

O kolik déle čekáme po delší erupci — a jak jistě to víme?

Od odhadu k nejistotě: výběrová variabilita a interval spolehlivosti

21. října 2026

Obsah

1 Úvod	2
2 Výsledky učení	3
3 Každý řádek je jedna erupce	3
4 Delší erupce, delší čekání?	10
4.1 Sedí přímka k datům?	11
5 Pepa: jeden měsíc, jeden odhad	13
6 Mařenka a Karel: jiné měsíce, jiné přímky	14
7 Sto hlášení v aplikaci GeyserWatch	15
8 Co měří standardní chyba?	18
9 Co dělá sklon stabilnějším?	19
10 Standardní chyba z jediného výběru	21
10.1 Typická velikost residua	21
10.2 Rozptýlení délek a počet erupcí	23
10.3 Souhlasí vzorec se simulací?	23
10.4 Obvyklý zápis vzorce	24
11 Kde ve výstupu najdeme velikost a přesnost?	24
11.1 Co říká standardní chyba o našem sklonu?	29
12 Jak vyjádřit rozsah nejistoty?	30
12.1 Připomenutí kvantilů	30
12.2 Od odhadu a standardní chyby k intervalu	31
12.3 Jak se interval mění s požadovanou hladinou spolehlivosti?	33
12.4 Jak číst hladinu spolehlivosti?	34
13 Jak přesně známe celou přímku?	36
13.1 Sto hlášení, sto přímek	37
13.2 Pás spolehlivosti kolem přímky	38
13.3 Kdy je pás široký?	39
13.4 Pás v číslech: <code>predict()</code>	40
13.5 Co pás neříká	44
14 Může jedna erupce změnit sklon?	45

14.1	Krajní erupce mají delší páku	47
14.2	Cookova vzdálenost: co kdyby erupce chyběla?	49
14.3	Mění některá z našich erupcí závěr?	53
14.4	Co s vlivným pozorováním?	55
15	Co můžeme říct o Old Faithful?	56
16	Co bude následovat dál	56
17	Shrnutí	56
17.1	Závěrečná otázka	57
	Slovníček pojmů	57

1 Úvod

Gejzír Old Faithful v Yellowstonském národním parku střídá erupce a období klidu. Podle [National Park Service](#) trvá erupce přibližně 1,5 až 5 minut a čekání na další erupci se obvykle pohybuje od 50 do 127 minut. Návštěvníci i správa parku se proto ptají, zda délka právě skončené erupce napoví, jak dlouho budeme čekat na další.

V předchozí lekci jsme se naučili odhadnout lineárním modelem, o kolik se v průměru liší odezva dvou pozorování, jejichž prediktor se liší o jednu jednotku. Odhad z jednoho souboru dat ale není celá odpověď. Kdyby gejzír pozoroval někdo jiný v jiných dnech, zaznamenal by jiné erupce a jeho odhad by mohl vyjít jinak. Potřebujeme proto znát nejen velikost efektu, ale také to, jak přesně jsme ji odhadli.

Hlavní otázka této lekce

O kolik déle čekáme po delší erupci — a jak jistě to víme?

Co čekáte ještě před prvním grafem?

Představte si, že gejzír pozorují další lidé v jiných dnech a každý z nich spočítá vlastní sklon přímk. Jaké sklony dostanou?

- **A:** Přesně stejný sklon jako my.
- **B:** Podobné, ale ne stejné sklony.
- **C:** Zcela odlišné sklony, které spolu nesouvisejí.

Než budete číst dál, vyberte možnost, která vám připadá nejpravděpodobnější, a zdůvodněte ji.

Odpověď: podobné, ale ne stejné sklony

Nejpravděpodobnější je **možnost B**. Každý pozorovatel zaznamená jiné erupce. Pokud vztah mezi délkou erupce a čekáním skutečně existuje, budou si odhady podobné; protože však body

nikdy neleží přesně na přímce, budou se navzájem trochu lišit. Tato lekce ukazuje, jak velký takový rozdíl bývá a jak jej popsat čísly.

2 Výsledky učení

Po prostudování této lekce budete schopni:

- vysvětlit, proč jiný soubor pozorování dává jiný odhad sklonu;
- najít ve výstupu `summary()` odhad sklonu a jeho standardní chybu, vysvětlit, co standardní chybu zmenšuje, a rozlišit velikost vztahu od přesnosti odhadu;
- získat 95% interval spolehlivosti sklonu funkcí `confint()` a pás spolehlivosti kolem přímky funkcí `predict()` a interpretovat je bez tvrzení, že obsahují 95 % pozorování nebo dokazují hypotézu;
- posoudit pomocí pákového efektu a Cookovy vzdálenosti, zda jedna erupce výrazně mění odhad sklonu;
- zapsat výsledek jednou větou s odhadem, standardní chybou, intervalem a jednotkami.

3 Každý řádek je jedna erupce

Používáme skutečná pozorování erupcí Old Faithful z léta 2024, která zaznamenali dobrovolníci prostřednictvím služby [GeyserTimes](#). Jeden řádek tabulky odpovídá jedné erupci a čekání na erupci následující; erupce je tedy jednotkou pozorování.

Soubor [old_faithful_2024.csv](#) je součástí materiálů. Pro následující kód jej uložte do složky `data` ve svém projektu v R. Délka erupce je v souboru v sekundách; pro výklad ji převedeme na minuty, aby obě měření měla stejnou jednotku:

```
# Načteme tabulku erupcí ze souboru CSV
data_gejzir <-
  read.csv(file = "data/old_faithful_2024.csv")

# Datum převedeme z textu na kalendářní datum
data_gejzir$date <-
  as.Date(x = data_gejzir$date)

# Délku erupce převedeme ze sekund na minuty
data_gejzir$delka_erupce_min <-
  data_gejzir$duration / 60

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

```
'data.frame':  117 obs. of  8 variables:
 $ eruption_id    : int  1460898 1460909 1461175 1461192 1461302 1461893 1461911
1461934 1462008 1462123 ...
 $ geyser          : chr   "Old Faithful" "Old Faithful" "Old Faithful"
"Old Faithful" ...
 $ date            : Date, format: "2024-06-02" "2024-06-02" ...
 $ time            : int   912 1157 1317 1458 1117 1130 1323 1455 1816 723 ...
 $ cekani_min      : int   68 102 101 94 67 113 92 102 107 85 ...
 $ webcam          : chr   "Yes" "Yes" "Yes" "Yes" ...
 $ duration        : int  119 264 259 258 116 222 230 241 215 136 ...
 $ delka_erupce_min: num   1.98 4.4 4.32 4.3 1.93 ...
```

V lekci budeme pracovat se dvěma numerickými proměnnými:

- `delka_erupce_min`: délka právě pozorované erupce v minutách — náš prediktor;
- `cekani_min`: čekání do následující erupce v minutách — naše odezva.

Sloupec `date` použijeme k rozdělení pozorování podle měsíců. Ostatní sloupce popisují identifikátor erupce, čas a to, zda erupci zachytila webkamera.

Kontrola dat pokračuje rozsahy obou měření, chybějícími hodnotami a počtem erupcí v jednotlivých měsících:

```
# Minimum, kvartily, průměr a maximum obou měření
summary(
  object = data_gejzir[, c("delka_erupce_min", "cekani_min")]
)
```

```
delka_erupce_min  cekani_min
Min.   :1.650    Min.    : 62.00
1st Qu.:3.000    1st Qu.: 88.00
Median :4.000    Median : 99.00
Mean   :3.625    Mean    : 94.26
3rd Qu.:4.317    3rd Qu.:104.00
Max.   :4.833    Max.    :116.00
```

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

eruption_id	geyser	date	time
0	0	0	0
cekani_min	webcam	duration	delka_erupce_min
0	0	0	0

```
# Počet erupcí v jednotlivých měsících (06 = červen)
table(format(x = data_gejzir$date, format = "%m"))
```

```
06 07 08
39 39 39
```

Tabulka obsahuje 117 erupcí od 02. 06. 2024 do 30. 08. 2024: v červnu, červenci a srpnu postupně 39, 39, 39. Erupce trvaly od 1,65 do 4,83 minuty a čekání se pohybovalo od 62 do 116 minut; žádný sloupec nemá chybějící hodnoty. Oba rozsahy leží uvnitř hodnot uváděných National Park Service, takže nic nenaznačuje chybu jednotek.

Datum	Délka erupce (min)	Čekání (min)
2. 6. 2024	1,98	68
2. 6. 2024	4,40	102
3. 6. 2024	4,32	101
3. 6. 2024	4,30	94
4. 6. 2024	1,93	67
6. 6. 2024	3,70	113

Doplňující: jak vznikl soubor `old_faithful_2024.csv`

Erupce Old Faithful zaznamenávají dobrovolníci v databázi [GeyserTimes](#). Tabulku z roku 2024 připravili autoři učebnice ModernDive jako tabulku `old_faithful_2024` v balíčku [modern-dive](#). Pro tuto lekci jsme stejný postup zopakovali přímo z [oficiálního archivu erupcí Old Faithful](#):

- ponechali jsme pouze přesně zaznamenané hlavní erupce z června, července a srpna 2024;
- čekání jsme spočítali jako rozdíl časů dvou po sobě jdoucích erupcí a ponechali intervaly od 50 do 127 minut, což odpovídá [rozsaahu uváděnému National Park Service](#); delší mezera v dobrovolnických záznamech pravděpodobně znamená, že některá erupce nebyla zaznamenána;
- odstranili jsme tři erupce, které vypadaly jako odlehlá pozorování, jednu z každého měsíce; jejich identifikátory jsou uvedeny přímo v kódu. Erupce dlouhá jen 41 sekund leží pod rozsahem 1,5 až 5 minut uváděným National Park Service. Zbylé dvě jsme odstranili jen kvůli přehlednějšímu výukovému příkladu; v kapitole „Může jedna erupce změnit sklon?“ je vrátíme zpět a ukážeme, že sklon téměř neovlivňují, takže bychom je po takovém ověření v datech ponechali.

Základní postup lekce si můžete vyzkoušet i s klasickou tabulkou `faithful`, která je součástí R. Neobsahuje však data erupcí, takže s ní nelze zopakovat rozdělení podle měsíců:

```
# Přejmenujeme sloupce klasické tabulky podle této lekce
data_faithful <-
  data.frame(
    delka_erupce_min = faithful$eruptions,
    cekani_min = faithful$waiting
  )
```

Úplný záznam stažení a přípravy dat používá i pokročilejší kroky, které pro lekci není nutné rozumět ani spouštět. Pokud GeyserTimes požádá o ověření v prohlížeči, uložte archiv ručně do složky `data` pod názvem `Old_Faithful_eruptions.tsv.gz`.

```

# Sestavíme adresu oficiálního archivu Old Faithful
archive_url <-
  paste0(
    "https://geysertimes.org/archive/geysers/",
    "Old_Faithful_eruptions.tsv.gz"
  )

# Určíme, kam se stažený soubor uloží v projektu
archive_file <-
  here::here(
    "data",
    "Old_Faithful_eruptions.tsv.gz"
  )

# Soubor stáhneme jen tehdy, když jej ještě nemáme
if (
  !file.exists(archive_file)
) {
  download.file(
    url = archive_url,
    destfile = archive_file,
    mode = "wb"
  )
}

# Načteme archiv, pojmenujeme potřebné sloupce a převedeme čas
data_archive <-
  readr::read_tsv(
    file = gzfile(description = archive_file),
    na = c("", "NULL"),
    quote = "",
    show_col_types = FALSE
  ) |>
  dplyr::transmute(
    eruption_id = as.numeric(eruptionID),
    primary_id = as.numeric(associated_primaryID),
    geyser,
    datetime = lubridate::as_datetime(
      x = as.numeric(eruption_time_epoch),
      tz = "America/Denver"
    ),
    exact = exact == 1,

```

```

    near_start = ns == 1,
    in_eruption = ie == 1,
    questionable = q == 1,
    webcam = wc == 1,
    duration = as.numeric(duration_seconds)
  ) |>
# Ponecháme pouze přesně zaznamenané hlavní erupce
dplyr::filter(
  is.na(primary_id) | eruption_id == primary_id,
  exact,
  !near_start,
  !in_eruption,
  !questionable
) |>
dplyr::arrange(datetime) |>
# Z časů dvou sousedních erupcí vypočítáme dobu čekání
dplyr::mutate(
  waiting = as.numeric(
    difftime(
      time1 = dplyr::lead(datetime),
      time2 = datetime,
      units = "mins"
    )
  )
)

# Vybereme tři měsíce a proměnné potřebné pro výuku
data_export <-
  data_archive |>
  dplyr::filter(
    datetime >= lubridate::ymd(
      "2024-06-01",
      tz = "America/Denver"
    ),
    datetime < lubridate::ymd(
      "2024-09-01",
      tz = "America/Denver"
    ),
    !is.na(duration),
    dplyr::between(waiting, left = 50, right = 127)
  ) |>
  dplyr::transmute(

```

```

    eruption_id,
    geyser,
    date = lubridate::as_date(datetime),
    time = lubridate::hour(datetime) * 100 +
      lubridate::minute(datetime),
    cekani_min = waiting,
    webcam = dplyr::if_else(webcam, "Yes", "No"),
    duration
  )

# Odstraníme tři předem prověřená odlehlá pozorování
outlier_ids <-
  c(
    1461337, # červen
    1468130, # červenec
    1471579 # srpen
  )

data_export <-
  data_export |>
  dplyr::filter(!eruption_id %in% outlier_ids)

# Ověříme, že příprava dat skončila očekávaným počtem řádků
expected_rows <- 117L

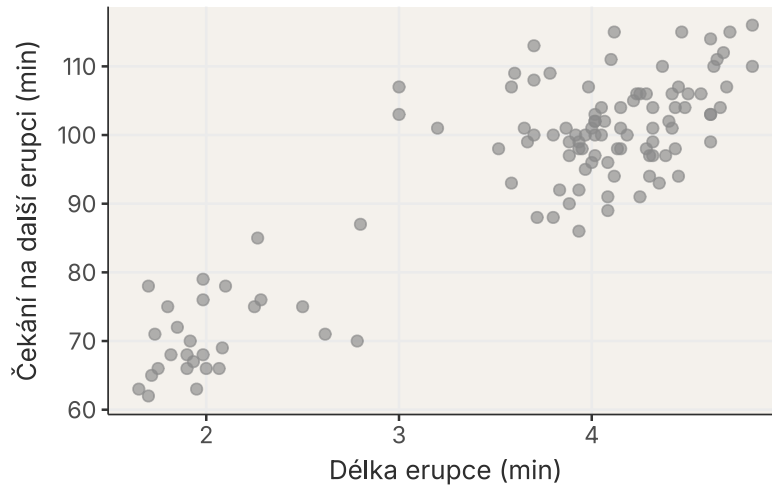
if (
  nrow(data_export) != expected_rows
) {
  stop("Zdrojový soubor má neočekávaný počet řádků.")
}

# Uložíme stabilní výukovou tabulku
readr::write_csv(
  x = data_export,
  file = here::here("data", "old_faithful_2024.csv"),
  na = ""
)

```

Pro výuku používáme samostatný soubor CSV, protože v praxi stačí načíst jediný soubor pomocí `read.csv()`. Původ dat a podmínky jejich použití shrnuje také soubor `data/README.md`.

4 Delší erupce, delší čekání?



Obrázek 1: Každý bod je jedna zaznamenaná erupce: její délka a čekání na další erupci.

Co v grafu vidíte?

1. Jaký směr má vztah mezi délkou erupce a čekáním?
2. Které části grafu obsahují nejvíce bodů?
3. Leží body na jediné přímce?

Odpověď: rostoucí vztah s rozptýlenými body

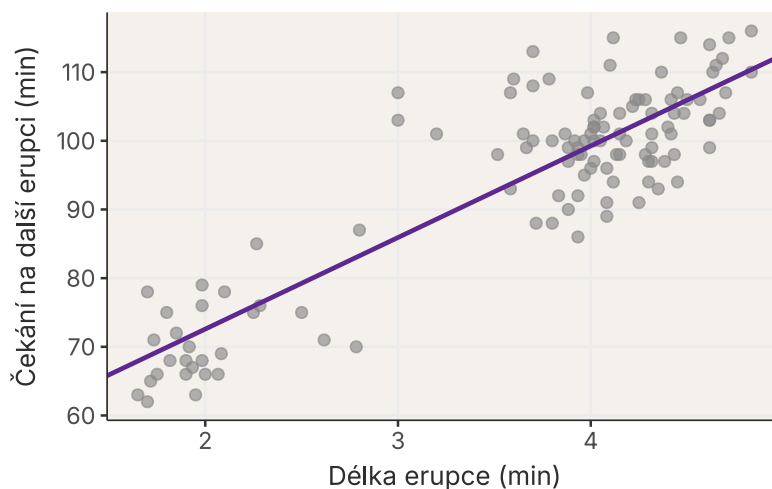
Vztah je rostoucí: po delších erupcích zpravidla čekáme déle. Body tvoří dvě skupiny – krátké erupce kolem dvou minut a dlouhé erupce kolem čtyř minut – a ani v jedné neleží na jediné přímce. Při stejné délce erupce se čekání liší o několik minut.

Stejný vztah shrneme lineárním modelem, jak jsme to dělali v předchozí lekci:

```
# Odhadneme čekání podle délky erupce ze všech erupcí
mod_gejzir <-
  lm(
    formula = cekani_min ~ delka_erupce_min,
    data = data_gejzir
  )

