

# Proč mají větší ostrovy více druhů rostlin?

## Tvar rozdělení, transformace a nelinearita — podkladové čtenářské materiály k přednášce

14. října 2026

### Obsah

|     |                                                                |    |
|-----|----------------------------------------------------------------|----|
| 1   | Co prozradil oblouk v residuích stromů?                        | 3  |
| 2   | Co se v této lekci naučíte                                     | 4  |
| 3   | Tvar rozdělení                                                 | 4  |
| 3.1 | Jeden řádek tabulky představuje jeden ostrov                   | 4  |
| 3.2 | Co v tabulce chybí a co je neobvyklé?                          | 7  |
| 3.3 | Jak vypadají počty druhů a plochy ostrovů?                     | 9  |
| 3.4 | Jak popsat tvar rozdělení slovy?                               | 10 |
| 3.5 | Proč je tolik biologických veličin pravostranně šikmých?       | 12 |
| 3.6 | K čemu slouží normální rozdělení?                              | 13 |
| 3.7 | Jak se tvar rozdělení projeví v Q-Q grafu?                     | 16 |
| 4   | Co lineární model předpokládá                                  | 17 |
| 4.1 | Jak počet druhů popíše přímka?                                 | 17 |
| 4.2 | Co přesně má být u lineárního modelu normální?                 | 18 |
| 4.3 | Co ukazují residua přímky?                                     | 24 |
| 4.4 | Jak jeden velký ostrov ovlivní celou přímku?                   | 27 |
| 4.5 | Co říká test normality residuí?                                | 28 |
| 4.6 | Proč výsledek testu závisí na počtu pozorování?                | 28 |
| 5   | Transformace: měřítko, nebo tvar?                              | 29 |
| 5.1 | Změní se tvar rozdělení, když plochu převedeme na hektary?     | 30 |
| 5.2 | Co se změní, když od plochy odečteme průměr?                   | 30 |
| 5.3 | Jak daleko od průměru leží ostrov?                             | 32 |
| 5.4 | Je Fernandina výjimečnější plochou, nebo výškou?               | 34 |
| 5.5 | Odstraní z-skóre šikmost?                                      | 36 |
| 5.6 | Jak logaritmus převádí násobky na stejné kroky?                | 37 |
| 5.7 | Co logaritmus udělá s tvarem rozdělení?                        | 38 |
| 5.8 | Které transformace mění jen měřítko a které i tvar?            | 42 |
| 6   | Tvar vztahu                                                    | 43 |
| 6.1 | Jak vypadá vztah plochy a počtu druhů na logaritmických osách? | 43 |
| 6.2 | Jaké tvary vztahu dávají biologický smysl?                     | 44 |
| 6.3 | Jak se kandidátní křivky shodují s ostrovy?                    | 46 |

|     |                                                                         |    |
|-----|-------------------------------------------------------------------------|----|
| 6.4 | Můžeme porovnat AIC modelu pro počet druhů a pro jeho logaritmus? ..... | 48 |
| 6.5 | Pomohl logaritmus také residuům? .....                                  | 49 |
| 7   | Interpretace po transformaci .....                                      | 52 |
| 7.1 | Kolikrát víc druhů má dvakrát větší ostrov? .....                       | 52 |
| 7.2 | Jak čist koeficient, když logaritmujeme jen jednu stranu? .....         | 54 |
| 7.3 | Jak rychle roste počet druhů s plochou ostrova? .....                   | 55 |
| 7.4 | Kdy transformaci použít a kdy ne? .....                                 | 57 |
|     | Shrnutí .....                                                           | 58 |
|     | Slovníček pojmů .....                                                   | 59 |



Obrázek 1: Souostroví Galapágy na družicovém snímku z 12. března 2002. Největší ostrov ve tvaru písmene L je Isabela, vlevo od něj leží Fernandina. Snímek: Jacques Descloitres, MODIS Rapid Response Project, NASA/GSFC (volné dílo).

Galapágy leží v Tichém oceánu asi tisíc kilometrů od pobřeží Ekvádoru. Na snímku je vidět, jak odlišné ostrovy souostroví tvoří: několik obrovských sopečných ostrovů a mnoho malých ostrůvků a skalek. Botanici na nich spočítali druhy cévnatých rostlin a zjistili, že větší ostrovy jich hostí více. Takový vztah ekologové znají z ostrovů, horských vrcholů i ostrůvků lesa v zemědělské krajině.

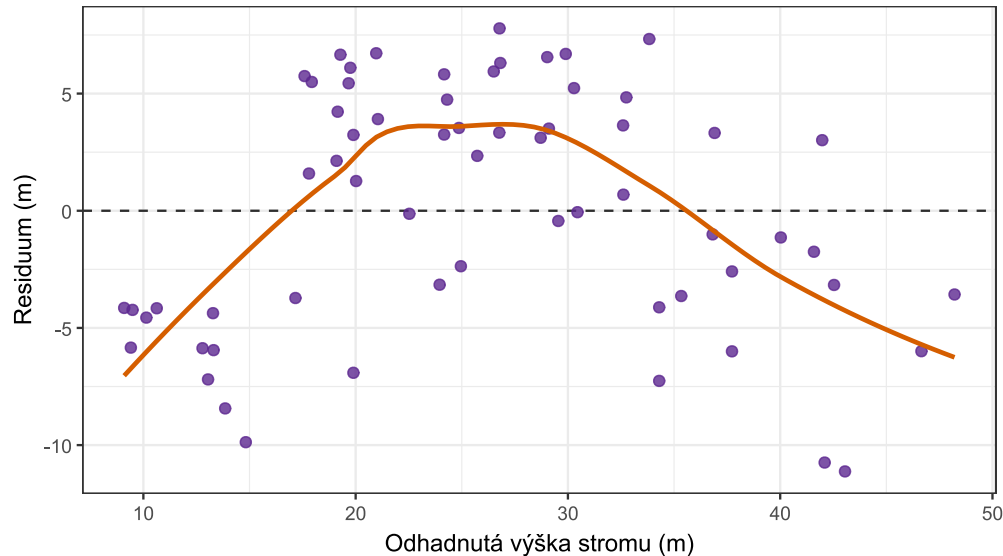
#### Hlavní otázka

Jak rychle roste počet druhů rostlin s plochou ostrova? Má ostrov, který je dvakrát větší, také dvakrát více druhů?

Než si odpovíte, zkuste odhadnout: kdyby ostrov měl desetkrát větší plochu, kolikrát více druhů byste na něm čekali? Na konci lekce svůj odhad porovnáte s tím, co ukážou data.

## 1 Co prozradil oblouk v residuích stromů?

Abychom na otázku o ostrovech mohli odpovědět, budeme potřebovat popsat zakřivený vztah. S takovým vztahem jsme se setkali už v předchozí lekci, kde jsme porovnávali kandidátní modely pro výšku jehličnanů. Nejlépe obhajitelný byl aditivní model s průměrem kmene a druhem stromu. U jeho residuů však zůstal jeden vzor, ke kterému se nyní vrátíme.



Obrázek 2: Residua aditivního modelu pro výšku stromů z předchozí lekce podle odhadnuté výšky. Oranžová křivka pouze vyhlazuje průběh bodů, aby byl vzor lépe vidět.

### Co residua říkají o přímce?

Kde leží residua převážně nad nulou a kde pod nulou? Co by to znamenalo pro tvar vztahu mezi průměrem kmene a výškou stromu?

U nejmenších a největších stromů leží residua převážně pod nulou, uprostřed nad nulou. Model tedy u krajních stromů výšku systematicky přeceňuje a u středně velkých podceňuje. **Takový oblouk v residuích naznačuje, že skutečný vztah je zakřivený, zatímco model předpokládá přímku.** Žádná z porovnávaných přímek tento vzor neodstranila. Zda a jak oblouk zmizí, zjistíme ve chvíli, kdy budeme umět měnit tvar vztahu; k těmto stromům se vrátíme na konci lekce.

Předchozí lekce proto skončila otevřenou otázkou: co dělat, když žádný kandidát složený z přímek nestačí? Tato lekce na ni odpoví. Uvidíme, že lineární model umí popsat i zakřivený vztah, pokud vhodně upravíme proměnné. Předtím se ale musíme naučit popisovat **tvar** – tvar rozdělení hodnot i tvar vztahu.

## 2 Co se v této lekci naučíte

Po prostudování lekce budete umět:

- popsat tvar rozdělení proměnné nebo residuí pomocí histogramu a Q–Q grafu a posoudit, zda je tento tvar pro lineární model podstatný;
- rozlišit transformace, které mění jen polohu a měřítko hodnot (změna jednotek, centrování, z-skóre), od transformací, které mění tvar (logaritmus, odmocnina);
- rozpoznat zakřivený vztah v datech a navrhnout jednoduché řešení;
- porovnat několik biologicky smysluplných tvarů vztahu jako kandidátní modely;
- vysvětlit, jak transformace mění interpretaci koeficientů modelu.

## 3 Tvar rozdělení

### 3.1 Jeden řádek tabulky představuje jeden ostrov

Budeme pracovat s klasickými daty o rostlinách Galapág, která publikovali M. P. Johnson a P. H. Raven (1973). Každý řádek tabulky představuje jeden ostrov. Pro každý ostrov známe:

| Proměnná                 | Význam                                           | Jednotka        |
|--------------------------|--------------------------------------------------|-----------------|
| ostrov                   | název ostrova                                    | —               |
| pocet_druhu              | počet druhů cévnatých rostlin                    | počet           |
| pocet_endemitu           | počet druhů, které nerostou nikde jinde na světě | počet           |
| plocha_km2               | plocha ostrova                                   | km <sup>2</sup> |
| nejvyssi_bod_m           | nadmořská výška nejvyššího bodu ostrova          | m n. m.         |
| vzdalenost_nejblizsi_km  | vzdálenost k nejbližšímu ostrovu                 | km              |
| vzdalenost_santa_cruz_km | vzdálenost k ostrovu Santa Cruz                  | km              |
| plocha_sousedniho_km2    | plocha nejbližšího sousedního ostrova            | km <sup>2</sup> |

Hlavní odezvou bude počet druhů rostlin a hlavním prediktorem plocha ostrova. Ostatní proměnné použijeme pro srovnání tvarů rozdělení a pro práci s různými jednotkami.

```
# Načteme připravenou výukovou tabulku
data_ostrovny <-
  read.csv(
```

```

file = here::here("data", "druhy_galapagy.csv"),
stringsAsFactors = FALSE
)

```

### Doplňující: původ dat a příprava tabulky

Tabulka pochází z datasetu `galamiss` v balíčku `faraway` verze 1.0.9 (J. Faraway), který je šířen pod licencí GPL. Hodnoty převzal z článku M. P. Johnson a P. H. Raven (1973): Species number and endemism: The Galapagos Archipelago revisited. *Science* 179: 893–895. Používáme verzi `galamiss`, která zachovává původní chybějící hodnoty. Sesterský dataset `gala` je pro pohodlí doplňuje, a proto jej nepoužíváme.

Následující kód stáhne zdrojový balíček z archivu CRAN, ověří jeho otisk SHA-256, načte z něj tabulku bez instalace balíčku a uloží výukovou verzi s českými názvy proměnných. Řádky ani hodnoty nemění.

```

# Stáhneme připnutou verzi zdrojového balíčku z archivu CRAN
cesta_archiv <-
  tempfile(fileext = ".tar.gz")

download.file(
  url = paste0(
    "https://cran.r-project.org/src/contrib/Archive/",
    "faraway/faraway_1.0.9.tar.gz"
  ),
  destfile = cesta_archiv,
  mode = "wb"
)

# Ověříme, že jsme stáhli přesně očekávaný soubor
stopifnot(
  digest::digest(
    object = cesta_archiv,
    algo = "sha256",
    file = TRUE
  ) == paste0(
    "1074ed85a0cf018fc2fd3e2a48bed666",
    "b2fa06b8461b7a3820aeb67b34a52d5d"
  )
)

# Rozbalíme z archivu jen soubor s daty

```

```

cesta_slozka <- tempfile()

untar(
  tarfile = cesta_archiv,
  files = "faraway/data/galamiss.rda",
  exdir = cesta_slozka
)

# Načteme původní tabulku galamiss
load(
  file = file.path(
    cesta_slozka,
    "faraway",
    "data",
    "galamiss.rda"
  )
)

# Názvy ostrovů jsou uloženy jen jako názvy řádků bez mezer
vec_nazvy_ostrovu <-
  gsub(
    pattern = "\\.",
    replacement = " ",
    x = rownames(galamiss)
  )

vec_nazvy_ostrovu <-
  gsub(
    pattern = "([a-z])([A-Z0-9])",
    replacement = "\\1 \\2",
    x = vec_nazvy_ostrovu
  )

# Sestavíme výukovou tabulku s českými názvy proměnných
data_ostrovypriprava <-
  data.frame(
    ostrov = vec_nazvy_ostrovu,
    pocet_druhu = as.integer(galamiss$Species),
    pocet_endemitu = as.integer(galamiss$Endemics),
    plocha_km2 = galamiss$Area,
    nejvyssi_bod_m = galamiss$Elevation,
    vzdalenost_nejbližsi_km = galamiss$Nearest,
  )

```

```

    vzdalenost_santa_cruz_km = galamiss$Scruz,
    plocha_sousedniho_km2 = galamiss$Adjacent
  )

# Uložíme tabulku; chybějící hodnoty zůstanou prázdné
write.csv(
  x = data_ostrovy_priprava,
  file = here::here("data", "druhy_galapagy.csv"),
  row.names = FALSE,
  na = ""
)

```

Kód potřebuje balíček `digest`. Stejný postup obsahuje skript `R/prepare_galapagos_data.R` v repozitáři lekce.

### 3.2 Co v tabulce chybí a co je neobvyklé?

Než začneme cokoli modelovat, tabulku prohlédneme. Zajímá nás struktura proměnných, chybějící hodnoty a rozsahy, které by mohly prozradit chybu nebo neobvyklé pozorování.

```

# Prohlédneme první řádky tabulky
head(
  x = data_ostrovy,
  n = 6
)

```

|   | ostrov       | pocet_druhu | pocet_endemitu | plocha_km2 | nejvyssi_bod_m |
|---|--------------|-------------|----------------|------------|----------------|
| 1 | Baltra       | 58          | 23             | 25.09      | NA             |
| 2 | Bartolome    | 31          | 21             | 1.24       | 109            |
| 3 | Caldwell     | 3           | 3              | 0.21       | 114            |
| 4 | Champion     | 25          | 9              | 0.10       | 46             |
| 5 | Coamano      | 2           | 1              | 0.05       | NA             |
| 6 | Daphne Major | 18          | 11             | 0.34       | NA             |

|   | vzdalenost_nejbлизsi_km | vzdalenost_santa_cruz_km | plocha_sousedniho_km2 |
|---|-------------------------|--------------------------|-----------------------|
| 1 | 0.6                     | 0.6                      | 1.84                  |
| 2 | 0.6                     | 26.3                     | 572.33                |
| 3 | 2.8                     | 58.7                     | 0.78                  |
| 4 | 1.9                     | 47.4                     | 0.18                  |
| 5 | 1.9                     | 1.9                      | 903.82                |
| 6 | 8.0                     | 8.0                      | 1.84                  |

```
# Ověříme datové typy proměnných
str(object = data_ostrov)
```

```
'data.frame': 30 obs. of 8 variables:
 $ ostrov          : chr  "Baltra" "Bartolome" "Caldwell" "Champion" ...
 $ pocet_druhu     : int   58 31 3 25 2 18 24 10 8 2 ...
 $ pocet_endemitu  : int   23 21 3 9 1 11 0 7 4 2 ...
 $ plocha_km2      : num   25.09 1.24 0.21 0.1 0.05 ...
 $ nejvyssi_bod_m  : int   NA 109 114 46 NA NA 93 168 NA 112 ...
 $ vzdalenost_nejbлизsi_km : num  0.6 0.6 2.8 1.9 1.9 8 6 34.1 0.4 2.6 ...
 $ vzdalenost_santa_cruz_km: num  0.6 26.3 58.7 47.4 1.9 ...
 $ plocha_sousedniho_km2   : num  1.84 572.33 0.78 0.18 903.82 ...
```

```
# Spočítáme chybějící hodnoty v každém sloupci
colSums(x = is.na(data_ostrov))
```

|                          |                       |                         |
|--------------------------|-----------------------|-------------------------|
| ostrov                   | pocet_druhu           | pocet_endemitu          |
| 0                        | 0                     | 0                       |
| plocha_km2               | nejvyssi_bod_m        | vzdalenost_nejbлизsi_km |
| 0                        | 6                     | 0                       |
| vzdalenost_santa_cruz_km | plocha_sousedniho_km2 |                         |
| 0                        | 0                     |                         |

```
# Zkontrolujeme rozsahy hlavních proměnných
range(x = data_ostrov$pocet_druhu)
```

```
[1] 2 444
```

```
range(x = data_ostrov$plocha_km2)
```

```
[1] 0.01 4669.32
```

Tabulka obsahuje 30 ostrovů. Počty druhů i plochy jsou kompletní; nejvyšší bod chybí u 6 ostrovů. Při prohlídce si všimneme tří věcí:

- **Plochy ostrovů se liší extrémně.** Nejmenší ostrov má 0.01 km<sup>2</sup>, největší ostrov Isabela 4 669 km<sup>2</sup>. Největší ostrov je tedy zhruba 466 932krát větší než nejmenší.
- **Ostrov Santa Cruz má vzdálenost k ostrovu Santa Cruz rovnou nule**, protože jde o tentýž ostrov. Hodnota není chybou, ale je to skutečná nula; bude důležitá, až budeme hodnoty logaritmovat.

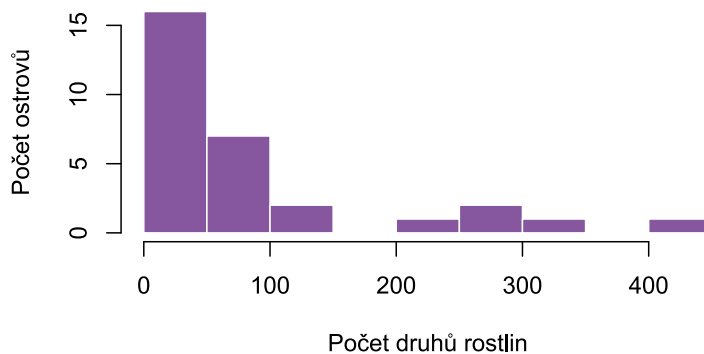
- **Ostrov Daphne Minor nemá žádný endemický druh.** Také tato nula je skutečná hodnota.

Záporné hodnoty ani zjevné chyby jednotek v tabulce nevidíme. Počty a plochy jsou biologicky věrohodné.

