nečteme jako dvě samostatné biologické odpovědi bez ohledu na jejich překryv.

Pomocným ukazatelem bývá VIF (*variance inflation factor*), který shrnuje, nakolik je nejistota koeficientu zvětšena překryvem s ostatními prediktory. VIF ale nemá univerzální hranici, po jejímž překročení musíme proměnnou automaticky smazat. Rozhodujeme podle biologické otázky, rozsahu společně pozorovaných hodnot a toho, zda chceme samostatně interpretovat koeficienty, nebo předpovídat podporované kombinace.

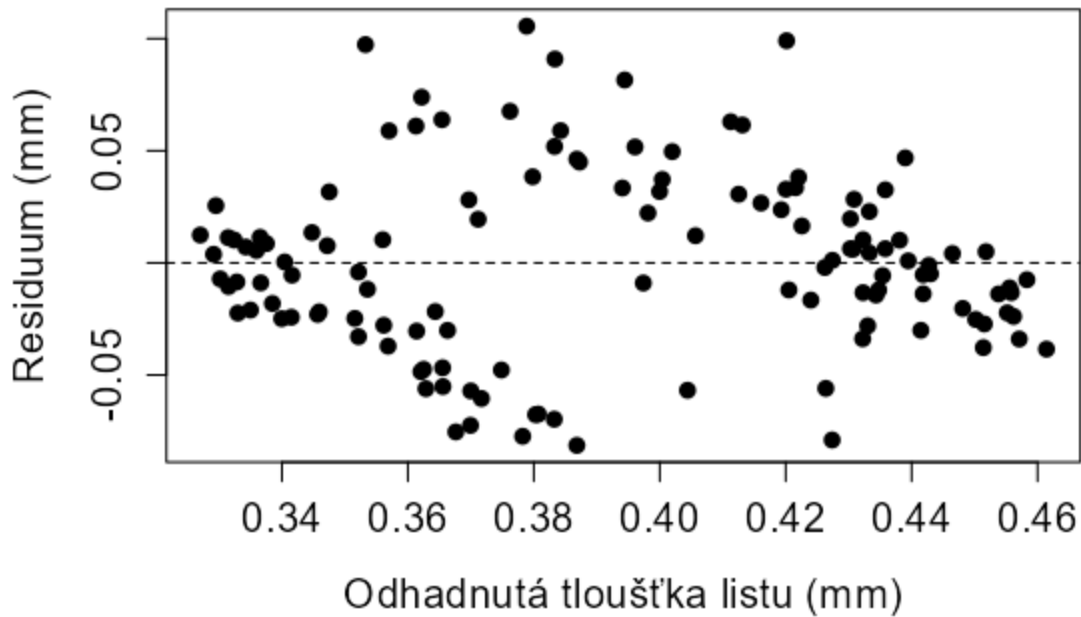
Doplňující: porovnání není výběr nejlepšího modelu

Zde porovnáváme koeficienty, protože chceme pochopit, jak přidání souvisejícího prediktoru změnilo statistickou otázku. O výběru modelu ale nerozhodujeme. V dalších lekcích budeme pracovat s principem úspornosti: mezi biologicky smysluplnými modely budeme hledat takový, který popisuje data dostatečně dobře bez zbytečné složitosti. K tomu nestačí vybrat model s nejvyšším R^2 .

9.1 Změnil druhý prediktor strukturu residuí?

Také model se dvěma teplotami zkontrolujeme grafem residuí:

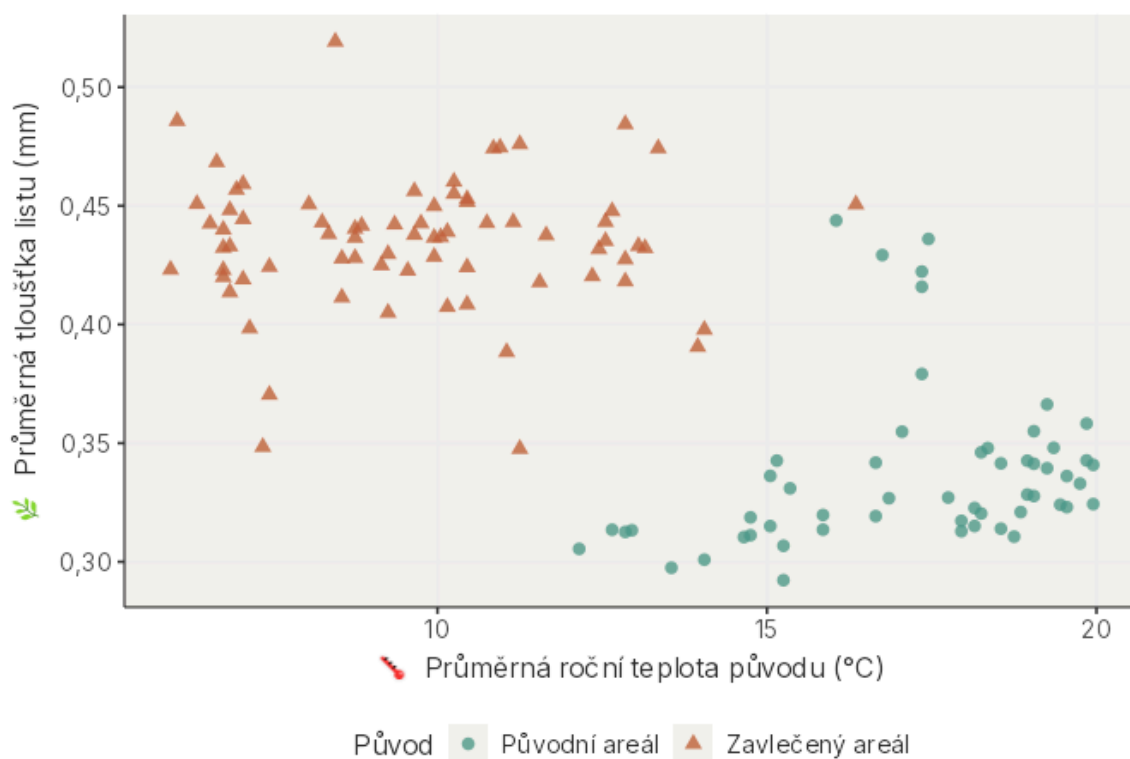
```
# Residua proti odhadnutým hodnotám modelu se dvěma teplotami
plot(
  x = fitted(object = mod_dve_teploty),
  y = resid(object = mod_dve_teploty),
  xlab = "Odhadnutá tloušťka listu (mm)",
  ylab = "Residuum (mm)",
  pch = 16,
  cex = 1.2,
  cex.axis = 1.3,
  cex.lab = 1.3
)
abline(h = 0, lty = 2)
```



Graf vypadá podobně jako u modelu s jednou teplotou: žádný výrazný oblouk ani trychtýř, ale stejné mírné seskupení residuí pod nulou a nad ní. Druhá teplota je tedy nevysvětlila. Změnu koeficientu proto nečteme jako opravu nevyhovujícího prvního modelu, ale jako změnu otázky a její nejistoty.

10 Původ populací mění příběh vztahu

Jeden lineární model může kombinovat číselné i kategoriální prediktory. Typ prediktoru nemění základní princip: znaménkem + přidáváme do stejného modelu další informaci. Rozlišíme pozorované populace podle původu, zatím bez modelových přímek:



Obrázek 6: Pozorované populace rozlišené podle původního a zavlečeného areálu. Jeden bod je jedna zdrojová populace.