# Zobrazíme odhadnutý intercept a sklon
coef(object = mod_gejzir)
```

```
(Intercept) delka_erupce_min
45.91218      13.33503
```



Obrázek 2: Stejně erupce s přímkou odhadnutou ze všech zaznamenaných erupcí.

Odhadnutý sklon je 13,34 minuty čekání na minutu erupce: dvě erupce, jejichž délky se liší o jednu minutu, se podle modelu v průměru liší čekáním na další erupci přibližně o 13,34 minuty. Stejná data a stejný vzorec modelu dávají vždy stejné odhady; funkce `lm()` nic náhodně nelosuje. Odhad se může změnit jen tehdy, když se změní pozorované erupce.

4.1 Sedí přímka k datům?

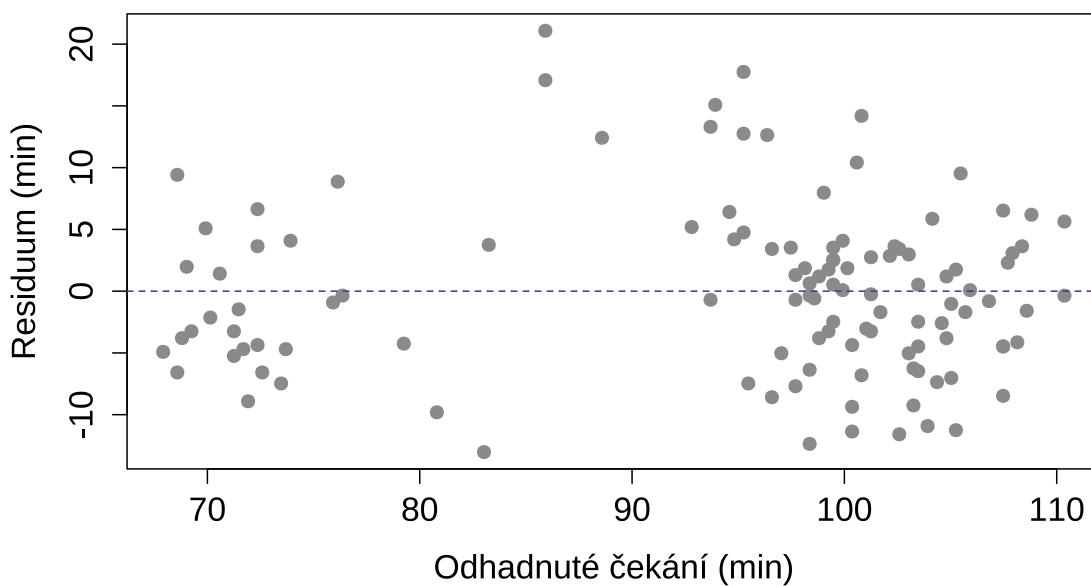
Než budeme odhadu věřit, zkontrolujeme jako v předchozí lekci graf residuí proti odhadnutým hodnotám:

```

plot(
  # Každý bod spojuje odhadnutou hodnotu s residuem stejné erupce
  x = fitted(object = mod_gejzir),
  y = resid(object = mod_gejzir),
  xlab = "Odhadnuté čekání (min)",
  ylab = "Residuuum (min)",
  pch = 16,
  col = "#8A8A8A",
  cex = 1.3,
  cex.axis = 1.4,
  cex.lab = 1.4
)

abline(
  # Body hledáme rozptýlené kolem nulového residua
  h = 0,
  lty = 2,
  col = "#5D2890"
)

```



Residua tvoří dvě skupiny podle krátkých a dlouhých erupcí a v obou leží kolem nuly bez zřejmého vějíře, tedy bez rozšiřování rozptylu residuí (heteroskedasticity). Několik erupcí střední délky ale leží výrazně nad nulou: po nich se čekalo déle, než přímka odhaduje. Přímka tedy

vztah dobře shrnuje, ale ne dokonale. Tuto výhradu si ponecháme, protože i standardní chyba a interval spolehlivosti, ke kterým lekce směřuje, platí jen za předpokladu, že model data přiměřeně popisuje.

Data jsou navíc skutečná, ale dobrovolníci nezaznamenali náhodně všechny erupce: pozorovali v dostupných termínech a blízké erupce mohou být časově propojené. Při interpretaci proto oddělujeme **vztah v zaznamenaných datech** od širšího procesu, který chceme popsat.

5 Pepa: jeden měsíc, jeden odhad

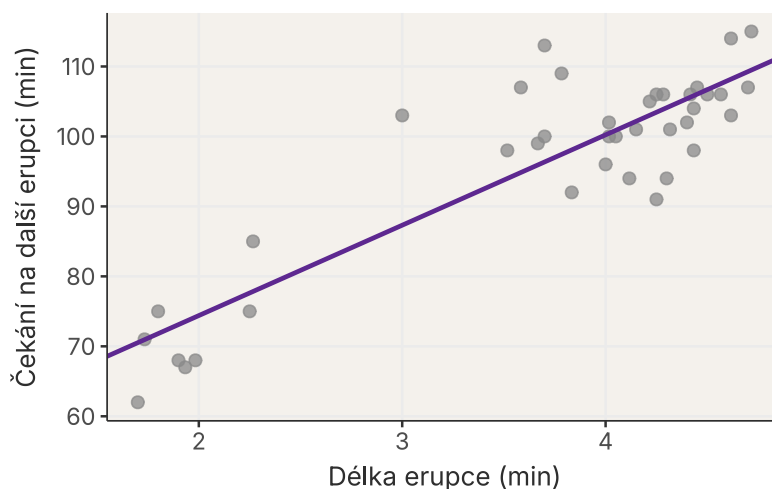
Představme si výzkumníka **Pepu**, který gejzír pozoroval pouze v červnu. Ze všech erupcí ponechá jen ty červnové a odhadne stejný model:

```
# Ponecháme jen řádky, jejichž datum patří do června
data_pepa <-
  data_gejzir[
    format(x = data_gejzir$date, format = "%m") == "06",
  ]

# Odhadneme stejný vztah pouze z Pepových dat
mod_pepa <-
  lm(
    formula = cekani_min ~ delka_erupce_min,
    data = data_pepa
  )