### 3.3 Jak vypadají počty druhů a plochy ostrovů?

Rozdělení číselné proměnné nejrychleji uvidíme v histogramu (anglicky *histogram*). Vodorovná osa ukazuje hodnoty proměnné rozdělené do košů, svislá osa četnost ostrovů v každém koši.

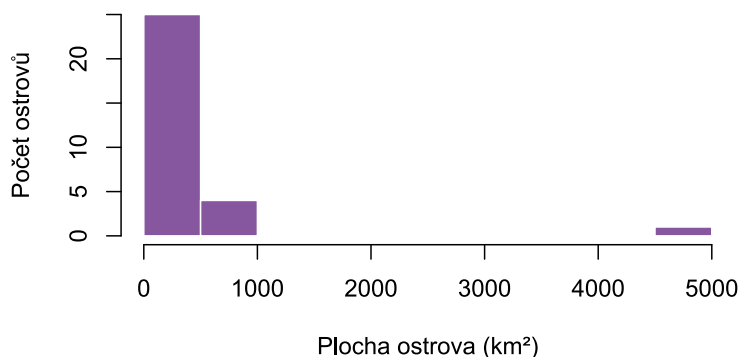
```
# Zobrazíme rozdělení počtu druhů
hist(
  x = data_ostrovny$pocet_druhu,
  breaks = 15,
  col = "#86579E",
  border = "white",
  main = "",
  xlab = "Počet druhů rostlin",
  ylab = "Počet ostrovů"
)
```



Obrázek 3: Histogram počtu druhů rostlin na 30 ostrovech.

```
# Zobrazíme rozdělení plochy ostrovů
hist(
  x = data_ostrovny$plocha_km2,
  breaks = 15,
  col = "#86579E",
  border = "white",
  main = "",
  xlab = "Plocha ostrova (km²)",
  ylab = "Počet ostrovů"
)
```

```
ylab = "Počet ostrovů"
)
```



Obrázek 4: Histogram plochy 30 ostrovů.

Kde leží většina ostrovů?

Kde v obou histogramech leží většina ostrovů? Co s obrázkem dělá několik největších ostrovů?

Většina ostrovů je malá a má málo druhů: 17 z 30 ostrovů má plochu menší než 10 km<sup>2</sup> a 23 ostrovů má méně než 100 druhů. Několik velkých ostrovů vytváří na pravé straně histogramu dlouhý řídký „ocas“. U plochy je to extrémní: téměř všechny ostrovy se vejdu do prvního koše a celý zbytek osy zabírá jen hrstka obrů. **Histogram plochy tak o malých ostrovech neříká téměř nic.**

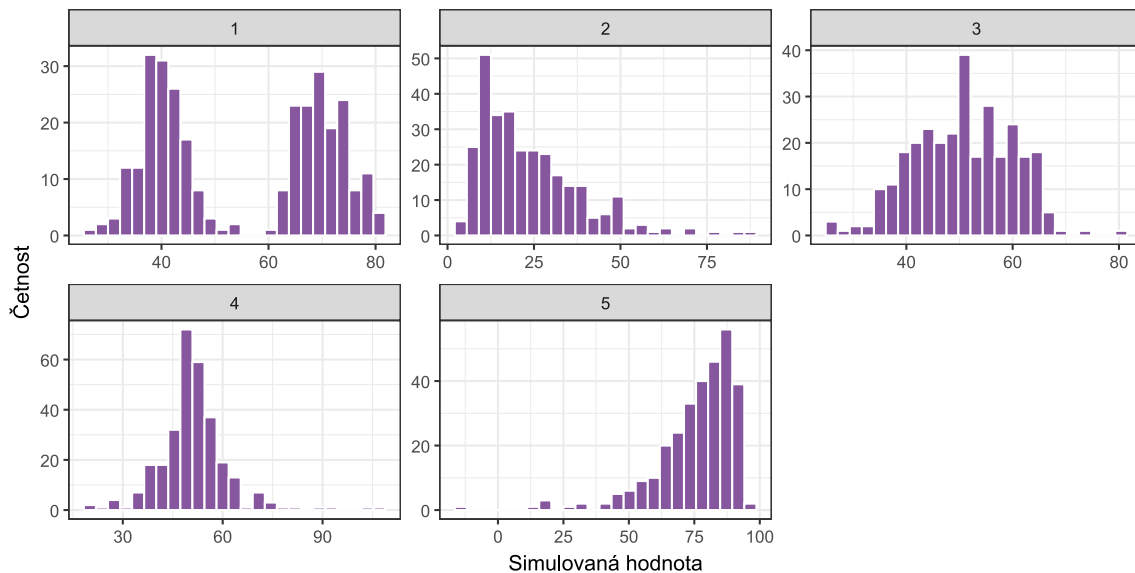
### 3.4 Jak popsat tvar rozdělení slovy?

Pro takové obrázky potřebujeme slova. Tvar rozdělení popisujeme podle několika vlastností:

- **Souměrnost.** Levá a pravá polovina rozdělení vypadají přibližně jako zrcadlový obraz.
- **Šikmost** (anglicky *skewness*). Rozdělení má jeden ocas delší než druhý. Rozdělení s dlouhým ocasem vpravo nazýváme **pravostranně šikmé** (*right-skewed*), s dlouhým ocasem vlevo **levostranně šikmé** (*left-skewed*). Šikmost pojmenováváme podle strany dlouhého ocasu, nikoli podle strany, kde leží většina hodnot.
- **Těžké konce** (*heavy tails*). Rozdělení je souměrné, ale velmi vzdálené hodnoty na obou stranách se objevují častěji, než bychom u souměrného zvonovitého tvaru čekali (tento referenční tvar pojmenujeme o několik kapitol dál). Stejnou vlastnost jsme v lekci o testování hypotéz popsali u t-rozdělení s malým počtem stupňů volnosti jako **těžší chvosty**.

- **Vícevrcholovost** (*multimodality*). Rozdělení má dva nebo více vrcholů, například proto, že data obsahují dvě odlišné skupiny. Tak tomu bylo u délek erupcí gejzíru Old Faithful, které tvořily skupinu krátkých a skupinu dlouhých erupcí.
- **Omezení hodnot**. Některé veličiny nemohou klesnout pod nulu (plocha, počet, hmotnost) nebo leží mezi nulou a jedničkou (podíly). Hranice často tvar rozdělení ovlivňuje.

Abychom tvary viděli čistě, připravili jsme pět **simulovaných** souborů po 300 hodnotách. Nejde o skutečná měření; každý soubor jsme vytvořili tak, aby měl jeden typický tvar.



Obrázek 5: Pět simulovaných rozdělení 1–5. Každý panel má vlastní vodorovnou osu.

#### Přiřadte tvary

Přiřadte panelům 1–5 slova *souměrné*, *pravostranně šikmé*, *levostranně šikmé*, *těžké konce* a *dva vrcholy*.

#### Řešení

1 = dva vrcholy, 2 = pravostranně šikmé, 3 = souměrné, 4 = těžké konce, 5 = levostranně šikmé. Panel 4 se od panelu 3 liší hlavně ojedinělými hodnotami daleko od středu; jeho střed je naopak užší.

Počet druhů i plocha ostrovů jsou tedy **pravostranně šikmé**: většina hodnot leží vlevo a dlouhý ocas míří doprava. U plochy je šikmost extrémní.

Šikmost poznáme i bez obrázku, porovnáním průměru a mediánu. V první lekci jsme u tělesné hmotnosti savců viděli, že dlouhý ocas táhne průměr k sobě, zatímco medián zůstává uprostřed seřazených hodnot. Tehdy jsme taková data nazývali „výrazně asymetrická“; přesnější název je

**pravostranně šikmá.** Průměrný počet druhů je 85.2, medián jen 42. Průměrná plocha je 261.7 km<sup>2</sup>, medián 2.59 km<sup>2</sup>. **Průměr plochy tak popisuje ostrov, jakých je v souostroví jen několik.**

Doplňující: Jak šikmost vyjádřit jedním číslem?

Šikmost lze shrnout jedním číslem, **koefficientem šikmosti**. Pro každou hodnotu spočítáme odchylku od průměru, vydělíme ji směrodatnou odchylkou a výsledek umocníme na třetí. Třetí mocnina zachová znaménko, takže dlouhý ocas vpravo převáží kladnými hodnotami a dlouhý ocas vlevo zápornými. Koefficient je průměr těchto třetích mocnin:

$$\text{koefficient šikmosti} = \text{průměr z } \left( \frac{\text{hodnota} - \text{průměr}}{\text{směrodatná odchylka}} \right)^3.$$

```
# Koefficient šikmosti počtu druhů
vec_druhy <- data_ostrovypocet_druhu

mean(
  x = ((vec_druhy - mean(x = vec_druhy)) / sd(x = vec_druhy))^3
)
```

```
[1] 1.707024
```

| Proměnná       | Koefficient šikmosti |
|----------------|----------------------|
| Počet druhů    | 1.71                 |
| Plocha ostrova | 4.42                 |
| Nejvyšší bod   | 1.44                 |

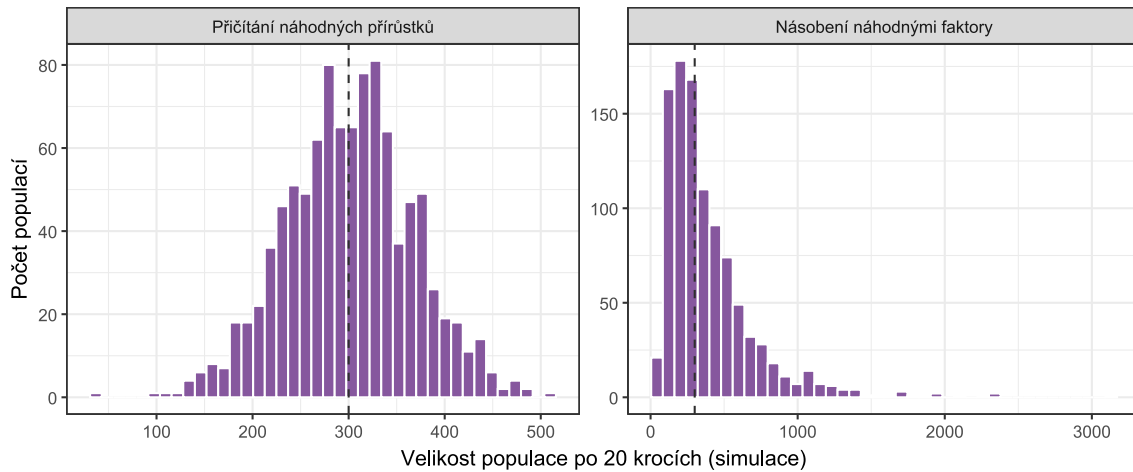
Kladné hodnoty odpovídají pravostranné šikmosti, záporné levostranné, hodnoty blízko nuly přibližně souměrnému tvaru. Číslo je užitečné pro stručný popis, **ale neexistuje hranice, od které by šikmost „vadila“**. O tom, zda tvar pro model něco znamená, rozhoduje graf, otázka a to, kterých hodnot se tvar týká. V literatuře najdete i varianty vzorce, které se pro malé vzorky mírně liší.

### 3.5 Proč je tolik biologických veličin pravostranně šikmých?

Pravostranná šikmost není u biologických dat výjimkou, ale pravidlem. Plochy, hmotnosti, koncentrace, velikosti populací nebo počty druhů bývají šikmé ze dvou důvodů.

**Hranice v nule.** Plocha ani počet druhů nemohou být záporné. Když je většina hodnot blízko nule, rozdělení se nemá kam rozšířit doleva, ale doprava ano.

**Růst násobením.** Mnoho biologických procesů nepřičítá stále stejný přírůstek, ale násobí. Populace roste o určité procento, buňka se dělí, strom přirůstá úměrně své velikosti. Simulace v následujícím obrázku ukazuje rozdíl. Celkem 1000 simulovaných populací začíná na stejné velikosti a 20krát se změní. Vlevo se v každém kroku k velikosti přičte náhodný přírůstek, vpravo se velikost vynásobí náhodným faktorem.



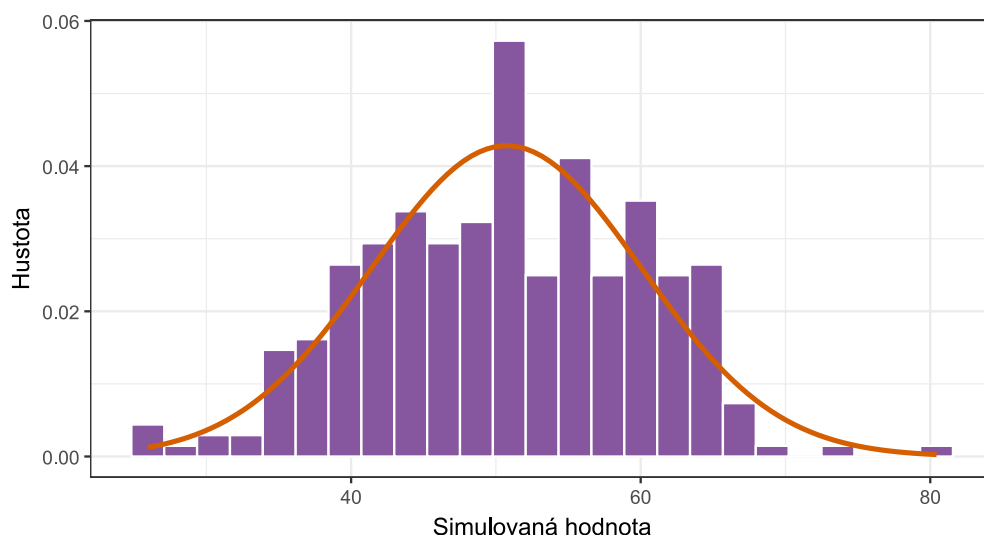
Obrázek 6: Simulované velikosti 1000 populací po 20 náhodných změnách. Vlevo se přírůstky přičítají, vpravo se velikost násobí.

Když se sčítá mnoho malých náhodných vlivů, kladné a záporné přírůstky se navzájem vyrovnávají a vzniká souměrný zvon. Když se náhodné vlivy násobí, populace, které už jsou velké, rostou ve stejných procentech o mnohem více jedinců než malé. Několik populací proto uteče daleko doprava a vznikne pravostranně šikmé rozdělení.

Na Galapágách se na šikmosti plochy podílí hlavně hranice v nule: malých ostrůvků a skalek je mnoho, velkých sopečných ostrovů jen několik a plocha nemůže klesnout pod nulu. Šikmý počet druhů pak zčásti kopíruje šikmou plochu, protože větší ostrovy nabízejí více prostředí pro více druhů. **Šikmost tedy sama o sobě není chybou dat. Je to vlastnost biologie, kterou musíme znát a zohlednit při volbě modelu.**

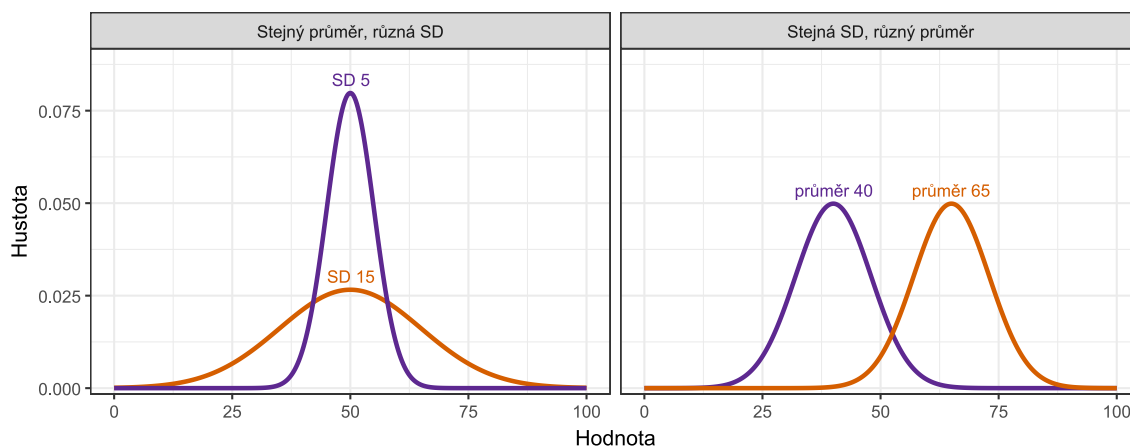
### 3.6 K čemu slouží normální rozdělení?

Souměrný zvon z levého panelu simulace už známe. V předchozích lekcích jsme jej používali jako referenční tvar v Q–Q grafech residuí a podobný tvar má také t-rozdělení, ze kterého jsme počítali p-hodnoty a intervaly spolehlivosti. Tento zvonovitý tvar se nazývá **normální rozdělení** (anglicky *normal distribution*, také Gaussovo rozdělení).



Obrázek 7: Simulovaný souměrný soubor 3 z galerie tvarů a křivka normálního rozdělení se stejným průměrem a směrodatnou odchylkou.

Normální rozdělení je souměrné a má jeden vrchol. Jeho tvar určují jen dvě čísla, která známe už z první lekce: průměr a směrodatná odchylka (SD). **Průměr říká, kde leží střed zvonu. SD říká, jak je zvon široký.**

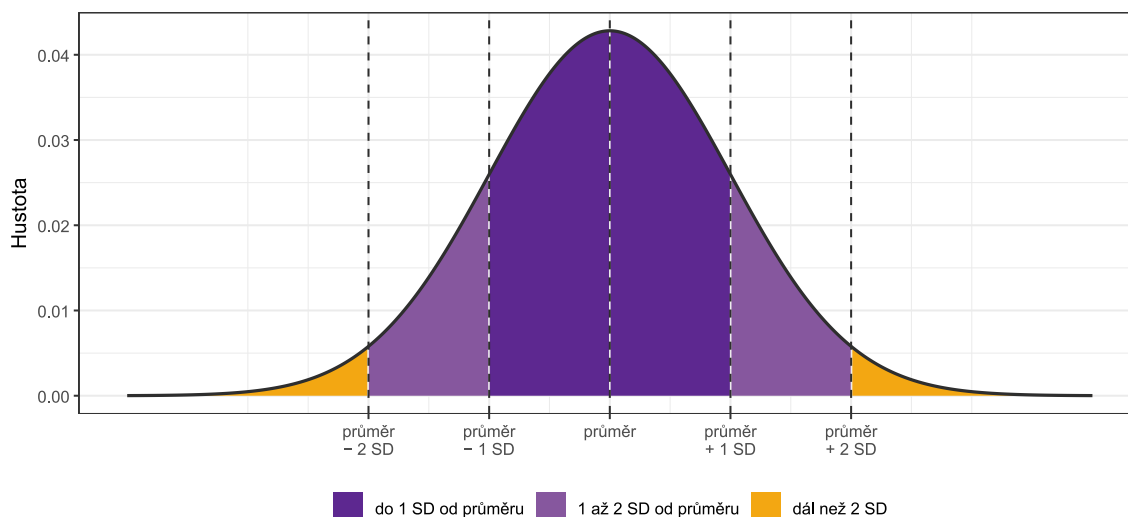


Obrázek 8: Normální rozdělení s různými průměry a směrodatnými odchylkami. Vlevo stejný průměr a různá SD, vpravo stejná SD a různý průměr.