Jak se liší obě skupiny?

Ve kterých teplotách pozorujeme původní a zavlečené populace? Která skupina má při podobné teplotě výše položené body? Je klesající vztah stejně zřetelný uvnitř obou skupin jako v celém souboru?

Možná odpověď

Původní populace pocházejí převážně z teplejších míst, zavlečené z chladnějších. Tam, kde se teploty obou skupin překrývají, leží zavlečené populace výše: mají silnější listy. Uvnitř každé skupiny už klesající vztah zřetelný není. Velká část celkového poklesu tak vzniká tím, že skupiny leží v různých částech teplotní osy.

Celkový vztah napříč všemi populacemi proto spojuje dvě věci: rozdíl mezi skupinami a případný vztah k teplotě uvnitř skupin. K roční teplotě přidáme kategoriální prediktor původu:

```
# Jeden model s původem a roční teplotou
mod_puvod <-
```

```
lm(
  formula = tloustka_listu_mm ~ puvod + teplota_rocni_C,
  data = data_kridlatka
)

# Zobrazíme celý přehled modelu
summary(object = mod_puvod)
```

```
Call:
lm(formula = tloustka_listu_mm ~ puvod + teplota_rocni_C, data = data_kridlatka)

Residuals:
    Min       1Q   Median       3Q      Max
-0.089547 -0.017475 -0.003619  0.009723  0.108468

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.311427   0.021253  14.653  <2e-16 ***
puvodZavleeny   0.108947   0.010651  10.228  <2e-16 ***
teplota_rocni_C 0.001491   0.001213   1.229   0.221
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03082 on 125 degrees of freedom
Multiple R-squared:  0.7172,    Adjusted R-squared:  0.7126
F-statistic: 158.5 on 2 and 125 DF,  p-value: < 2.2e-16
```

Jak číst model s teplotou a původem?

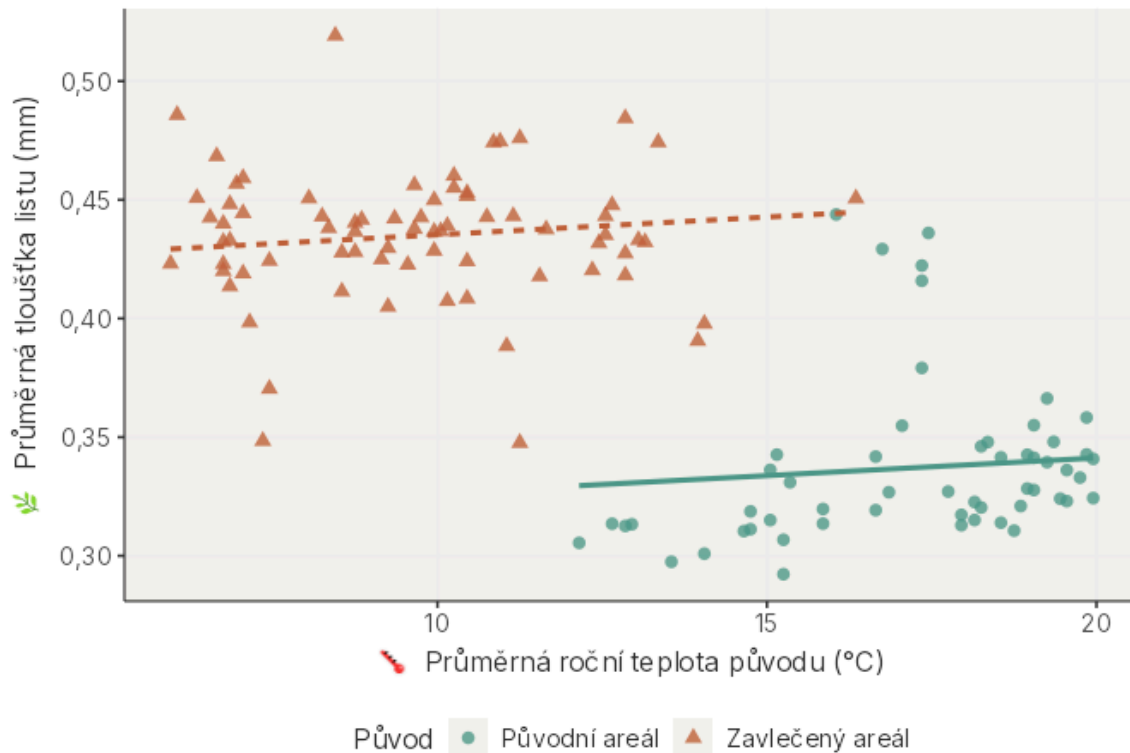
Která skupina je ve výstupu referenční? Najděte společný sklon roční teploty a posun zavlečených populací vůči původním. Jak byste oba koeficienty popsali v biologických jednotkách?

Možná odpověď

Referenční skupinou jsou původní populace, protože ve výstupu je jen řádek `puvodZavleeny`. Jeho koeficient, 0,109 mm, říká, o kolik silnější list mají zavlečené populace než původní **při stejné roční teplotě**. Koeficient `teplota_rocni_C`, +0,00149 mm na 1 °C, je sklon roční teploty **při stejném původu**; model předpokládá, že je v obou skupinách stejný.

Model se znaménkem + je i zde aditivní. Koeficient původu posouvá jednu skupinu vůči druhé, zatímco koeficient roční teploty určuje společný sklon obou skupin. V grafu proto vidíme dvě

rovnoběžné přímky. Každou kreslíme jen v rozsahu teplot, ve kterém danou skupinu skutečně pozorujeme:

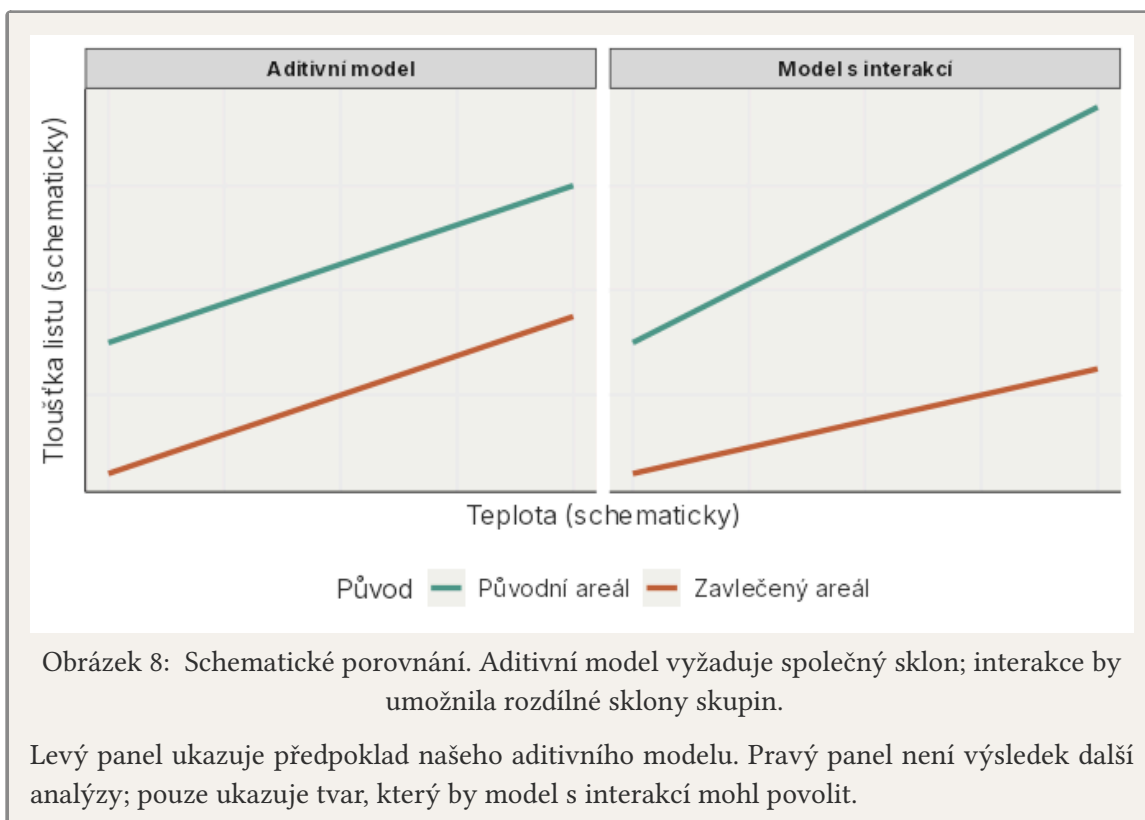


Obrázek 7: Body jsou pozorované populace. Rovnoběžné přímky jsou odhady aditivního modelu se společným sklonem roční teploty a rozdílnou výškou pro každý původ; každá přímka pokrývá jen pozorované teploty své skupiny.

Přímky jsou téměř vodorovné. Rozdíl mezi nimi, tedy posun zavlečených populací, je zato zřetelný.

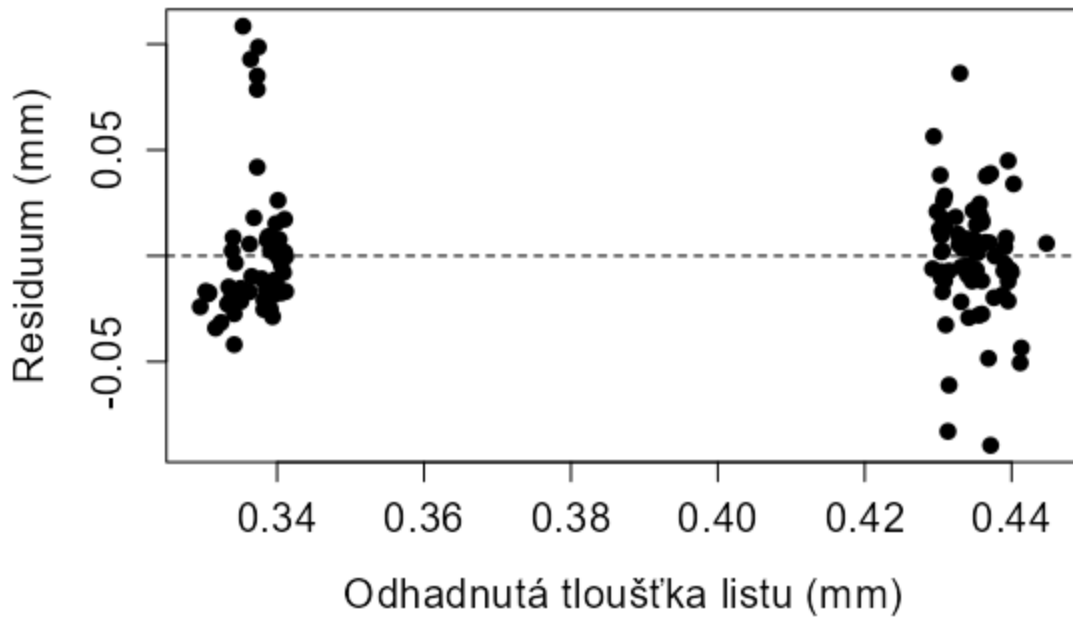
Doplňující: co kdyby skupiny neměly společný sklon?

Graf může vyvolat otázku, zda se vztah mezi roční teplotou a tloušťkou listu liší podle původu populace. Takovou otázku řeší interakce prediktorů: dovolila by každé skupině jiný sklon, takže přímky by nemusely být rovnoběžné. Interakce patří do následující lekce; zde ji nefitujeme ani neinterpretujeme.



10.1 Sedí model s teplotou a původem k datům?

```
# Residua proti odhadnutým hodnotám modelu s původem a teplotou
plot(
  x = fitted(object = mod_puvod),
  y = resid(object = mod_puvod),
  xlab = "Odhadnutá tloušťka listu (mm)",
  ylab = "Residuum (mm)",
  pch = 16,
  cex = 1.2,
  cex.axis = 1.3,
  cex.lab = 1.3
)
abline(h = 0, lty = 2)
```



Odhadnuté hodnoty tvoří dva shluky, protože model dává každé skupině jinou výšku přímky. V obou shlucích leží residua kolem nuly bez oblouku nebo zřetelného trychtýře. Seskupení residuí, které jsme viděli u modelů s teplotou, tak vysvětlil původ populací. Diagnostický graf neukazuje důvod měnit strukturu modelu.

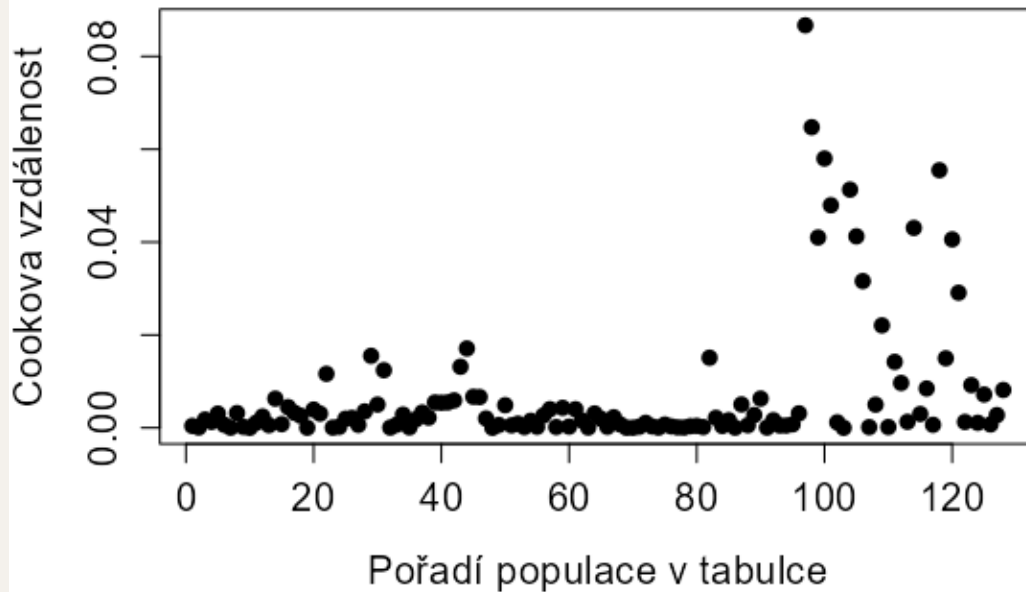
Doplňující: které populace mají na model velký vliv?