# Zobrazíme odhadnutý intercept a sklon
coef(object = mod_pepa)
```

```
(Intercept) delka_erupce_min
48.56511      12.91533
```



Obrázek 3: Pepova červnová pozorování a přímka odhadnutá pouze z nich.

Pepa zaznamenal 39 erupcí. Jeho odhad sklonu je 12,92 minuty čekání na minutu erupce: v jeho datech se dvě erupce, jejichž délky se liší o jednu minutu, v průměru liší čekáním přibližně o 12,92 minuty.

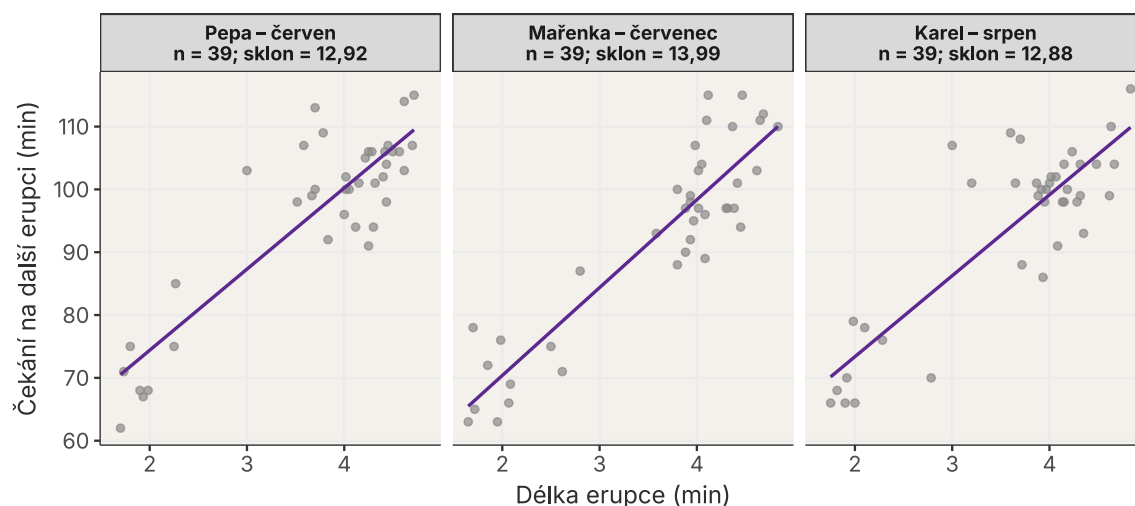
Tento výsledek správně popisuje **Pepova data**. Neznamená však, že by každý další pozorovatel musel získat stejný sklon.

6 Mařenka a Karel: jiné měsíce, jiné přímky

Dalšími pozorovateli jsou **Mařenka**, která sledovala gejzír v červenci, a **Karel**, který jej sledoval v srpnu. Oba použijí stejné proměnné a stejný model jako Pepa.

Co čekáte?

Budou mít Pepa, Mařenka a Karel stejný sklon? Pokud ne, čím se jejich výpočty liší?



Obrázek 4: Tři lineární modely odhadnuté stejným postupem na erupcích ze tří měsíců. V záhlaví každého panelu je počet erupcí a odhad sklonu.

Výzkumník a měsíc	Počet erupcí	Intercept (min)	Sklon (min/min)
Pepa – červen	39	48,57	12,92
Mařenka – červenec	39	42,41	13,99
Karel – srpen	39	47,61	12,88

Pepa, Mařenka i Karel použili **stejný postup**: stejné proměnné, stejný model a stejnou funkci `lm()`. Změnila se pouze pozorovaná data — a s nimi přímka i sklon. Rozdíly mezi měsíci mohou zahrnovat náhodnou proměnlivost i skutečné změny chování gejzíru; tři měsíce proto nejsou dokonale kontrolovaný experiment. Jsou ale prvním konkrétním důkazem, že **odhad není neměnné číslo oddělené od dat, ze kterých vznikl**.

Jednomu souboru pozorování, ze kterého počítáme odhad, se odborně říká **výběr** (anglicky *sample*). Česky se používá také **výběrový soubor** nebo méně formálně **vzorek**. Proměnlivost odhadu mezi možnými výběry nazýváme **výběrová variabilita** (anglicky *sampling variability*): jiné pozorované erupce mohou vytvořit jinou přímku.

7 Sto hlášení v aplikaci GeyserWatch

Tři měsíce nám daly pouze tři přímky. Abychom viděli, jak se odhady chovají při mnoha opakováních, zavedeme fiktivní aplikaci **GeyserWatch**. Do ní pošle hlášení 100 návštěvníků; každé hlášení obsahuje 40 erupcí. Z každého hlášení odhadneme stejný model a uložíme jeho sklon.

Pomůže oddělit tři pojmy. Rozdíl mezi výběrem a cílovou populací jsme zmínili v první lekci; přibývá k nim odhad:

- **Skutečnost** (širší proces, cílová populace) je to, jak se gejzír chová u všech erupcí. Skutečný sklon nikdy přímo nevidíme.

- **Výběr** jsou erupce, které někdo skutečně zaznamenal, například našich 117.
- **Odhad** je číslo spočítané z výběru, například sklon 13,34.

Ve výzkumu máme jeden výběr a jeden odhad, skutečnost neznáme. V simulaci si naopak skutečnost zvolíme sami a vytvoříme z ní mnoho výběrů. Proto u simulace známe skutečný (referenční) sklon a můžeme sledovat, jak daleko od ní odhady z jednotlivých výběrů padají.

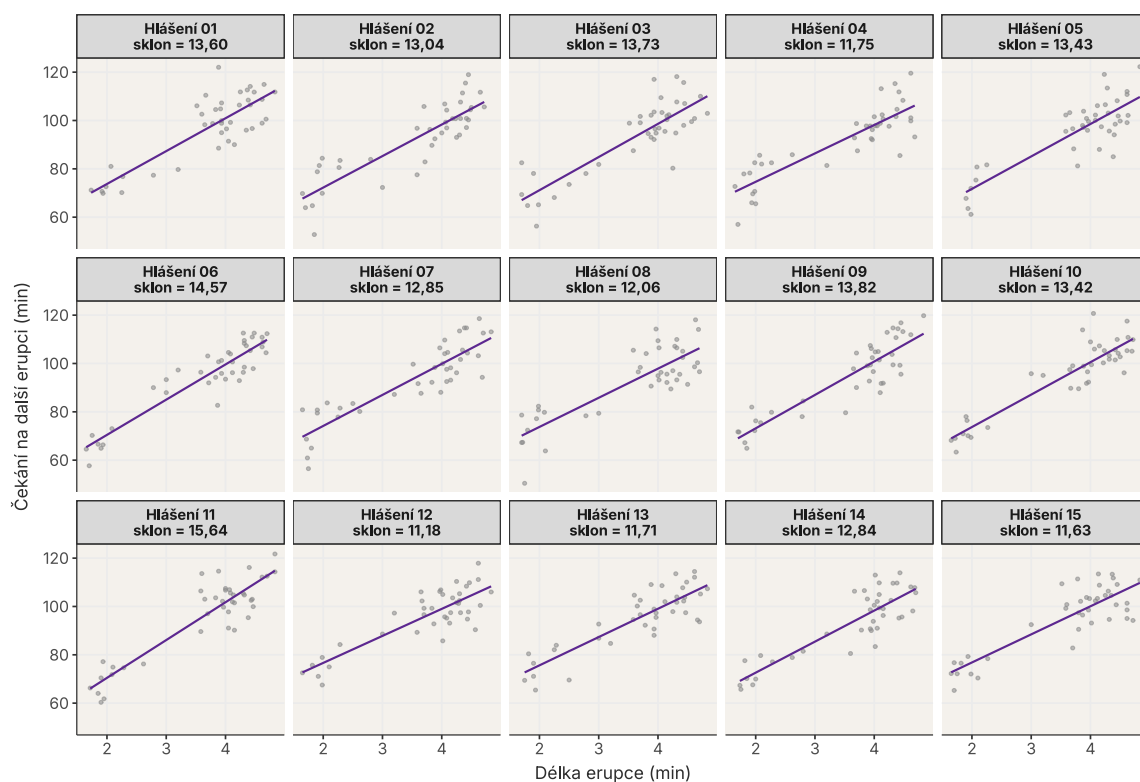
Hlášení jsou **simulovaná**. Délky erupcí v každém hlášení vybereme z našich skutečných erupcí. Čekání k nim vytvoříme podle přímky odhadnuté ze všech erupcí a k přímce přidáme náhodné odchylky, které jsou typicky stejně velké jako residua kolem ní. Každé hlášení tak přináší nová čekání, jako by šlo o nové erupce. Stejný zůstává model, přímka, ze které čekání vytváříme, i typická velikost odchylek. Sklon této přímky (13,34 minuty na minutu) je proto v simulaci známý **referenční sklon**; skutečný sklon všech erupcí Old Faithful neznáme.

Doplňující: proč hlášení simulujeme a co z toho plyne

Sto skutečných nezávislých pozorovatelů nemáme. Simulace vychází z přímky odhadnuté ze všech 117 erupcí. **V této ukázce proto známe „správný“ sklon** — je to sklon 13,34 minuty na minutu, ze kterého jsme čekání vytvořili. Tato **referenční hodnota** platí jen pro simulaci; skutečný sklon všech erupcí Old Faithful neznáme.

Mezi hlášeními se mění vybrané délky erupcí a náhodné odchylky čekání. Stejný zůstává model, referenční přímka i typická velikost odchylek. Proč nepoužíváme přímo výběry ze 117 skutečných erupcí? Všechna hlášení by pak sdílela tytéž erupce, byla by si podobnější než skutečně nové výběry a jejich sklony by se lišily méně, než odpovídá novým pozorováním.

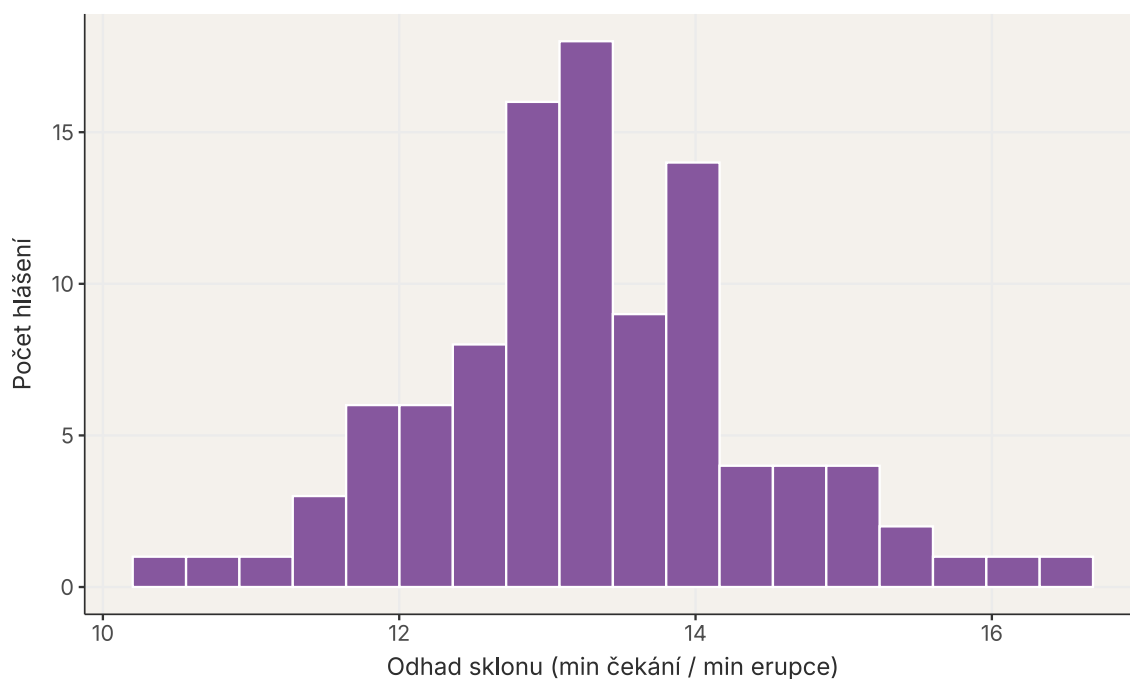
Takto vypadá 15 ze 100 hlášení:



Obrázek 5: Prvních 15 simulovaných hlášení GeyserWatch. Každý panel obsahuje 40 erupcí, vlastní přímkou a vlastní sklon.

Které přímky jsou strmější a které plošší? Dokázali byste z jediného panelu poznat, jak moc by se mohl lišit panel další?

Každý panel opakuje stejný postup, ale body, přímky i sklony se mění. Kdybychom měli pouze jedno hlášení, neviděli bychom, jak široká je množina dalších rozumně možných odhadů. Všechna 100 hlášení nám dává 100 sklonoů a jejich histogram je ukazuje společně:



Obrázek 6: Všechny 100 simulovaných hlášení GeyserWatch vytvořilo 100 odhadů sklonu.

Většina sklonů leží ve společné oblasti, ale nejsou totožné. Odhady z různých výběrů tvoří vlastní rozdělení: některé hodnoty jsou časté, jiné vzácné.

8 Co měří standardní chyba?

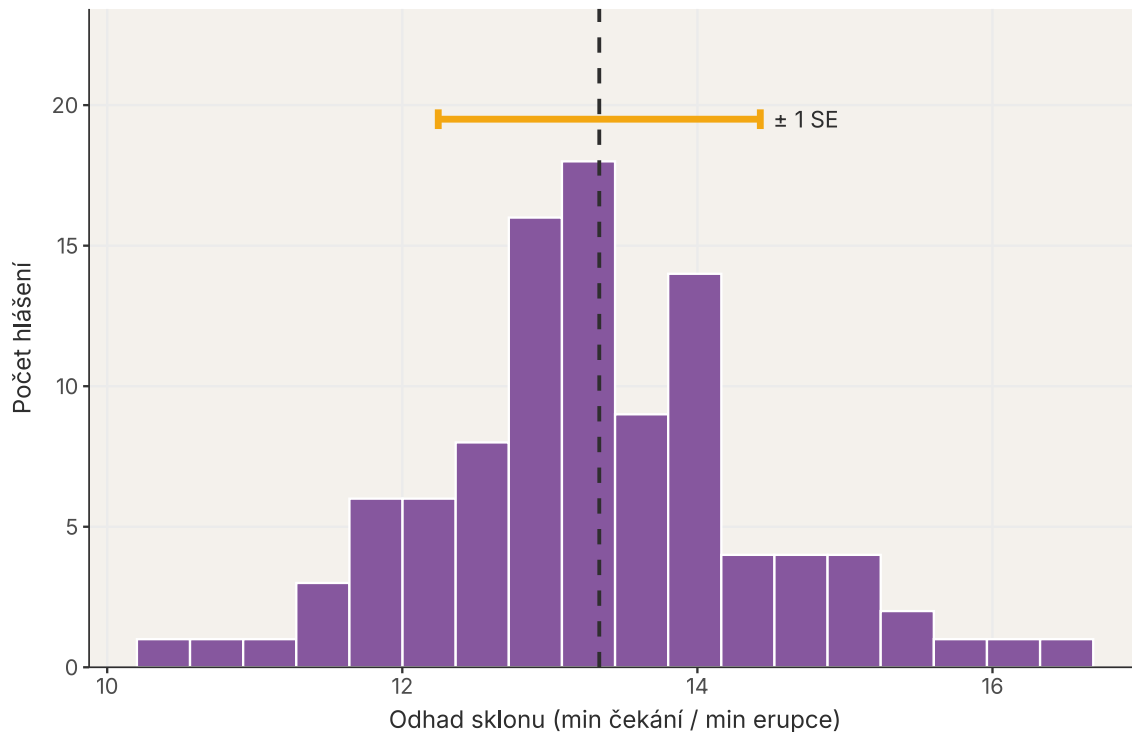
V první lekci jsme odlišili dvě otázky. Směrodatná odchylka (SD) odpovídá na otázku **jak rozdílná jsou pozorování?** Druhá otázka zní **jak přesný je náš odhad?** Histogram sklonů nám na ni dává první odpověď: kdyby se odhady z různých výběrů téměř nelišily, byl by histogram úzký; čím víc se liší, tím je širší.

Šířku histogramu můžeme shrnout stejně jako jakékoli jiné hodnoty, směrodatnou odchylkou. Všechny 100 sklonů z GeyserWatch má směrodatnou odchylku **1,09 minuty čekání na minutu erupce**.

Skutečný výzkumník ale tolik hlášení nemá — má jediné. Přesto mu R tuto velikost odhadne. Z jediného hlášení vypočítá u sklonu číslo, které ve výstupu `summary()` nazývá **Std. Error** (celý výstup si prohlédneme později). Porovnejme je se směrodatnou odchylkou všech sklonů:

Co počítáme	Hodnota
Směrodatná odchylka 100 sklonů (potřebuje všech 100 hlášení)	1,095
Std. Error z jednoho hlášení, typická hodnota (medián)	1,091
Std. Error z jednoho hlášení, rozmezí u prostředních 80 % hlášení	0,91 až 1,33

Obě čísla jsou si velmi blízká, i když nejsou stejná. **Std. Error** je standardní chyba odhadu sklonu (anglicky *standard error*, zkratka **SE**). **Standardní chyba odhaduje, jak moc by se odhady sklonu typicky lišily mezi výběry stejné velikosti – a odhaduje to z jediného výběru.** Má stejné jednotky jako sklon (minuty čekání na minutu délky erupce). Hodnota z jednoho hlášení se od hlášení k hlášení trochu liší, protože i ona je odhadem.



Obrázek 7: Histogram simulovaných sklonů s referenčním sklonem simulace (tmavá čára) a rozsahem plus minus jedna typická standardní chyba (oranžová úsečka).

Do vzdálenosti jedné typické standardní chyby od referenčního sklonu padlo 69 ze 100 sklonů, tedy asi 69 %. Rozsah $\pm$ **jedna standardní chyba** ale ještě není interval spolehlivosti. Je to měřítko přesnosti, ze kterého se interval vytváří.

Jak ale může R poznat šířku histogramu z jediného hlášení, když ostatní hlášení nikdy neviděl?

9 Co dělá sklon stabilnějším?

Šířka histogramu sklonů ale nemusí být vždy stejná. Představme si tři srovnání dvou variant aplikace GeyserWatch. V každém srovnání se obě varianty liší jen jednou věcí:

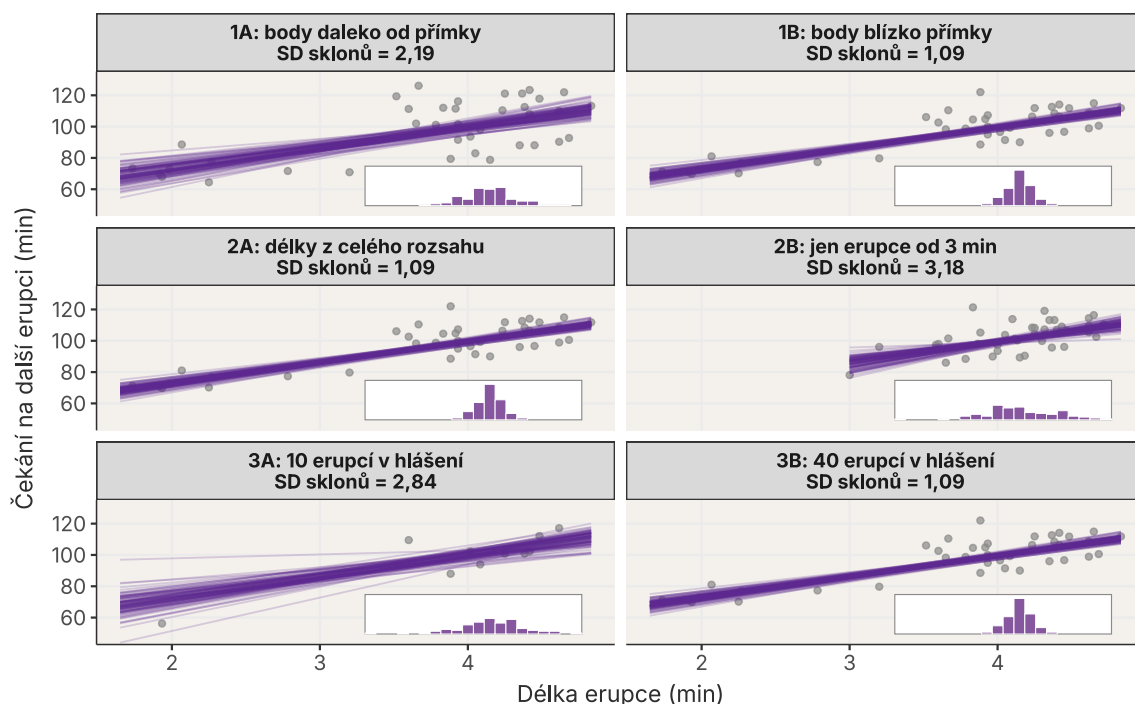
Co čekáte?

Ve které variantě, A, nebo B, se budou sklony jednotlivých hlášení lišit méně?

Srovnání	Co se mění	Varianta A	Varianta B
1	body	daleko od přímky	blízko přímky
2	délky erupcí	z celého rozsahu	jen od 3 min
3	počet erupcí	10	40

Odpověď: 1B, 2A, 3B

Stabilnější sklony dávají **body blízko přímky (1B)**, **délky erupcí z celého rozsahu (2A)** a **více erupcí v hlášení (3B)**. Následující graf to ukazuje na simulovaných hlášeníh.



Obrázek 8: Tři srovnání simulovaných variant aplikace, každá varianta se 100 hlášeními. Fialové přímky patří k jednotlivým hlášením, šedé body ukazují erupce jednoho z nich. V záhlaví panelu je směrodatná odchylka (SD) sklonů všech hlášení, vpravo dole histogram těchto sklonů; jeho vodorovná osa je ve všech panelech stejná, užší histogram tedy znamená menší SD.

Vždy jsme změnili jen jednu věc a vějíř přímek se rozšířil nebo zúžil:

1. **Body daleko od přímky** dávají širší vějíř. Když se čekání kolem přímky hodně rozptylují, může každé hlášení náhodou zachytit body, které přímku víc nakloní. Při dvojnásobném rozptýlení bodů vychází i směrodatná odchylka sklonů zhruba dvojnásobná (2,19 místo 1,09).
2. **Délky erupcí z celého rozsahu** dávají užší vějíř. Přímku si můžeme představit jako houpačku: body daleko od sebe po vodorovné ose ji drží pevně, body natlačené blízko sebe ji nechají snadno naklonit. Když hlášení obsahují jen erupce od 3 minut, je směrodatná odchylka sklonů 3,18 místo 1,09.
3. **Více erupcí v hlášení** dává užší vějíř. Hlášení s 10 erupcemi má směrodatnou odchylku sklonů 2,84, hlášení se 40 erupcemi 1,09.

Tyto tři vlastnosti výzkumník zná i z jediného souboru dat: vidí, jak daleko leží body od přímky, jak široce jsou rozprostřené délky erupcí a kolik erupcí zaznamenal. Právě z nich R standardní chybu počítá.

10 Standardní chyba z jediného výběru

Standardní chyba sklonu spojuje právě tři vlastnosti dat z předchozí kapitoly; budeme jim říkat ingredience standardní chyby:

1. jak daleko leží body od přímky — typická velikost residua;
2. jak široce jsou rozprostřené délky erupcí — jejich směrodatná odchylka;
3. kolik erupcí máme.

První vlastnost stojí v čitateli: větší residua standardní chybu zvětšují. Zbylé dvě stojí ve jmenovateli: širší rozptýlení délek a více erupcí ji zmenšují.

$$\text{standardní chyba sklonu} = \frac{\text{typická velikost residua}}{\text{směrodatná odchylka délek erupcí} \times \sqrt{\text{počet erupcí} - 1}}.$$

Odečtení jedničky pod odmocninou souvisí s tím, jak se počítá směrodatná odchylka (vysvětlíme níže u obvyklého zápisu vzorce). Při desítkách erupcí je to prakticky odmocnina z jejich počtu.

10.1 Typická velikost residua

Residua známe z předchozí lekce: jsou to svislé odchylky erupcí od přímky. Jejich typickou velikost shrneme podobně jako směrodatnou odchylku. Residua umocníme a sečteme. Tento součet čtverců residuí (anglicky *residual sum of squares*, **RSS**) známe z minulé lekce, kde jsme ho značili SSE: funkce `lm()` vybírá přímku právě tak, aby byl co nejmenší. Z něj odhadneme typickou velikost residua. Součet vydělíme počtem erupcí zmenšeným o 2 a odmocníme, abychom se vrátili k minutám čekání:

$$\text{residuální směrodatná odchylka} = \sqrt{\frac{\text{součet čtverců residuí}}{\text{počet erupcí} - 2}}.$$

V R to spočítáme takto:

```

# Residua: svislé odchylky erupcí od přímky (v minutách)
vec_residua <-
  resid(object = mod_gejzir)

# Počet erupcí v modelu
pocet_erupci <-
  nrow(x = data_gejzir)

# Součet čtverců residuí vydělíme počtem erupcí minus 2 a odmocníme
rezidualni_sd <-
  sqrt(x = sum(vec_residua^2) / (pocet_erupci - 2))

rezidualni_sd

```

```
[1] 6.832579
```

Výsledek se nazývá residuální směrodatná odchylka (ve výstupu R *Residual standard error*): erupce leží typicky asi 6,83 minuty nad přímkou nebo pod ní. V R ji vrátí funkce `sigma()`. Výraz počet erupcí minus 2 se nazývá stupně volnosti (anglicky *degrees of freedom*); proč se odečítá právě 2, vysvětlíme v následující lekci. Po dosazení zaokrouhlených hodnot:

$$\text{residuální směrodatná odchylka} = \sqrt{\frac{5368,7}{115}} = \sqrt{46,68} = 6,83 \text{ min} .$$

Obecně zapisujeme:

$$\hat{\sigma} = \sqrt{\frac{\text{RSS}}{n - 2}}.$$

Zde $\hat{\sigma}$ označuje residuální směrodatnou odchylku, RSS součet čtverců residuí a n počet erupcí.

10.2 Rozptýlení délek a počet erupcí

Zbylé dvě ingredience spočítáme funkcemi, které známe z první lekce:

```
# Jak široce jsou rozprostřené délky erupcí
sd_delky <-
  sd(x = data_gejzir$delka_erupce_min)

sd_delky
```

```
[1] 0.9649201
```

```
# Standardní chyba sklonu ze tří ingrediencí
standardni_chyba_rucne <-
  rezidualni_sd / (sd_delky * sqrt(x = pocet_erupci - 1))

standardni_chyba_rucne
```

```
[1] 0.6574524
```

Po dosazení zaokrouhlených hodnot dostaneme:

$$\begin{aligned}\text{standardní chyba sklonu} &= \frac{6,83}{0,965 \times \sqrt{117 - 1}} \\ &= \frac{6,83}{0,965 \times 10,77} = \frac{6,83}{10,39} = 0,66 .\end{aligned}$$

Standardní chyba sklonu ze všech 117 erupcí je 0,66 minuty čekání na minutu erupce. Stejně číslo vypisuje i funkce `summary()`.

10.3 Souhlasí vzorec se simulací?

Ve vzorci můžeme dosadit i počet erupcí jednoho hlášení GeyserWatch. Kdyby hlášení mělo stejně rozptýlená residua i délky erupcí jako naše data a obsahovalo 40 erupcí, vyšla by standardní chyba 1,13. Směrodatná odchylka sta simulovaných sklonů byla 1,09. Vzorec tedy z jediného výběru odhadne šířku histogramu, který bychom jinak museli vytvořit z mnoha hlášení.

Vzorec také vysvětluje, co jsme viděli ve třech srovnáních. Dvojnásobná typická velikost residua zdvojnásobí standardní chybu. Délky erupcí rozprostřené jen v úzkém rozsahu mají malou směrodatnou odchylku, a standardní chyba proto roste. Počet erupcí stojí ve jmenovateli pod odmocninou: **standardní chyba klesá s odmocninou z počtu pozorování**, ne s počtem samotným. Čtyřnásobek erupcí ji proto zmenší jen na polovinu, ne na čtvrtinu.

Ve třetím srovnání vyšel rozdíl o něco větší: směrodatná odchylka sklonů 2,84 u deseti erupcí proti 1,09 u čtyřiceti. Vzorec se stejným rozprostřením délek dává pro deset erupcí 2,36. Hlášení s deseti

erupcemi ale mívají rozprostření délek velmi různé: někdy všechny délky padnou do jednoho shluku a přímka se pak opírá o úzký rozsah. Vzorec s rozprostřením délek z našich dat proto tento rozdíl trochu podcení; standardní chyba každého jednotlivého hlášení ale vychází z jeho vlastních délek.

10.4 Obvyklý zápis vzorce

V učebnicích se tentýž vzorec obvykle zapisuje takto:

$$\text{SE}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = (n-1) s_x^2.$$

Zde $\hat{\beta}_1$ označuje odhad sklonu, $\text{SE}(\hat{\beta}_1)$ jeho standardní chybu, $\hat{\sigma}$ residuální směrodatnou odchylku, n počet erupcí, x_i délku i -té erupce, $\bar{x}$ průměrnou délku a s_x směrodatnou odchylku délek. S_{xx} je součet čtverců odchylek délek erupcí od jejich průměru. Protože se směrodatná odchylka počítá z tohoto součtu vyděleného $n-1$, platí $S_{xx} = (n-1) s_x^2$ a oba zápisy dávají stejné číslo.

11 Kde ve výstupu najdeme velikost a přesnost?

Funkce `summary()` vypíše pro model ze všech 117 erupcí odhady i jejich standardní chyby najednou. Celý výstup vypadá tak, jak jej R skutečně vypisuje:

```
# Souhrn obsahuje odhad sklonu i jeho standardní chybu
summary(object = mod_gejzir)
```

```
Call:
lm(formula = cekani_min ~ delka_erupce_min, data = data_gejzir)

Residuals:
    Min       1Q   Median       3Q      Max
-13.0280  -4.4756  -0.3648   3.6354  21.0827

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    45.9122     2.4658  18.62  <2e-16 ***
delka_erupce_min 13.3350     0.6575  20.28  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.833 on 115 degrees of freedom
Multiple R-squared:  0.7815,    Adjusted R-squared:  0.7796
F-statistic: 411.4 on 1 and 115 DF,  p-value: < 2.2e-16
```

Najděte ve výstupu dvě čísla

1. Které číslo popisuje **velikost vztahu** mezi délkou erupce a čekáním?
2. Které číslo popisuje **přesnost tohoto odhadu**?

Odpověď: řádek `delka_erupce_min`, sloupec `Estimate` a `Std. Error`

V tabulce `Coefficients` najdeme řádek `delka_erupce_min`. Sloupec `Estimate` obsahuje odhad sklonu 13,34 – velikost vztahu. Sloupec `Std. Error` obsahuje jeho standardní chybu 0,66 – přesnost odhadu. Obě čísla mají jednotku minuty čekání na minutu erupce.

Výstup můžeme číst shora:

1. `Call` připomíná vzorec modelu `cekani_min ~ delka_erupce_min`: čekání odhadujeme z délky předchozí erupce.
2. `Residuals` shrnují residua, tedy svislé odchylky bodů od přímky z předchozí lekce. Pět čísel (minimum, dolní kvartil, medián, horní kvartil, maximum) čteme stejně jako `summary()` dat v první lekci: residua mají být rozložená kolem nuly zhruba souměrně.
3. V tabulce `Coefficients` obsahuje sloupec `Estimate` stejné odhady, které vrací `coef(mod_gejzir)`. Sloupec `Std. Error` vedle nich ukazuje jejich standardní chyby.
4. Řádek `Residual standard error` je residuální směrodatná odchylka 6,83 z předchozí kapitoly, tedy první ingredience standardní chyby. Za ní následuje počet stupňů volnosti. Přestože se tento řádek v angličtině jmenuje *standard error*, není to standardní chyba odhadu sklonu, ale typická velikost residuí.

Sloupce `t value` a `Pr(>|t|)` a poslední řádky výstupu zatím přeskočíme. Souvisí s posuzováním konkrétních tvrzení o parametrech modelu, kterému se budeme věnovat v následující lekci.

`Std. Error` sklonu je přesně číslo, které jsme spočítali ze tří ingrediencí: 0,66. Je menší než typická standardní chyba jednoho hlášení (1,09), protože model ze všech erupcí vychází z 117 pozorování, hlášení jen ze 40. **Velikost vztahu a přesnost odhadu jsou dvě různé informace** – velký sklon může být odhadnutý nepřesně a malý sklon velmi přesně.

Vzor zápisu: odhad a standardní chyba

„Dvě erupce, jejichž délky se lišily o jednu minutu, se podle lineárního modelu v průměru lišily čekáním na další erupci o 13,34 minuty (odhad sklonu = 13,34, SE = 0,66 minuty čekání na minutu erupce).“

Věta nejprve interpretuje vztah v běžném jazyce a až za ní uvádí kompaktní statistický zápis. Samotná dvojice čísel bez proměnných a jednotek by nestačila.

Doplňující: standardní chyba průměru

Standardní chybu nemá jen sklon. Má ji každý odhad, například i průměr. Průměrné čekání na další erupci je v našich datech 94,26 minuty. Kdyby někdo zaznamenal jiných 117 erupcí, vyšel by mu trochu jiný průměr. Jak moc by se takové průměry typicky lišily, odhadne standardní chyba průměru:

$$\text{standardní chyba průměru} = \frac{\text{směrodatná odchylka čekání}}{\sqrt{\text{počet erupcí}}}.$$

Po dosazení:

$$\text{standardní chyba průměru} = \frac{14,55}{\sqrt{117}} = 1,35 \text{ min}.$$

Obecně zapisujeme $SE(\bar{y}) = s/\sqrt{n}$, kde $\bar{y}$ je průměr, s směrodatná odchylka hodnot a n jejich počet.

Logika je stejná jako u sklonu: větší rozptýlení jednotlivých čekání standardní chybu zvětšuje a více erupcí ji zmenšuje, opět s odmocninou z jejich počtu. Chybí jen rozprostření délek erupcí, protože průměr žádnou přímku nenaklání.

Průměr je ve skutečnosti také lineární model, jen bez prediktoru, tzv. nulový model. Ve vzorci modelu jej zapíšeme jako ~ 1 ; jediný odhadnutý koeficient, (*Intercept*), je pak průměrné čekání:

```
# Standardní chyba průměru podle vzorce
sd(x = data_gejzir$cekani_min) / sqrt(x = nrow(x = data_gejzir))

[1] 1.345608

# Stejný průměr jako model bez prediktoru
mod_prumer <-
  lm(
    formula = cekani_min ~ 1,
    data = data_gejzir
  )

# Odhad a standardní chyba z tabulky Coefficients
summary(object = mod_prumer)$coefficients[, c("Estimate", "Std. Error")]

Estimate Std. Error
94.256410  1.345608
```

Model bez prediktoru vrací stejný průměr i stejnou standardní chybu jako vzorec.

Doplňující: výpočet standardní chyby sklonu krok za krokem

Vzorec z kapitoly „Standardní chyba z jediného výběru“ můžeme zpomalit tak, aby každý krok šel spočítat na kalkulačce. Použijeme malý soubor prvních 5 erupcí. Všechny hodnoty v tabulkách níže jsou zaokrouhlené na dvě desetinná místa a v každém kroku počítáme s nimi. První tabulka ukazuje, jak daleko leží jednotlivé erupce od přímky odhadnuté z těchto pěti erupcí:

Erupce	Čekání	Odhadnuté čekání	Residuum	Čtverec residua
1	68	67,78	0,22	0,05
2	102	99,85	2,15	4,62
3	101	98,74	2,26	5,11
4	94	98,52	-4,52	20,43
5	67	67,11	-0,11	0,01

Sloupec **Residuum** ukazuje residuum (anglicky *residual*), tedy svistou odchylku bodu od přímky odhadnuté z těchto pěti erupcí. Slovo „chyba“ by zde bylo méně přesné, protože skutečnou chybu měření neznáme.

Standardní chyba sklonu porovnává dvě informace: jak daleko leží body od přímky a jak široce jsou rozprostřené délky erupcí kolem svého průměru.

Krok 1: součet čtverců residuí (RSS). Druhé mocniny všech residuí sečteme.

$$\text{součet čtverců residuí} = (\text{residuum erupce 1})^2 + \dots + (\text{residuum poslední erupce})^2$$

Dosadíme residua ze sloupce **Residuum**, umocníme je (sloupec **Čtverec residua**) a sečteme:

$$\begin{aligned}\text{součet čtverců residuí} &= (0,22)^2 + (2,15)^2 + (2,26)^2 \\ &\quad + (-4,52)^2 + (-0,11)^2 \\ &= 0,05 + 4,62 + 5,11 + 20,43 + 0,01 \\ &= 30,22 .\end{aligned}$$

Tentýž výpočet můžeme zapsat kratším vzorcem:

$$\text{RSS} = \sum_{i=1}^n e_i^2.$$

Zde e_i označuje residuum i -té erupce, i postupně prochází všechna pozorování a n je jejich počet. Značka $\sum$ znamená „sečti následující hodnotu pro všechna pozorování“.

Krok 2: residuální směrodatná odchylka. Součet čtverců vydělíme počtem erupcí minus 2 a odmocníme, čímž se vrátíme k minutám čekání.

$$\text{residuální směrodatná odchylka} = \sqrt{\frac{\text{součet čtverců residuí}}{\text{počet erupcí} - 2}}.$$

$$\begin{aligned}\text{residuální směrodatná odchylka} &= \sqrt{\frac{30,22}{5 - 2}} \\ &= \sqrt{10,07} \\ &= 3,17 .\end{aligned}$$

Obecně zapisujeme:

$$\hat{\sigma} = \sqrt{\frac{\text{RSS}}{n - 2}}.$$

Zde $\hat{\sigma}$ označuje residuální směrodatnou odchylku, RSS právě vypočtený součet čtverců residuí a n počet erupcí; $n - 2$ jsou stupně volnosti.

Krok 3: součet čtverců odchylek délek erupcí. Druhou informací je, jak daleko jsou délky erupcí od své průměrné hodnoty 3,39 minuty. V tabulce jsou délky i jejich odchylky v minutách erupce, jejich čtverce v min^2 .

Erupce	Délka erupce	Odchylka od průměru	Čtverec odchylky
1	1,98	-1,41	1,99
2	4,40	1,01	1,02
3	4,32	0,93	0,86
4	4,30	0,91	0,83
5	1,93	-1,46	2,13

$$\begin{aligned}\text{součet čtverců odchylek délek} &= (\text{délka 1} - \text{průměr})^2 + \dots \\ &\quad + (\text{délka 5} - \text{průměr})^2.\end{aligned}$$

Dosadíme odchylky ze sloupce **Odchylka od průměru** a jejich čtverce ze sloupce **Čtverec odchylky**:

$$\begin{aligned}\text{součet čtverců odchylek délek erupcí} &= (-1,41)^2 + (1,01)^2 + (0,93)^2 \\ &\quad + (0,91)^2 + (-1,46)^2 \\ &= 1,99 + 1,02 + 0,86 + 0,83 + 2,13 \\ &= 6,83 .\end{aligned}$$

Obecně zapisujeme:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Zde x_i označuje délku i -té erupce, $\bar{x}$ průměrnou délku a S_{xx} součet čtverců odchylek délek erupcí od průměru (anglicky *sum of squares of x*). Čím rozmanitější délky erupcí máme a čím víc jich je, tím větší je S_{xx} .

Krok 4: standardní chyba sklonu. Standardní chyba sklonu spojí obě informace: je to podíl residuální směrodatné odchylky a odmocniny ze součtu čtverců odchylek délek erupcí.

$$\text{standardní chyba sklonu} = \frac{\text{residuální směrodatná odchylka}}{\sqrt{\text{součet čtverců odchylek délek erupcí}}}.$$

$$\begin{aligned} \text{standardní chyba sklonu} &= \frac{3,17}{\sqrt{6,83}} \\ &= \frac{3,17}{2,61} \\ &= 1,21. \end{aligned}$$

R počítá s nezaokrouhlenými hodnotami a pro těchto pět erupcí vypíše 1,22; případný rozdíl v poslední číslici způsobuje jen zaokrouhlování.

Je to tentýž vzorec $SE(\hat{\beta}_1) = \hat{\sigma} / \sqrt{S_{xx}}$ jako v kapitole o standardní chybě z jediného výběru, jen dosazený pro pět erupcí. Pět erupcí dává mnohem větší standardní chybu než všech 117 erupcí (0,66): S_{xx} má jen pět členů.

11.1 Co říká standardní chyba o našem sklonu?

Pro otázku, o kolik déle čekáme po delší erupci, potřebujeme z tabulky `Coefficients` modelu ze všech 117 erupcí jen řádek `delka_erupce_min` a jeho první dva sloupce:

```
# Odhad sklonu a jeho standardní chyba z tabulky Coefficients
summary(object = mod_gejzir)$coefficients[
  "delka_erupce_min",
  c("Estimate", "Std. Error")
]
```