Změna průměru posune celý zvon doleva nebo doprava, aniž by změnila jeho tvar. Větší SD zvon roztáhne do šířky, a protože plocha pod křivkou zůstává stejná, zvon zároveň sníží. Hodnoty daleko od středu jsou u normálního rozdělení vzácné. Jak vzácné, můžeme vyjádřit právě v jednotkách SD.

Kolik hodnot leží blízko průměru?

Představte si přibližně normální veličinu, například délku křídla u velkého vzorku ptáků jednoho druhu. Kolik procent jedinců odhadujete, že bude mít délku křídla nejvýš o jednu SD menší nebo větší než průměr? A kolik do dvou SD?



Obrázek 9: Normální rozdělení se stejným průměrem a SD jako simulovaný souměrný soubor 3. Barva ukazuje, jak daleko od průměru hodnoty leží, měřeno v SD.

#### Průměr a SD u normálního rozdělení

- Přibližně 68% hodnot leží nejvýš **jednu SD** od průměru.
- Přibližně 95% hodnot leží nejvýš **dvě SD** od průměru.
- Přibližně 99.7% hodnot leží nejvýš **tři SD** od průměru; vzdálenější hodnoty jsou u normálního rozdělení velmi vzácné.

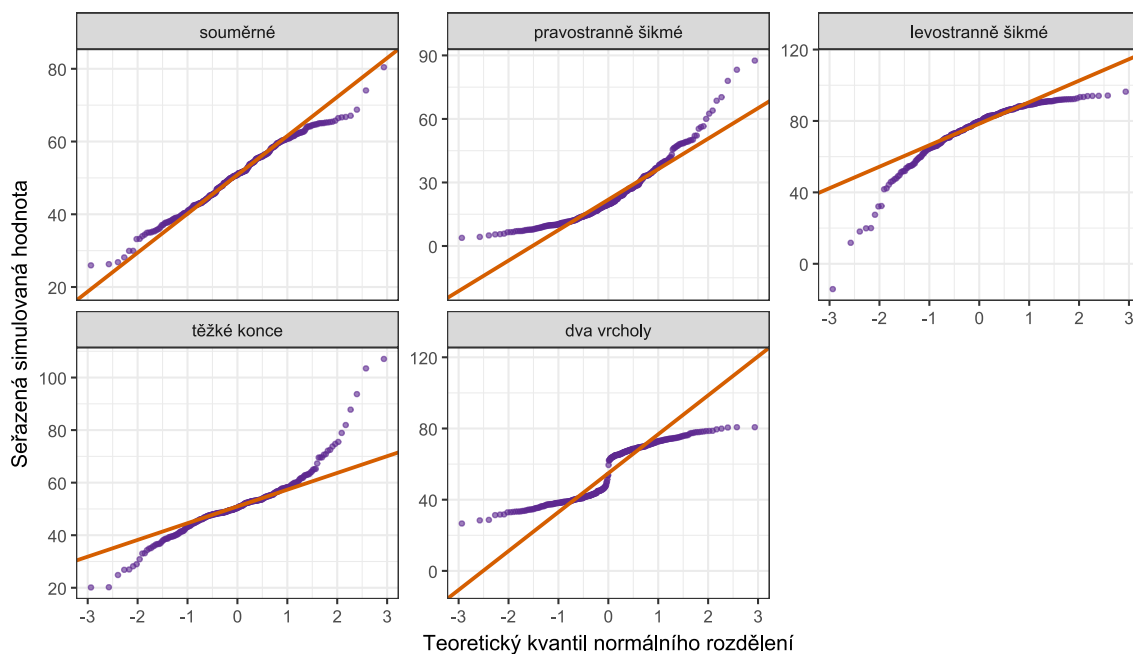
V simulovaném souměrném souboru 3 leží do jedné SD od průměru 64% hodnot a do dvou SD 97% hodnot, tedy blízko pravidlu; v souboru o 300 hodnotách se podíly od teoretických trochu liší náhodou. Podobnou myšlenku jsme použili u intervalů spolehlivosti. Odhady sklonu z opakovaných výběrů kolísaly kolem skutečné hodnoty přibližně zvonovitě a jejich „směrodatnou odchylkou“ byla standardní chyba. Proto 95% interval sahal od odhadu přibližně dvě standardní chyby na každou stranu (přesněji o kritickou hodnotu  $t$ , u malých souborů o něco větší než 2). Interval tedy nepokrývá 95 % pozorování; s 95% spolehlivostí zachycuje skutečnou hodnotu sklonu.

**U šikmých dat toto pravidlo neplatí.** Průměrná plocha ostrovů je  $261.7 \text{ km}^2$  a SD  $864.1 \text{ km}^2$ . Průměr minus jedna SD dává  $-602.4 \text{ km}^2$ , tedy zápornou plochu, která nemůže existovat. Do jedné SD od průměru přitom padá 97% ostrovů místo přibližně 68%. Dvojice průměr a SD tedy dobře popisuje hlavně přibližně souměrná data; proto jsme v první lekci u výrazně asymetrických dat doporučili medián a mezikvartilové rozpětí.

**Normální rozdělení je užitečné měřítko, nikoli výchozí vlastnost biologických dat.** Jak jsme viděli, mnoho biologických veličin normální není a ani být nemá. Normální rozdělení potřebujeme proto, že s ním lineární model počítá u jedné konkrétní části modelu. Kterou částí to je, zjistíme v další kapitole.

### 3.7 Jak se tvar rozdělení projeví v Q–Q grafu?

Histogram je názorný, ale u malého počtu pozorování silně závisí na volbě šířky koše (anglicky *bin width*), jak jsme viděli v první lekci. Proto jsme při kontrole residuí používali Q–Q graf (anglicky *quantile–quantile plot*). Ten porovnává seřazené hodnoty s kvantily, které bychom čekali u normálního rozdělení. Pokud mají data normální tvar, body leží přibližně na přímce. Každý jiný tvar má svůj typický vzor odchylek.



Obrázek 10: Q–Q grafy stejných pěti simulovaných souborů jako v galerii tvarů, nyní seřazené podle názvu tvaru. Přímka prochází prvním a třetím kvantilem.

Najděte typický vzor

Porovnejte Q–Q grafy s histogramy v galerii tvarů. Jak se v Q–Q grafu projeví dlouhý ocas vpravo? A jak dva vrcholy?

Jak číst odchylky v Q–Q grafu

- **Souměrné, přibližně normální:** body leží blízko přímky, jen na krajích mírně kolísají.

- **Pravostranně šikmé:** body vpravo nahoře se zvedají nad přímkou a celý graf se prohne do oblouku. Největší hodnoty jsou mnohem větší, než by normální rozdělení čekalo.
- **Levostranně šikmé:** zrcadlový obraz; body vlevo dole klesají pod přímkou.
- **Těžké konce:** body na obou koncích se od přímky odklánějí opačnými směry, levý dolů a pravý nahoru, takže graf připomíná natažené písmeno S.
- **Dva vrcholy:** body tvoří schod, protože mezi dvěma skupinami je méně hodnot, než by normální rozdělení čekalo.

I u dokonale normálních dat body kolem přímky trochu kolísají, nejvíce na krajích. S malým počtem pozorování je toto kolísání větší. Proto hledáme **systematický vzor**, ne jednotlivé body mimo přímkou.

Nový slovník pomáhá přecházet i grafy z předchozích lekcí. U residuí kosatců v lekci o první přímce se oba konce Q–Q grafu mírně odkláněly opačnými směry: residua měla **slabě těžké konce**.

## 4 Co lineární model předpokládá

### 4.1 Jak počet druhů popíše přímkou?

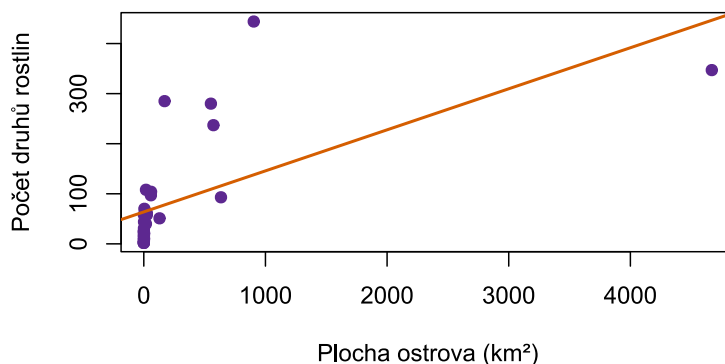
Vraťme se k hlavní otázce. Nejjednodušší model, který známe, je přímkou: počet druhů jako lineární funkce plochy ostrova. Vztah zobrazíme v bodovém grafu (anglicky *scatter plot*) a odhadneme lineární model.

```
# Odhadneme přímkou pro počet druhů podle plochy
mod_primka <-
  lm(
    formula = pocet_druhu ~ plocha_km2,
    data = data_ostrovny
  )

# Zobrazíme ostrovy a odhadnutou přímkou
plot(
  x = data_ostrovny$plocha_km2,
  y = data_ostrovny$pocet_druhu,
  pch = 19,
  col = "#5D2890",
  xlab = "Plocha ostrova (km²)",
  ylab = "Počet druhů rostlin"
)

abline(
  reg = mod_primka,
  col = "#D55E00",
```

```
lwd = 2  
)
```



Obrázek 11: Počet druhů rostlin a plocha ostrova s odhadnutou přímkou.

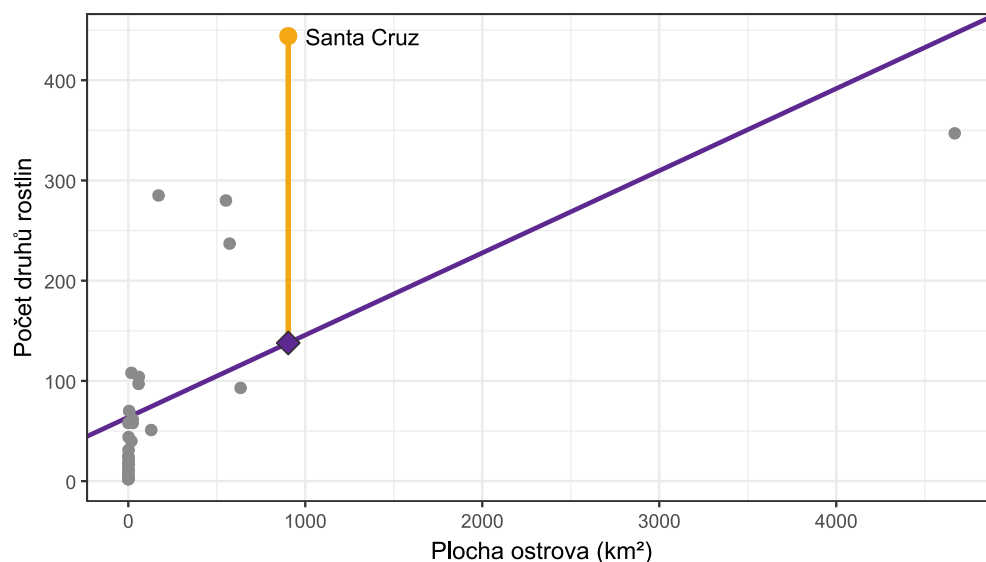
```
# Vypíšeme odhadnutý intercept a sklon  
coef(object = mod_primka)
```

```
(Intercept)  plocha_km2  
63.78286147  0.08196317
```

Přímka říká, že každý další kilometr čtvereční přidá v průměru 0.082 druhu, tedy asi 8.2 druhu na 100 km<sup>2</sup>. Obrázek ale ukazuje problém: 17 ostrovů je namačkaných u levého okraje grafu a o sklonu přímky rozhoduje hlavně několik velkých ostrovů vpravo. Jak dobře přímka data popisuje, zjistíme z residuů. Předtím si ale ujasníme, co přesně u nich kontrolujeme.

## 4.2 Co přesně má být u lineárního modelu normální?

Viděli jsme, že počet druhů i plocha jsou silně pravostranně šikmé. Znamená to, že lineární model pro tato data nelze použít? Odpověď začíná otázkou, z čeho se počet druhů jednoho ostrova podle modelu skládá.



Obrázek 12: Počet druhů ostrova Santa Cruz rozložený na odhad přímky (kosočtverec na přímce) a residuum (svislá úsečka).

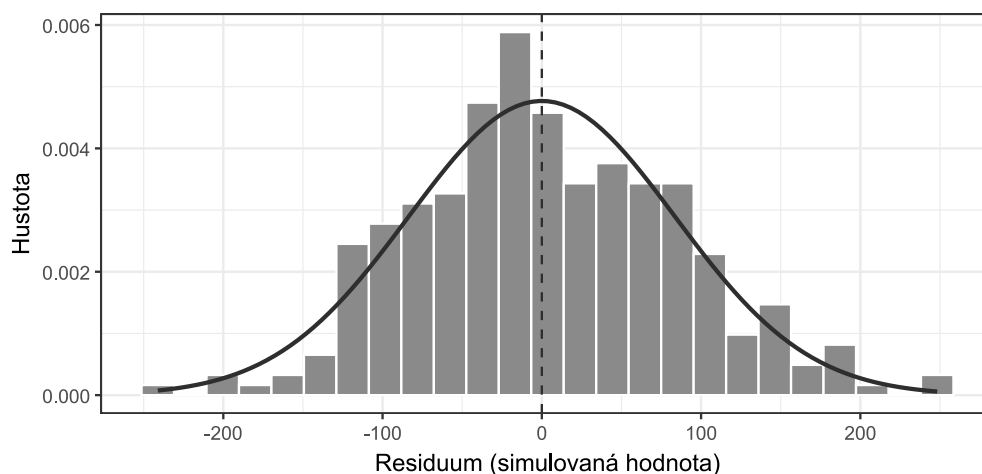
Ostrov Santa Cruz má 444 druhů. Přímka pro jeho plochu odhaduje 138 druhů a zbytek, 306 druhů, je jeho residuum:

$$444 = 138 + 306.$$

Totéž platí pro každý ostrov: **pozorovaná hodnota = odhad přímky + residuum**. Přímka nese systematickou část vztahu, residuum to, co přímka nevysvětlila.

Jak by vypadala residua, kdyby přímka seděla?

Představte si, že by přímka zachytila všechno systematické, co plocha o počtu druhů prozrazuje. Co by v residuích zbylo? Načrtněte, jak by vypadal jejich histogram.



Obrázek 13: Simulace: 300 residuí přímky, která vztah vystihuje. Směrodatná odchylka je stejná jako u residuí ostrovů; tmavá křivka je normální rozdělení.

Kdyby přímka vystihovala vztah, zbyly by v residuích jen náhodné odchylky: kladné i záporné, malé častěji než velké, rozložené souměrně kolem nuly. Tvar se blíží zvonu normálního rozdělení z předchozí kapitoly. **Právě takový tvar residuí lineární model předpokládá.** O tvaru samotné odezvy nebo prediktoru přitom nic netvrdí. Ověřte si, že je to jasné:

Ověřte si: co musí mít přibližně normální rozdělení?

Lineární model předpokládá přibližně normální rozdělení:

- a) odezvy (počtu druhů),
- b) prediktoru (plochy ostrova),
- c) residuí modelu,
- d) odezvy i prediktoru zároveň.

#### Řešení

**Správně je c).** Lineární model popisuje odezvu jako součet systematické části (přímky) a náhodných odchylek kolem ní. Předpoklad normality se týká právě těchto odchylek, tedy residuí.

- a) Odezva může být šikmá, například proto, že šikmý je prediktor. Pokud model vysvětlí, proč jsou některé hodnoty velké, residua mohou být souměrná i u šikmé odezvy.
- b) O rozdělení prediktoru model nepředpokládá nic. Hodnoty prediktoru bereme jako dané.
- d) Tento požadavek je ještě přísnější než a) a model jej nemá.

Častá chyba: kontrolovat normalitu dat místo residuí

Histogram nebo test normality samotné odezvy či prediktoru nerozhoduje o tom, zda je lineární model vhodný. Rozhodují residua odhadnutého modelu. Šikmý prediktor ani šikmá odezva samy o sobě předpoklad neporušují. Totéž jsme poprvé uvedli u přímky pro kosatce: lineární model nevyžaduje, aby samotná odezva měla normální rozdělení. Příklad modelu, jehož odezva je silně šikmá, a residua přesto vyhovující, uvidíme v části o tvaru vztahu.

Doplňující: měly modely z předchozích lekcí přibližně normální residua?

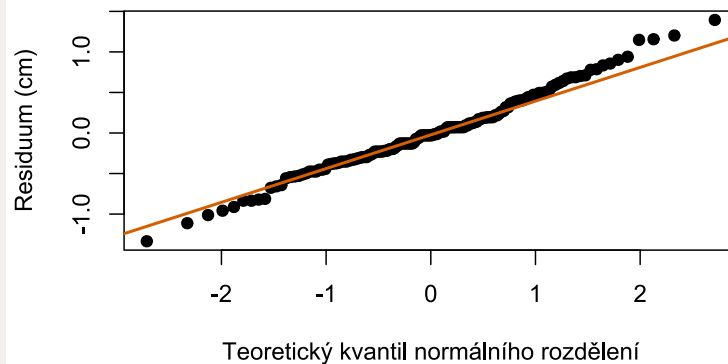
Předpoklad o tvaru residuí platil pro každý lineární model v kurzu, ale ne vždy jsme jej kontrolovali stejně důkladně. Vyzkoušejte to zpětně. Pro dva modely máme data k dispozici přímo zde.

**Kosatce (lekce o první přímce).** Model délky korunního lístku podle jeho šířky lze odhadnout z dat `iris`, která jsou součástí R.

```
# Model z lekce o první přímce, odhadnutý z vestavěných dat iris
mod_kosatce <-
  lm(
    formula = Petal.Length ~ Petal.Width,
    data = iris
  )

qqnorm(
  y = residuals(object = mod_kosatce),
  pch = 19,
  main = "",
  xlab = "Teoretický kvantil normálního rozdělení",
  ylab = "Residuum (cm)"
)

qqline(
  y = residuals(object = mod_kosatce),
  col = "#D55E00",
  lwd = 2
)
```



Obrázek 14: Q-Q graf residuí přímky pro kosatce z lekce o první přímce.

