Proč ♥ víceúrovňové modely

Aleš Vomáčka

30. 10. 2025

Filozofická fakulta Univerzity Karlovy, STEM

Kdo jsem?

Kdo jsem?

Aleš Vomáčka

Kvantitativní metodologie

Environmentální sociologie

Analytik ve STEM

To jsem já!

Motivace

Motivace

Andrew Gelman

Richard McElreath

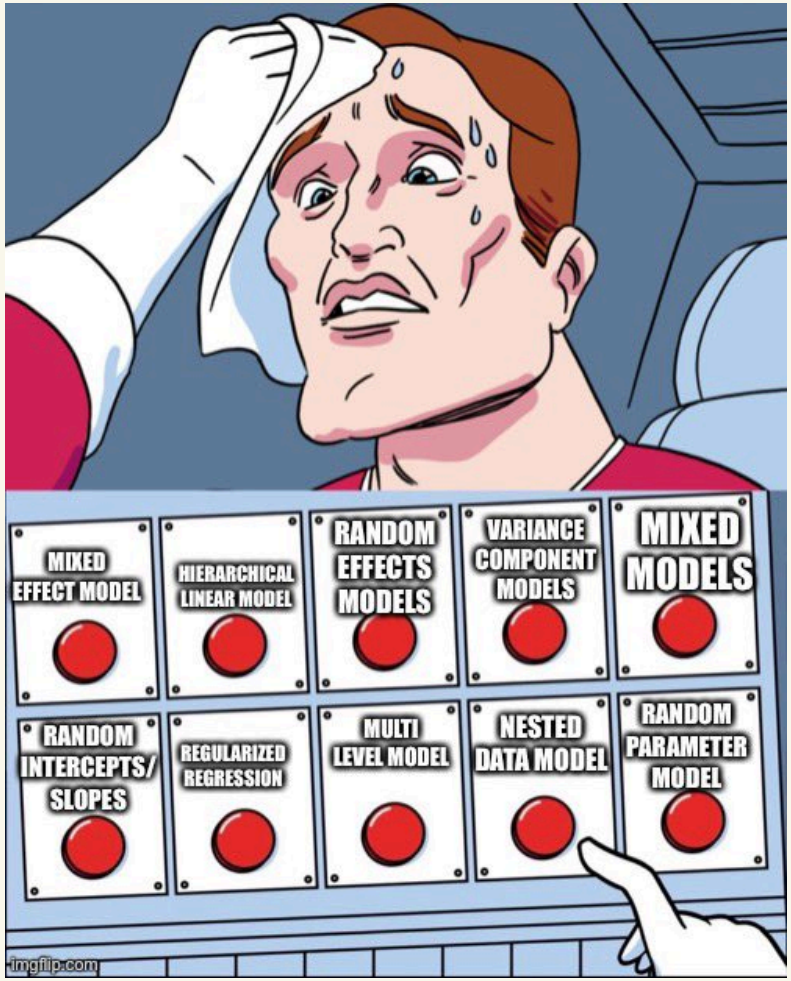
Plán

Plán

1. Snížení nároků na velikost vzorku
2. Meziskupinové a vnitroskupinové efekty
3. Technické detaily a FAQ

Snížení nároků na velikost vzorku

Snížení nároků na velikost vzorku



Zdroj: Chelsea Parlett-Pelleriti

Snížení nároků na velikost vzorku

Modelový příklad:

Naším cílem je odhadnout podíl obyvatel ČR, kteří byli v posledním roce obětí kriminální činnosti.

Data pochází z reprezentativního dotazníkového šetření s 500 respondenty.

Zkušenost s kriminální činností měřená jako binární proměnná.

Snížení nároků na velikost vzorku

Odhad pro celou republiku je přímočarý:

```
glm(crime_experience ~ 1, family = binomial) # logistický model
```

Výsledek: $47\%,\ CI_{95}(42\%; 51\%)$

Snížení nároků na velikost vzorku

Co když ale chceme odhad pro každý kraj?