```
Estimate Std. Error
13.3350280  0.6574524
```

Obě čísla čteme společně a obě mají jednotku minuty čekání na minutu erupce:

- **Estimate** = 13,34 je velikost vztahu: dvě erupce, jejichž délky se liší o jednu minutu, se podle modelu v průměru liší čekáním na další erupci o 13,34 minuty.

- **Std. Error = 0,66** je přesnost tohoto odhadu. Kdyby někdo za stejných podmínek zaznamenal jiný výběr 117 erupcí a odhadl stejný model, dostal by jiný odhad sklonu. Takové odhady by typicky kolísaly zhruba o 0,66 minuty kolem skutečného sklonu – stejně jako kolísaly sklony hlášení GeyserWatch, jen méně, protože jedno hlášení mělo jen 40 erupcí (typická standardní chyba 1,09).

Ve srovnání s velikostí sklonu je standardní chyba malá: odpovídá asi 5 % odhadu. Náš sklon je tedy odhadnutý poměrně přesně, i když 13,34 není jediná hodnota, kterou data připouštějí. Vlastní standardní chybu má i intercept (2,47 minuty) a čte se stejně; pro naši otázku je ale podstatný sklon.

Kolik erupcí na poloviční standardní chybu?

Kolik erupcí bychom museli zaznamenat, aby standardní chyba sklonu klesla na polovinu, tedy asi na 0,33? Předpokládejte, že residua i délky erupcí budou rozptýlené stejně jako dosud.

Odpověď: čtyřikrát víc

Počet erupcí stojí ve vzorci pod odmocninou. Aby se jmenovatel zdvojnásobil, musí se počet erupcí zhruba zčtyřnásobit: místo 117 erupcí bychom jich potřebovali asi 468. Přesnost se tedy s dalšími daty zlepšuje stále pomaleji.

Doplňující: co standardní chyba nezahrnuje

Standardní chyba popisuje výběrovou variabilitu **za předpokladu, že použitý model a způsob zacházení s pozorováními jsou přiměřené**. Sama automaticky neopraví systematickou chybu měření, nereprezentativní výběr, opakovaná měření vydávaná za nezávislá ani zřetelné zakřivení (nelinearitu) či změnu rozptýlení residuí.

Malá SE tedy neznamena, že studie nemůže být systematicky zavádějící. Můžeme velmi přesně odhadnout vztah v nereprezentativním souboru nebo získat příliš optimistickou SE, pokud jsme přehlédli závislost pozorování. Proto se nejistota z tabulky modelu čte společně s diagnostikou z předchozí lekce a s popisem získání dat. Zda odhad nestojí na jediném pozorování, zkontrolujeme v kapitole „Může jedna erupce změnit sklon?“.

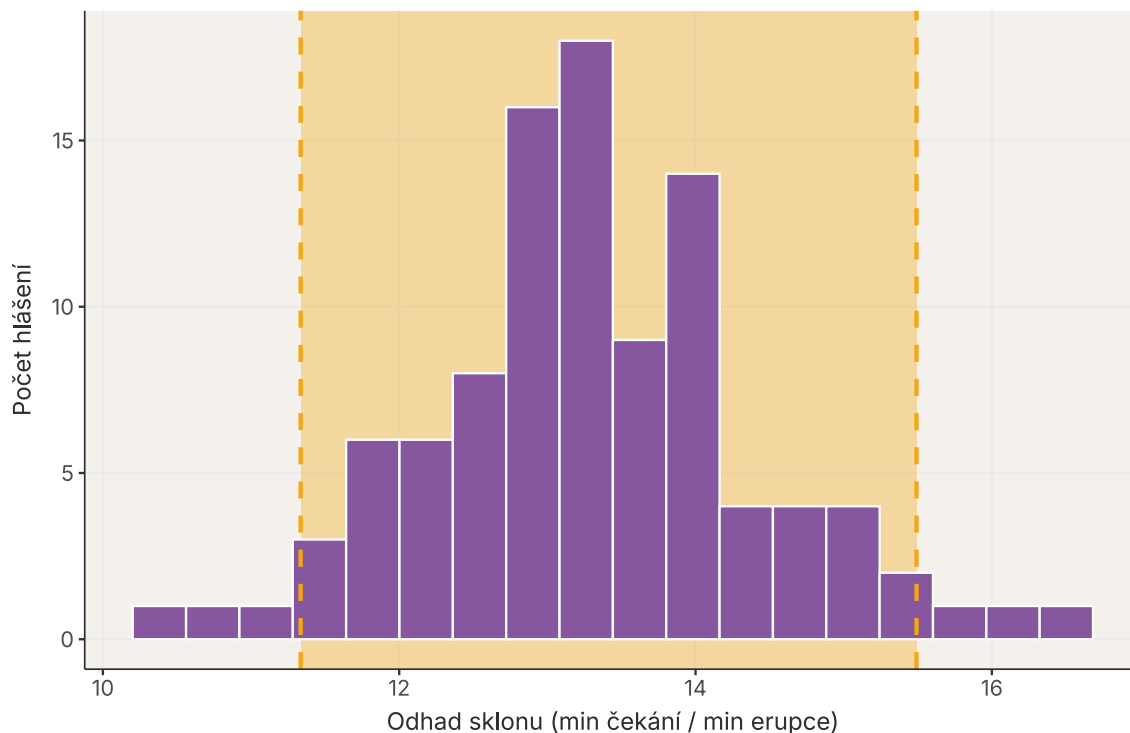
Jaký rozsah sklonů je tedy s našimi daty slučitelný?

12 Jak vyjádřit rozsah nejistoty?

12.1 Připomenutí kvantilů

V první lekci jsme pomocí kvantilů (anglicky *quantiles*) rozdělovali seřazené hodnoty na části. Například 2,5% kvantil odděluje nejnižších 2,5 % hodnot a 97,5% kvantil nejvyšších 2,5 %. Mezi

těmito dvěma hranicemi proto leží prostředních 95 % rozdělení. Na histogramu sta sklonů z GeyserWatch vyznačíme právě tyto dva kvantily:



Obrázek 9: Mezi svislými čarami leží prostředních 95 % simulovaných sklonů.

Kvantily vždy patří k určitému rozdělení hodnot; zde popisují 100 simulovaných sklonů. Skutečný výzkumník ale tolik hlášení nemá. Interval spolehlivosti (anglicky *confidence interval*, **CI**; setkáte se i s označením konfidenční interval) proto vychází z jediného souboru: použije odhad sklonu, jeho standardní chybu a předpoklady modelu, aby vyjádřil rozsah hodnot efektu slučitelných s daty.

12.2 Od odhadu a standardní chyby k intervalu

Interval vytvoříme ze tří čísel: odhadu sklonu, jeho standardní chyby a kritické hodnoty (anglicky *critical value*). Kritická hodnota určuje, kolik standardních chyb přidáme na každou stranu od odhadu. Závisí na zvolené hladině spolehlivosti (anglicky *confidence level*), nejčastěji 95 %. Co toto číslo přesně znamená, ukážeme na simulaci v oddílu „Jak číst hladinu spolehlivosti?“. Pro 95% interval je kritická hodnota blízko čísla 2; její přesná hodnota závisí také na počtu pozorování a podrobně ji vysvětlíme v následující lekci.

1. Vzdálenost od odhadu k hranici intervalu

Vzdálenost každé hranice intervalu od odhadu se nazývá **polovina šířky intervalu** nebo **mez chyby** (anglicky *margin of error*). Samotným hranicím se v odborných textech říká také **meze intervalu**, anglicky *confidence limits*. V této lekci používáme slovo **hranice**.

polovina šířky intervalu = kritická hodnota $\times$ standardní chyba sklonu .

Pro 95% interval je kritická hodnota 1,981. Mezivýsledky zde uvádíme na tři desetinná místa, aby výpočet přesně navazoval:

$$\begin{aligned}\text{polovina šířky intervalu} &= 1,981 \times 0,657 \\ &= 1,302 .\end{aligned}$$

Obecně zapisujeme:

$$M = t_{1-\alpha/2, n-2} \text{ SE } (\hat{\beta}_1).$$

Zde M označuje polovinu šířky intervalu, $t_{1-\alpha/2, n-2}$ kritickou hodnotu a $\text{SE } (\hat{\beta}_1)$ standardní chybu odhadu sklonu. Písmeno α je doplněk hladiny do jedné: $\alpha = 1 - \text{hladina}$, takže pro 95% interval je $\alpha = 0,05$. Písmeno t označuje rozdělení, ze kterého kritická hodnota pochází, a $n - 2$ jsou opět stupně volnosti; obojí vysvětlíme v následující lekci.

2. Dolní hranice

dolní hranice = odhad sklonu – polovina šířky intervalu .

$$\begin{aligned}\text{dolní hranice} &= 13,335 - 1,302 \\ &= 12,033 .\end{aligned}$$

Obecně zapisujeme $L = \hat{\beta}_1 - M$. Zde L označuje dolní hranici, $\hat{\beta}_1$ odhad sklonu a M právě vypočtenou polovinu šířky.

3. Horní hranice

horní hranice = odhad sklonu + polovina šířky intervalu .

$$\begin{aligned}\text{horní hranice} &= 13,335 + 1,302 \\ &= 14,637 .\end{aligned}$$

Obecně zapisujeme $U = \hat{\beta}_1 + M$. Zde U označuje horní hranici.

4. Celý interval

$$95\% \text{ interval} = [\text{dolní hranice}; \text{horní hranice}].$$

Po dosazení obou vypočtených hranic:

$$95\% \text{ interval} = [12,03; 14,64].$$

Celý interval můžeme zapsat kratším vzorcem:

$$I_{1-\alpha} = [L; U] = \hat{\beta}_1 \pm t_{1-\alpha/2, n-2} \text{ SE } (\hat{\beta}_1).$$

Značka $I_{1-\alpha}$ označuje interval se zvolenou hladinou spolehlivosti $1 - \alpha$; pro 95% interval platí $1 - \alpha = 0,95$.

Stejný interval v R

V praxi interval nepočítáme ručně. Funkce `confint()` vypíše intervaly pro všechny koeficienty modelu:

```
# 95% intervaly spolehlivosti interceptu a sklonu  
confint(object = mod_gejzir)
```