**Stromy (předchozí lekce).** Aditivní model výšky stromu podle průměru kmene a druhu:

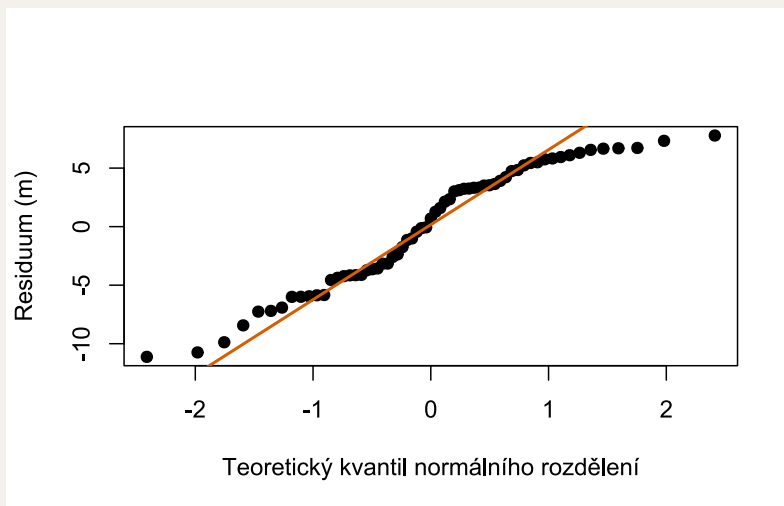
```
# Stromy z předchozí lekce
data_stromy_kontrola <-
  read.csv(
    file = here::here("data", "allometrie_stromu.csv"),
    stringsAsFactors = TRUE
  )

mod_stromy_kontrola <-
  lm(
    formula = vyska_stromu_m ~ prumer_kmene_cm + druh_stromu,
    data = data_stromy_kontrola
  )

qqnorm(
  y = residuals(object = mod_stromy_kontrola),
  pch = 19,
  main = "",
  xlab = "Teoretický kvantil normálního rozdělení",
  ylab = "Residuum (m)"
)

qqline(
  y = residuals(object = mod_stromy_kontrola),
  col = "#D55E00",
```

```
lwd = 2
)
```



Obrázek 15: Q-Q graf residuí aditivního modelu pro stromy z předchozí lekce.

U kosatců se konce odklánějí jen mírně. U stromů je odchylka výraznější; souvisí s obloukem v residuích, ke kterému se vrátíme na konci lekce.

**Ostatní modely zkontrolujte sami** v R skriptech z praktických cvičení příslušných lekcí. Ke každému modelu stačí doplnit `qqnorm()` a `qqline()` s jeho residui:

| Lekce                   | Model                                                 | Co jsme o residuích tehdy zkontrolovali                                                                       | Na co se zaměřit                                                                                                                                               |
|-------------------------|-------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| gejzír Old Faithful     | čekání podle délky erupce                             | graf residuí podle odhadnutých hodnot: dvě skupiny erupcí, obě kolem nuly, a několik velkých kladných residuí | zda velká kladná residua neprodlužují pravý konec Q-Q grafu; schod by se objevil jen tehdy, kdyby residua jedné skupiny ležela soustavně výš než residua druhé |
| kraby a zeměpisná šířka | šířka krunýře podle zeměpisné šířky (průměry lokalit) | diagnostické grafy bez zřetelného vzoru, ale jen z několika lokalit                                           | jak málo lze o tvaru říct z tak malého počtu bodů                                                                                                              |

| Lekce                  | Model                                                       | Co jsme o residuích tehdy zkontrolovali                      | Na co se zaměřit                              |
|------------------------|-------------------------------------------------------------|--------------------------------------------------------------|-----------------------------------------------|
| tučňáci a jejich druhy | hmotnost podle druhu                                        | Q-Q graf: body většinou na přímce                            | zda se konce chovají ve všech druzích podobně |
| křídlatka              | tloušťka listu podle teploty a původu                       | graf residuí podle odhadnutých hodnot: dva shluky kolem nuly | Q-Q graf, který jsme tehdy neukázali          |
| kraby a pohlaví        | zadní šířka podle délky krunýře, pohlaví a jejich interakce | graf residuí podle odhadnutých hodnot                        | Q-Q graf, který jsme tehdy neukázali          |

Nepřekvapí vás, když residua nebudou dokonale normální. Smyslem je posoudit, zda je odchylka natolik velká, aby zpochybnila nejistotu, kterou jsme v dané lekci vyvozovali.

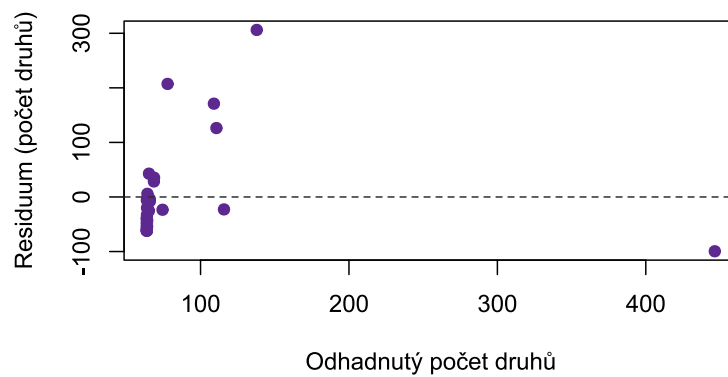
Jak se tedy chovají residua přímky pro ostrovy?

### 4.3 Co ukazují residua přímky?

Residua kontrolujeme stejnými grafy jako v předchozích lekcích: residui podle odhadnutých hodnot, histogramem residuí a Q-Q grafem residuí.

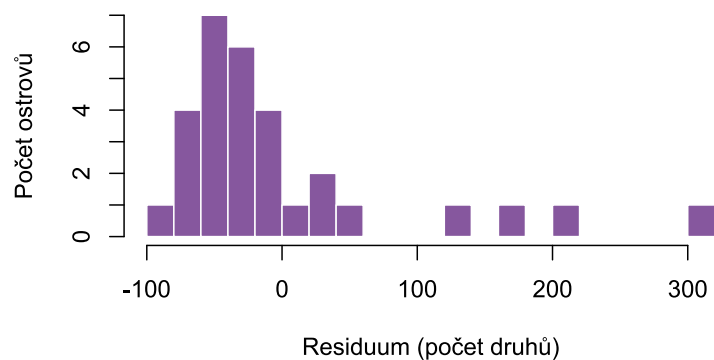
```
# Residua podle odhadnutého počtu druhů
plot(
  x = fitted(object = mod_primka),
  y = residuals(object = mod_primka),
  pch = 19,
  col = "#5D2890",
  xlab = "Odhadnutý počet druhů",
  ylab = "Residuum (počet druhů)"
)

abline(
  h = 0,
  col = "#2E2E2E",
  lty = 2
)
```



Obrázek 16: Residua přímky podle odhadnutého počtu druhů.

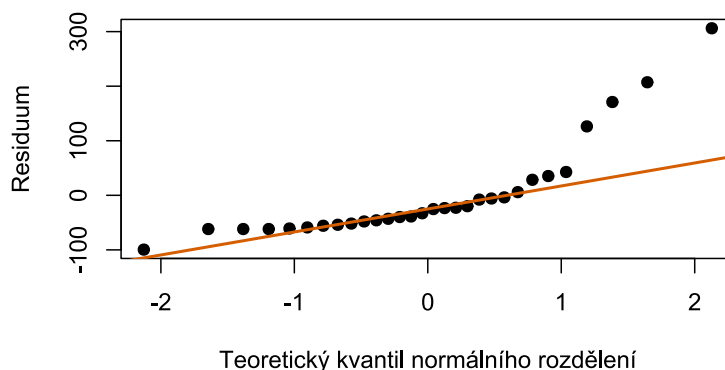
```
# Histogram residuí přímky
hist(
  x = residuals(object = mod_přímka),
  breaks = 15,
  col = "#86579E",
  border = "white",
  main = "",
  xlab = "Residuum (počet druhů)",
  ylab = "Počet ostrovů"
)
```



Obrázek 17: Histogram residuí přímky.

```
# Q-Q graf residuí přímky
qqnorm(
  y = residuals(object = mod_přímka),
  pch = 19,
  main = "",
  xlab = "Teoretický kvantil normálního rozdělení",
  ylab = "Residuum"
)

qqline(
  y = residuals(object = mod_přímka),
  col = "#D55E00",
  lwd = 2
)
```



Obrázek 18: Q-Q graf residuí přímky.

#### Co residua prozrazují?

Najděte v obou grafech alespoň dva vzory, které by u dobře padnoucího modelu neměly být vidět.

V grafu residuí podle odhadnutých hodnot netvoří body vodorovný pás kolem nuly. Většina ostrovů je nahuštěná vlevo, kde má přímka malé odhady, a **rozptyl residuí se směrem doprava zvětšuje** (heteroskedasticita). Histogram residuí má dlouhý ocas vpravo a Q-Q graf se vpravo prudce zvedá: několik kladných residuí je mnohem větších, než by normální rozdělení čekalo. Residua přímky jsou tedy pravostranně šikmá, nemají přibližně normální tvar a jejich rozptyl není konstantní. **Problém se netýká jen tvaru rozdělení, ale i tvaru vztahu: přímka nesleduje, jak se počet druhů s plochou opravdu mění.**

## 4.4 Jak jeden velký ostrov ovlivní celou přímku?

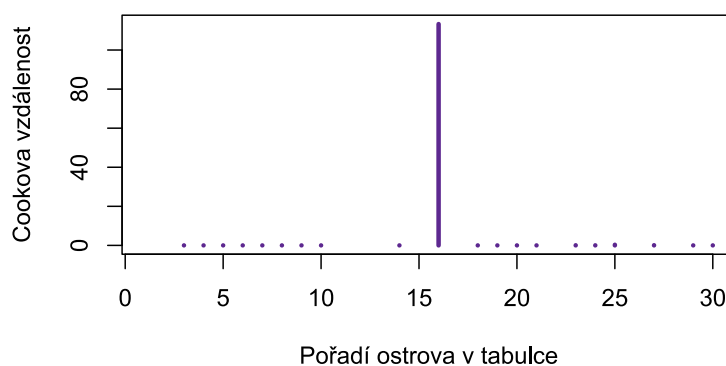
Šikmý prediktor předpoklad modelu neporušuje, ale má praktický důsledek. Několik extrémně velkých ostrovů leží daleko od ostatních na ose x. Taková pozorování mají velký pákový efekt (anglicky *leverage*): přímka se k nim natáčí podobně jako dlouhá páka. Jejich vliv na model měří Cookova vzdálenost, kterou známe z předchozích lekcí.

Který ostrov přímku ovládá?

Podívejte se znovu na bodový graf s přímkou. Který ostrov by podle vás po vynechání změnil přímku nejvíc a proč?

```
# Vypočítáme vliv jednotlivých ostrovů na přímku
vec_cook_primka <-
  cooks.distance(model = mod_primka)

plot(
  x = vec_cook_primka,
  type = "h",
  lwd = 3,
  col = "#5D2890",
  xlab = "Pořadí ostrova v tabulce",
  ylab = "Cookova vzdálenost"
)
```



Obrázek 19: Cookovy vzdálenosti jednotlivých ostrovů v modelu přímky.

Ostrov Isabela má Cookovu vzdálenost 113.3, zatímco druhá nejvyšší hodnota je jen 0.32. Hodnota větší než 1 je podle všech běžných orientačních pravidel neobvykle vysoká. **Sklon přímky tak**

**z velké části určuje jediný ostrov.** Ostrov přitom není chybou měření a nemáme důvod jej odstranit. Problém je v modelu: přímka neumí popsat ostrovy, jejichž plochy se liší o několik řádů.

## 4.5 Co říká test normality residuí?

Q–Q graf posuzujeme očima. Někdy chceme odchylku od normálního tvaru shrnout také jedním číslem. K tomu slouží **test normality**, nejčastěji Shapiro–Wilkův test (*Shapiro–Wilk test*). Jak se jeho statistika počítá, nepotřebujeme; důležité je umět výsledek přečíst.

Test používá stejnou logiku jako ostatní testy v kurzu. Nulová hypotéza tvrdí, že residua pocházejí z normálního rozdělení. P-hodnota říká, jak neobvyklá by byla tak velká odchylka od normálního tvaru, kdyby nulová hypotéza platila.

```
# Shapiro–Wilkův test normality residuí přímky  
shapiro.test(x = residuals(object = mod_primka))
```

Shapiro-Wilk normality test

```
data: residuals(object = mod_primka)  
W = 0.73771, p-value = 5.8e-06
```

Test vrátil  $p < 0.001$ . Kdyby residua pocházela z normálního rozdělení, tak velká odchylka by byla velmi neobvyklá. Test tedy potvrzuje to, co jsme viděli v Q–Q grafu.

### Jak číst test normality

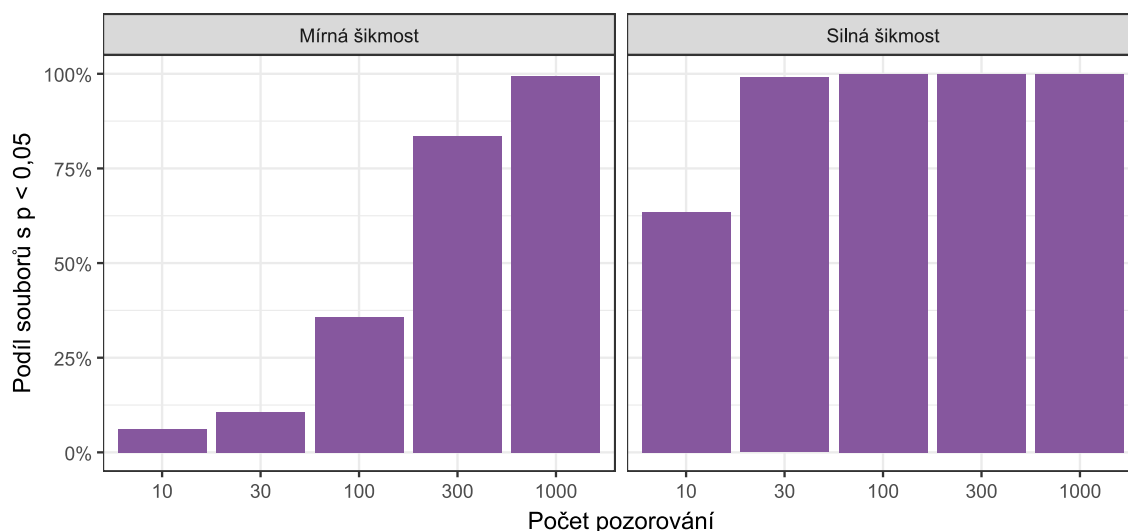
- **Malá p-hodnota** znamená, že residua jen stěží pocházejí z normálního rozdělení. Neříká, jak velká odchylka je ani zda je pro naši otázku podstatná.
- **Velká p-hodnota** neznamená, že residua *jsou* normální. Znamená jen, že test odchylku neprokázal.
- Test vždy čteme **až po Q–Q grafu** a spolu s ním, nikdy místo něj.

## 4.6 Proč výsledek testu závisí na počtu pozorování?

Test normality odpovídá na otázku „lze odchylku od normálního tvaru odlišit od náhody?“, nikoli „je odchylka pro model podstatná?“. Odpověď na první otázku silně závisí na počtu pozorování. Ukážeme to simulací: opakovaně vytvoříme soubory residuí s mírnou a se silnou pravostrannou šikmostí a pro každý soubor provedeme Shapiro–Wilkův test.

### Předpovězte výsledek

Odchylka residuí od normálního tvaru bude stejná, jen se zvětší počet pozorování. Bude test takovou odchylku odhalovat častěji, nebo méně často?



Obrázek 20: Simulace: jak často Shapiro–Wilkův test na hladině 0.05 odhalí stejnou odchylku od normálního tvaru při různém počtu pozorování (400 opakování pro každou kombinaci).

Při mírné šikmosti test odchylku odhalil jen v 11% souborů se 30 pozorováními, ale v 100% souborů s 1000 pozorováními. Silnou šikmost při 10 pozorováních odhalil jen v 64% souborů.

V lekci o první přímce jsme upozornili, že u velkého souboru může i nepatrná odchylka vypadat „statisticky zjistitelně“. Simulace tento jev ukazuje přímo. Podíl souborů, ve kterých test skutečně existující odchylku odhalil, jsme v lekci o testování nazvali **síla testu**: roste s počtem pozorování a s velikostí odchylky.

Z toho plynou dvě praktická pravidla:

- U **velkých vzorků** test označí i drobné odchylky, které pro lineární model nemají praktický význam. U velkých vzorků navíc mírná odchylka od normálního tvaru residuí odhady sklonů a intervaly spolehlivosti obvykle téměř neovlivní; podrobné zdůvodnění přesahuje tento kurz.
- U **malých vzorků** může test nevýznamným výsledkem „propustit“ i velkou odchylku. Naše data mají jen 30 ostrovů, takže velká p-hodnota by sama o sobě mnoho nedokazovala.

Rozhodnutí proto stojí na Q–Q grafu, na grafu residuí podle odhadnutých hodnot a na otázce, zda odchylka může změnit biologický závěr. Test je doplňující číslo, nikoli brána, kterou musí model projít.

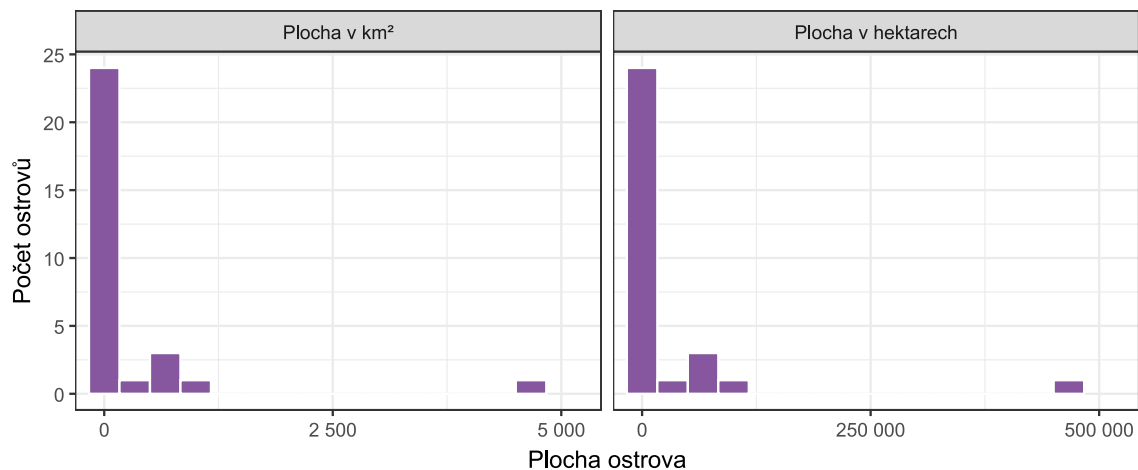
## 5 Transformace: měřítko, nebo tvar?