Problém: Musíme balancovat systematickou a náhodnou chybu odhadu

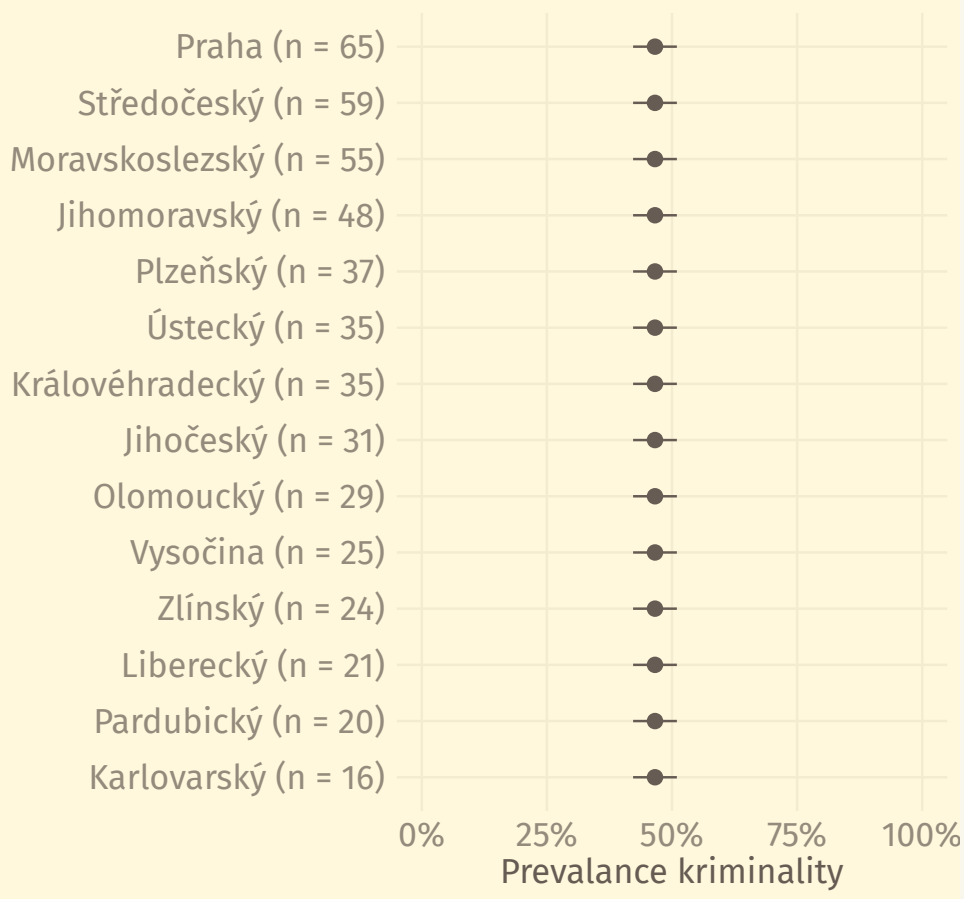
3 možná řešení

Snížení nároků na velikost vzorku

Můžeme předpokládat, že kraje jsou zaměnitelné.

„Slijeme“ všechny informace do jednoho odhadu (**complete pooling**).

```
glm(crime_experience ~ 1,
    family = binomial)
```



Snížení nároků na velikost vzorku

Výhoda: Malá náhodná chyba

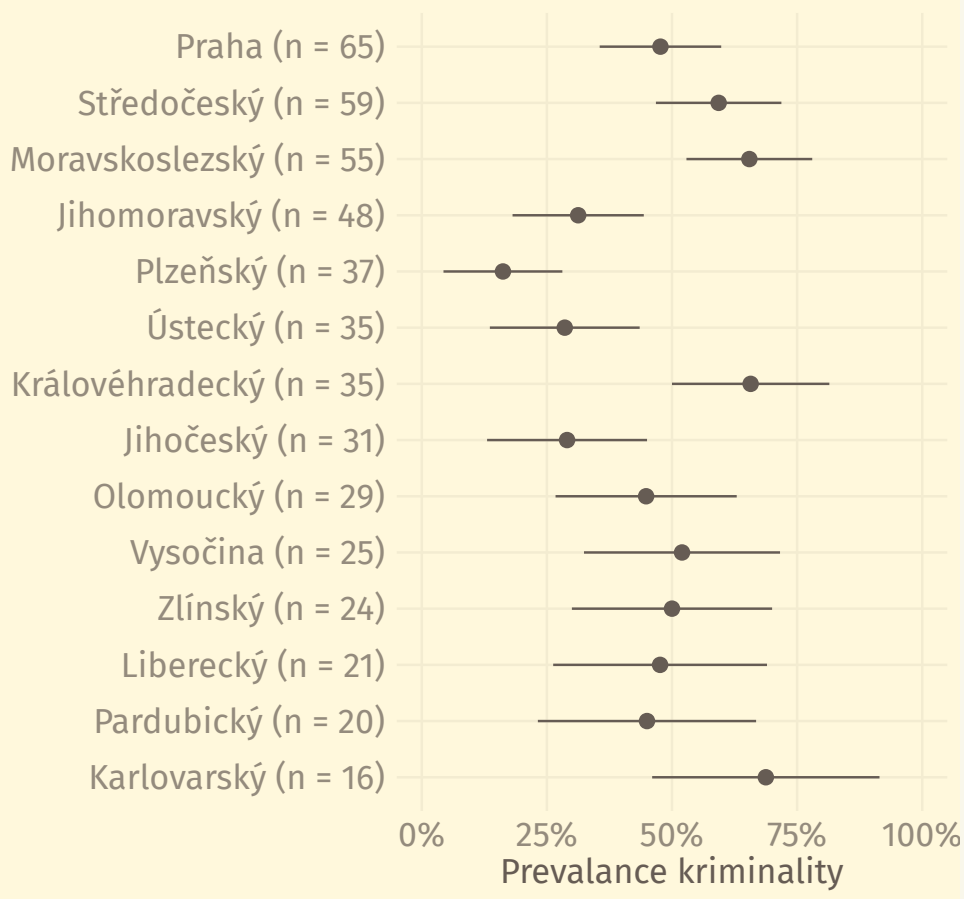
Nevýhoda: Velká systematická chyba

Snížení nároků na velikost vzorku

Můžeme předpokládat, že každý kraje je unikátní.