Cookova vzdálenost doplňuje kontrolu residuí o otázku, zda některá populace výrazně ovlivňuje odhadnuté koeficienty:

```
# Cookova vzdálenost každé populace
vec_cook <- cooks.distance(model = mod_puvod)

plot(
  x = vec_cook,
  xlab = "Pořadí populace v tabulce",
  ylab = "Cookova vzdálenost",
  pch = 16,
  cex = 1.2,
  cex.axis = 1.3,
```

```
cex.lab = 1.3  
)
```



```
# Populace s nejvyšší hodnotou  
data_kridlatka$populace[which.max(x = vec_cook)]
```

```
[1] "JA2"
```

Nejvyšší hodnotu má populace JA2 (0,087), ale ostatní populace mají hodnoty podobně malé. Žádná populace tedy od ostatních výrazně nevybočuje a odhady nestojí na jediném pozorování. Kdyby některá populace vybočovala, nemazali bychom ji automaticky: nejprve bychom ověřili její zdrojová data a zjistili, proč má na model velký vliv.

11 Při stejném původu sklon teploty téměř mizí

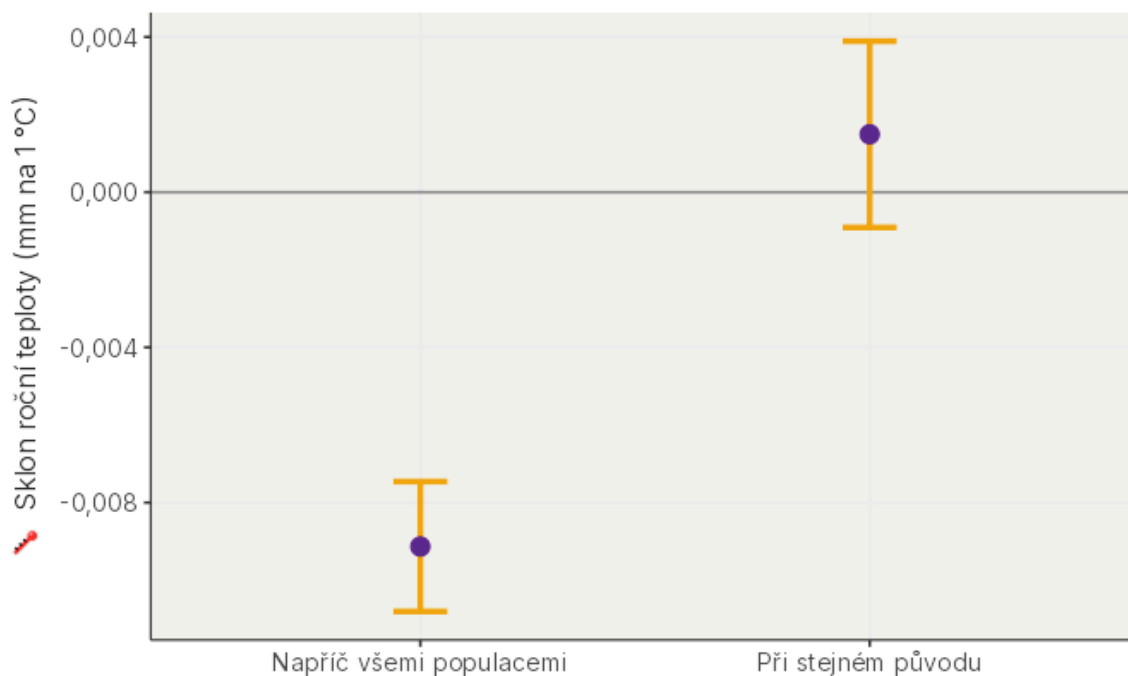
Porovnejme dva odhady sklonu roční teploty. V prvním modelu, kde byla teplota jediným prediktorem, tloušťka listu klesala o 0,00913 mm na 1 °C. V modelu s původem je sklon +0,00149 mm na 1 °C a jeho 95% interval spolehlivosti sahá od -0,00091 do 0,00389, takže zahrnuje nulu. Oba intervaly vypíše funkce `confint()`, kterou známe z předchozích lekcí:

```
# 95% intervaly spolehlivosti koeficientů modelu s jednou teplotou
confint(object = mod_rocni_teplota)
```

	2.5 %	97.5 %
(Intercept)	0.48776083	0.533330010
teplota_rocni_C	-0.01080769	-0.007458633

```
# 95% intervaly spolehlivosti koeficientů modelu s původem a teplotou
confint(object = mod_puvod)
```

	2.5 %	97.5 %
(Intercept)	0.2693633321	0.353489741
puvodZavleceny	0.0878668511	0.130027751
teplota_rocni_C	-0.0009100173	0.003891958



Obrázek 9: Sklon roční teploty se dvěma odlišnými otázkami. Body jsou odhady, úsečky jejich 95% intervaly spolehlivosti, šedá čára nulový sklon.

Jak je možné, že přidání jednoho prediktoru téměř smaže zřetelný klesající trend? Odpověď jsme už viděli v grafu s barevně rozlišeným původem. Původní populace pocházejí z míst s průměrnou roční teplotou 12,15 až 19,95 °C, zavlečené z míst s teplotou 5,95 až 16,35 °C. Příslušnost k areálu tedy souvisí s teplotou. Zároveň souvisí s tloušťkou listu: zavlečené populace mají při stejné

teplotě silnější listy. Celkový sklon proto mísí dvě cesty: případný vztah teploty k tloušťce a rozdíl mezi původními a zavlečenými populacemi, které leží v různých částech teplotní osy.

Když třetí proměnná souvisí s prediktorem i s odezvou a vytváří tak část jejich celkové souvislosti, mluvíme o konfundování. Zde konfundující proměnnou představuje původ populace. Po jeho zohlednění odpovídá koeficient teploty na užší otázku: liší se tloušťka listu podle teploty mezi populacemi **stejného původu**? Podle našich dat je takový vztah malý a jeho směr nedokážeme spolehlivě určit.

Ani jeden koeficient není „ten jediný pravý“. První popisuje celkový vztah napříč oblastmi, druhý vztah uvnitř skupin se stejným původem. Výsledek také neříká, že teplota nemá žádný biologický význam. Původ v sobě nese mnoho dalších rozdílů, například genetickou a evoluční historii populací, a teploty obou skupin se překrývají jen v rozsahu 12,15 až 16,35 °C. Podmíněný sklon proto stojí na omezené části dat.

Doplňující: přidat další prediktor není vždy správná kontrola

Další prediktor má smysl přidat podle biologické otázky a představy o tom, jak proměnné vznikly. Při konfundování ovlivňuje třetí proměnná prediktor i odezvu, takže jejich celková asociace mísí více cest. Zohlednění takové proměnné může otázku zpřesnit.

Jiná situace nastane, když doplníme **prostřední článek** mechanismu. Pokud hnojení mění obsah dusíku v listu a ten následně růst, kontrola dusíku odstraní část účinku hnojení a položí užší otázku o přímé cestě. Ještě opatrnější musíme být u **společného následku** dvou proměnných: výběr pouze přeživších jedinců může vytvořit asociaci mezi vlastnostmi, které ovlivňují přežití, i když spolu v původní populaci nesouvisely.

Pravidlo „přidejte do modelu vše, co máte“ proto nechrání automaticky před zkreslením. Model popisuje podmíněné vztahy; kauzální význam vyžaduje navíc znalost uspořádání studie a biologického mechanismu.