Přímka selhala dvakrát: její residua mají šikmý tvar a rostoucí rozptyl a o jejím sklonu rozhoduje jediný ostrov. Než začneme měnit model, podíváme se na nástroj, který ke změně potřebujeme: **transformaci** proměnné. Transformace znamená, že na každou hodnotu proměnné použijeme stejný výpočet, například vynásobení stem nebo logaritmus. Uvidíme, že transformace se dělí do dvou velmi odlišných skupin.

## 5.1 Změní se tvar rozdělení, když plochu převedeme na hektary?

Nejjednodušší transformací je změna jednotek. Jeden kilometr čtvereční má 100 hektarů.

```
# Převedeme plochu z km² na hektary
data_ostrovny$plocha_ha <-
  data_ostrovny$plocha_km2 * 100
```



Obrázek 21: Plocha ostrovů v km² a v hektarech. Mění se jen popisky vodorovné osy.

Oba histogramy mají stejný tvar. Změnila se jen čísla na ose. Vynásobení konstantou natáhne nebo stlačí osu, ale **pořadí ostrovů ani poměry vzdáleností mezi nimi se nezmění**, a proto se nezmění ani tvar rozdělení.

### Kdy měníme jednotky?

- **Spojujeme data z různých zdrojů.** Jedna studie uvádí plochu v hektarech, druhá v km²; hmotnost jednou v gramech, jindy v kilogramech. Před společnou analýzou je musíme převést na stejné jednotky.
- **Chceme čitelný koeficient.** Příмка v části o předpokladech modelu přidávala 0.082 druhu na 1 km². Totéž vyjádřené na 100 km² (8.2 druhu) se čte snáz. Změna jednotek prediktoru změní sklon stejným násobkem, ale nezmění žádný závěr modelu.

## 5.2 Co se změní, když od plochy odečteme průměr?

Druhou jednoduchou transformací je **centrování** (anglicky *centring*): od každé hodnoty odečteme průměr. Centrovaná plocha říká, o kolik km² je ostrov větší nebo menší než průměrný ostrov.

### Který koeficient se změní?

Odhadneme stejnou přímku jednou s původní a jednou s centrovanou plochou. Změní se po centrování sklon, intercept, nebo oba koeficienty?

```
# Od plochy každého ostrova odečteme průměrnou plochu
data_ostrovny$plocha_centrovana <-
  data_ostrovny$plocha_km2 - mean(x = data_ostrovny$plocha_km2)

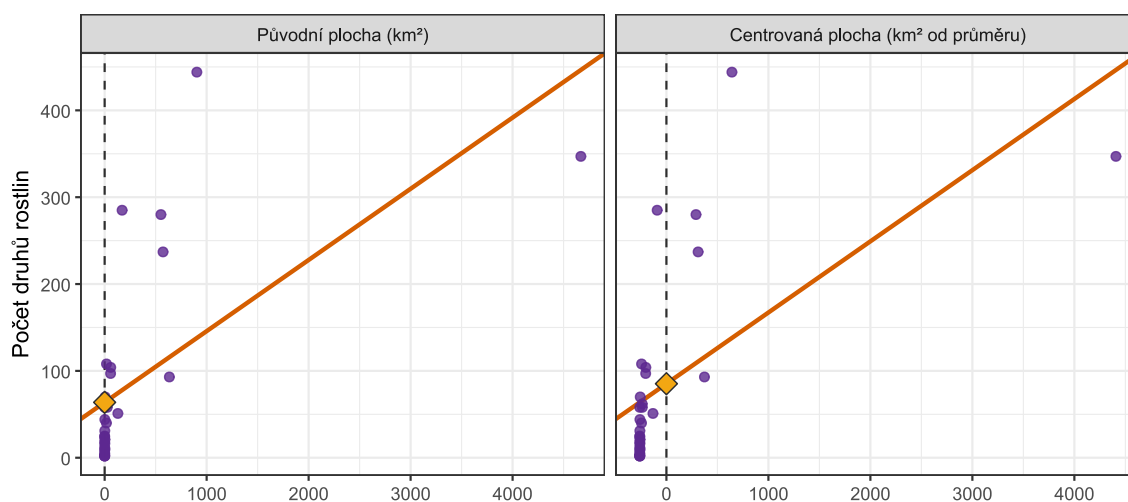
# Odhadneme stejnou přímku s centrovanou plochou
mod_primka_centrovana <-
  lm(
    formula = pocet_druhu ~ plocha_centrovana,
    data = data_ostrovny
  )

# Porovnáme koeficienty obou přímek
coef(object = mod_primka)
```

```
(Intercept)  plocha_km2
63.78286147  0.08196317
```

```
coef(object = mod_primka_centrovana)
```

```
(Intercept) plocha_centrovana
85.23333333      0.08196317
```



Obrázek 22: Stejná přímka nad původní a nad centrovanou plochou. Svislá přerušovaná čára označuje nulu na vodorovné ose; kosočtverec je intercept, tedy výška přímky v této nule.

Body i přímka jsou v obou panelech stejné; centrování jen posune osu tak, že nula leží v průměru. Tvar rozdělení zůstane stejný. V modelu se **sklon nezmění** (v obou modelech je 0.0820), protože jeden km<sup>2</sup> navíc znamená totéž bez ohledu na to, odkud plochu měříme. **Změní se význam interceptu.** V původním modelu je intercept 63.8 odhadnutým počtem druhů na ostrově s nulovou plochou, tedy na ostrově, který neexistuje. V centrovaném modelu je intercept 85.2 odhadnutým počtem druhů na ostrově s průměrnou plochou 261.7 km<sup>2</sup>. U přímky odhadnuté metodou nejmenších čtverců se tato hodnota rovná průměrnému počtu druhů.

#### Kdy centrujeme?

- **Intercept by jinak popisoval nesmysl.** V lekci o interakcích jsme narazili na to, že koeficient pohlaví u krabů se vztahoval k délce krunýře 0 mm, tedy mimo data. Po centrování délky by popisoval rozdíl mezi pohlavími u kraba průměrné délky.
- **Prediktor je letopočet nebo nadmořská výška.** Trend „od roku 0“ nebo počet druhů „u hladiny moře“ často nemají biologický smysl. Odečteme proto smysluplnou referenční hodnotu: průměr, začátek sledování nebo průměrnou nadmořskou výšku lokalit.
- **Model obsahuje interakci.** Hlavní koeficienty se pak vztahují k nulové hodnotě druhé proměnné; po centrování ke středu dat, kde se nejsnáze vysvětlují.

Centrování se proto hodí, když chceme, aby intercept popisoval smysluplnou, typickou hodnotu prediktoru.

### 5.3 Jak daleko od průměru leží ostrov?

Třetí transformace spojuje centrování se změnou jednotek. Odchylku od průměru nevyjádříme v km<sup>2</sup> nebo metrech, ale v počtu směrodatných odchylek. Výsledek se nazývá **z-skóre** (anglicky *z-score*) a převod na z-skóre **standardizace**.

Myšlenku už známe: t-statistika v předchozích lekcích také měřila vzdálenost odhadu od nulové hodnoty v jednotkách jeho standardní chyby. Z-skóre dělá totéž s jednotlivou hodnotou. Slovy:

$$z\text{-skóre} = \frac{\text{hodnota} - \text{průměr}}{\text{směrodatná odchylka}}.$$

Spočítejme z-skóre nejvyššího bodu ostrova Fernandina. Průměr a směrodatnou odchylku počítáme z 24 ostrovů, u kterých je nejvyšší bod známý.

| Veličina               | Fernandina | Průměr | Odchylka | SD     | z    |
|------------------------|------------|--------|----------|--------|------|
| Nejvyšší bod (m n. m.) | 1494       | 424.46 | 1069.54  | 452.96 | 2.36 |

Do slovního vzorce dosadíme hodnoty z tabulky:

$$\begin{aligned}\text{z-skóre výšky Fernandiny} &= \frac{1494 - 424.46}{452.96} \\ &= \frac{1069.54}{452.96} \\ &= 2.36.\end{aligned}$$

Nejvyšší bod Fernandiny leží 2.36 směrodatné odchylky nad průměrem ostrovů. Obecně pro hodnotu  $x_i$  píšeme:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Symbol  $z_i$  je z-skóre  $i$ -tého pozorování,  $x_i$  jeho původní hodnota,  $\bar{x}$  průměr všech hodnot a  $s$  jejich směrodatná odchylka. Kladné z-skóre znamená hodnotu nad průměrem, záporné pod průměrem a z-skóre 0 přesně průměr. U přibližně normální proměnné proto podle pravidla z kapitoly o normálním rozdělení leží asi 95% z-skóre mezi -2 a 2.

#### Spočítejte z-skóre plochy

Fernandina má plochu 634.49 km<sup>2</sup>. Průměrná plocha ostrovů je 261.71 km<sup>2</sup> a směrodatná odchylka ploch 864.11 km<sup>2</sup>. Spočítejte z-skóre plochy Fernandiny.

#### Řešení

$$\frac{634.49 - 261.71}{864.11} = \frac{372.78}{864.11} = 0.43.$$

Plocha Fernandiny leží 0.43 směrodatné odchylky nad průměrnou plochou.

V R lze z-skóre spočítat přímo podle vzorce, nebo funkcí `scale()`:

```
# z-skóre nejvyššího bodu podle vzorce
data_ostrovny$vyska_z <-
  (data_ostrovny$nejvyssi_bod_m -
    mean(x = data_ostrovny$nejvyssi_bod_m, na.rm = TRUE)) /
    sd(x = data_ostrovny$nejvyssi_bod_m, na.rm = TRUE)

# Stejný výsledek funkcí scale(); as.numeric() odstraní pomocné atributy
data_ostrovny$vyska_z_scale <-
  as.numeric(x = scale(x = data_ostrovny$nejvyssi_bod_m))

# Porovnáme oba výpočty u prvních ostrovů
head(
  x = data_ostrovny[, c("ostrov", "vyska_z", "vyska_z_scale")],
```

```
n = 4
)
```

|   | ostrov    | vyska_z    | vyska_z_scale |
|---|-----------|------------|---------------|
| 1 | Baltra    | NA         | NA            |
| 2 | Bartolome | -0.6964372 | -0.6964372    |
| 3 | Caldwell  | -0.6853987 | -0.6853987    |
| 4 | Champion  | -0.8355222 | -0.8355222    |

```
# Průměr a směrodatná odchylka všech z-skóre
round(
  x = mean(x = data_ostrovny$vyska_z, na.rm = TRUE),
  digits = 10
)
```

```
[1] 0
```

```
sd(x = data_ostrovny$vyska_z, na.rm = TRUE)
```

```
[1] 1
```

Když převedeme na z-skóre **všechny hodnoty proměnné**, má výsledná proměnná vždy průměr 0 a směrodatnou odchylku 1. Průměr je nulový, protože jsme od každé hodnoty odečetli původní průměr, a směrodatná odchylka je 1, protože jsme každou odchylku vydělili původní směrodatnou odchylkou. Z-skóre tedy všechny proměnné převádí na stejnou „měrnou jednotku“: počet směrodatných odchylek od průměru.

## 5.4 Je Fernandina výjimečnější plochou, nebo výškou?

Z-skóre nemá jednotky. Díky tomu můžeme porovnat, jak neobvyklá je hodnota u veličin měřených v úplně jiných jednotkách. Plochu v km<sup>2</sup> a výšku v metrech přímo porovnat nelze; jejich z-skóre ano.

Nejvyšší bod Fernandiny leží 2.36 směrodatné odchylky nad průměrnou výškou, její plocha jen 0.43 směrodatné odchylky nad průměrnou plochou. Podle z-skóre by tedy byla výjimečná výškou, ale plochou spíš průměrná.

Pořadí ostrovů ale ukazuje něco jiného: plochou je Fernandina 3. největší a výškou 2. nejvyšší ostrov souostroví. Výjimečná je tedy v obou ohledech.

Z-skóre u šikmé veličiny klame

Průměr i směrodatná odchylka jsou citlivé na extrémní hodnoty. Obrovská plocha Isabely zvětšila směrodatnou odchylku ploch natolik, že i 3. největší ostrov vypadá v z-skóre téměř průměrně a většina ostrovů má z-skóre plochy těsně pod nulou. **Z-skóre je užitečné hlavně u přibližně souměrných veličin.** U silně šikmé veličiny samo z-skóre nestačí.

#### Doplňující: z-skóre v korelaci a v modelu

Při výpočtu Pearsonovy korelace jsme dělili součet součinů odchylek odmocninami součtů čtverců odchylek obou proměnných. Jednodušeji to vyjádříme takto: **Pearsonova korelace je průměrný součin z-skóre obou proměnných** (s dělením  $n - 1$  místo  $n$ , stejně jako u směrodatné odchylky). Protože z-skóre nemá jednotky, nemá je ani korelace. A protože změna jednotek ani centrování z-skóre nezmění, nezmění ani korelaci.

Z-skóre lze použít i v modelu: pokud před odhadem standardizujeme prediktory, jejich sklony udávají změnu odezvy při posunu o jednu směrodatnou odchylku. Takové *standardizované koeficienty* usnadňují porovnání prediktorů měřených v různých jednotkách. Neříkají ale samy o sobě, který prediktor je biologicky důležitější.

```
# Pearsonova korelace plochy a počtu druhů
cor(
  x = data_ostrovy$plocha_km2,
  y = data_ostrovy$pocet_druhu
)
```

```
[1] 0.6178431
```

```
# Stejná korelace s plochou v hektarech
cor(
  x = data_ostrovy$plocha_ha,
  y = data_ostrovy$pocet_druhu
)
```

```
[1] 0.6178431
```

```
# Stejně číslo jako součet součinů z-skóre vydělený n - 1
sum(
  scale(x = data_ostrovy$plocha_km2) *
  scale(x = data_ostrovy$pocet_druhu)
) / (nrow(x = data_ostrovy) - 1)
```

[1] 0.6178431

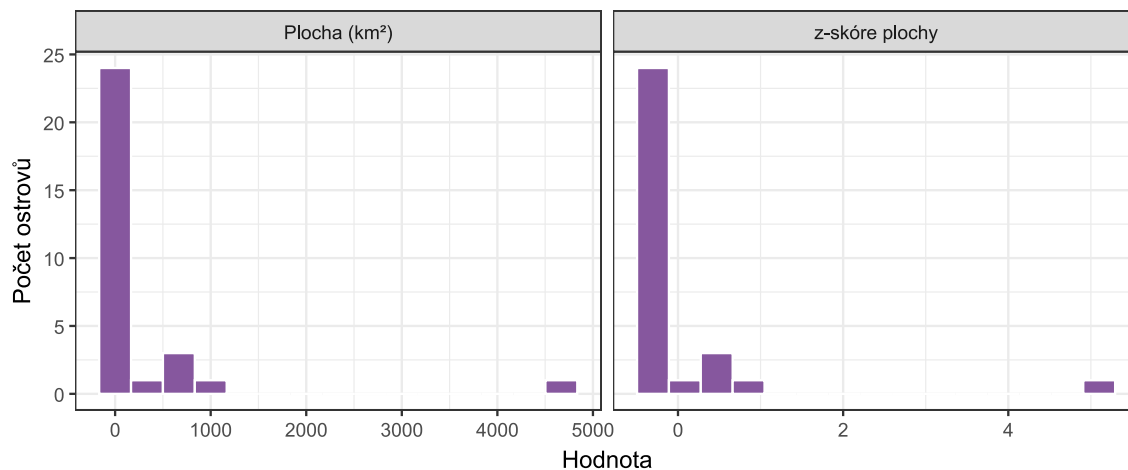
### Kdy používáme z-skóre?

- **Porovnáváme, jak neobvyklá je hodnota u veličin v různých jednotkách**, například zda je jedinec výjimečný spíše hmotností, nebo délkou křídla.
- **Skládáme několik veličin do jednoho ukazatele**. Chceme-li například shrnout teplotu, srážky a nadmořskou výšku lokalit do jednoho „gradientu prostředí“, převedeme je nejprve na z-skóre, aby žádná veličina nepřevládala jen kvůli svým jednotkám.
- **Porovnáváme sklony prediktorů v různých jednotkách** pomocí standardizovaných koeficientů (doplňující rámeček „z-skóre v korelaci a v modelu“ výše).
- Vždy pamatujeme, že u silně šikmých veličin z-skóre klame (rámeček „Z-skóre u šikmé veličiny klame“).

## 5.5 Odstraní z-skóre šikmost?

### Předpovězte tvar

Plocha ostrovů je extrémně pravostranně šikmá. Bude histogram z-skóre plochy souměrnější?



Obrázek 23: Plocha ostrovů v km² a její z-skóre.

Nebude. Histogram z-skóre má přesně stejný tvar jako histogram plochy. (Kdo četl doplňující rámeček o koeficientu šikmosti: ten je před standardizací 4.42 a po ní 4.42.)

Transformace, které mění jen polohu a měřítko

Změna jednotek, centrování i z-skóre jsou **lineární transformace**: každou hodnotu vynásobí stejným číslem a/nebo k ní přičtou stejné číslo. Posunou a natáhnou osu, ale **tvar rozdělení nezmění**. Šikmost, těžké konce i dva vrcholy po nich zůstanou. Hodí se pro změnu významu interceptu nebo pro porovnání veličin v různých jednotkách, nikoli pro opravu tvaru.