Pro každý krajský odhad použijeme pouze data z daného kraje (**no pooling**).

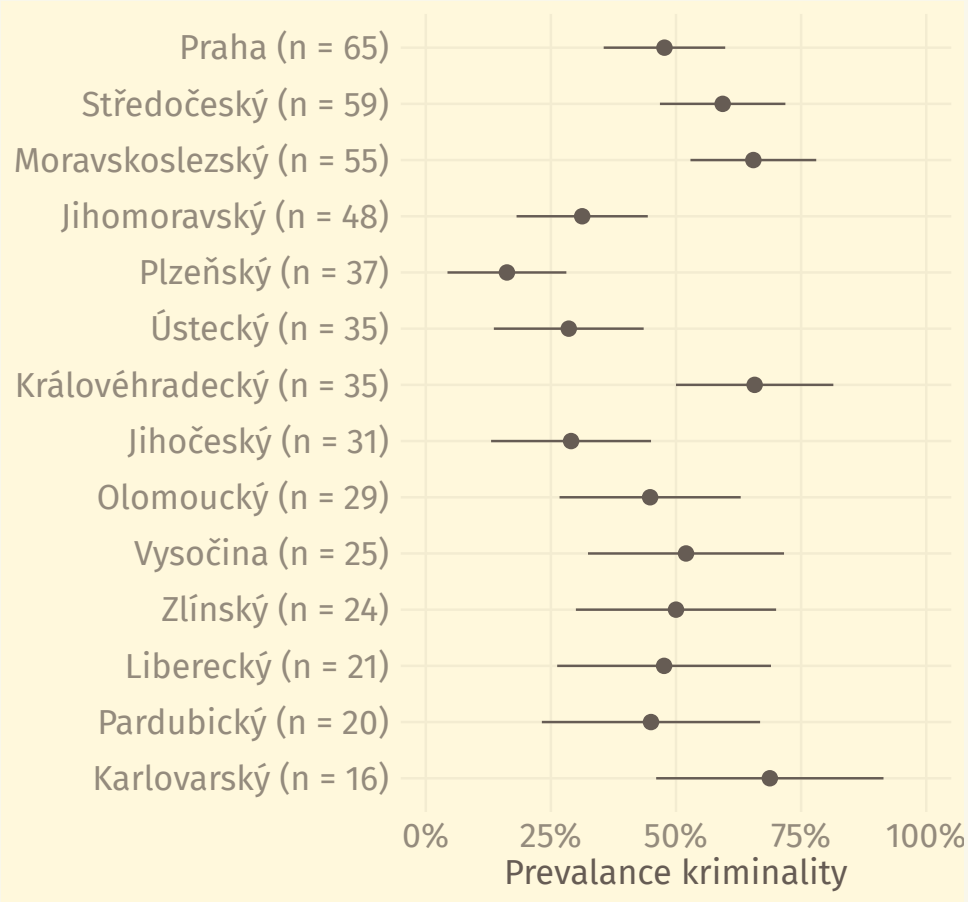
```
glm(crime_experience ~ region,
    family = binomial)
```