12 Jak porovnat skupiny při stejné teplotě?

Koeficient `puvodZavleceny` porovnává skupiny při stejné roční teplotě. Proč nestačí porovnat prosté průměry tloušťky obou skupin? Spočítejme je:

```
# Prostý průměr tloušťky listu původních populací
prumer_puvodni <-
  mean(x = data_kridlatka$tloustka_listu_mm[data_kridlatka$puvod == "Puvodni"])
prumer_puvodni
```

```
[1] 0.3370428
```

```
# Prostý průměr tloušťky listu zavlečených populací
prumer_zavleceeny <-
  mean(x = data_kridlatka$tloustka_listu_mm[data_kridlatka$puvod ==
"Zavleceeny"])
prumer_zavleceeny
```

```
[1] 0.4347821
```

Rozdíl prostých průměrů je 0,098 mm. Tyto průměry popisují skutečně pozorované skupiny, jenže skupiny se liší i rozložením ročních teplot. Rozdíl prostých průměrů proto spojuje rozdíl původu s rozdílem klimatu. Chceme-li porovnat původ samotný, musíme obě skupiny porovnat **při stejné teplotě**, a to takové, při které máme data z obou skupin. Spočítáme rozsahy teplot obou skupin a jejich společnou část:

```
# Rozsah ročních teplot v každé skupině
vec_rozsah_puvodni <-
  range(data_kridlatka$teplota_rocni_C[data_kridlatka$puvod == "Puvodni"])
vec_rozsah_zavleceeny <-
  range(data_kridlatka$teplota_rocni_C[data_kridlatka$puvod == "Zavleceeny"])

# Společná část: vyšší z dolních mezí a nižší z horních mezí
c(
  max(vec_rozsah_puvodni[1], vec_rozsah_zavleceeny[1]),
  min(vec_rozsah_puvodni[2], vec_rozsah_zavleceeny[2])
)
```

```
[1] 12.15 16.35
```

Teploty obou skupin se překrývají přibližně od 12,15 do 16,35 °C. Zvolíme srovnávací teplotu uvnitř tohoto překryvu:

```
# Srovnávací teplota uvnitř společného rozsahu obou skupin
teplota_pro_porovnani <- 14
```

Vytvoříme dva řádky, které se liší pouze původem. Jde o modelové scénáře při stejné roční teplotě, nikoli o nově naměřené populace:

```
# Dva scénáře: stejná teplota, jiný původ
data_scenare <-
  data.frame(
    puvod = factor(
```

```

    x = c("Puvodni", "Zavleceeny"),
    levels = levels(x = data_kridlatka$puvod)
  ),
  teplota_rocni_C = c(
    teplota_pro_porovnani,
    teplota_pro_porovnani
  )
)

data_scenare

```

	puvod	teplota_rocni_C
1	Puvodni	14
2	Zavleceeny	14

Model skládá odhadnutou tloušťku ze tří částí:

odhadnutá tloušťka = intercept + posun zavlečeného areálu + sklon roční teploty × roční teplota .

U původní populace je posun nulový, protože jde o referenční skupinu. Dosazení pro první řádek je

$$\begin{aligned}
 &0,3114 + 0 + (0,00149) \times 14 \\
 &= 0,3114 + 0 + 0,0209 \\
 &= 0,332 \text{ mm.}
 \end{aligned}$$

U zavlečené populace se přidá koeficient `puvodZavleceeny`:

$$\begin{aligned}
 &0,3114 + 0,1089 + (0,00149) \times 14 \\
 &= 0,3114 + 0,1089 + 0,0209 \\
 &= 0,441 \text{ mm.}
 \end{aligned}$$

Ruční výpočet zkontrolujeme funkcí `predict()`. Výsledky přidáme do stejné dvouřádkové tabulky:

```

# Odhadnutá tloušťka pro oba scénáře
data_scenare$odhad_mm <-
  predict(
    object = mod_puvod,
    newdata = data_scenare
  )

data_scenare

```

	puvod	teplota_rocni_C	odhad_mm
1	Puvodni	14	0.3323001
2	Zavleceeny	14	0.4412474

Rozdíl obou odhadů při stejné teplotě je 0,109 mm. Je přesně roven koeficientu `puvodZavleceeny`, protože ostatní části rovnice jsou v obou scénářích stejné. Rozdíl prostých průměrů, 0,098 mm, odpovídá na jinou otázku. Zde jsou obě čísla přesto blízko, protože uvnitř jednoho původu se tloušťka s teplotou téměř nemění. Kdyby teplota uvnitř skupin s tloušťkou výrazně souvisela, mohla by se obě porovnání lišit mnohem víc.

Obecný zápis pro původní populaci je

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

zatímco pro zavlečenou populaci přidáme skupinový koeficient:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_{\text{zavlečený}} + \hat{\beta}_1 X_i.$$

$\hat{Y}_i$ je odhadnutá tloušťka populace i , X_i její průměrná roční teplota, $\hat{\beta}_0$ intercept, tedy výchozí hodnota původní skupiny, $\hat{\beta}_{\text{zavlečený}}$ rozdíl zavlečené skupiny při stejné teplotě a $\hat{\beta}_1$ změna s průměrnou roční teplotou při stejném původu.

Tabulka zatím obsahuje dva bodové odhady, ale neukazuje, jak přesně je model určit.

13 Jak přesně známe modelové průměry?

Funkci `emmeans()` známe z porovnávání skupin v minulé lekci. Zde jí navíc řekneme, při které hodnotě číselného prediktoru má skupiny porovnat. Vrátí stejné bodové odhady, které jsme sestavili ručně, a k nim jejich nejistotu:

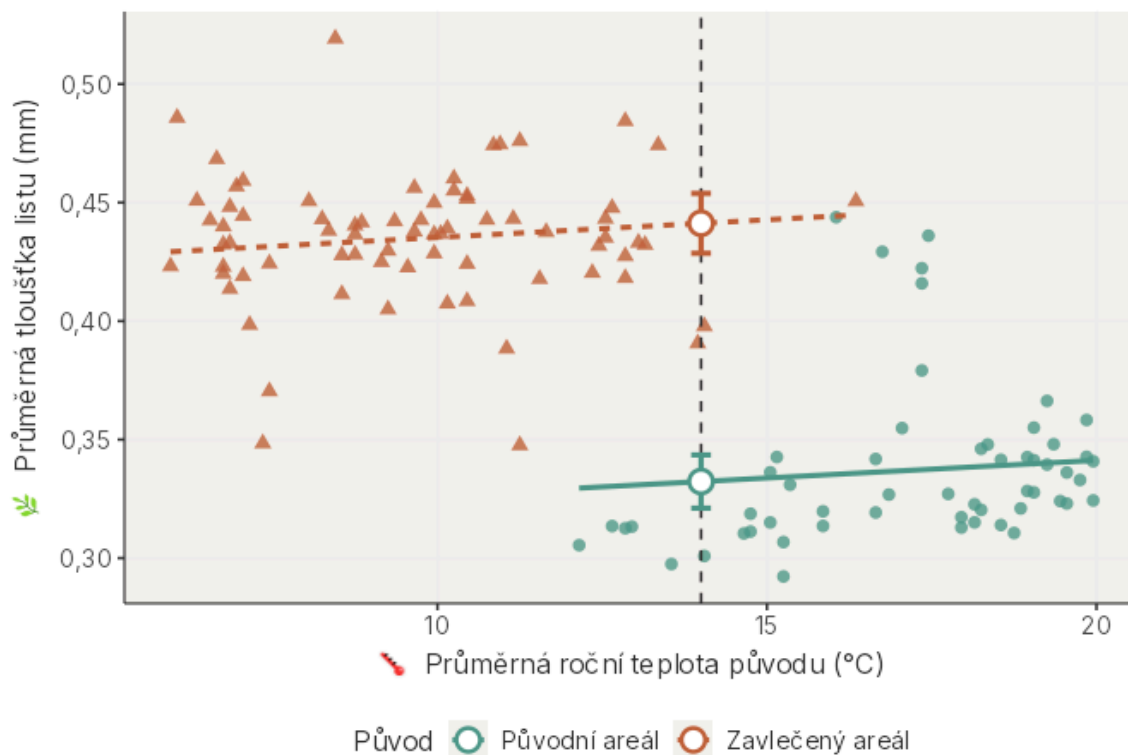
```
# Modelové průměry obou skupin při srovnávací teplotě
res_modelove_prumery <-
  emmeans::emmeans(
    object = mod_puvod,
    specs = ~ puvod,
    at = list(teplota_rocni_C = teplota_pro_porovnani)
  )

res_modelove_prumery
```