## 5.6 Jak logaritmus převádí násobky na stejné kroky?

Na začátku kurzu jsme logaritmickou osu použili, když se hodnoty v grafu lišily o několik řádů. Na takové ose leží 1, 10, 100 a 1 000 ve stejných vzdálenostech: každý krok o stejnou délku znamená vynásobení desíti. Logaritmus dělá totéž s čísly. V první lekci jsme zdůraznili, že logaritmická osa mění jen rozestupy na grafu a průměr i medián zůstávají vypočtené z kilogramů. Krok, který jsme tehdy záměrně neudělali, je použít logaritmus na samotné hodnoty proměnné.

| Plocha (km <sup>2</sup> ) | Desítkový logaritmus | Přirozený logaritmus |
|---------------------------|----------------------|----------------------|
| 1                         | 0.000                | 0.000                |
| 2                         | 0.301                | 0.693                |
| 4                         | 0.602                | 1.386                |
| 8                         | 0.903                | 2.079                |
| 10                        | 1.000                | 2.303                |
| 100                       | 2.000                | 4.605                |
| 1000                      | 3.000                | 6.908                |

Každé zdvojnásobení plochy ( $1 \rightarrow 2 \rightarrow 4 \rightarrow 8$ ) zvýší přirozený logaritmus o stejný krok 0.693. Každé vynásobení desíti zvýší desítkový logaritmus přesně o 1 a přirozený logaritmus o 2.303. **Logaritmus tedy převádí stejné poměry na stejné rozdíly.**

V modelech používáme **přirozený logaritmus** (funkce `log()` v R), protože s ním se koeficienty nejnázne interpretují. Desítkový logaritmus (`log10()`) se hodí hlavně pro popis os grafu.

Proč to funguje, ukazuje základní vlastnost logaritmu. Slovy:

$$\text{logaritmus dvojnásobné plochy} = \text{logaritmus původní plochy} + \text{logaritmus čísla } 2.$$

Pro ostrov s plochou 10 km<sup>2</sup>, jehož plochu zdvojnásobíme:

$$\begin{aligned}\ln(20) &= \ln(10) + \ln(2) \\ &= 2.303 + 0.693 \\ &= 2.996.\end{aligned}$$

Obecně pro dvě kladná čísla  $a$  a  $b$  platí:

$$\ln(a \cdot b) = \ln(a) + \ln(b).$$

Symbol  $\ln$  označuje přirozený logaritmus,  $a$  je výchozí hodnota (zde plocha) a  $b$  násobek (zde 2). Násobení se v logaritmech mění na sčítání.

```
# Přirozený logaritmus řady zdvojnásobených ploch
log(x = c(1, 2, 4, 8))
```

```
[1] 0.0000000 0.6931472 1.3862944 2.0794415
```

```
# Logaritmus nuly není konečné číslo
log(x = 0)
```

```
[1] -Inf
```

```
# Opakem logaritmu je funkce exp(); vrátí původní hodnotu
exp(x = log(x = 8))
```

```
[1] 8
```

### Logaritmus jen pro kladné hodnoty

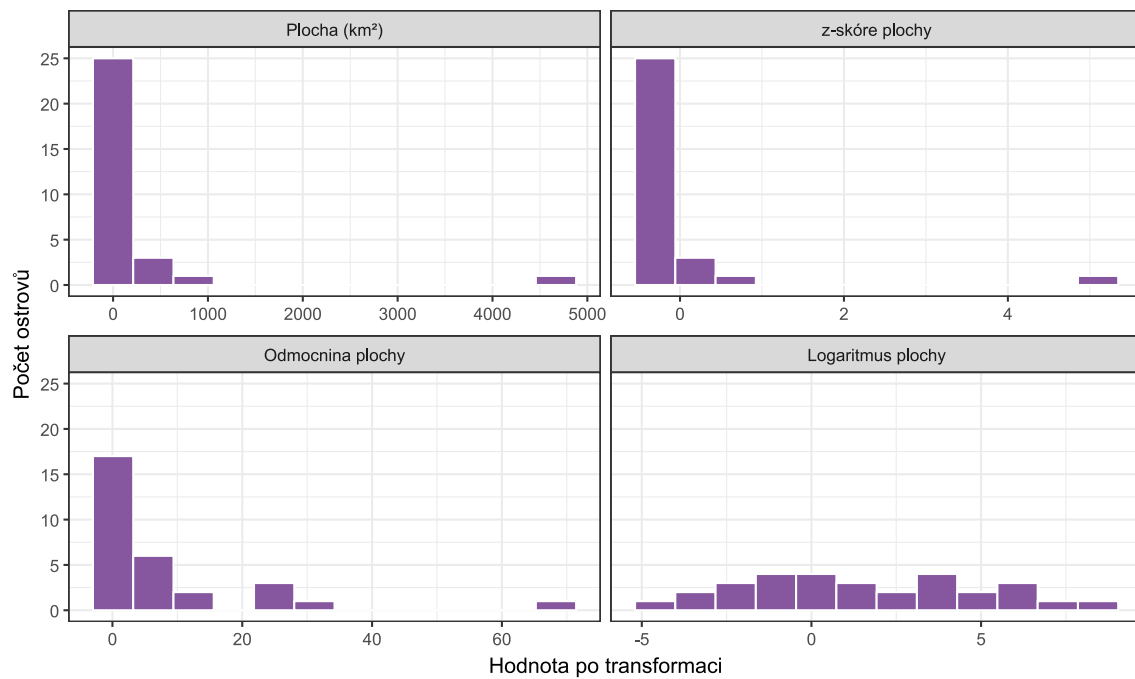
Logaritmus existuje jen pro kladná čísla. `log(0)` vrátí `-Inf` (minus nekonečno) a logaritmus záporného čísla vrátí `NaN`. Plocha i počet druhů jsou na Galapágách vždy kladné. Proměnné se skutečnými nulami, které jsme našli při prohlídce dat (vzdálenost Santa Cruz od sebe sama, ostrov bez endemitů), bychom však logaritmovat nemohli.

## 5.7 Co logaritmus udělá s tvarem rozdělení?

Logaritmus není lineární transformace. Co to znamená pro tvar rozdělení?

### Který histogram změní tvar?

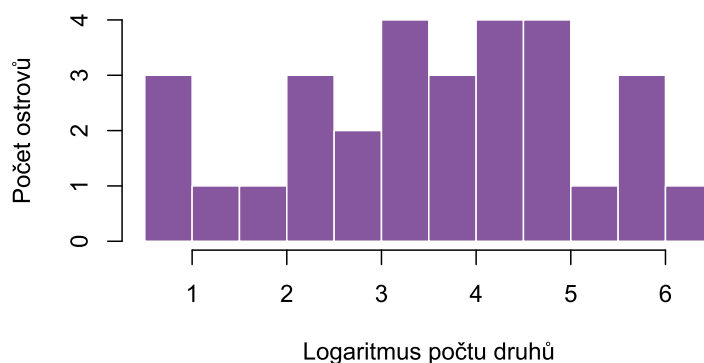
Následující obrázek ukáže plochu ostrovů čtyřikrát: v  $\text{km}^2$ , jako z-skóre, po odmocnění a po logaritmování. U kterých panelů čekáte stejný tvar jako u původní plochy?



Obrázek 24: Plocha ostrovů v km<sup>2</sup>, jako z-skóre, po odmocnění a po logaritmování.

Z-skóre zachovalo původní tvar. Logaritmus naopak malé hodnoty od sebe roztáhl a velké k sobě stlačil: rozdíl mezi 1 a 10 km<sup>2</sup> je na logaritmické škále stejně velký jako rozdíl mezi 1 000 a 10 000 km<sup>2</sup>. **Odmocnina šikmost zmírnila, ale neodstranila. Logaritmus z extrémně šikmé plochy udělal přibližně souměrné rozdělení.** Malé ostrovy, které byly v původním histogramu namačkané v prvním koši, se rozprostřely přes celou osu.

```
# Rozdělení logaritmu počtu druhů
hist(
  x = log(x = data_ostrovny$pocet_druhu),
  breaks = 10,
  col = "#86579E",
  border = "white",
  main = "",
  xlab = "Logaritmus počtu druhů",
  ylab = "Počet ostrovů"
)
```



Obrázek 25: Histogram logaritmu počtu druhů.

Podobně se chová i počet druhů: po logaritmování již nemá dlouhý ocas vpravo. Veličinám, které jsou pravostranně šikmé, ale po logaritmování mají přibližně normální tvar, se říká **log-normální**. Typicky vznikají právě násobením mnoha náhodných vlivů, jako v simulaci populací.

#### Kdy používáme logaritmus a kdy odmocninu?

**Logaritmus** použijeme u kladných veličin, které se mění násobně nebo se liší o několik řádů:

- velikost populací, buněčné kultury a další procesy, které rostou nebo klesají o určité procento;
- koncentrace živin, toxinů nebo hormonů, které mezi vzorky kolísají o celé řády;
- velikost a hmotnost organismů napříč druhy, například hmotnost savců z první lekce;
- vztahy mezi rozměry těla (alometrie), například výška a průměr kmene stromu nebo počet druhů a plocha ostrova;
- situace, kdy nás zajímají relativní změny („o kolik procent“, „kolikrát“), nikoli absolutní rozdíly.

Logaritmickou škálu znáte i z běžného života: pH je záporný desítkový logaritmus koncentrace vodíkových iontů a decibely měří hlasitost na logaritmické škále.

**Odmocninu** použijeme u nezáporných veličin s mírnou pravostrannou šikmostí, zvláště u menších počtů, kde se objevují nuly (například počet semenáček na ploše nebo počet hmyzu v pasti). Na rozdíl od logaritmu si s nulou poradí.

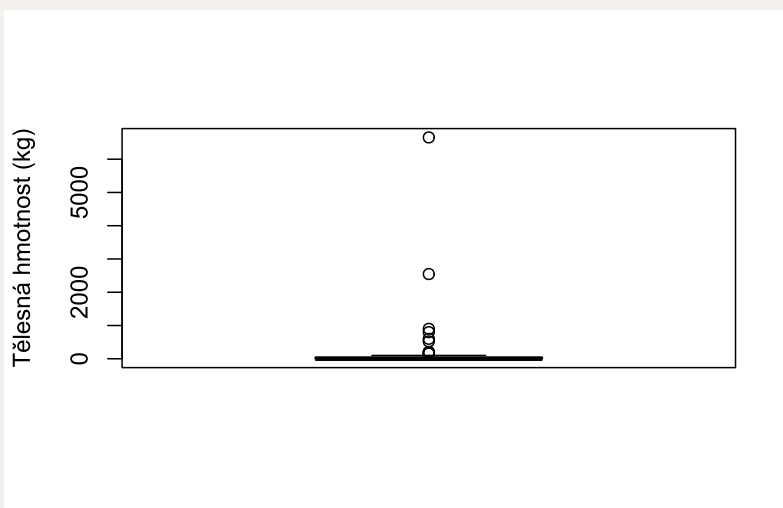
**Kdy ne:** pokud veličina obsahuje nuly nebo záporné hodnoty, logaritmus přímo použít nelze. U podílů mezi 0 a 1 a u počtů s mnoha nulami existují vhodnější modely, ke kterým se dostaneme později. A pokud jsou residua modelu v pořádku a biologické tvrzení je aditivní, transformace není potřeba.

### Doplňující: odlehlí savci z první lekce

V první lekci krabicový graf tělesné hmotnosti 83 druhů savců označil 11 druhů jako možná odlehlá pozorování. Hranice jsme tehdy počítali v kilogramech a logaritmickou osu jsme použili jen pro čitelnost. Co se stane, když hranice spočítáme až po logaritmování?

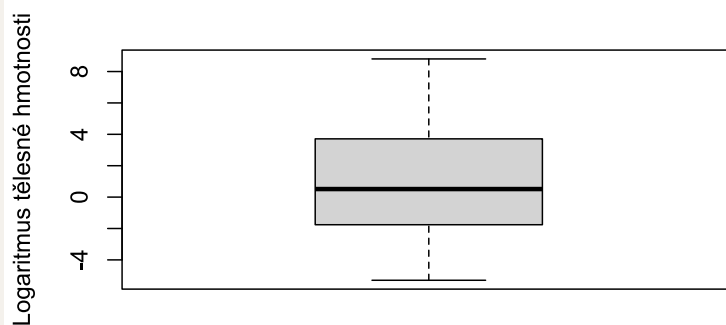
```
# Hmotnost savců z první lekce (balíček ggplot2)
vec_hmotnost <- ggplot2::msleep$bodywt

# Krabicový graf v kilogramech
boxplot(
  x = vec_hmotnost,
  ylab = "Tělesná hmotnost (kg)"
)
```



Obrázek 26: Tělesná hmotnost savců z první lekce v kilogramech.

```
# Krabicový graf logaritmu hmotnosti
boxplot(
  x = log(x = vec_hmotnost),
  ylab = "Logaritmus tělesné hmotnosti"
)
```



Obrázek 27: Stejní savci po logaritmování hmotnosti.

Po logaritmování neleží za vousy žádný druh. Savci označení v první lekci jako možná odlehlá pozorování tedy na logaritmické škále neobvyklí nejsou; leželi jen v dlouhém ocasu pravostranně šikmého rozdělení. Slon nebo velryba jsou skutečně velcí, ale na násobné škále, na které se velikost organismů mění, tolik nevybočují. **Označení odlehlé hodnoty proto závisí na tom, na jaké škále se díváme.**

## 5.8 Které transformace mění jen měřítko a které i tvar?

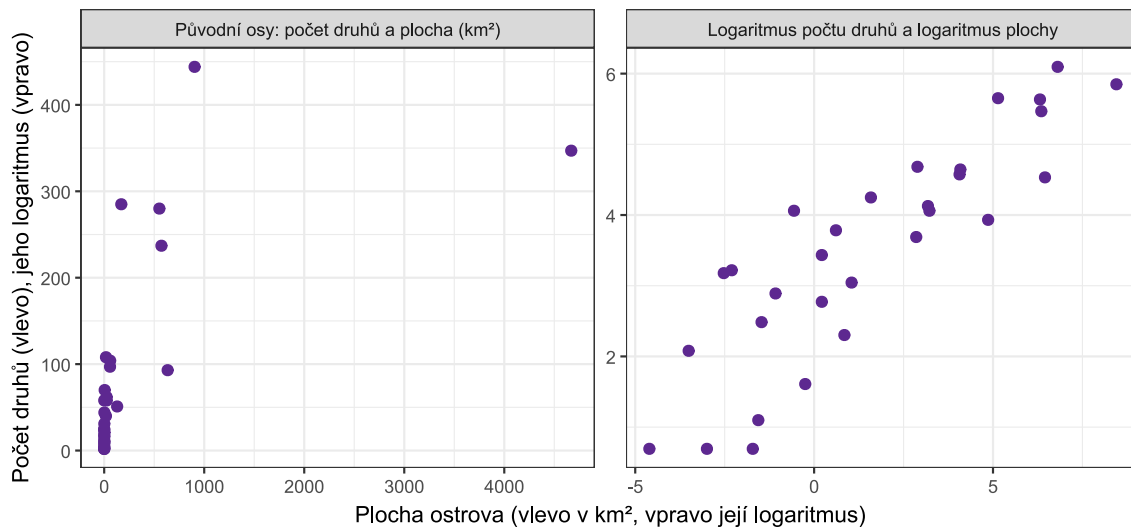
| Transformace         | Výpočet v R                                                      | Mění tvar?                                | Kdy se hodí                               |
|----------------------|------------------------------------------------------------------|-------------------------------------------|-------------------------------------------|
| změna jednotek       | $x * 100$                                                        | ne                                        | převod mezi jednotkami                    |
| centrování           | $x - \text{mean}(x)$                                             | ne                                        | intercept pro typickou hodnotu prediktoru |
| z-skóre              | $(x - \text{mean}(x)) / \text{sd}(x)$ nebo <code>scale(x)</code> | ne                                        | porovnání veličin v různých jednotkách    |
| odmocnina            | <code>sqrt(x)</code>                                             | ano, mírně zmenší pravostrannou šikmost   | mírně šikmé nezáporné veličiny            |
| přirozený logaritmus | <code>log(x)</code>                                              | ano, výrazně zmenší pravostrannou šikmost | kladné veličiny, které se mění násobně    |

Logaritmus má ještě jednu vlastnost, která je pro nás důležitější než souměrnost histogramu. Mění totiž nejen tvar rozdělení, ale i **tvar vztahu** mezi proměnnými. K tomu se dostaneme v další části.

## 6 Tvar vztahu

### 6.1 Jak vypadá vztah plochy a počtu druhů na logaritmických osách?

Vraťme se k ostrovům. Příмка selhala hlavně proto, že malé ostrovy byly namačkané u počátku a o sklonu rozhodoval jediný obr. Co s tímto vztahem udělá logaritmus? Zobrazme stejné ostrovy dvakrát: na původních osách a na osách, kde jsou obě proměnné logaritmované.



Obrázek 28: Vztah počtu druhů a plochy ostrova. Vlevo na původních osách, vpravo po logaritmování obou proměnných.

Stejný obrázek pravého panelu získáme i v základním R:

```
# Bodový graf logaritmu počtu druhů podle logaritmu plochy
plot(
  x = log(x = data_ostrovy$plocha_km2),
  y = log(x = data_ostrovy$pocet_druhu),
  pch = 19,
  col = "#5D2890",
  xlab = "Logaritmus plochy",
  ylab = "Logaritmus počtu druhů"
)
```

#### Co znamená příмка na logaritmických osách?

Na logaritmických osách leží body přibližně podél přímk. Co to znamená pro vztah mezi plochou a počtem druhů? Nápvěda: stejný krok na logaritmické ose odpovídá stejnému násobku původní hodnoty.

Vlevo je vztah zakřivený: u malých ostrovů počet druhů roste rychle, u velkých už pomaleji. Vpravo je přibližně přímkový. Na logaritmických osách znamená stejný krok doprava stejný násobek plochy a stejný krok nahoru stejný násobek počtu druhů. **Přímka na logaritmických osách tedy říká: když plochu vynásobíme určitým číslem, počet druhů se vynásobí jiným, stále stejným číslem.** Takovému vztahu se říká **mocninný vztah** (anglicky *power law*). V ekologii se pro ostrovy nazývá *vztah druhů a plochy* (*species–area relationship*).

Doplňující: co na to Pearsonova a Spearmanova korelace?

V lekci o korelaci jsme u zakřiveného vztahu viděli, že Spearmanova korelace zůstala vysoká, zatímco Pearsonova byla nižší. Galapágy ukazují proč. Pearsonova korelace plochy a počtu druhů je na původních osách 0.62 a po logaritmování obou proměnných 0.87: měří těsnost kolem **přímky**, a přímka sedí lépe na logaritmických osách. Spearmanova korelace je v obou případech 0.88, protože pracuje s pořadím ostrovů a logaritmus pořadí nezmění. **Rozdíl mezi Pearsonovou a Spearmanovou korelací byl tedy upozorněním, že vztah je monotónní, ale není přímkový.** V lekci o korelaci jsme uvedli i druhou možnou příčinu takového rozdílu, vliv několika vzdálených bodů; na Galapágách působí obě, protože na původních osách přímku silně ovlivňuje Isabela.

## 6.2 Jaké tvary vztahu dávají biologický smysl?

Stejně jako v předchozí lekci sestavíme malou sadu kandidátních modelů a každý z nich zdůvodníme biologicky dříve, než uvidíme jakékoli srovnávací skóre.