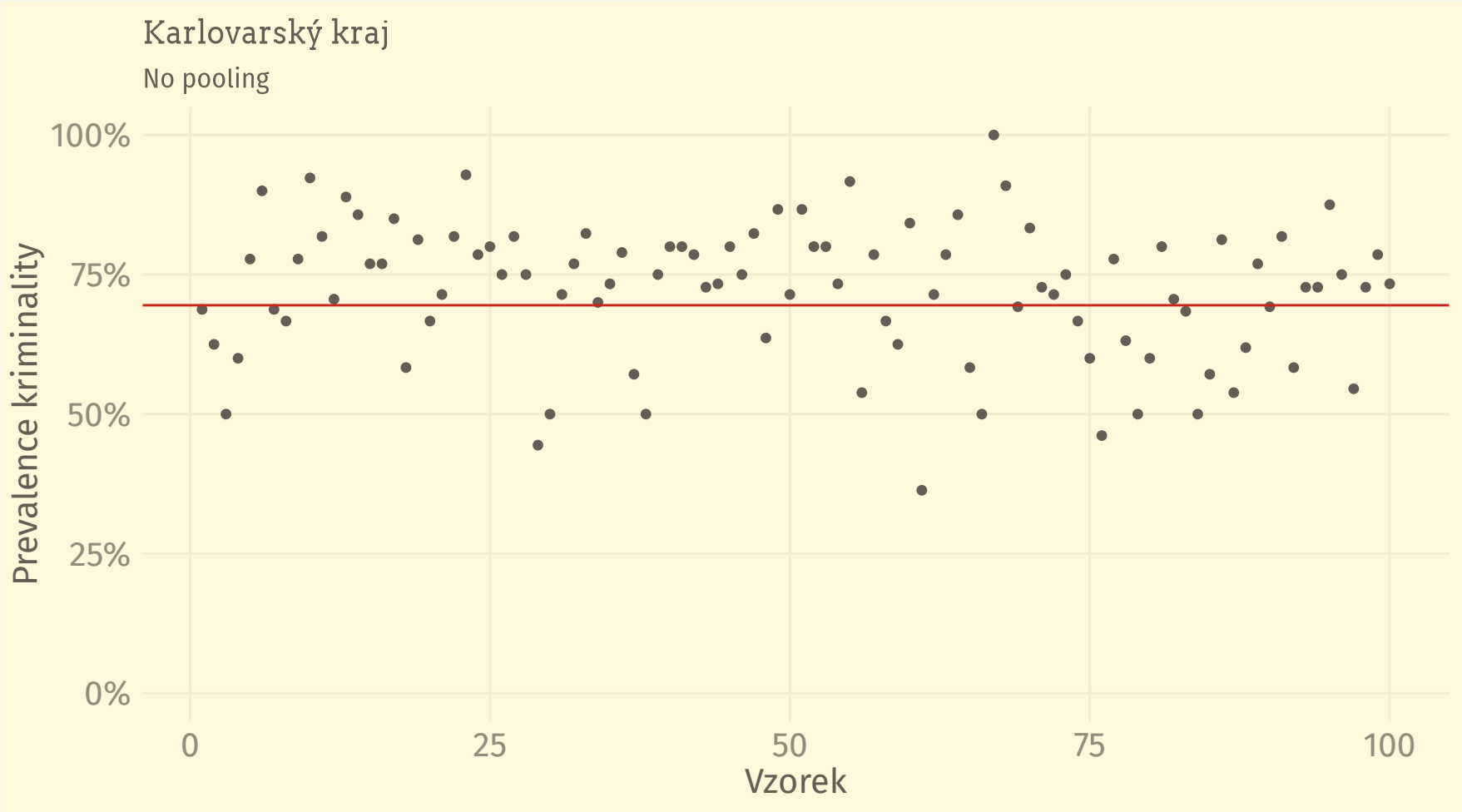
Snížení nároků na velikost vzorku

Výhoda: Malá systematická chyba

Nevýhoda: Velká náhodná chyba



Snížení nároků na velikost vzorku



Snížení nároků na velikost vzorku

Complete pooling ← ??? → No pooling

Snížení nároků na velikost vzorku

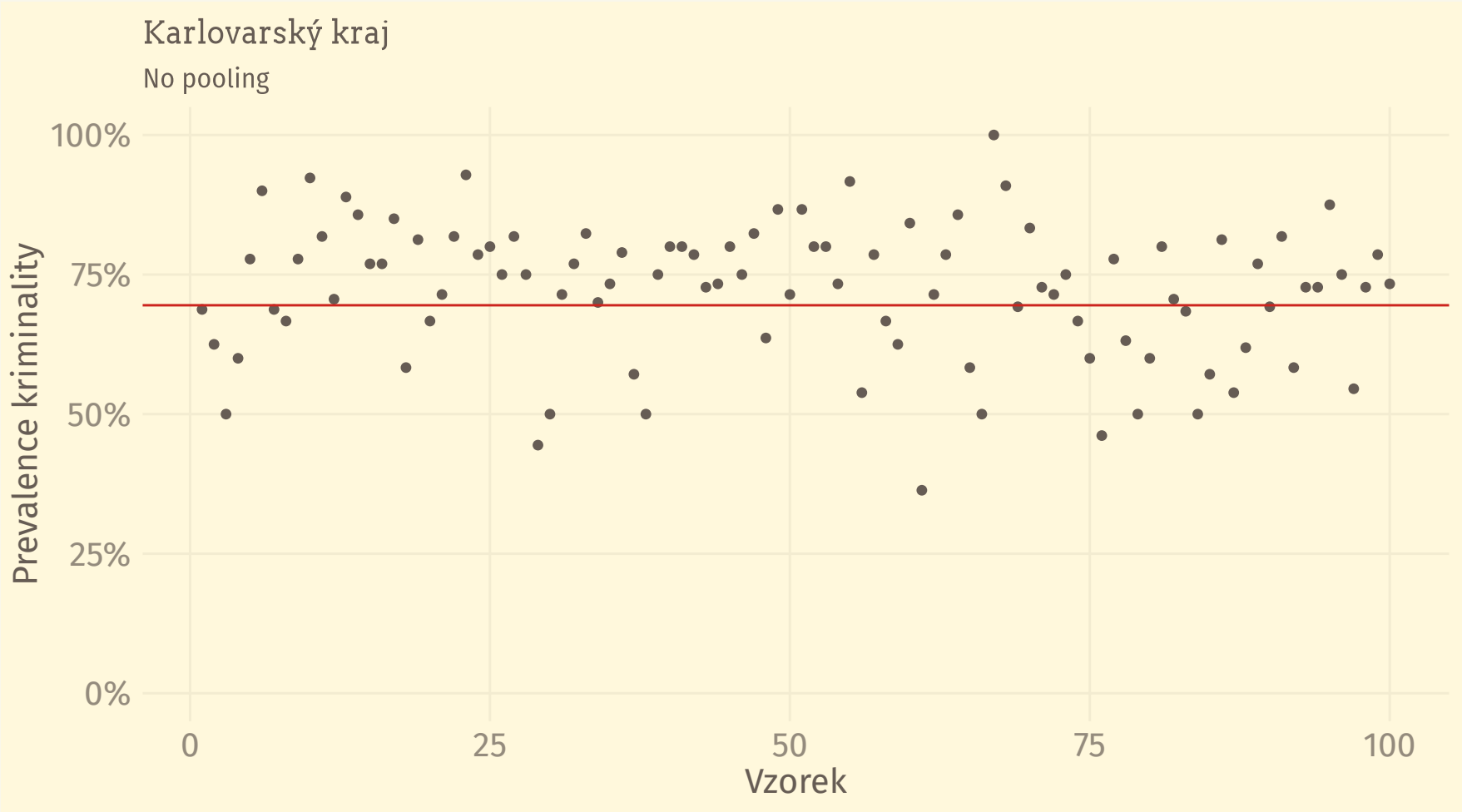
Můžeme předpokládat, že kraje jsou si příbuzné, ale distiktivní (**partial pooling**).