puvod	emmean	SE	df	lower.CL	upper.CL
Puvodni	0.332	0.00567	125	0.321	0.344
Zavleceeny	0.441	0.00638	125	0.429	0.454

Confidence level used: 0.95

Sloupec `emmean` obsahuje modelové průměry a sloupce `lower.CL` a `upper.CL` meze jejich 95% intervalu spolehlivosti. Interval platí pro **průměrnou modelovou tloušťku** populací daného původu při 14 °C, nikoli pro jednu novou populaci. Pro původní populace sahá od 0,321 do 0,344 mm, pro zavlečené od 0,429 do 0,454 mm. Intervaly se nepřekrývají.



Obrázek 10: Modelové průměry původních a zavlečených populací při stejné roční teplotě. Svislé úsečky jsou 95% intervaly spolehlivosti.

Doplňující: jak bezpečně predikovat z více prediktorů

Stejné intervaly vrátí také `predict()` s argumentem `interval = "confidence"`:

```
# Stejné scénáře, k odhadům přidáme 95% intervaly spolehlivosti
predict(
  object = mod_puvod,
  newdata = data_scenare,
  interval = "confidence"
)
```

	fit	lwr	upr
1	0.3323001	0.3210759	0.3435244
2	0.4412474	0.4286232	0.4538716

Tabulka `newdata` musí obsahovat každý prediktor modelu se stejnými názvy, jednotkami a úrovněmi faktorů. U modelu se dvěma teplotami navíc nestačí, aby byla každá hodnota samostatně v pozorovaném rozsahu. Jejich **kombinace** musí odpovídat reálnému klimatu.

Predikce není stavebnice libovolných hodnot

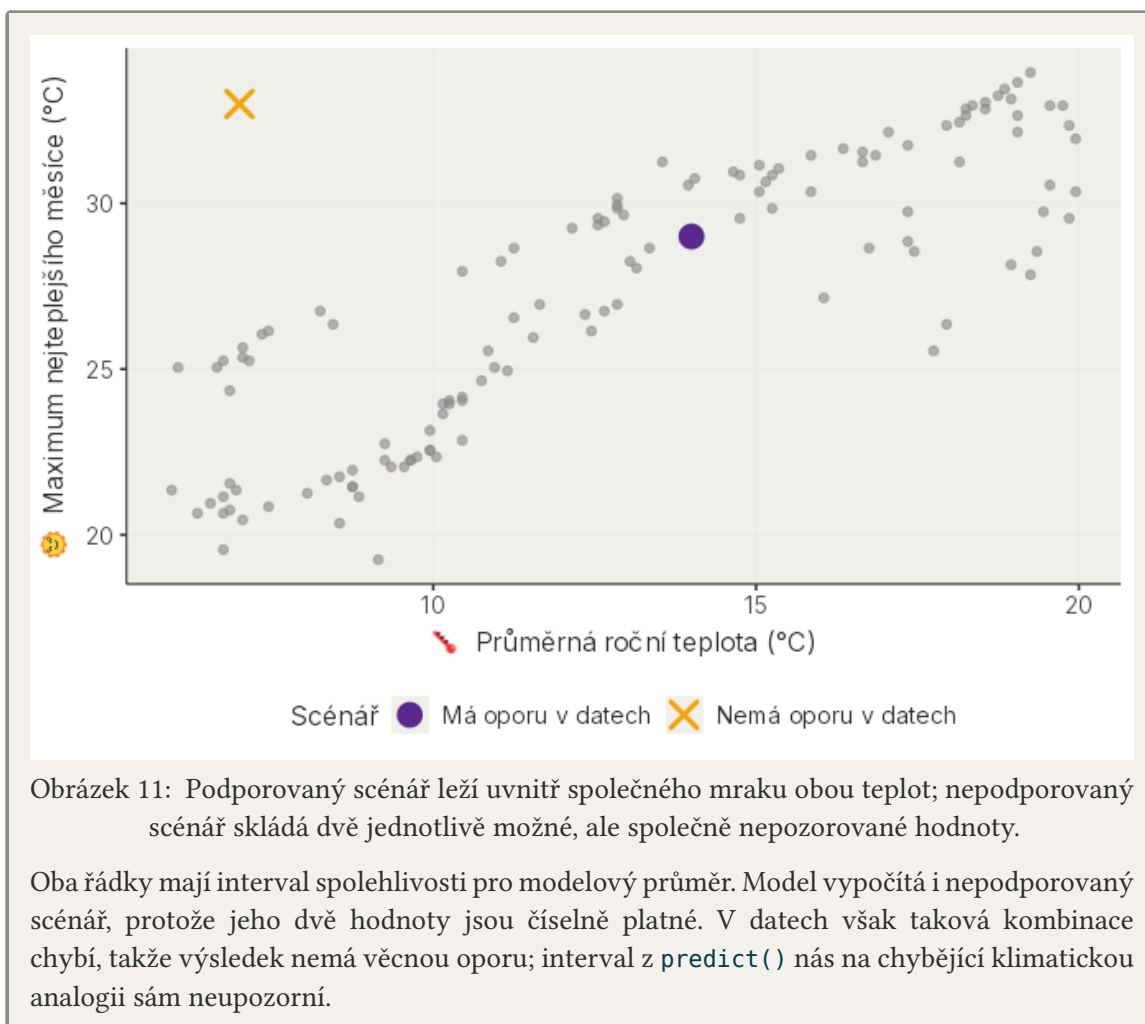
Kombinace nízké roční teploty a velmi vysokého maxima nejteplejšího měsíce může ležet mimo společný mrak dat. Model číslo vypočítá, ale data takovou klimatickou situaci nemusí podporovat.

Porovnáme dva scénáře pro model se dvěma teplotami:

```
# Podporovaný a nepodporovaný klimatický scénář
data_klima_scenare <-
  data.frame(
    teplota_rocni_C = c(teplota_pro_porovnani, 7),
    teplota_max_C = c(29, 33)
  )

# Odhady s 95% intervaly spolehlivosti
predict(
  object = mod_dve_teploty,
  newdata = data_klima_scenare,
  interval = "confidence"
)
```

	fit	lwr	upr
1	0.3782561	0.3703263	0.3861858
2	0.3962052	0.3591626	0.4332477



14 Záleží u testu na pořadí prediktorů?

V předchozích lekcích jsme testovali, zda je koeficient slučitelný s nulou, a funkcí `anova()` jsme pokládali celkovou otázku o rozdílech mezi skupinami. U modelu s několika prediktory musíme rozlišit **test koeficientu** a **sekvenční test prediktoru**. Oba uvádějí p-hodnotu, ale odpovídají na jinak formulované otázky.