Důležitá vlastnost lineárního modelu: „**lineární**“ **znamená lineární v koeficientech, nikoli nutně přímka v grafu.** Model zůstává lineární, pokud je odezva součtem koeficientů násobených nějakými hodnotami. Těmi hodnotami mohou být i  $\log(\text{plocha\_km}^2)$  nebo druhá mocnina plochy. Funkce `lm()` proto umí popsat i zakřivený vztah.

| Kandidát              | Biologické tvrzení                                                  | Formule v R                                             |
|-----------------------|---------------------------------------------------------------------|---------------------------------------------------------|
| A – přímka            | Každý další $\text{km}^2$ přidá stejný počet druhů.                 | <code>pocet_druhu ~ plocha_km2</code>                   |
| B – logaritmus plochy | Každé zdvojnásobení plochy přidá stejný počet druhů.                | <code>pocet_druhu ~ log(plocha_km2)</code>              |
| C – mocninný vztah    | Každé zdvojnásobení plochy vynásobí počet druhů stejným číslem.     | <code>log(pocet_druhu) ~ log(plocha_km2)</code>         |
| D – kvadratický člen  | Počet druhů roste s plochou stále pomaleji, případně se pak obrátí. | <code>pocet_druhu ~ plocha_km2 + I(plocha_km2^2)</code> |
| E – logaritmus odezvy | Každý další $\text{km}^2$ zvýší počet druhů o stejné procento.      | <code>log(pocet_druhu) ~ plocha_km2</code>              |

Funkce `I()` v kandidátovi D říká R, aby výraz `plocha_km2^2` spočítal jako obyčejnou druhou mocninu. Bez ní by symbol `^` uvnitř formule měl jiný význam.

Který kandidát zní biologicky nejméně věrohodně?

Pro každého kandidáta zkuste říct, zda jeho tvrzení zní rozumně pro ostrov o ploše 0,01 km<sup>2</sup> i pro ostrov o ploše 4 000 km<sup>2</sup>. Který z nich vám připadá nejméně věrohodný?

Kandidát A předpokládá, že první km<sup>2</sup> malého ostrova přidá stejně druhů jako tisíce km<sup>2</sup> velkého. Kandidát D umí vztah ohnout, ale kvadratická křivka je parabola: po vrcholu musí klesat, i když biologický důvod, proč by větší ostrov měl mít méně druhů, nemáme. Kandidáti B a C popisují zpomalující růst pomocí násobků plochy. Kandidát E předpokládá zrychlující exponenciální růst, který by u největších ostrovů předpovídal obrovské počty druhů.

```
# A: stejná přímka jako v části o předpokladech modelu
mod_a_primka <-
  lm(
    formula = pocet_druhu ~ plocha_km2,
    data = data_ostrovy
  )

# B: logaritmus plochy
mod_b_log_plochy <-
  lm(
    formula = pocet_druhu ~ log(plocha_km2),
    data = data_ostrovy
  )

# C: mocninový vztah, logaritmus obou proměnných
mod_c_mocninny <-
  lm(
    formula = log(pocet_druhu) ~ log(plocha_km2),
    data = data_ostrovy
  )

# D: přímka s kvadratickým členem
mod_d_kvadraticky <-
  lm(
    formula = pocet_druhu ~ plocha_km2 + I(plocha_km2^2),
    data = data_ostrovy
  )

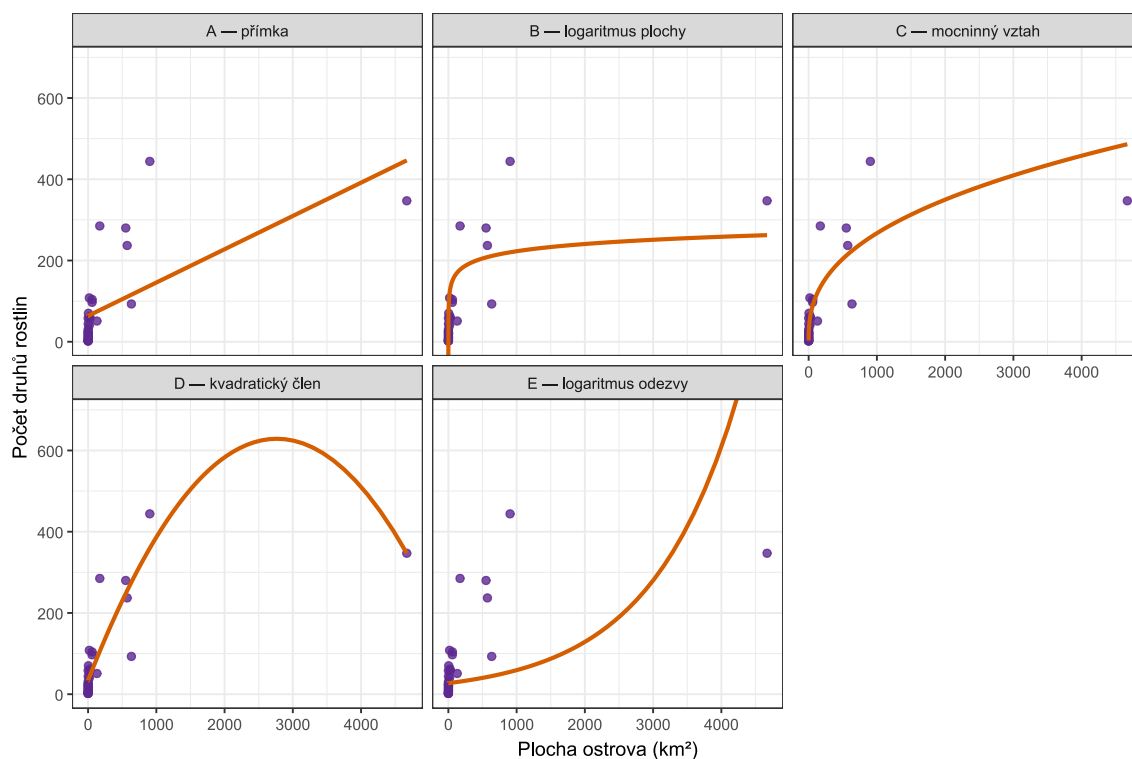
# E: logaritmus odezvy
```

```
mod_e_log_odezvy <-
  lm(
    formula = log(pocet_druhu) ~ plocha_km2,
    data = data_ostrovu
  )
```

Počet druhů je počet, tedy celé nezáporné číslo. Pro taková data existují modely, které s počty pracují přímo; poznáme je v jedné z následujících lekcí. Pro otázku tvaru vztahu nám zde lineární modely s transformacemi stačí.

### 6.3 Jak se kandidátní křivky shodují s ostrovy?

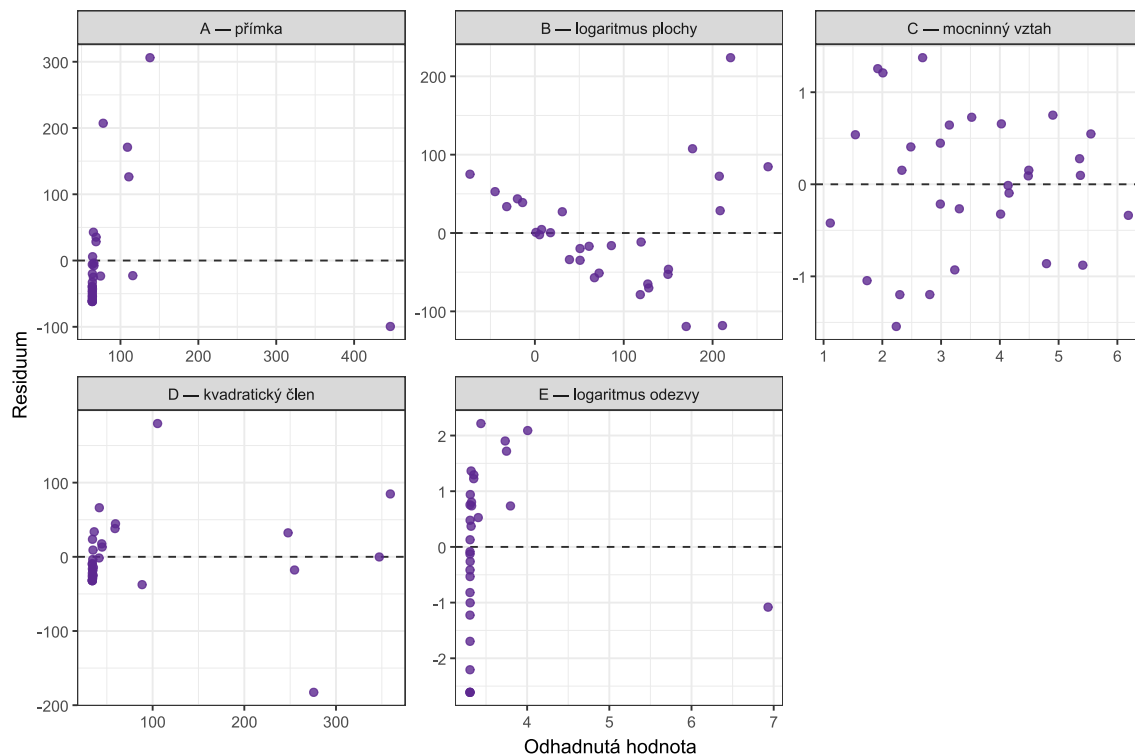
Modely C a E předpovídají logaritmus počtu druhů. Abychom všech pět kandidátů viděli na stejných osách, převedeme jejich odhady zpět na počet druhů funkcí `exp()`, která je opakem logaritmu.



Obrázek 29: Pět kandidátních modelů převedených zpět na počet druhů. Body jsou stejné ostrovy ve všech panelech.

Přímka A stejně jako dříve vede nad většinou malých ostrovů. Křivky B a C rostou rychle u malých ostrovů a u velkých zpomalují. Parabola D prochází blízko velkých ostrovů, ale má vrchol kolem 2 766 km<sup>2</sup> a dál klesá. Pro větší plochy tedy předpovídá *méně* druhů. Za vrcholem leží 1 ostrov (Isabela) a parabola se k němu ohýbá hlavně proto, že tento ostrov na ni má obrovský vliv

(Cookova vzdálenost 141.8). Exponenciála E zůstává téměř plochá a pak prudce stoupá, opačně než data.



Obrázek 30: Residua pěti kandidátů podle jejich odhadnutých hodnot. U kandidátů C a E jsou residua i odhady na logaritmické škále, protože tyto modely předpovídají logaritmus počtu druhů.

Nejvyrovnanější pás residuí kolem nuly má zatím kandidát C; podrobně jeho residua zkontrolujeme po srovnání AIC. U ostatních zůstávají vějíře nebo oblouky.

#### Šikmá odezva nemusí znamenat šikmá residua

Model B používá jako odezvu přímo počet druhů, který je silně pravostranně šikmý (Shapiro–Wilkův test samotného počtu druhů:  $p < 0.001$ ). U residuí modelu B test odchylku od normálního tvaru přesto neprokázal ( $p = 0.15$ ). Šikmost odezvy totiž z velké části vysvětluje šikmý prediktor. **Proto kontrolujeme residua, nikoli samotnou odezvu.** Za lepší model to model B nedělá: jeho residua se směrem doprava rozšiřují.

Grafy residuí však samy neříkají, o kolik je jeden model lepší než druhý. K tomu jsme v předchozí lekci používali AIC.

## 6.4 Můžeme porovnat AIC modelu pro počet druhů a pro jeho logaritmus?

Který model má nejnižší AIC?

Kandidát A má AIC 360.2 a kandidát C má AIC 74.5. Znamená to, že je model C mnohem lepší?

Ne. Rozdíl 285.7 vypadá obrovsky, ale nic neříká. Pravidlo z předchozí lekce zní: **AIC lze porovnávat jen mezi modely se stejnou odezvou na stejných pozorováních**. Model A popisuje počet druhů, model C jeho logaritmus. Logaritmus stlačí velké hodnoty, takže residua modelu C jsou čísla v úplně jiném měřítku než residua modelu A. AIC by se změnilo i tehdy, kdybychom použili logaritmus o jiném základu, aniž by se model jakkoli zlepšil. Stejně omezení platí pro  $R^2$ : u modelu A popisuje podíl variability počtu druhů, u modelu C podíl variability jeho logaritmu. Modely pro různé škály odezvy proto porovnáváme jinak: podle průběhu křivek převedených na původní osy a podle vzorů v residuích (nikoli podle velikosti residuí, která jsou v různých jednotkách).

Kandidáty proto porovnáme ve dvou oddělených skupinách:

| Kandidát              | Odezva                 | AIC   |
|-----------------------|------------------------|-------|
| C — mocninný vztah    | logaritmus počtu druhů | 74.5  |
| E — logaritmus odezvy | logaritmus počtu druhů | 110.8 |
| D — kvadratický člen  | počet druhů            | 334.5 |
| B — logaritmus plochy | počet druhů            | 345.9 |
| A — přímka            | počet druhů            | 360.2 |

**Mezi modely pro počet druhů** (A, B, D) má nejnižší AIC kvadratický model D (334.5). Z předchozí kapitoly ale víme, že jeho tvar biologicky nedává smysl, jeho residua netvoří vyrovnaný pás a jeho průběh určuje jediný ostrov. **Nejnižší AIC neznamená, že model je biologicky obhajitelný**. AIC porovnává jen shodu a složitost v rámci zvolených kandidátů.

**Mezi modely pro logaritmus počtu druhů** (C, E) má mocninný vztah C mnohem nižší AIC (74.5) než model E (110.8).

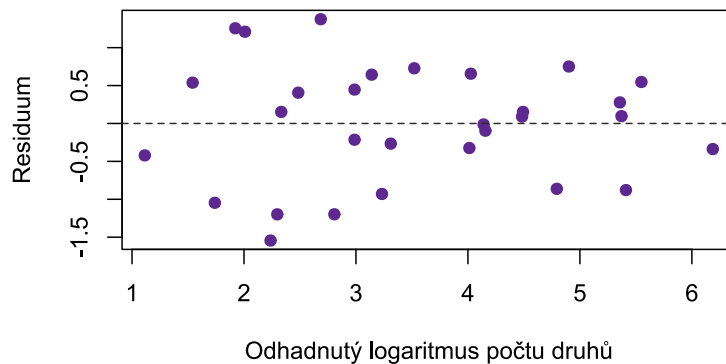
Jako kontrolu, nikoli jako dalšího kandidáta, se můžeme ještě zeptat, zda mocninný vztah sám nepotřebuje zakřivení, tedy zda body na logaritmických osách neleží spíš podél oblouku než přímky. Přidáme-li k modelu C kvadratický člen  $I(\log(\text{plocha\_km2})^2)$ ,  $R^2$  se změní jen z 0.757 na 0.758 a AIC vzroste na 76.3. Dodatečný koeficient si své místo nezaslouží.

## 6.5 Pomohl logaritmus také residuům?

Mocninný vztah C dává biologický smysl a mezi modely pro logaritmus počtu druhů vychází nejlépe. Mřížka residuů všech kandidátů naznačila, že jeho residua jsou vyrovnaná. Podrobně je zkontrolujeme stejnými nástroji jako u přímky.

```
# Residua podle odhadnutých hodnot
plot(
  x = fitted(object = mod_c_mocninny),
  y = residuals(object = mod_c_mocninny),
  pch = 19,
  col = "#5D2890",
  xlab = "Odhadnutý logaritmus počtu druhů",
  ylab = "Residuum"
)

abline(
  h = 0,
  col = "#2E2E2E",
  lty = 2
)
```



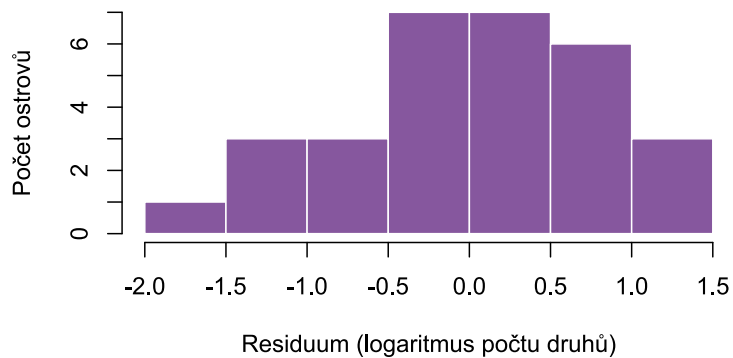
Obrázek 31: Residua mocninného vztahu podle odhadnutého logaritmu počtu druhů.

```
# Histogram residuů mocninného vztahu
hist(
  x = residuals(object = mod_c_mocninny),
  breaks = 10,
  col = "#86579E",
  border = "white",
)
```

```

main = "",
xlab = "Residuum (logaritmus počtu druhů)",
ylab = "Počet ostrovů"
)

```



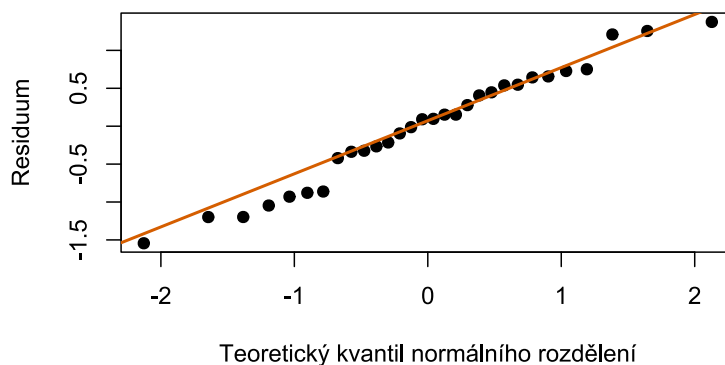
Obrázek 32: Histogram residuí mocninného vztahu.

```

# Q-Q graf residuí
qqnorm(
  y = residuals(object = mod_c_mocninny),
  pch = 19,
  main = "",
  xlab = "Teoretický kvantil normálního rozdělení",
  ylab = "Residuum"
)

qqline(
  y = residuals(object = mod_c_mocninny),
  col = "#D55E00",
  lwd = 2
)

```



Obrázek 33: Q–Q graf residuí mocninného vztahu.

```
# Doplnující test normality residuí
shapiro.test(x = residuals(object = mod_c_mocninny))
```

Shapiro-Wilk normality test

```
data:  residuals(object = mod_c_mocninny)
W = 0.97193, p-value = 0.5932
```

```
# Největší vliv jednoho ostrova
max(cooks.distance(model = mod_c_mocninny))
```

```
[1] 0.1401268
```

Zůstal v residuích nějaký vzor?

Porovnejte oba grafy s grafy residuí přímky. Vidíte ještě vějíř, oblouk nebo dlouhý ocas?