Každý krajský odhad je založený částečně na informacích o daném kraji a částečně o informacích od zbytku země.

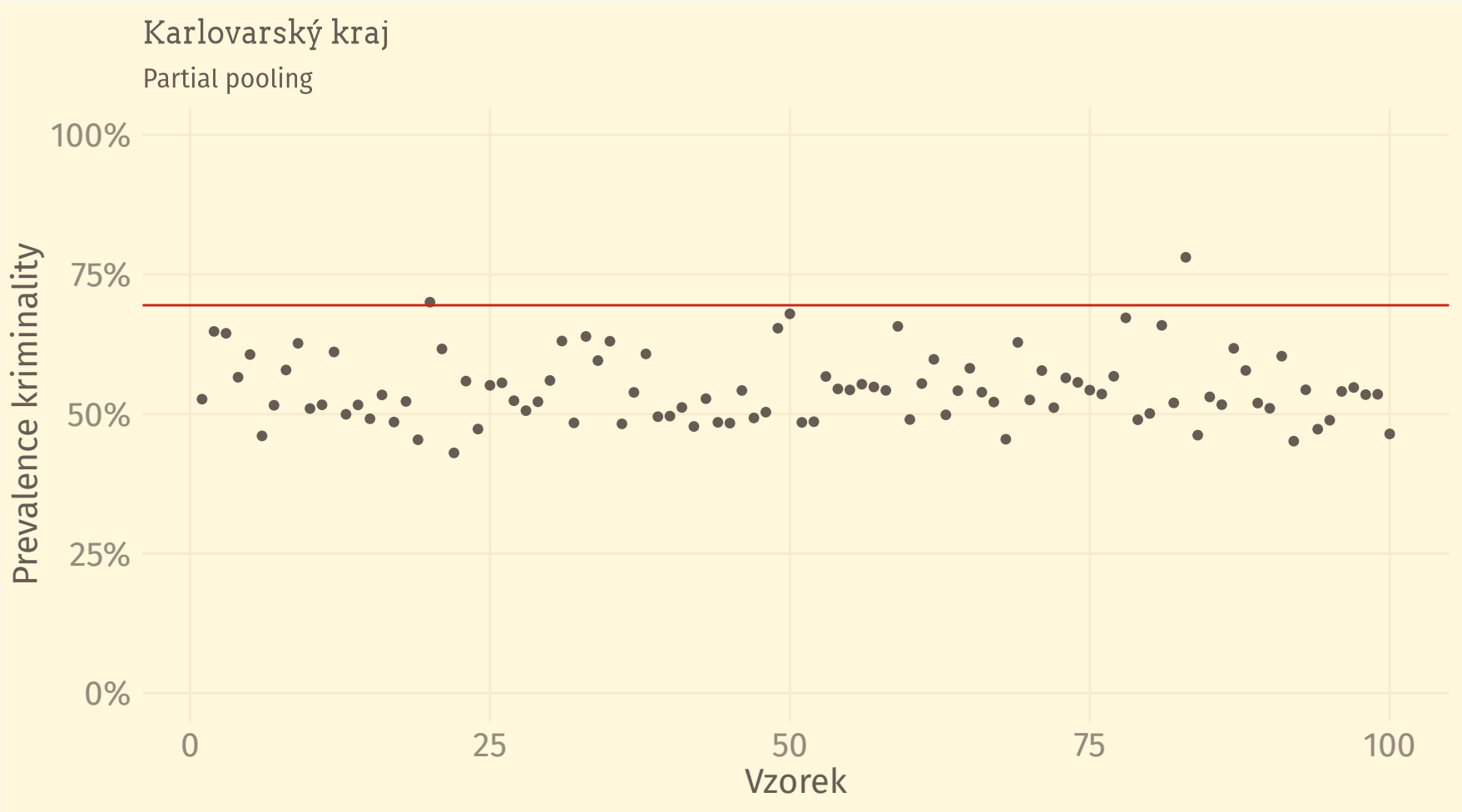
A to dělají víceúrovňové modely!

```
stan_glmer(crime_experience ~ (1|region), family = binomial)
```

Snížení nároků na velikost vzorku



Snížení nároků na velikost vzorku



Snížení nároků na velikost vzorku

Víceúrovňové modely využívají **partial pooling** - kombinují informace o dané skupině s informacemi o ostatních (podobných) skupinách.