```
                2.5 %    97.5 %  
(Intercept)    41.02795  50.79642  
delka_erupce_min 12.03274  14.63731
```

V řádku `delka_erupce_min` najdeme stejné hranice, jaké jsme vypočítali ručně: 12,03 a 14,64 minuty čekání na minutu erupce. Názvy sloupců `2.5 %` a `97.5 %` připomínají, že 95% interval nechává na každé straně 2,5 %.

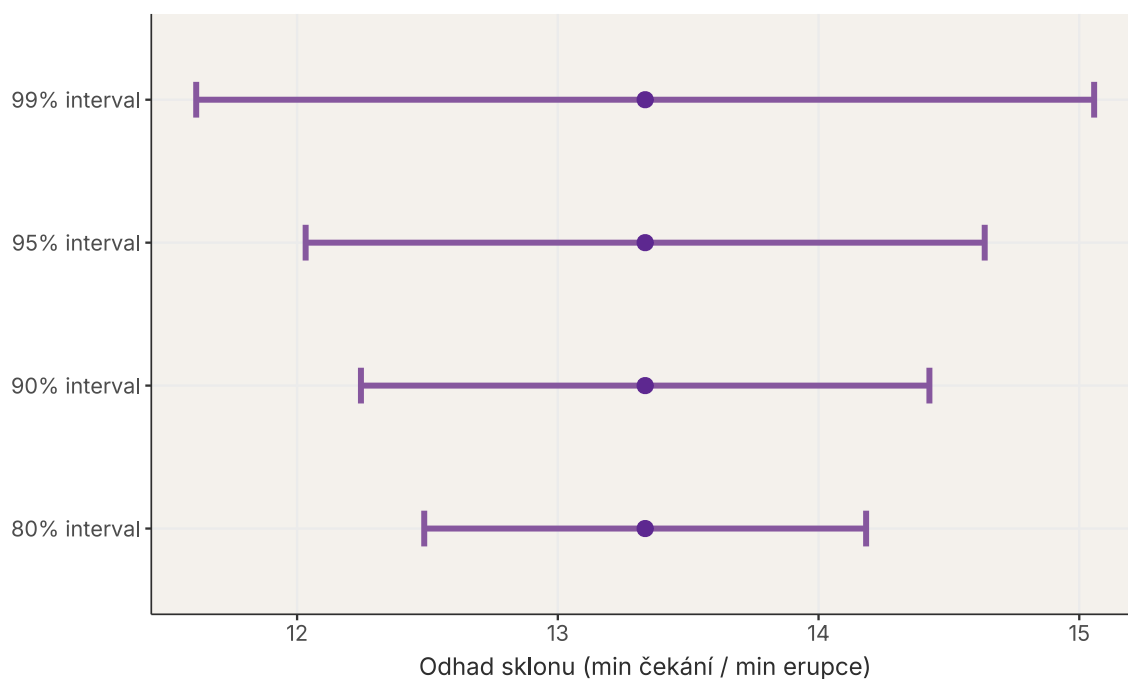
12.3 Jak se interval mění s požadovanou hladinou spolehlivosti?

Hodnota 95 % není jediná možnost. Stejným postupem můžeme vytvořit například 80%, 90% nebo 99% interval; v `confint()` stačí nastavit argument `level`, například `level = 0.80`.

Co čekáte?

Seřaďte 80%, 90%, 95% a 99% interval od nejužšího po nejširší. Musí interval s vyšší hladinou spolehlivosti být širší, nebo užší?

Hladina	Kritická hodnota	Dolní hranice	Odhad	Horní hranice
80 %	1,29	12,49	13,34	14,18
90 %	1,66	12,24	13,34	14,43
95 %	1,98	12,03	13,34	14,64
99 %	2,62	11,61	13,34	15,06



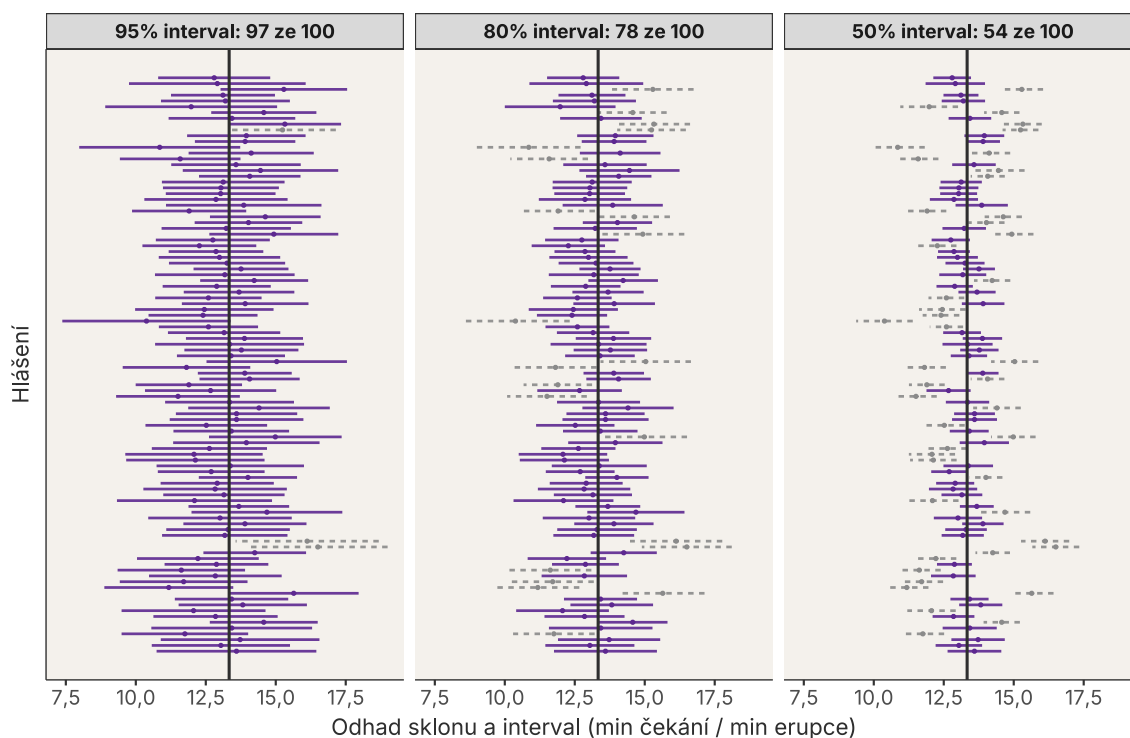
Obrázek 10: Vyšší požadovaná hladina vede k širšímu intervalu kolem stejného odhadu.

Vyšší hladina spolehlivosti používá větší kritickou hodnotu, a proto přidá na každou stranu odhadu více standardních chyb. Proto je 80% interval nejúžší a 99% interval nejširší. Co přesně vyšší hladina spolehlivosti získá, ukáže následující oddíl.

12.4 Jak číst hladinu spolehlivosti?

Vraťme se ke všem 100 hlášením GeyserWatch. Každý návštěvník by ze svého hlášení získal **jeden sklon, jednu standardní chybu a jeden interval** — přesně tak, jak to děláme s našimi daty. Protože hlášení jsou simulovaná, známe referenční sklon 13,34, ze kterého vznikla. Můžeme proto u každého intervalu ověřit, zda referenční sklon zachytil.

Následující trojice panelů používá **stejná hlášení** a mění pouze hladinu spolehlivosti.



Obrázek 11: Stejných 100 simulovaných hlášení s 95%, 80% a 50% intervaly. Plné fialové intervaly zachytily referenční sklon (tmavá čára), šedé přerušované jej minuly.

Z 100 intervalů zachytilo referenční sklon 97 při hladině 95 %, 78 při hladině 80 % a 54 při hladině 50 %. Při dalším opakování simulace by počty vyšly trochu jinak, ale kolísaly by kolem hodnot odpovídajících hladině. Podíl intervalů, které hledanou hodnotu zachytí, se nazývá *pokrytí* (anglicky *coverage*). **Hladina spolehlivosti popisuje, jak často postup při mnoha opakováních zachytí skutečný sklon** – v simulaci referenční sklon, ve skutečném výzkumu neznámý sklon celého procesu. U 95% intervalů jej asi každý dvacátý mine, a výzkumník, který má jediný interval, nepozná, zda je jeho interval mezi nimi.

Je důležité rozlišit dvě věty:

- U kvantilů histogramu jsme vyznačili oblast, ve které leží prostředních 95 % simulovaných **odhadů sklonu**.
- U 95% intervalu spolehlivosti číslo 95 % říká, jaký podíl opakovaně vytvořených **intervalů** dlouhodobě zachytí skutečný sklon.

Jeden vypočtený interval tedy neříká, že uvnitř musí ležet 95 ze 100 sklonů. Neříká ani, že skutečný sklon má po výpočtu 95% pravděpodobnost být uvnitř. Číslo 95 % popisuje úspěšnost postupu při mnoha opakováních.

Nižší hladina spolehlivosti, například 80 %, vytvoří z týchž dat užší interval. To ale neznamená, že jsme z dat získali přesnější informaci: pouze jsme přijali, že postup bude častěji minout.

Užší interval znamená vyšší přesnost pouze tehdy, porovnááme-li intervaly se stejnou hladinou spolehlivosti, například dva 95% intervaly získané z různě informativních dat.

Proč právě 95 %?

Hodnota 95 % není dána přírodním zákonem ani našimi daty. Je to ustálená vědecká konvence a kompromis: vyšší hladina spolehlivosti dává širší interval, nižší užší interval, který ale častěji mine. V následující lekci uvidíme, proč se 95 % často objevuje společně s posuzováním statistických tvrzení. Pro tuto lekci je důležité vědět, že **95 % je standardní volba, nikoli jediná správná volba.**

Co interval spolehlivosti neříká

- **Neobsahuje prostředních 95 % erupcí ani čekacích dob.** Popisuje nejistotu odhadu sklonu, nikoli rozdělení jednotlivých pozorování.
- **Nedokazuje, že některá hodnota je pravdivá nebo nepravdivá.** Ukazuje, které velikosti efektu jsou s daty a modelem slučitelné.
- **Neříká, že skutečný sklon má po výpočtu tohoto intervalu 95% pravděpodobnost ležet uvnitř.** Číslo 95 % patří k dlouhodobému postupu vytváření intervalů.

Vzor zápisu: odhad, standardní chyba a interval

„Dvě erupce, jejichž délky se lišily o jednu minutu, se podle lineárního modelu v průměru lišily čekáním na další erupci o 13,34 minuty (odhad sklonu = 13,34, SE = 0,66, 95% CI [12,03; 14,64] minuty čekání na minutu erupce).“

Tento zápis obsahuje velikost vztahu i jeho přesnost. Stále musí být doprovázen popisem dat, modelu a jednotek.

13 Jak přesně známe celou přímku?

Zatím jsme nejistotu vyjadřovali pro jediné číslo — sklon. Přímka ale dává ještě něco dalšího: pro každou délku erupce odhadnuté průměrné čekání. I výška přímky je tedy odhad, a proto má vlastní nejistotu. To, jak se průměrné čekání mění s délkou erupce, budeme nazývat **průměrný trend**; přímka modelu je jeho odhadem.

Co čekáte?

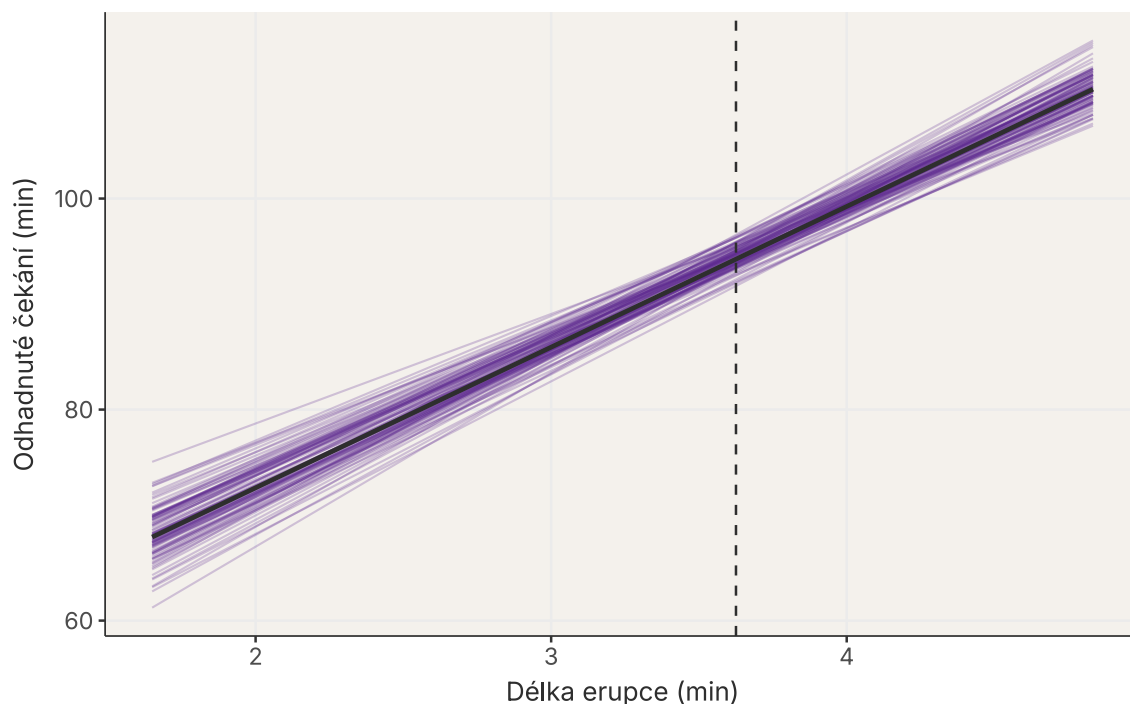
Kdybychom do jednoho grafu nakreslili přímky všech 100 hlášení GeyserWatch, kde by se od sebe lišily nejméně?

- **A:** U krátkých erupcí.
- **B:** U průměrně dlouhých erupcí.
- **C:** U dlouhých erupcí.

Odpověď: u průměrně dlouhých erupcí

Správně je **možnost B**. Přímky se nejméně liší u průměrně dlouhých erupcí a směrem ke krátkým i dlouhým erupcím se rozcházejí. Proč, ukáže následující graf.

13.1 Sto hlášení, sto přímek



Obrázek 12: Přímky všech 100 simulovaných hlášení GeyserWatch (fialově), referenční přímka ze všech erupcí (tmavě) a průměrná délka erupce (přerušovaná čára).

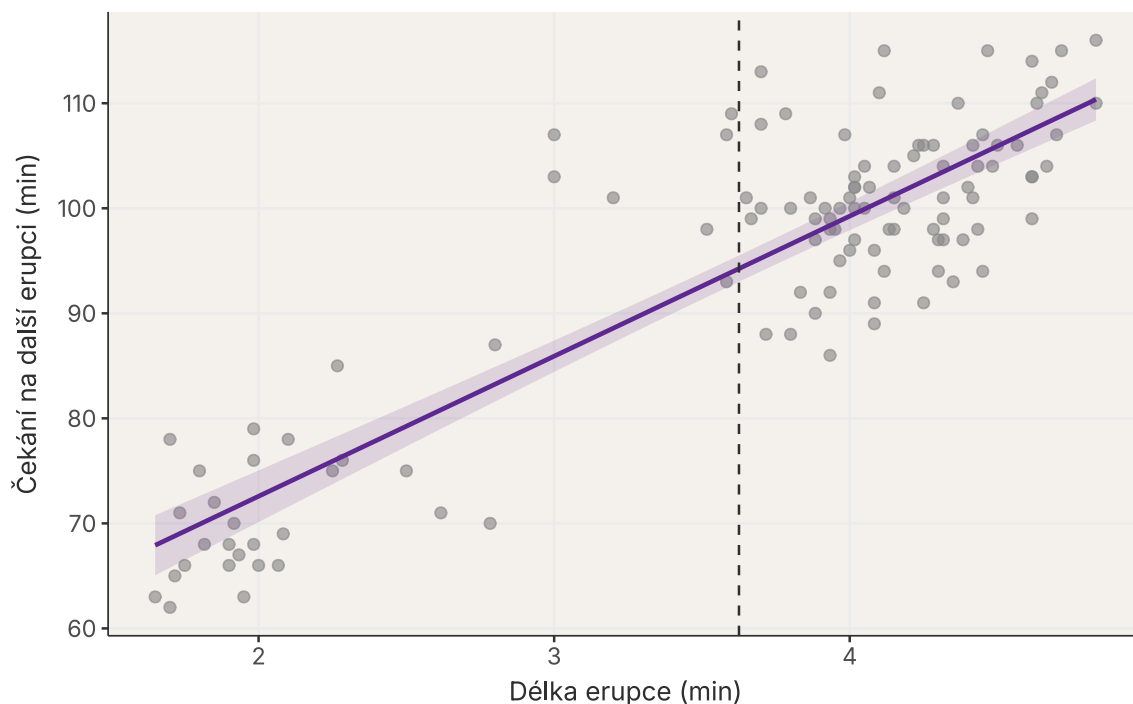
Každá fialová přímka patří k jednomu hlášení. Přímky mají různé sklony, ale všechny procházejí blízko středu dat. Přímka odhadnutá funkcí `lm()` má totiž tu vlastnost, že vždy prochází bodem, jehož souřadnicemi jsou průměrná délka erupce a průměrné čekání v datech, ze kterých byla odhadnuta. Tyto průměry se mezi hlášeními liší jen málo. Přímky s různým sklonem se proto kříží blízko průměrné délky erupce (3,63 minuty) a směrem k okrajům se rozbíhají jako vějíř. Po

erupcích dlouhých 2 minuty odhadují přímky hlášení průměrné čekání od 67,0 do 78,7 minuty, u průměrné délky erupce jen od 91,7 do 96,6 minuty. **Nejistota sklonu se projeví tím víc, čím dál jsme od průměrné délky erupce.**

Každé hlášení obsahuje jen 40 erupcí. Vějíř je proto širší než nejistota přímky odhadnuté ze všech 117 erupcí, stejně jako byla standardní chyba sklonu v jednom hlášení větší než ve výstupu pro všechny erupce.

13.2 Pás spolehlivosti kolem přímky

Skutečný výzkumník sto hlášení nemá. Nejistotu přímky ale umí odhadnout ze svých dat, podobně jako odhadl standardní chybu sklonu: kolem přímky nakreslí pás spolehlivosti. Kód, kterým pás nakreslíte sami, ukážeme o kousek níže.



Obrázek 13: Všechny erupce, přímka modelu a její 95% pás spolehlivosti. Přerušovaná čára označuje průměrnou délku erupce.

Pás spolehlivosti (anglicky *confidence band*) ukazuje pro každou délku erupce 95% interval spolehlivosti pro **průměrné čekání** po erupcích této délky. Skládá se tedy z mnoha intervalů, jednoho pro každou délku erupce. Má podobný tvar jako vějíř přímek: je nejužší u průměrné délky erupce a k okrajům se rozšiřuje. Je ale užší než vějíř, protože vychází ze všech 117 erupcí.

Nejužší místo pásu neleží tam, kde je nejvíc erupcí. Průměrná délka erupce (3,63 minuty) leží v řidší oblasti mezi shluky krátkých a dlouhých erupcí, a pás je přesto právě tam nejužší. Rozhoduje vzdálenost od průměrné délky erupce, ne počet bodů v okolí.

Jeden soubor dat dává jeden pás, stejně jako jeden interval z `confint()`. Číslo 95 % čteme stejně jako u sklonu: popisuje úspěšnost postupu při mnoha opakováních, ne jistotu jednoho nakresleného pásu.

13.3 Kdy je pás široký?

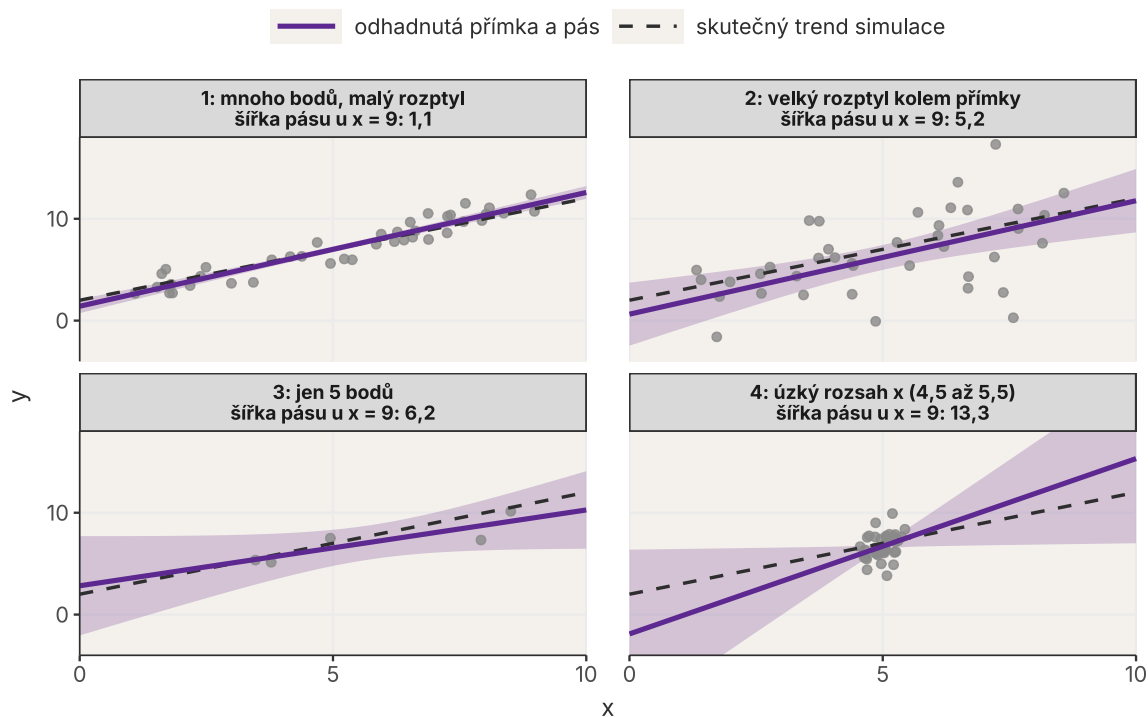
Co čekáte?

Představte si čtyři simulované soubory dat se stejným skutečným trendem. Ve kterém z nich bude pás spolehlivosti u $x = 9$ nejširší?

- **1:** 40 bodů blízko přímky, x od 1 do 9.
- **2:** 40 bodů daleko od přímky, x od 1 do 9.
- **3:** jen 5 bodů blízko přímky, x od 1 do 9.
- **4:** 40 bodů blízko přímky, x jen od 4,5 do 5,5.

Odpověď: soubor 4

Nejširší je pás v souboru 4, přestože jeho body leží blízko přímky. Hodnota $x = 9$ je od nich daleko. Široké jsou ale i pásy v souborech 2 a 3. Všechny čtyři pásy ukazuje následující graf.



Obrázek 14: Čtyři simulované soubory dat se stejným skutečným trendem (přerušovaná čára). Fialová přímka je odhadnutá z bodů, světle fialový pás je její 95% pás spolehlivosti, nakreslený i mimo rozsah bodů. V záhlaví panelu je šířka pásu u $x = 9$.

Šířku pásu určují stejné tři ingredience jako standardní chybu sklonu. Soubor 1 je má všechny příznivé a jeho pás je u $x = 9$ široký jen 1,1. Body daleko od přímky (soubor 2) ho rozšíří na 5,2, pouhých 5 bodů (soubor 3) na 6,2. Ve všech souborech se navíc pás rozšiřuje se vzdáleností od průměru x , stejně jako vějíř přímek.

Soubor 4 ukazuje, co se děje mimo rozsah dat. Mezi $x = 4,5$ a $5,5$, kde body leží, je pás úzký. Kdo by ale přímku prodloužil až k $x = 9$, dostane pás široký 13,3: data tam o poloze přímky skoro nic neříkají. Použití modelu mimo rozsah pozorovaných hodnot se nazývá extrapolace. I široký pás přitom ukazuje jen nejistotu odhadnuté přímky. Nezaručuje, že vztah mimo data vůbec zůstává přímkový.

Naše erupce trvaly od 1,65 do 4,83 minuty. Krátkých i dlouhých erupcí máme dost, a proto je náš pás v celém tomto rozsahu úzký. Odhady průměrného čekání ale platí jen pro erupce v tomto rozsahu.

13.4 Pás v číslech: `predict()`

Interval pro jednu zvolenou délku erupce spočítáme funkcí `predict()` s argumentem `interval = "confidence"`. Zvolíme krátkou erupci, průměrnou délku a dlouhou erupci:

```
# Tři délky erupce: krátká, průměrná a dlouhá
data_tri_delky <-
  data.frame(
    delka_erupce_min = c(2, mean(x = data_gejzir$delka_erupce_min), 4.5)
  )

# Odhad průměrného čekání a jeho 95% interval spolehlivosti
predict(
  object = mod_gejzir,
  newdata = data_tri_delky,
  interval = "confidence"
)
```

	fit	lwr	upr
1	72.58224	70.12340	75.04108
2	94.25641	93.00519	95.50763
3	105.91981	104.22778	107.61184

Sloupec `fit` je odhadnuté průměrné čekání — stejné číslo, jaké leží na přímce. Sloupce `lwr` a `upr` (z anglického *lower* a *upper*) jsou dolní a horní hranice intervalu. Po dvouminutové erupci je interval široký 4,9 minuty, u průměrné délky erupce jen 2,5 minuty a po erupci dlouhé 4,5 minuty 3,4 minuty. Prostřední řádek tedy odpovídá nejužšímu místu pásu.

Když stejný výpočet zopakujeme pro mnoho délek erupce v celém pozorovaném rozsahu a hranice spojíme čarami, dostaneme celý pás. Délky erupce a hranice pásu spočítáme takto:

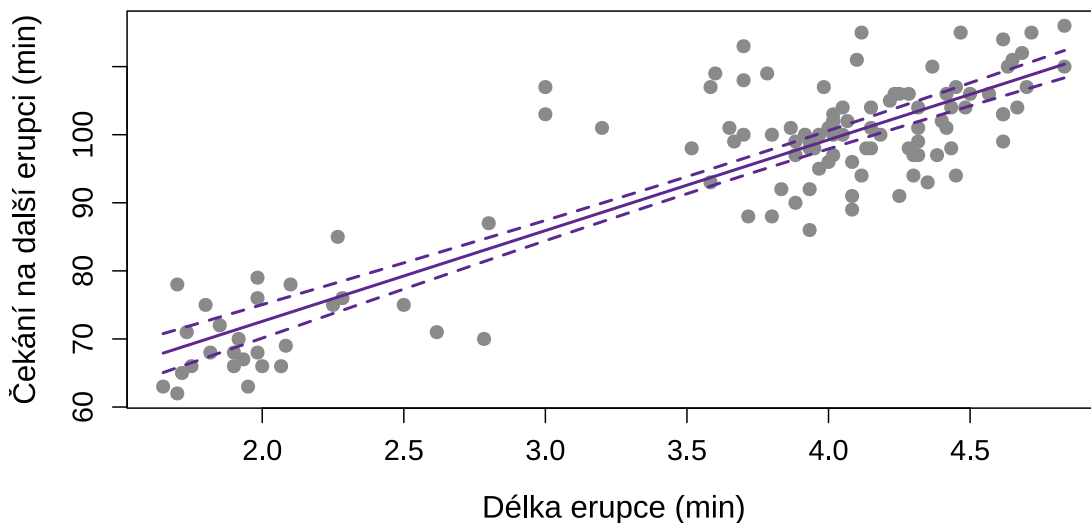
```
# Padesát délek erupce od nejkratší po nejdelší pozorovanou
data_mrizka <-
  data.frame(
    delka_erupce_min = seq(
      from = min(data_gejzir$delka_erupce_min),
      to = max(data_gejzir$delka_erupce_min),
      length.out = 50
    )
  )

# Pro každou délku odhadnuté čekání (fit) a hranice intervalu (lwr, upr)
mat_pas <-
  predict(
    object = mod_gejzir,
    newdata = data_mrizka,
    interval = "confidence"
  )
```

Body, přímku a obě hranice pásu nakreslíme takto:

```
# Body erupcí
plot(
  x = data_gejzir$delka_erupce_min,
  y = data_gejzir$cekani_min,
  xlab = "Délka erupce (min)",
  ylab = "Čekání na další erupci (min)",
  pch = 16,
  col = "#8A8A8A",
  cex = 1.3,
  cex.axis = 1.2,
  cex.lab = 1.3
)

# Odhadnutá přímku plnou čarou, hranice pásu přerušovaně
lines(
  x = data_mrızka$delka_erupce_min,
  y = mat_pas[, "fit"],
  col = "#5D2890",
  lwd = 2
)
lines(
  x = data_mrızka$delka_erupce_min,
  y = mat_pas[, "lwr"],
  col = "#5D2890",
  lty = 2,
  lwd = 2
)
lines(
  x = data_mrızka$delka_erupce_min,
  y = mat_pas[, "upr"],
  col = "#5D2890",
  lty = 2,
  lwd = 2
)
```



Funkce `lines()` přidá do již nakresleného grafu čáru spojující zadané body. Pás nekreslíme mimo pozorovaný rozsah délek erupce: tam model nemá data, o která by se opřel.

Doplňující: kratší alternativa pomocí `geom_smooth()`

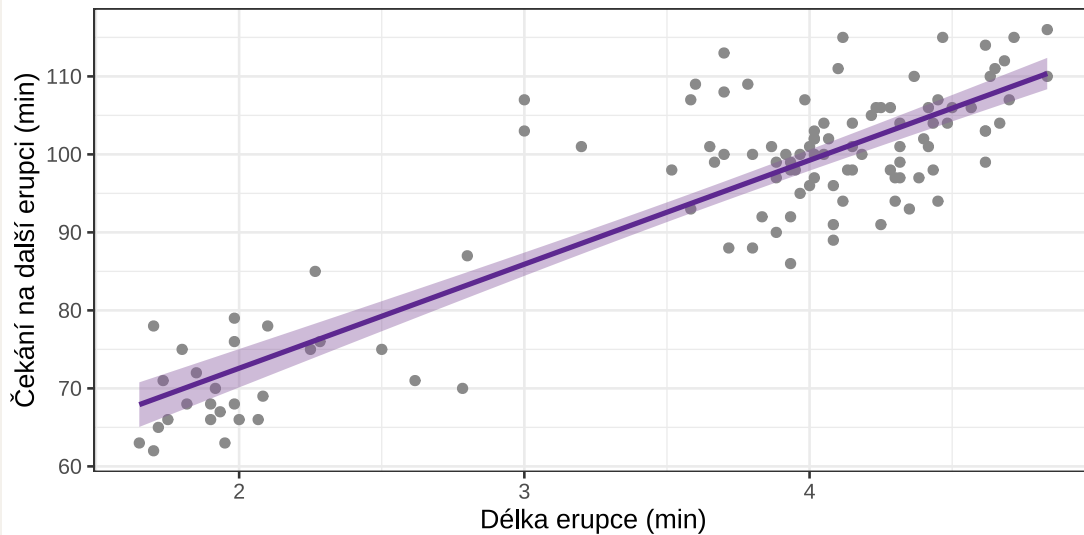
V balíčku `{ggplot2}` nakreslí přímku i pás jediná vrstva `geom_smooth(method = "lm")`. Argument `se = TRUE` přidá pás spolehlivosti, `se = FALSE` nakreslí jen přímku. Zkratka se hodí pro rychlý graf, ale hranice pásu jako čísla nevrátí; ty získáme z `predict()`.