Residua tvoří přibližně stejně široký pás kolem nuly, Q–Q graf je blízko přímky a Shapiro–Wilkův test odchylku neprokázal ( $p = 0.59$ ). Žádný ostrov už model neovládá: největší Cookova vzdálenost je 0.14, oproti 113.3 u přímky. Protože máme jen 30 ostrovů, čteme velkou  $p$ -hodnotu opatrně; přesvědčivé je hlavně to, že všechny tři kontroly ukazují stejným směrem.

Zmizel i vějíř, ve kterém se residua přímky směrem doprava rozšiřovala. Logaritmus převádí stejné *relativní* rozptýlení (například o několik desítek procent nahoru i dolů) na přibližně stejné *absolutní* rozptýlení na logaritmické škále. V lekci o porovnání více skupin jsme u rozdílné variability

ve skupinách upozornili, že diagnostický graf nemá být automatickým příkazem k transformaci. Zde transformaci nepoužíváme kvůli grafu, ale proto, že biologie předpovídá násobné změny; vyrovnaná residua jsou toho důsledkem.

## 7 Interpretace po transformaci

### 7.1 Kolikrát víc druhů má dvakrát větší ostrov?

Mocninný vztah dává odpověď na hlavní otázku. Odpověď se skrývá v jeho koeficientech.

```
# Koeficienty a jejich 95% intervaly spolehlivosti  
coef(object = mod_c_mocninny)
```

```
(Intercept) log(plocha_km2)  
2.9037031      0.3885968
```

```
confint(  
  object = mod_c_mocninny,  
  level = 0.95  
)
```

```
                2.5 %    97.5 %  
(Intercept)    2.5819216 3.2254846  
log(plocha_km2) 0.3033903 0.4738033
```

Model říká:

$$\text{logaritmus počtu druhů} = \text{intercept} + \text{sklon} \times \text{logaritmus plochy}.$$

Sklon 0.389 neudává počet druhů na km<sup>2</sup>. Na logaritmických osách znamená, že **když se logaritmus plochy zvětší o jednotku, logaritmus počtu druhů se zvětší o 0.389**. To je pro biologa málo názorné. Přeložíme to proto zpět do násobků. Zdvojnásobení plochy zvětší logaritmus plochy o  $\ln(2)$ , a tedy logaritmus počtu druhů o  $\text{sklon} \times \ln(2)$ . Funkce `exp()`, opak logaritmu, převede tento přírůstek zpět na násobek:  $\exp(0.389 \times 0.693) = \exp(0.269) = 1.31$ . Stejné číslo dostaneme jednodušeji jako  $2^{0.389}$ . Slovy:

$$\text{násobek počtu druhů} = \text{násobek plochy}^{\text{sklon}}.$$

Pro dvojnásobnou plochu:

$$\begin{aligned}\text{násobek počtu druhů} &= 2^{0.389} \\ &= 1.31.\end{aligned}$$

**Ostrov s dvojnásobnou plochou má podle modelu typicky asi 1.31krát více druhů**, tedy zhruba o 31 % více, nikoli dvakrát více. 95% interval spolehlivosti sklonu (0.30 až 0.47) odpovídá násobku 1.23 až 1.39.

Obecně mocninný vztah zapisujeme:

$$S = c \cdot A^b \quad \Leftrightarrow \quad \ln(S) = \ln(c) + b \cdot \ln(A).$$

Symbol  $S$  je typický počet druhů,  $A$  plocha ostrova,  $b$  **exponent** (sklon na logaritmických osách) a  $c$  počet druhů očekávaný na ostrově o ploše 1 km<sup>2</sup> (jeho logaritmus je intercept). Při vynásobení plochy číslem  $k$  se počet druhů vynásobí číslem  $k^b$ . V ekologické literatuře se tento exponent obvykle značí  $z$  (nezaměňovat se  $z$ -skóre); pro souostroví se jeho odhady nejčastěji pohybují zhruba mezi 0,2 a 0,4 (Rosenzweig 1995, *Species Diversity in Space and Time*, Cambridge University Press).

#### Desetkrát větší ostrov

Spočítejte z exponentu, kolikrát více druhů má podle modelu typicky ostrov s desetinásobnou plochou. Porovnejte výsledek se svým odhadem ze začátku lekce.

#### Řešení

$10^{0.389} = 2.45$ . Desetkrát větší ostrov má typicky asi 2.45krát více druhů.

```
# Typické počty druhů pro vybrané plochy ostrovů
data_nove_plochy <-
  data.frame(plocha_km2 = c(1, 10, 100, 1000))

# Model předpovídá logaritmus; exp() jej převede zpět na počet druhů
exp(
  x = predict(
    object = mod_c_mocninny,
    newdata = data_nove_plochy
  )
)
```

| 1        | 2        | 3         | 4         |
|----------|----------|-----------|-----------|
| 18.24157 | 44.63330 | 109.20835 | 267.20995 |

Zpětný převod dává typickou, ne průměrnou hodnotu

Model C odhaduje průměr *logaritmu* počtu druhů. Po převodu funkcí `exp()` proto nedostaneme průměrný počet druhů, ale hodnotu blízkou mediánu: typický ostrov dané plochy, kterého polovina ostrovů přesahuje a polovina nedosahuje. U pravostranně šikmých dat je průměr vyšší než tato typická hodnota. Číslo, které vznikne zpětným převodem průměru logaritmu, se nazývá **geometrický průměr**. U počtu druhů na Galapágách je geometrický průměr 33.4, medián 42 a aritmetický průměr 85.2. V první lekci jsme pro asymetrická data doporučili medián; u dat, která jsou po logaritmování přibližně souměrná (log-normální), leží geometrický průměr blízko mediánu a přirozeně vzniká na logaritmické škále.

## 7.2 Jak číst koeficient, když logaritmujeme jen jednu stranu?

Mocninný vztah logaritmujeme obě strany. Při logaritmování jen jedné strany se mění i význam koeficientu. Ukážeme to na kandidátech B, E a D, i když nevyšli jako nejlepší.

### Stejný přírůstek, nebo stejný násobek?

Model B tvrdí, že každé zdvojnásobení plochy přidá stejný počet druhů. Je takový přírůstek stejně velká změna pro ostrov s několika druhy a pro ostrov se stovkami druhů?

**Logaritmus jen prediktoru (B: `pocet_druhu ~ log(plocha_km2)`).** Sklon 25.70 udává počet druhů přidáný při zvětšení logaritmu plochy o 1. Názornější je přepočtení na zdvojnásobení plochy:

$$\text{přírůstek druhů na zdvojnásobení plochy} = \text{sklon} \times \ln(2).$$

$$25.70 \times 0.693 = 17.8.$$

Podle modelu B přidá každé zdvojnásobení plochy asi 17.8 druhů, ať jde o ostrov malý nebo velký. Pro ostrov s několika druhy je to velká relativní změna, pro ostrov se stovkami druhů malá. Mocninný vztah C naopak předpovídá při každém zdvojnásobení stejnou *relativní* změnu. Obecně: přírůstek odezvy na  $k$ -násobek prediktoru je  $b \cdot \ln(k)$ , kde  $b$  je sklon.

**Logaritmus jen odezvy (E: `log(pocet_druhu) ~ plocha_km2`).** Sklon 0.00078 udává změnu logaritmu počtu druhů na 1 km<sup>2</sup>. Po zpětném převodu se z něj stává násobek:

$$\text{násobek počtu druhů na 100 km}^2 = \exp(100 \times \text{sklon}).$$

$$\exp(100 \times 0.00078) = 1.081.$$

Podle modelu E zvýší každých 100 km<sup>2</sup> počet druhů asi o 8.1 %. Obecně: při změně prediktoru o  $\Delta x$  se odezva vynásobí číslem  $\exp(b \cdot \Delta x)$ . Pro malé sklony je  $100 \cdot b$  přibližně procentní změna na jednotku prediktoru, protože pro malá  $b$  platí  $\exp(b) \approx 1 + b$ .

**Kvadratický člen (D).** Model D nemá jeden sklon. Jak rychle počet druhů s plochou roste, se mění podél křivky: u malých ploch rychle, pak stále pomaleji, až kolem 2 766 km<sup>2</sup> přestane růst a začne klesat. Koeficienty kvadratického modelu proto nečteme jednotlivě. Čteme predikce pro konkrétní plochy a místo vrcholu z grafu. Předpovědi mimo rozsah pozorovaných ploch jsou u

paraboly obzvlášť riskantní, protože křivka pokračuje dolů, i když k tomu data nedávají žádný důvod.

#### Jak číst koeficienty po transformaci

| Model                  | Co znamená sklon                                                                                                    |
|------------------------|---------------------------------------------------------------------------------------------------------------------|
| $y \sim x$             | změna $y$ o sklon na jednotku $x$                                                                                   |
| $y \sim \log(x)$       | změna $y$ o sklon $\times \ln(2)$ na každé zdvojnásobení $x$                                                        |
| $\log(y) \sim x$       | vynásobení $y$ číslem $\exp(\text{sklon})$ na jednotku $x$ ; přibližně $100 \times \text{sklon} \%$ pro malé sklony |
| $\log(y) \sim \log(x)$ | vynásobení $y$ číslem $2^{\text{sklon}}$ na každé zdvojnásobení $x$                                                 |
| $y \sim x + I(x^2)$    | sklon se mění podél křivky; čteme predikce, ne jednotlivé koeficienty                                               |

### 7.3 Jak rychle roste počet druhů s plochou ostrova?

#### Odpověď na hlavní otázku

Počet druhů rostlin na Galapágách roste s plochou ostrova podle mocninného vztahu s exponentem asi 0.39 (95% interval spolehlivosti 0.30 až 0.47). Dvojnásobná plocha znamená typicky asi 1.31krát více druhů a desetinásobná plocha asi 2.45krát více. **Počet druhů tedy roste s plochou, ale mnohem pomaleji než plocha samotná.**

Tento závěr podporují tři argumenty, které ale vycházejí ze stejných ostrovů: biologicky zdůvodněný tvar vztahu, residua bez systematického vzoru a nejnižší AIC mezi modely pro logaritmus počtu druhů. S modely pro samotný počet druhů (A, B, D) model C pomocí AIC porovnat nelze; od nich jej odlišují biologický smysl a residua. Závěr má také jasná omezení:

- Máme jen 30 ostrovů, takže diagnostické grafy i test normality mají omezenou citlivost.
- Odezvou je počet. Model, který s počty pracuje přímo, by mohl lépe popsat malé ostrovy s několika druhy.
- Ostrovy jednoho souostroví nejsou úplně nezávislé: sdílejí dodavatelské ostrovy, klima i historii osídlení.
- Plocha souvisí s dalšími vlastnostmi ostrova, například s nadmořskou výškou a rozmanitostí prostředí. Mocninný vztah popisuje souvislost, nikoli to, že plocha sama o sobě druhy „vytváří“.

Napište vlastní závěr

Shrňte ve dvou větách, co jste se o vztahu plochy a počtu druhů dozvěděli. Uveďte jedno číslo s interpretací a jedno omezení.

### Příklad odpovědi

Na Galapágách má ostrov s dvojnásobnou plochou typicky asi o 31 % více druhů rostlin, jak ukazuje mocninný vztah s exponentem 0.39. Závěr vychází z pouhých 30 ostrovů jednoho souostroví, takže jeho přenos na jiná souostroví je nejistý.

Úvodní otázku o stromech z předchozí lekce zbývá uzavřít. Stejný postup můžeme vyzkoušet i na ní.

### Návrat ke stromům: zmizí oblouk v residuích?

Na začátku lekce jsme si připomněli oblouk v residuích aditivního modelu pro výšku jehličnanů. Stejně jako u ostrovů zvažme biologicky věrohodný tvar: výška stromu roste s průměrem kmene stále pomaleji, protože strom nemůže růst do výšky donekonečna. Takové zpomalení vyjadřuje logaritmus průměru kmene.

```
# Načteme stromy z předchozí lekce
data_stromy_l <-
  read.csv(
    file = here::here("data", "allometrie_stromu.csv"),
    stringsAsFactors = TRUE
  )

# Aditivní model z předchozí lekce
mod_stromy_primka <-
  lm(
    formula = vyska_stromu_m ~ prumer_kmene_cm + druh_stromu,
    data = data_stromy_l
  )

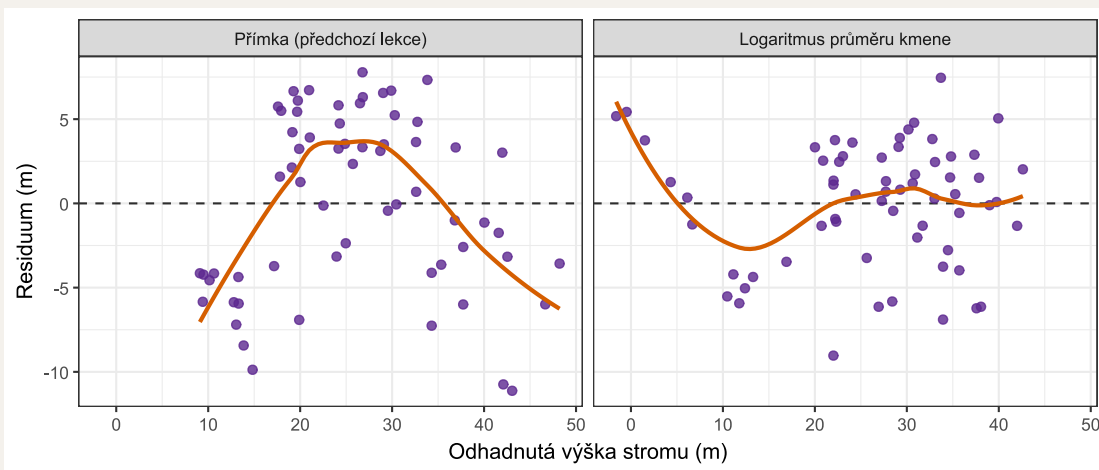
# Stejný model s logaritmem průměru kmene
mod_stromy_log_prumer <-
  lm(
    formula = vyska_stromu_m ~ log(prumer_kmene_cm) + druh_stromu,
    data = data_stromy_l
  )
```

```
# Oba modely mají stejnou odezvu, takže AIC můžeme porovnat  
AIC(object = mod_stromy_primka)
```

```
[1] 395.4366
```

```
AIC(object = mod_stromy_log_prumer)
```

```
[1] 350.6625
```



Obrázek 34: Residua dvou modelů výšky stromů podle odhadnuté výšky. Oranžová křivka pouze vyhlazuje průběh bodů.

Oblouk se zploštil, ale nezmizel úplně: stromy s odhadnutou výškou zhruba 10–20 m model stále systematicky přeceňuje (jejich residua leží převážně pod nulou), zatímco u vyšších stromů leží residua kolem nuly. AIC kleslo z 395.4 na 350.7. Odpověď na otázku z předchozí lekce tedy zní: **tvar vztahu byl hlavní, i když ne jediný problém**. Nižší stromy by si zasloužily další pozornost, například jiný tvar vztahu nebo kontrolu jejich měření. Podle nového modelu přidá každé zdvojnásobení průměru kmene stromu asi 10.3 m výšky.

Kvadratický model by měl AIC ještě o něco nižší (341.3). Jeho parabola má ale vrchol u průměru kmene asi 61 cm a pro 6 nejširších stromů předpovídá, že jsou *nižší* než užší stromy. Je to stejná lekce jako u ostrovů: nejnižší AIC nenahradí biologickou úvahu o tvaru vztahu.

## 7.4 Kdy transformaci použít a kdy ne?

Postup z této lekce lze použít i na jiná data:

1. Položte biologickou otázku a prohlédněte data, včetně tvaru rozdělení hlavních proměnných.

2. Odhadněte nejjednodušší rozumný model a zkontrolujte jeho residua: graf podle odhadnutých hodnot, Q–Q graf a vliv jednotlivých pozorování. Test normality čtete až po grafech a s ohledem na počet pozorování.
3. Pokud residua ukazují systematický vzor, zeptejte se, jaký tvar vztahu předpovídá biologie: stejný přírůstek, stejný násobek, zpomalující růst, nebo optimum?
4. Převeďte biologická tvrzení na malou sadu kandidátních formulí s  $\log()$  nebo  $I(x^2)$ .
5. Porovnejte kandidáty na původní škále v grafu, pomocí residuí a pomocí AIC, ale AIC jen mezi modely se stejnou odezvou.
6. Výsledek interpretujte v původních jednotkách: násobky, přírůstky na zdvojnásobení, procenta.

### Transformace není rituál

Transformaci volíme proto, že vyjadřuje biologické tvrzení o vztahu, nikoli proto, aby test normality vrátil velkou p-hodnotu. Lineární transformace (z-skóre, centrování) tvar rozdělení neopraví nikdy. A pokud šikmá je jen odezva nebo prediktor, ale residua jsou v pořádku, není co opravovat kvůli normalitě. Vlivná pozorování, jako byla Isabela u přímky, ale sledujeme vždy.

### Jsou ostrovy jednoho souostroví nezávislá pozorování?

Všech 30 ostrovů leží v jednom souostroví. Sousední ostrovy si vyměňují semena, sdílejí podnebí a mají společnou historii. Jsou to opravdu 30 nezávislých kousků informace? Co když data obsahují měření ze stejných lokalit, bloků nebo jedinců? Těmito otázkami se budeme zabývat v následující lekci. Později se také vrátíme k počtům druhů a ukážeme modely, které s počty a podíly pracují přímo, bez transformace.

## Shrnutí

- Tvar rozdělení popisujeme slovy (souměrné, pravostranně a levostranně šikmé, těžké konce, dva vrcholy, omezené hodnoty), histogramem a Q–Q grafem.
- Mnoho biologických veličin je pravostranně šikmých, protože jsou zdola omezené nulou nebo vznikají násobením.
- Lineární model předpokládá přibližně normální rozdělení **residuí**, nikoli odezvy nebo prediktorů. Test normality residuí doplňuje Q–Q graf a jeho výsledek závisí na počtu pozorování.
- Změna jednotek, centrování a z-skóre mění jen polohu a měřítko; tvar rozdělení zachovají. Logaritmus a odmocnina mění tvar.
- Logaritmus převádí stejné násobky na stejné kroky, a proto umí z mocninného vztahu udělat přímku.
- Kandidátní tvary vztahu odvozujeme z biologie a porovnáváme grafem, residuí a AIC, ale AIC jen mezi modely se stejnou odezvou.

- Po transformaci se mění význam koeficientů: sklon mocninného vztahu udává, kolikrát se změni odezva při zdvojnásobení prediktoru.

## Slovníček pojmů

Pojmy použité v těchto materiálech a jejich definice:

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