Výsledkem je (malý) nárust systematické chyby, ale (velké) snížení náhodné chyby.

Celková chyba odhadu je menší!

Snížení nároků na velikost vzorku

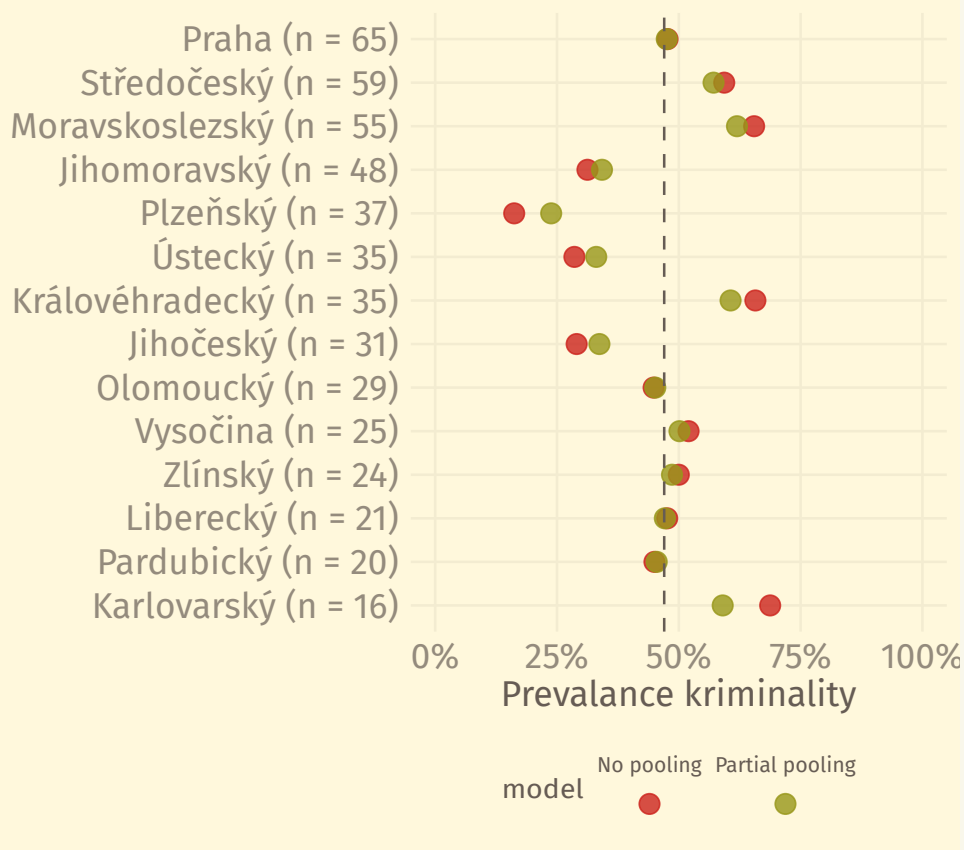
Partial pooling je vlastně způsob, jak dělat **shrinkage** nebo **regularization** (populární v prediktivním modelování.)

Souvisí s tzv. **bias-variance tradeoff**.

Snížení nároků na velikost vzorku

Čím více informací máme o daném krají, tím méně je jeho odhad ovlivněn ostatními.

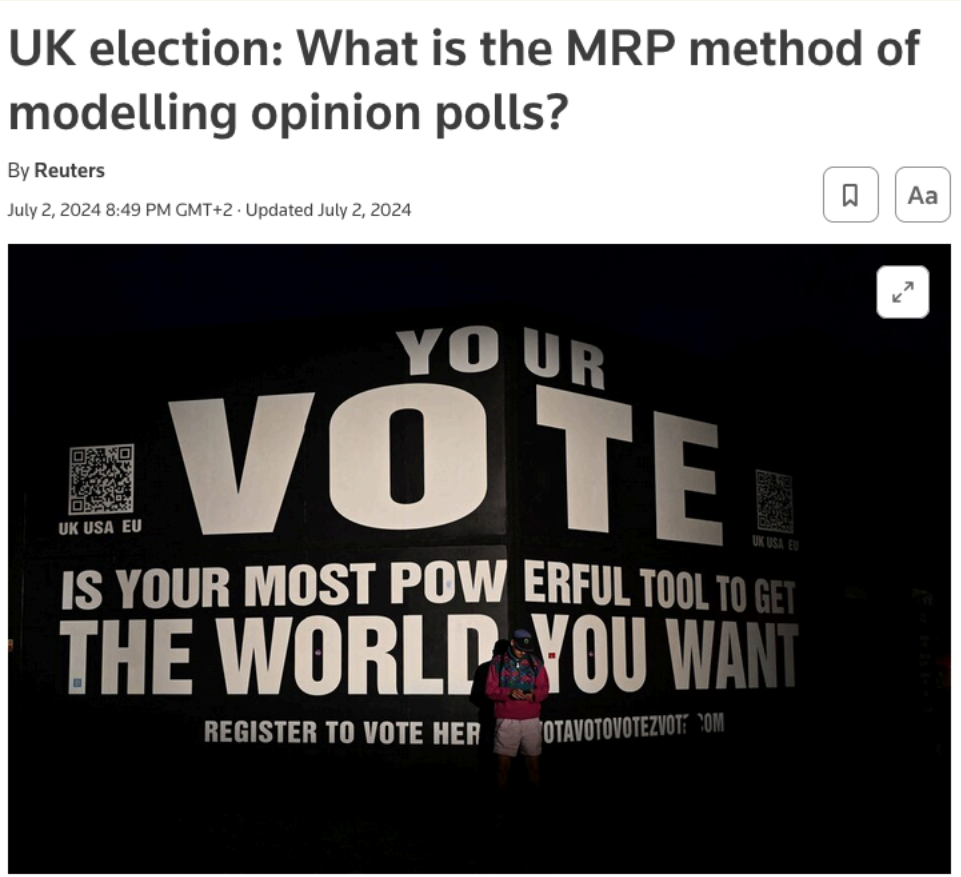
Čím méně informací o kraji a čím dále od celkového průměru, tím větší vliv budou mít ostatní.