```
# Body erupcí, přímka modelu a 95% pás spolehlivosti
ggplot2::ggplot(
  data = data_gejzir,
  mapping = ggplot2::aes(
    x = delka_erupce_min,
    y = cekani_min
  )
) +
  ggplot2::geom_point(colour = "#8A8A8A") +
  ggplot2::geom_smooth(
    method = "lm",
    formula = y ~ x,
    se = TRUE,
    colour = "#5D2890",
    fill = "#86579E"
```

```

) +
ggplot2::labs(
  x = "Délka erupce (min)",
  y = "Čekání na další erupci (min)"
) +
ggplot2::theme_bw(base_size = 14)

```



13.5 Co pás neříká

Na co si dát pozor

- **Pás neobsahuje 95 % erupcí.** Mimo pás leží 81 % erupcí (95 ze 117). Pás popisuje, kde může ležet průměrný trend, nikoli kde leží jednotlivá čekání. To, že většina bodů leží mimo pás, proto samo o sobě není známkou špatného modelu.
- **Pás neříká, jak dlouho budeme čekat po jedné konkrétní příští erupci.** Jedna erupce se od průměrného trendu odchýlí, stejně jako se od přímky odchylují už pozorované body. Rozsah pro jedno nové čekání musí zahrnout dvě nejistoty současně: nejistotu polohy trendu a rozptýlení jednotlivých čekání kolem něj. Takový rozsah se nazývá predikční interval (anglicky *prediction interval*) a je širší než pás; jeho výpočet zde nepotřebujeme.
- **Pás neověřuje, že přímka je vhodný model.** Platí jen tehdy, když přímka data přiměřeně popisuje (proto kontrolujeme residua), a jen v rozsahu pozorovaných délek erupce.

Vzor zápisu: průměrné čekání a pás spolehlivosti

„Po erupcích dlouhých 4 minuty model odhaduje průměrné čekání na další erupci 99,25 minuty (95% CI [97,91; 100,60] minuty).“

K obrázku: „Plná čára ukazuje přímku lineárního modelu, světlý pás je 95% pás spolehlivosti pro průměrné čekání po erupcích dané délky.“

14 Může jedna erupce změnit sklon?

Pás spolehlivosti se k okrajům rozšiřuje: čím dál od průměrné délky erupce, tím méně přesně známe výšku přímk. Na okrajích rozsahu má přímk. nejmeně opory, a jedna erupce ji tam proto může snáz naklonit. Zatím jsme pracovali se 117 erupcemi, ale při přípravě dat jsme tři erupce vyřadili (viz rámeček o vzniku souboru). Vrátime je zpět, označíme je písmeny A, B a C a podíváme se, co s přímkou udělají:

```
# Tři erupce vyřazené při přípravě dat (hodnoty z archivu GeyserTimes);
# v každém sloupci jsou v pořadí erupce A, B a C
data_vyrazene <-
  data.frame(
    oznaceni = c("A", "B", "C"),
    eruption_id = c(1461337, 1468130, 1471579),
    date = as.Date(x = c("2024-06-04", "2024-07-08", "2024-08-01")),
    cekani_min = c(94, 121, 106),
    duration = c(121, 256, 41)
  )

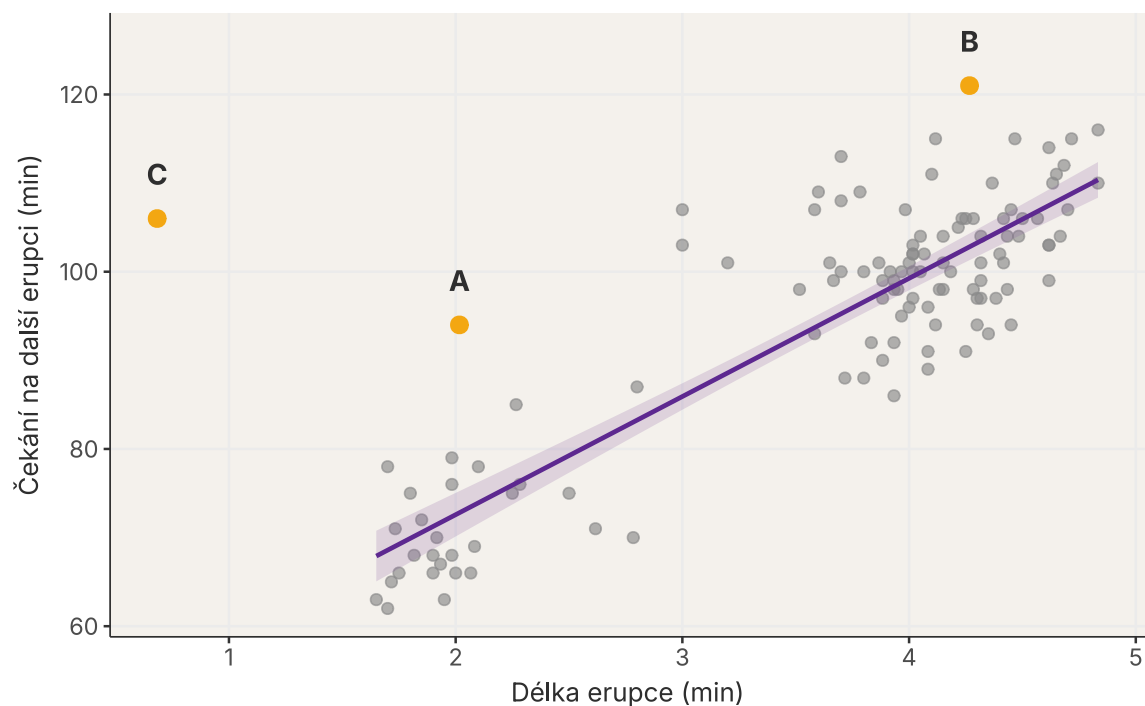
# Délku erupce převedeme ze sekund na minuty
data_vyrazene$delka_erupce_min <-
  data_vyrazene$duration / 60

# Připojíme je za erupce lekce; stačí dva sloupce modelu
vec_sloupce_modelu <- c("delka_erupce_min", "cekani_min")

data_gejzir_120 <-
  rbind(
    data_gejzir[, vec_sloupce_modelu],
    data_vyrazene[, vec_sloupce_modelu]
  )

nrow(x = data_gejzir_120)
```

```
[1] 120
```



Obrázek 15: Erupce lekce (šedě) s přímkou a pásem spolehlivosti odhadnutými z nich a tři vyřazené erupce A, B a C (oranžově).

Co čekáte?

Která z vyřazených erupcí by přímkou po vrácení do dat naklonila nejvíc?

- **A:** Erupce dlouhá asi 2,0 minuty, po které se čekalo 94 minut.
- **B:** Erupce dlouhá asi 4,3 minuty, po které se čekalo 121 minut.
- **C:** Erupce dlouhá jen 41 sekund, po které se čekalo 106 minut.

Odpověď: erupce C

Nejvíc přímku nakloní **erupce C**. Leží nejdál od průměrné délky erupce a zároveň daleko od přímky. Proč právě tato kombinace rozhoduje, ukážou následující dvě kapitoly.

Stejný model odhadneme ze všech 120 erupcí:

```
# Stejný vzorec modelu, data se 120 erupcemi
mod_gejzir_120 <-
  lm(
    formula = cekani_min ~ delka_erupce_min,
    data = data_gejzir_120
  )

coef(object = mod_gejzir_120)
```

```
(Intercept) delka_erupce_min
51.75863      11.91734
```

Tři erupce navíc změnilý sklon z 13,34 na 11,92 minuty čekání na minutu erupce, tedy o 1,42. To je asi 2,2násobek standardní chyby sklonu (0,66). Tři pozorování z 120 tak posunula odhad víc, než kolik odpovídá jeho výběrové variabilitě vyjádřené standardní chybou.

14.1 Krajní erupce mají delší páku

Ve srovnání variant aplikace jsme přímku přirovnali k houpačce: body rozprostřené daleko od sebe ji drží pevně. Platí to i obráceně: jedna erupce daleko od průměrné délky má **delší páku** a může přímku naklonit víc než erupce uprostřed. Tuto vlastnost každého pozorování měří pákový efekt (anglicky *leverage*); mohli jste se s ním setkat už v doplňujícím rámečku minulé lekce. V R jej vrátí funkce `hatvalues()`:

```
# Pákový efekt každé ze 120 erupcí
vec_pakovy_efekt <-
  hatvalues(model = mod_gejzir_120)

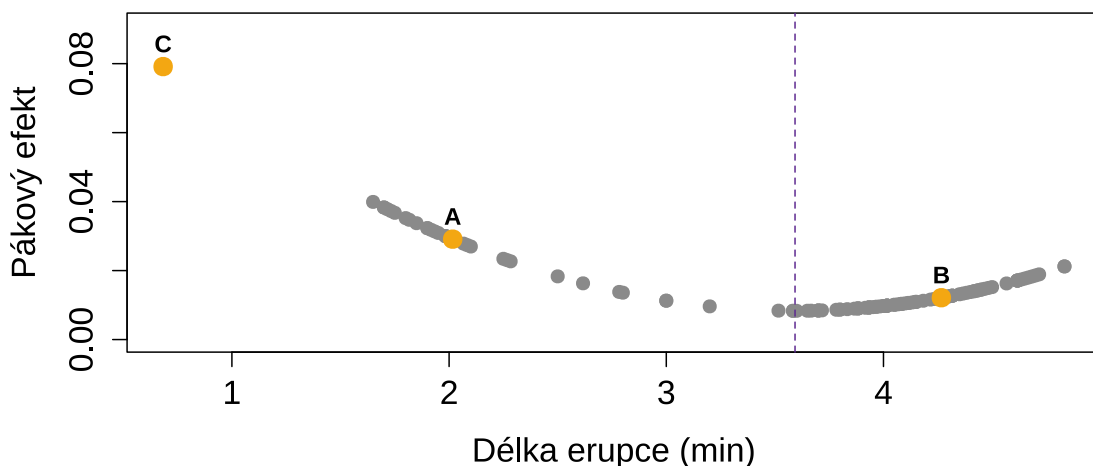
# Pákový efekt erupcí A, B a C (poslední tři řádky)
data.frame(
  oznaceni = data_vyrazene$oznaceni,
  pakovy_efekt = tail(x = vec_pakovy_efekt, n = 3)
)
```

```
   oznaceni pakovy_efekt
118      A    0.02911552
119      B    0.01213255
120      C    0.07915036
```

```
# Rozsah pákového efektu všech erupcí
range(vec_pakovy_efekt)
```

```
[1] 0.00833377 0.07915036
```

```
plot(  
  # Pákový efekt závisí jen na délce erupce  
  x = data_gejzir_120$delka_erupce_min,  
  y = vec_pakovy_efekt,  
  # Místo nad nejvyšším bodem pro jeho popis  
  ylim = c(0, 1.15 * max(vec_pakovy_efekt)),  
  xlab = "Délka erupce (min)",  
  ylab = "Pákový efekt",  
  pch = 16,  
  col = "#8A8A8A",  
  cex = 1.3,  
  cex.axis = 1.4,  
  cex.lab = 1.4  
)  
  
# Tři vrácené erupce zvýrazníme oranžově  
points(  
  x = tail(x = data_gejzir_120$delka_erupce_min, n = 3),  
  y = tail(x = vec_pakovy_efekt, n = 3),  
  pch = 16,  
  col = "#F3A712",  
  cex = 1.8  
)  
  
# Písmena A, B a C nad vrácenými erupcemi  
text(  
  x = tail(x = data_gejzir_120$delka_erupce_min, n = 3),  
  y = tail(x = vec_pakovy_efekt, n = 3),  
  labels = data_vyrazene$oznaceni,  
  pos = 3,  
  font = 2  
)  
  
# Průměrná délka erupce  
abline(  
  v = mean(x = data_gejzir_120$delka_erupce_min),  
  lty = 2,  
  col = "#5D2890"  
)
```

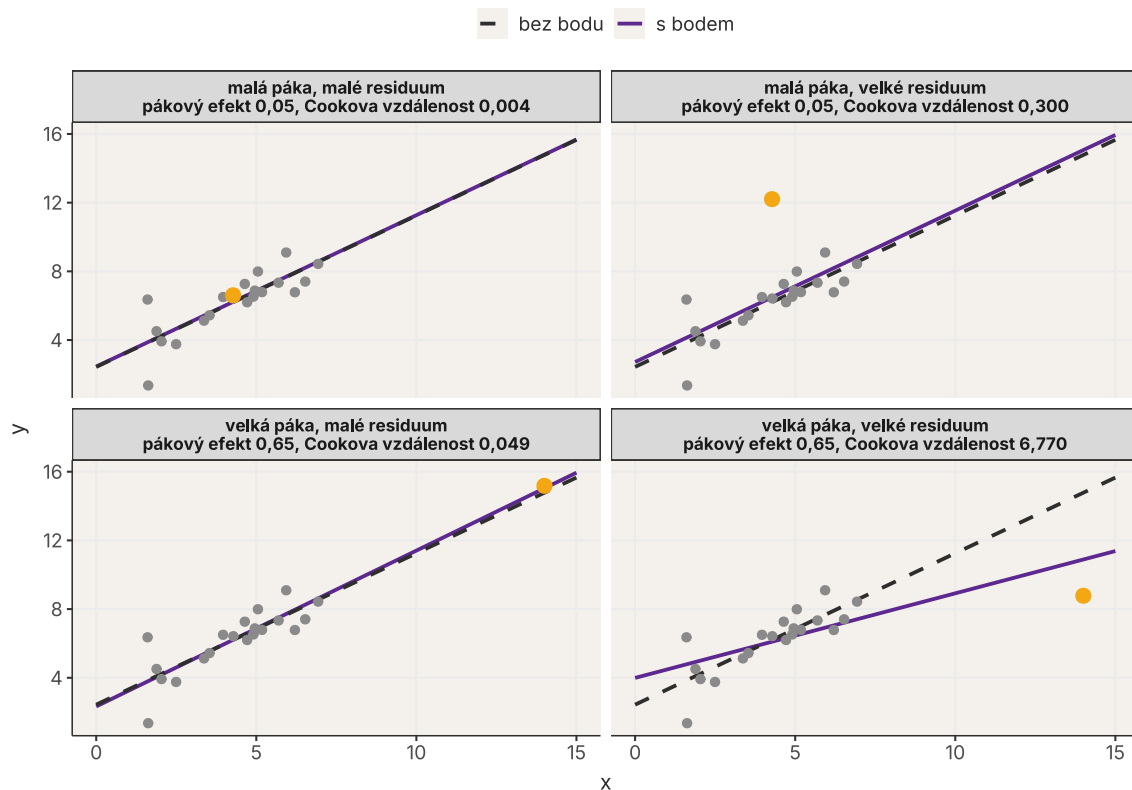


Pákový efekt nemá jednotky. Nejmenší je u erupcí blízko průměrné délky (0,008) a roste na obě strany, stejně jako se rozšiřuje pás spolehlivosti. Vyjadřuje, jak daleko leží délka jedné erupce od průměrné délky, poměřeno rozptřením všech délek, tedy druhou ingrediencí standardní chyby. Čekání na další erupci ani residuum na pákový efekt vliv nemají: erupce C má největší pákový efekt (0,079) jen proto, že trvala nejkratší dobu. Erupce A (0,029) leží dál od průměrné délky než erupce B (0,012).

14.2 Cookova vzdálenost: co kdyby erupce chyběla?

Dlouhá páka sama o sobě přímku nenakloní. Erupce s velkým pákovým efektem, která leží přímo na přímce, by sklon nezměnila. Vliv vzniká teprve tehdy, když se dlouhá páka potká s velkým residuem. Obojí spojuje do jednoho čísla Cookova vzdálenost (anglicky *Cook's distance*). Pro každou erupci vyjadřuje, jak moc by se změnily odhadnuté hodnoty všech erupcí, kdybychom právě tuto erupci vynechali a přímku odhadli znovu.

Jak spolu páka a residuum působí, ukazují simulovaná data na obrázku níže. Ve všech čtyřech panelech jsou stejné šedé body a jeden přidáný oranžový bod, který má malé nebo velké residuum a malý nebo velký pákový efekt. Plná přímká je odhadnutá s přidáním bodem, přerušovaná bez něj.



Obrázek 16: Simulovaná data: stejné šedé body a jeden přidaný oranžový bod se čtyřmi kombinacemi pákového efektu a residua. Plná přímka je odhadnutá s přidaným bodem, přerušovaná bez něj. V záhlaví je pákový efekt a Cookova vzdálenost přidaného bodu.

Bod uprostřed s malým residuem přímku téměř nezmění (Cookova vzdálenost 0,004). Velké residuum uprostřed ji jen posune nahoru (0,3) a bod s dlouhou pákou, který leží blízko přímky, ji jen potvrdí (0,05). Teprve když se dlouhá páka potká s velkým residuem, přímka se výrazně natočí (6,8).