Vytvoříme druhý zápis stejného modelu. Obsahuje stejné dva prediktory a liší se pouze jejich pořadím ve formuli:

```
# Stejně prediktory zapíšeme v opačném pořadí
mod_dve_teploty_opacne <-
  lm(
    formula = tloustka_listu_mm ~ teplota_max_C + teplota_rocni_C,
```

```
data = data_kridlatka
)
```

14.1 Co se změnil v tabulce koeficientů?

Tabulka `summary()` testuje u každého koeficientu nulovou hypotézu, že je při zohlednění ostatních prediktorů roven nule. V obou zápisech porovnáme odhady, standardní chyby, t-statistiky a p-hodnoty:

```
# Tabulka koeficientů: nejdřív roční teplota, potom maximum
coef(object = summary(object = mod_dve_teploty))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.589935700	0.030071033	19.618073	5.666443e-40
teplota_rocni_C	-0.005278927	0.001585346	-3.329828	1.142457e-03
teplota_max_C	-0.004750850	0.001669974	-2.844864	5.193218e-03

```
# Tabulka koeficientů: nejdřív maximum, potom roční teplota
coef(object = summary(object = mod_dve_teploty_opacne))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.589935700	0.030071033	19.618073	5.666443e-40
teplota_max_C	-0.004750850	0.001669974	-2.844864	5.193218e-03
teplota_rocni_C	-0.005278927	0.001585346	-3.329828	1.142457e-03

Co se po obrácení pořadí nezměnilo?

Porovnejte řádky `teplota_rocni_C` a `teplota_max_C` v obou tabulkách. Zůstaly stejné sloupce `Estimate`, `Std. Error`, `t value` a `Pr(>|t|)`?

Možná odpověď

Ano. Změnilo se pouze pořadí řádků. Oba zápisy mají stejné koeficienty a stejné testy koeficientů, protože každý koeficient stále popisuje vztah při stejné hodnotě druhého prediktoru.

Stejně jsou také odhadnuté hodnoty a residua:

```
# Shodují se odhadnuté hodnoty obou zápisů?
all.equal(
  target = fitted(object = mod_dve_teploty),
```

```
current = fitted(object = mod_dve_tploty_opacne)
)
```

```
[1] TRUE
```

```
# Shodují se residua obou zápisů?
all.equal(
  target = resid(object = mod_dve_tploty),
  current = resid(object = mod_dve_tploty_opacne)
)
```

```
[1] TRUE
```

Výsledek `TRUE` potvrzuje, že jde o tentýž fitovaný model. Nemusíme proto opakovat diagnostiku residuí.

14.2 Sekvenční analýza rozptylu klade otázku zleva doprava

Základní `anova()` v R používá **sekvenční testy**. První řádek se ptá, kolik variability popíše první prediktor oproti modelu bez klimatického prediktoru. Druhý řádek se ptá, kolik další informace přidá druhý prediktor po zohlednění prvního.

Co očekáváte od dvou tabulek?

Obě teploty jsou silně korelované. Který prediktor převezme jejich společnou informaci, když jej zapíšeme jako první?

Možná odpověď

Společnou informaci převezme vždy ten prediktor, který je ve formuli první. Druhý prediktor dostane jen to, co přidává navíc. Proto čekáme, že řádek stejné teploty bude v obou tabulkách vypadat jinak podle toho, zda stojí první, nebo druhý.

```
# Roční teplota vstupuje do sekvenčního testu jako první
anova(object = mod_dve_tploty)
```

Analysis of Variance Table

Response: tloustka_listu_mm

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
--	----	--------	---------	---------	--------

```
teplota_rocni_C    1 0.201696 0.201696 123.0616 < 2.2e-16 ***
teplota_max_C      1 0.013265 0.013265   8.0933  0.005193 **
Residuals          125 0.204873 0.001639
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Maximální teplota vstupuje do sekvenčního testu jako první
anova(object = mod_dve_tploty_opacne)
```

Analysis of Variance Table

```
Response: tloustka_listu_mm
      Df  Sum Sq Mean Sq F value    Pr(>F)
teplota_max_C    1 0.196788 0.196788 120.067 < 2.2e-16 ***
teplota_rocni_C    1 0.018173 0.018173  11.088  0.001142 **
Residuals        125 0.204873 0.001639
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Když je první roční teplota, její řádek převezme také velkou část informace sdílené s maximální teplotou. Řádek maxima nejteplejšího měsíce zachycuje pouze to, co maximum přidá nad rámec roční teploty. Při opačném pořadí převezme sdílenou část maximum.

Rozdíl je v tabulkách konkrétní. P-hodnota sekvenčního testu roční teploty je při jejím prvním místě menší než 10^{-16} , ale po předchozím zohlednění maximální teploty 0,0011. U maximální teploty se p-hodnota změní z hodnoty menší než 10^{-16} při prvním místě na 0,0052 po zohlednění roční teploty.

Proto se mezi tabulkami mění sekvenční součty čtverců, F-statistiky a p-hodnoty. Nemění se však data ani fitovaný model. P-hodnota koeficientu v `summary()` odpovídá podmíněné otázce „přidává tento prediktor informaci při zahrnutí druhého prediktoru?“. Protože má každý číselný prediktor jeden stupeň volnosti, shoduje se tato p-hodnota se sekvenčním testem prediktoru zapsaného jako druhý.

Z předchozích lekcí známe obecný význam vysvětlené a nevysvětlené variability, součtů čtverců a F-statistiky. Zde je používáme k pochopení důsledku pořadí prediktorů. K podrobnějšímu rozdělení sdílené variability mezi několik prediktorů a k dalším způsobům testování modelových členů se vrátíme v dalších lekcích.

Co je hlavní zpráva?

Pořadí prediktorů ve formuli mění **p-hodnoty sekvenčních testů** z `anova()`, ale nemění p-hodnoty koeficientů v `summary()`. Přesouvání prediktorů ve formuli proto není způsob, jak vybrat vhodné prediktory nebo získat příznivější výsledek.

Doplňující: test celého modelu

Poslední řádek výstupu `summary()` obsahuje ještě jeden test: **F-statistic** a jeho p-hodnotu. Ten se neptá na jeden prediktor, ale na celý model: popisují obě teploty dohromady tloušťku listu lépe než model bez jakéhokoli prediktoru, který pro každou populaci odhaduje jen celkový průměr? Je to stejná celková otázka, jakou jsme v minulé lekci kladli funkcí `anova()` u skupin.

Stejný test dostaneme, když model bez prediktorů porovnáme s modelem se dvěma teplotami:

```
# Model bez prediktorů odhaduje jen celkový průměr tloušťky
mod_nulovy <-
  lm(
    formula = tloustka_listu_mm ~ 1,
    data = data_kridlatka
  )

# Přidají obě teploty dohromady informaci o tloušťce listu?
anova(
  object = mod_nulovy,
  mod_dve_teploty
)
```

Analysis of Variance Table

```
Model 1: tloustka_listu_mm ~ 1
Model 2: tloustka_listu_mm ~ teplota_rocni_C + teplota_max_C
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     127 0.41983
2     125 0.20487  2    0.21496 65.577 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-statistika 65,58 je stejná jako v posledním řádku `summary(object = mod_dve_teploty)`. Na rozdíl od sekvenčních testů nezávisí na pořadí prediktorů, protože porovnává celý model naráz. Velmi malá p-hodnota ale neříká, který z korelovaných prediktorů je důležitý, ani zda je model vhodný. Porovnávání více předem zvolených modelů se budeme věnovat v dalších lekcích.