Otázky?

InteRmezzo!

Snížení nároků na velikost vzorku



Multilevel regression with poststratification

How YouGov's MRP model works for the 2024 U.S. presidential and congressional elections

Snížení nároků na velikost vzorku

Častý problém:

Může se stát, že odhady jednotlivých regionů budou staženy příliš k celkovému průměru.

Řešení - Přidejte do modelu regionální prediktory (míra nezaměstnatnosti, volební účast)

Meziskupinové a vnitroskupinové efekty

Meziskupinové a vnitroskupinové efekty

Modelový příklad:

V oblasti vzdělávání se často vedou debaty o vztahu mezi socioekonomickým zázemím a kognitivním výkonem.

Také se vedou debaty o tom, do jaké míry hrají roli zdroje školy versus zdroje jedince (míchání žáků s různým SES?)

Otázka: Může žákům z chudších rodin pomoc, pokud budou chodit na stejné školy, jako ti majetnější?

Meziskupinové a vnitroskupinové efekty

Modelový příklad:

Data z 1982 o SES a matematické gramotnosti žáků středních škol v USA.

Výzkumná otázka: Hraje větší roli SES školy nebo SES individuálních žáků?

Meziskupinové a vnitroskupinové efekty

Víceúrovňové modely umožňují efektivně rozkládat mezi-skupinové a vnitro-skupinové efekty.

Vašem případě:

1. Vztah mezi průměrným SES školy a průměrnými mat. znalostmi (meziskupinový)
2. Vztah mezi SES a a mat. znalostmi žáků uvnitř školy (vnitroskupinový)

Meziskupinové a vnitroskupinové efekty

Příprava dat - Spočítáme a) průměrný SES každé školy a b) odchylku SES žáka od průměru jeho školy.

```
schools |>
  mutate(ses_between = mean(ses),
         ses_within = ses - ses_between,
        .by = school_id)
```

Meziskupinové a vnitroskupinové efekty

Příprava dat - Spočítáme a) průměrný SES každé školy a b) odchylku SES žáka od průměru jeho školy.

Alternativně:

```
datawizard::demean(schools, ~ses, by = ~school_id)
```

Meziskupinové a vnitroskupinové efekty

Dvě možnosti vytvoření modelu:

Random Intercept model - školy můžou mít různý průměrný SES, ale vztah mezi SES a mat. gramotností žáků je na každé škole stejný.

Jednoduší na výpočet, ale méně realistické.

```
stan_glmer( math ~ ses_within + ses_between + (1 | school_id))
```

Meziskupinové a vnitroskupinové efekty

Dvě možnosti vytvoření modelu:

Random Slopes model - školy můžou mít různý průměrný SES, a vztah mezi SES a mat. gramotností žáků se může lišit napříč školami.