```
# Cookova vzdálenost každé ze 120 erupcí
vec_cookova_vzdalenost <-
  cooks.distance(model = mod_gejzir_120)

# Erupce A, B a C: délka, residuum, pákový efekt a Cookova vzdálenost
data.frame(
  oznaceni = data_vyrazene$oznaceni,
  delka_erupce_min = tail(x = data_gejzir_120$delka_erupce_min, n = 3),
  residuum = tail(x = resid(object = mod_gejzir_120), n = 3),
  pakovy_efekt = tail(x = vec_pakovy_efekt, n = 3),
  cookova_vzdalenost = tail(x = vec_cookova_vzdalenost, n = 3)
)
```

	oznaceni	delka_erupce_min	residuum	pakovy_efekt	cookova_vzdalenost
118	A	2.0166667	18.20806	0.02911552	0.07159345
119	B	4.2666667	18.39403	0.01213255	0.02940799
120	C	0.6833333	46.09785	0.07915036	1.38673477

Erupce A a B leží nad přímkou zhruba stejně daleko (residuum 18,2 a 18,4 minuty). Erupce A je ale dál od průměrné délky, má delší páku, a proto vyšší Cookovu vzdálenost (0,072 proti 0,029). Erupce C spojuje největší pákový efekt s největším residuem a její Cookova vzdálenost (1,39) je mnohonásobně vyšší než u všech ostatních.

Tři pojmy proto držíme odděleně. Odlehlé pozorování má velké residuum, leží daleko od přímky. Pozorování s velkým pákovým efektem leží daleko od průměrné hodnoty prediktoru. **Vlivné pozorování** výrazně mění odhad, a to obvykle tehdy, když se obojí potká. Erupce A a B jsou odlehlé, ale vlivná je jen erupce C.

Cookovy vzdálenosti všech erupcí zobrazíme v pořadí řádků tabulky. Přerušovaná čára ukazuje orientační hodnotu $4/n$, kde n je počet erupcí v modelu:

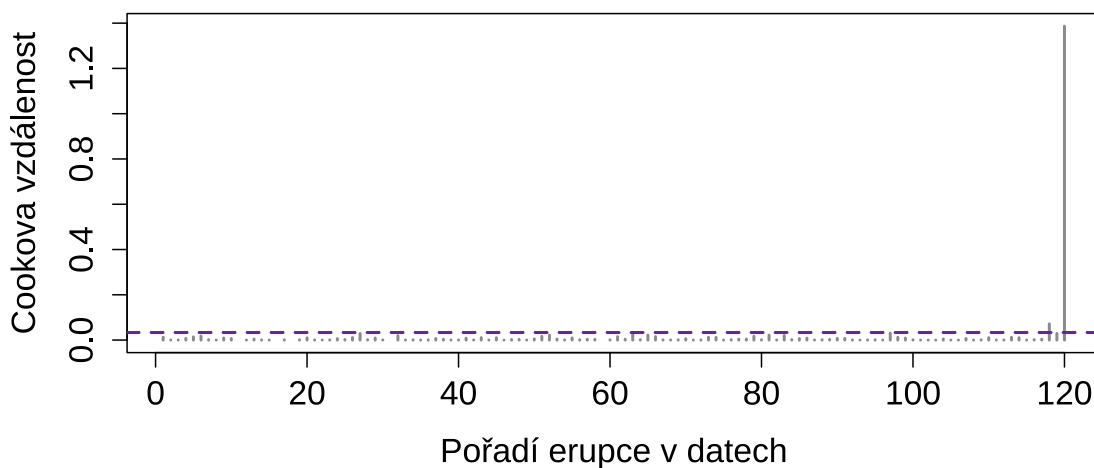
```

# Orientační hodnota 4/n
hranice_cook <- 4 / nrow(x = data_gejzir_120)

plot(
  # Každá svislá čárka představuje jednu erupci
  x = vec_cookova_vzdalenost,
  type = "h",
  lwd = 2,
  col = "#8A8A8A",
  xlab = "Pořadí erupce v datech",
  ylab = "Cookova vzdálenost",
  cex.axis = 1.4,
  cex.lab = 1.4
)

abline(
  # Orientační čára, nikoli hranice pro mazání dat
  h = hranice_cook,
  lty = 2,
  lwd = 2,
  col = "#5D2890"
)

```



Vedle erupce C se všechny ostatní čárky i čára 4/n krčí u nuly. Hodnotu 4/n přesahuje Cookova vzdálenost jen u 2 erupcí: kromě C také u erupce A. Hodnota 4/n není přísná hranice. Označuje

pozorování, na která se vyplatí podívat blíž. Jak velký vliv erupce C má, zjistíme přímo: přímku odhadneme znovu bez ní.

```
# Pořadí erupce s největší Cookovou vzdáleností
index_nejvlivnejsi <-
  which.max(x = vec_cookova_vzdalenost)

index_nejvlivnejsi
```

```
120
120
```

```
# Stejný model bez této jediné erupce
mod_gejzir_bez_c <-
  lm(
    formula = cekani_min ~ delka_erupce_min,
    data = data_gejzir_120[-index_nejvlivnejsi, ]
  )

coef(object = mod_gejzir_bez_c)
```

```
(Intercept) delka_erupce_min
46.96372      13.13583
```

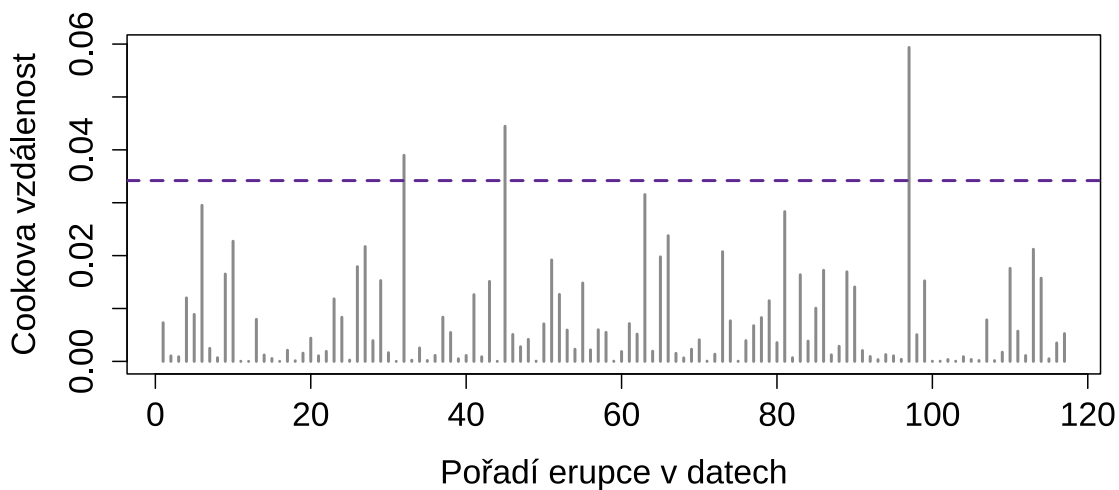
Bez jediné erupce C se sklon změnil z 11,92 na 13,14 minuty čekání na minutu erupce. Tato jedna erupce tedy způsobila většinu rozdílu mezi modelem se 120 a se 117 erupcemi; zbylé dvě erupce sklon změnily už jen o 0,20.

14.3 Mění některá z našich erupcí závěr?

Stejnou kontrolu uděláme pro model, se kterým pracuje celá lekce. Bez erupce C už je čára 4/n v grafu dobře vidět:

```
plot(
  # Cookova vzdálenost erupcí, se kterými pracuje lekce
  x = cooks.distance(model = mod_gejzir),
  type = "h",
  lwd = 2,
  col = "#8A8A8A",
  xlab = "Pořadí erupce v datech",
  ylab = "Cookova vzdálenost",
  cex.axis = 1.4,
  cex.lab = 1.4
)

abline(
  # Orientační hodnota 4/n pro 117 erupcí
  h = 4 / nrow(x = data_gejzir),
  lty = 2,
  lwd = 2,
  col = "#5D2890"
)
```



Funkce `dfbeta()` vrátí pro každou erupci, o kolik by se změnil každý koeficient, kdybychom právě tuto erupci vynechali. Tím se podíváme blíže na všechny erupce najednou:

```
# Největší Cookova vzdálenost a orientační hodnota 4/n pro 117 erupcí  
max(cooks.distance(model = mod_gejzir))
```

```
[1] 0.05936129
```

```
4 / nrow(x = data_gejzir)
```

```
[1] 0.03418803
```

```
# Změna sklonu po vynechání každé jednotlivé erupce  
vec_zmena_sklonu <-  
  dfbeta(model = mod_gejzir)[, "delka_erupce_min"]
```

```
# Největší změna sklonu bez ohledu na směr  
max(abs(x = vec_zmena_sklonu))
```

```
[1] 0.1754164
```

Hodnotu $4/n$ překračuje Cookova vzdálenost u 3 erupcí. Největší z těchto hodnot je ale 0,059, tedy mnohem méně než u erupce C. A hlavně: vynechání kterékoli jednotlivé erupce, včetně těchto 3, změní sklon nejvýš o 0,18 minuty čekání na minutu erupce. To je asi 27 % standardní chyby sklonu (0,66). **Žádná jednotlivá erupce našeho souboru závěr nenese.** Vliv jednoho pozorování proto čteme na stejném měřítku jako nejistotu: změnu sklonu po vynechání erupce porovnáme se standardní chybou.

14.4 Co s vlivným pozorováním?

Vlivné pozorování nejdřív prověříme, nemažeme

- Velká Cookova vzdálenost není důvod pozorování smazat. Je to důvod podívat se na ně blíž.
- O vyřazení rozhoduje znalost dat: chyba záznamu, jiný druh události nebo pozorování mimo proces, který chceme popsat. Samotná vlivnost důvodem není.
- Když si nejsme jistí, uvedeme výsledek s pozorováním i bez něj.

U erupce C máme důvod, který nezávisí na tom, jak nakloní přímku. Trvala jen 41 sekund, tedy méně, než je rozsah 1,5 až 5 minut uváděný National Park Service. Mohlo jít o chybu záznamu nebo o neobvyklou událost mimo běžné erupce, které chceme popsat. Je to stejný druh důvodu, kvůli kterému jsme už při přípravě dat ponechali jen čekání od 50 do 127 minut.

Erupce A a B takový důvod nemají. Odstranili jsme je jen proto, aby výukový příklad vypadal přehledněji. Jak jsme ukázali, tyto dvě erupce sklon téměř neovlivňují, a podle pravidla výše

bychom je po takovém ověření v datech ponechali. Závěry lekce se tím nemění: sklon by se posunul jen o 0,20, mnohem méně, než je jeho standardní chyba.

15 Co můžeme říct o Old Faithful?

Vraťme se k hlavní otázce: **o kolik déle čekáme po delší erupci — a jak jistě to víme?**

V zaznamenaných datech od června do srpna 2024 se dvě erupce, jejichž délky se liší o jednu minutu, podle modelu v průměru liší čekáním na další erupci o 13,34 minuty. To je **velikost vztahu**. 95% interval spolehlivosti sahá od 12,03 do 14,64 minuty čekání na minutu erupce. To je **přesnost odhadu**: data a model jsou slučitelné s tímto rozsahem kladných sklonů, ne pouze s jediným číslem.

Interval se vztahuje ke **sklonu**, tedy k průměrnému rozdílu čekání mezi erupcemi různé délky. Pro zvolenou délku erupce můžeme odhadnout i samotné průměrné čekání: po erupcích dlouhých 4 minuty je to podle modelu 99,25 minuty (95% CI [97,91; 100,60] minuty). Ani tento interval ale neříká, jak dlouho budeme čekat po jedné konkrétní budoucí erupci: ta se od průměrného trendu odchýlí o své vlastní rozptýlení, jak ukázal pás spolehlivosti a body mimo něj.

Odhad přitom nestojí na jediné erupci: vynechání kterékoli z 117 erupcí změni sklon nejvýš o 0,18, tedy asi o 27 % jeho standardní chyby. Erupci dlouhou jen 41 sekund jsme vyřadili z důvodu, který nesouvisí s tím, jak nakloní přímkou: leží mimo rozsah délek, které chceme popsat.

Interpretace stále závisí na použitém modelu a na způsobu získání dat. Dobrovolnická pozorování nejsou náhodný výběr všech erupcí. Rozdělení podle měsíců ukázalo, že se odhad sklonu mezi obdobími liší; ze tří měsíců ale nepoznáme, zda jde jen o výběrovou variabilitu, nebo i o změnu chování gejzíru. Výsledek proto čteme jako opatrný odhad vztahu v zachyceném období, nikoli jako neměnný přírodní zákon.

16 Co bude následovat dál

V této lekci jsme k odhadu sklonu přidali jeho nejistotu: standardní chybu, interval spolehlivosti a pás spolehlivosti kolem přímky. Ověřili jsme také, že odhad nestojí na jediné erupci. V následující lekci se zeptáme, jak pomocí dat posoudit konkrétní tvrzení o parametru modelu — například zda jsou data slučitelná s tím, že délka erupce s čekáním vůbec nesouvisí. Vrátime se přitom ke sloupcům `t value` a `Pr(>|t|)`, které jsme zatím přeskočili.

17 Shrnutí

- Jiný výběr pozorování dává jiný odhad sklonu; tuto proměnlivost mezi výběry nazýváme výběrová variabilita.
- Ve výstupu `summary()` popisuje **Estimate** velikost vztahu a **Std. Error** jeho přesnost. Standardní chyba odhaduje z jediného výběru, jak moc by se odhady typicky lišily mezi výběry stejné velikosti. U sklonu ji zvětšují velká residua a zmenšují ji široce rozprostřené hodnoty prediktoru a více pozorování; klesá s odmocninou z jejich počtu.

- 95% interval spolehlivosti z `confint()` ukazuje rozsah sklonů slučitelných s daty a modelem. Hladina spolehlivosti 95 % popisuje úspěšnost postupu při opakování, nikoli podíl pozorování uvnitř intervalu ani důkaz hypotézy.
- Pás spolehlivosti kolem přímky ukazuje nejistotu průměrného trendu pro každou délku erupce; hranice získáme z `predict()` s `interval = "confidence"`. Je nejužší u průměrné délky erupce a neobsahuje 95 % erupcí.
- Pákový efekt (`hatvalues()`) ukazuje, jak daleko leží pozorování od průměru prediktoru; Cookova vzdálenost (`cooks.distance()`) spojuje pákový efekt s velikostí residua. Změnu sklonu po vynechání pozorování porovnáváme se standardní chybou. Vlivné pozorování prověříme dřív, než o něm rozhodneme; jen kvůli vlivu je nemažeme.
- Vědecký zápis spojuje interpretaci v běžném jazyce s odhadem, standardní chybou, intervalem a jednotkami.

17.1 Závěrečná otázka

Někdo vám ukáže jen titulky „Po minutě erupce navíc čekáme o 13,34 minuty déle“. Co potřebujete vidět, než titulku uvěříte?

Možné otázky

Jak vypadá bodový graf a graf residuí — sedí přímka k datům? Je v grafu kolem přímky pás spolehlivosti, aby bylo vidět, jak přesně známe průměrný trend? Kolik erupcí autor měl, kdy a jak byly zaznamenány? Jak přesný je odhad — jaká je standardní chyba a 95% interval spolehlivosti? Nemění závěr jedna erupce, a vyřadil autor některé erupce, případně proč? Mění se sklon mezi obdobími? A tvrdí titulky jen průměrný rozdíl mezi erupcemi různé délky, nebo i to, že delší erupce čekání způsobuje?

Slovníček pojmů

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

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