Doplňující: testy nezávislé na pořadí prediktorů

Pokud chceme každý prediktor testovat po zohlednění všech ostatních prediktorů v aditivním modelu, můžeme použít testy typu II z funkce `car::Anova()`. Na rozdíl od sekvenčních testů z `anova()` nepřidělují sdílenou informaci podle pořadí členů ve formuli.

```
# Každý prediktor testujeme po zohlednění druhého prediktoru
car::Anova(
  mod = mod_dve_teploty,
  type = 2
)
```

Anova Table (Type II tests)

```
Response: tloustka_listu_mm
              Sum Sq Df F value    Pr(>F)
teplota_rocni_C 0.018173  1 11.0878 0.001142 **
teplota_max_C   0.013265  1  8.0933 0.005193 **
Residuals      0.204873 125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Obrácení pořadí ve formuli odpověď nezmění
car::Anova(
  mod = mod_dve_teploty_opacne,
  type = 2
)
```

Anova Table (Type II tests)

```
Response: tloustka_listu_mm
              Sum Sq Df F value    Pr(>F)
teplota_max_C   0.013265  1  8.0933 0.005193 **
teplota_rocni_C 0.018173  1 11.0878 0.001142 **
Residuals      0.204873 125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Řádky se v tabulkách objeví v pořadí podle formule, ale součet čtverců, F-statistika a p-hodnota každého prediktoru zůstávají stejné. V tomto aditivním modelu má každý číselný

prediktor jeden stupeň volnosti, takže jeho test typu II odpovídá testu příslušného koeficientu v `summary()`; platí $F = t^2$.

Testy typu II jsou vhodné pro náš aditivní model bez interakce. U modelů s interakcemi je nejprve potřeba přesně určit věcnou hypotézu; automatická volba typu testu ji nenahradí.

14.3 Výběr prediktorů patří do dalších lekcí

Výběr prediktorů budeme formulovat jako porovnání předem smysluplných kandidátních modelů. V dalších lekcích využijeme princip úspornosti a také AIC, které zohledňuje shodu modelu s daty i počet jeho parametrů. Takový postup poskytuje pro výběr modelu lepší oporu než měnění pořadí prediktorů nebo hledání nejmenší p-hodnoty.

15 Jak postupovat u modelu s více prediktory

1. Začněte biologickou otázkou a určete jednotku pozorování.
2. Prohlédněte vztah odezvy ke každému prediktoru i vztahy mezi prediktory.
3. Porovnávané modely fitujte na stejných řádcích.
4. Přeložte každý koeficient do věty „při stejných hodnotách ostatních prediktorů“.
5. Porovnejte odhady i jejich nejistotu, nejen p-hodnoty.
6. Rozlište test koeficientu od sekvenčního testu a u sekvenčního testu zdůvodněte pořadí prediktorů.
7. Predikujte pouze pro podporované kombinace hodnot.
8. Zkontrolujte residua a vlivná pozorování; nemažte je automaticky.

Tento postup není jednorázový seznam na konci analýzy. U každého nového modelu se vracíme k interpretaci koeficientů, kontrole residuí a vlivných pozorování, jak jsme to dělali u modelů výše.

16 Co tedy můžeme říct o křídlatce?

Vraťme se k hlavní otázce: **Mění klima původu tloušťku listů křídlatky?**

Napříč všemi populacemi mají populace z teplejších míst v průměru tenčí listy. Tento celkový vztah se ale téměř celý překrývá s rozdílem mezi původními a zavlečenými populacemi. Při stejném původu je sklon roční teploty malý a jeho směr nedokážeme spolehlivě určit, zatímco zavlečené populace mají při stejné teplotě zřetelně silnější listy.

Společná zahrada omezuje rozdíly v aktuálním prostředí, v němž byly listy změřeny. Neznamená však, že klima původu bylo populacím náhodně přiřazeno. Původní a zavlečené populace mají odlišnou geografickou i evoluční historii a klimatické proměnné spolu silně souvisejí. Naše jednoduché modely tak popisují asociace v těchto 128 populacích; samy nedokazují klimatickou adaptaci ani přímý příčinný účinek teploty.

Příklad úplného výsledku

Ve společné zahradě v Šanghaji měly populace křídlatky z teplejších míst původu v průměru tenčí listy: o 0,0091 mm na každý stupeň průměrné roční teploty (95% interval spolehlivosti $-0,0108$ až $-0,0075$ mm). Původní populace však pocházely převážně z teplejších míst než zavlečené. Při stejném původu byl sklon roční teploty $+0,0015$ mm na $1\text{ }^{\circ}\text{C}$ ($-0,0009$ až $0,0039$ mm), zatímco zavlečené populace měly při $14\text{ }^{\circ}\text{C}$ list silnější o 0,109 mm. Celkový klimatický trend tak z velké části odráží rozdíl mezi původním a zavlečeným areálem. Jde o asociace mezi 128 populacemi, nikoli o doklad příčinného účinku klimatu.

17 Co bude následovat dál

Aditivní model předpokládal, že sklon roční teploty je v obou skupinách stejný; proto byly přímky rovnoběžné. V následující lekci se budeme ptát, co dělat, když efekt jednoho prediktoru závisí na hodnotě nebo skupině jiného prediktoru. Později se vrátíme k otázce, jak mezi několika předem smysluplnými modely vybrat ten, který data popisuje dostatečně dobře bez zbytečné složitosti.

18 Shrnutí

- Koeficient ve víceprediktorovém modelu je podmíněný: popisuje rozdíl odezvy při stejných hodnotách ostatních prediktorů v rovnici.
- Přidání prediktoru může výrazně změnit velikost i směr odhadu, protože model odpovídá na jinou otázku. Když třetí proměnná souvisí s prediktorem i s odezvou, celkový vztah mísí více cest (konfundování).
- Korelované prediktory nesou překrývající se informaci; jejich samostatné příspěvky odhadujeme s větší nejistotou.
- Skupiny porovnáváme a predikce vytváříme jen pro kombinace hodnot, které mají v datech oporu.
- Pořadí prediktorů mění p-hodnoty sekvenčních testů z `anova()`, nikoli koeficienty a jejich testy v `summary()`.

18.1 Závěrečná otázka

Aditivní model kreslí pro původní i zavlečené populace rovnoběžné přímky. Jaký biologický vzorec by vyžadoval, aby měly obě skupiny různý sklon?

Možné odpovědi

Například kdyby zavlečené populace na chladnějších místech vytvářely silnější listy, ale u původních populací by teplota tloušťku listu téměř neovlivňovala. Nebo kdyby se populace s jinou historií přizpůsobovaly teplotě odlišně rychle. V obou případech by se vztah tloušťky k teplotě lišil podle původu a jedna společná přímka by ho nepopsala. Takové situace rozebereme v následující lekci.

Slovníček pojmů

Pojmy použité v těchto materiálech a jejich definice. Úplný průběžně aktualizovaný slovníček je dostupný také [online](#).

Interaktivní slovníček je dostupný online: <https://cuni-natur-biostatistics.github.io/slovník/>