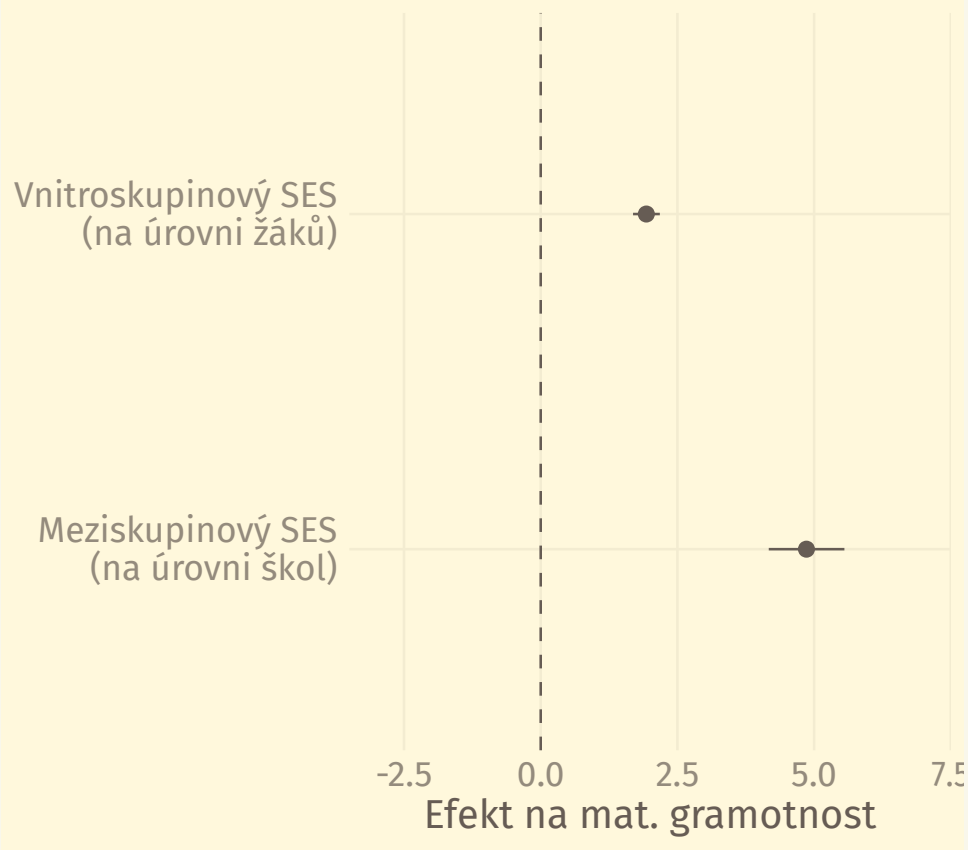
Výpočetně náročnější, ale realističtější

```
stan_glmer(math ~ ses_within + ses_between +
           (1 + ses_within| school_id))
```

Meziskupinové a vnitroskupinové efekty

Meziskupinový SES souvisí s mat. gramotností více, než vnitroskupiný!

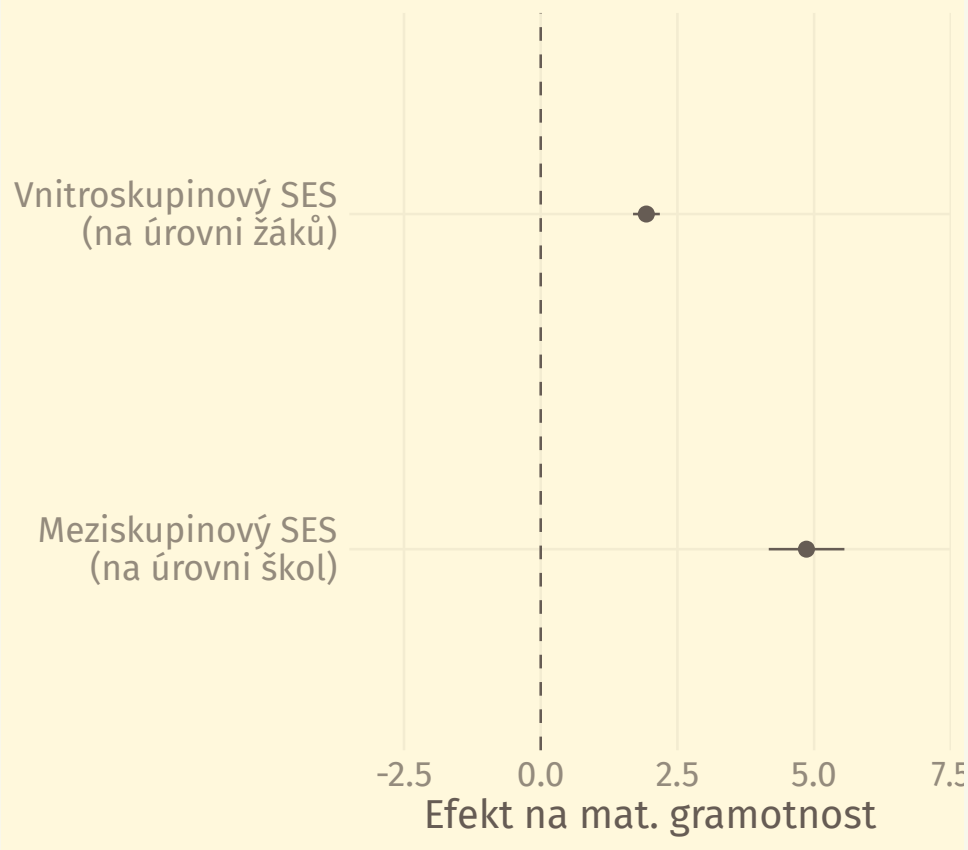
Rozdíly v mat. gramotnosti podle SES napříč školami jsou větší, než uvnitř škol.



Meziskupinové a vnitroskupinové efekty

Optimistická interpretace - i žáci s nízkým SES mohou benefitovat ze smíšených škol.

Negativní interpretace - velké regionální nerovnosti, chodit do chudé školy je velký handikep.



Otázky?

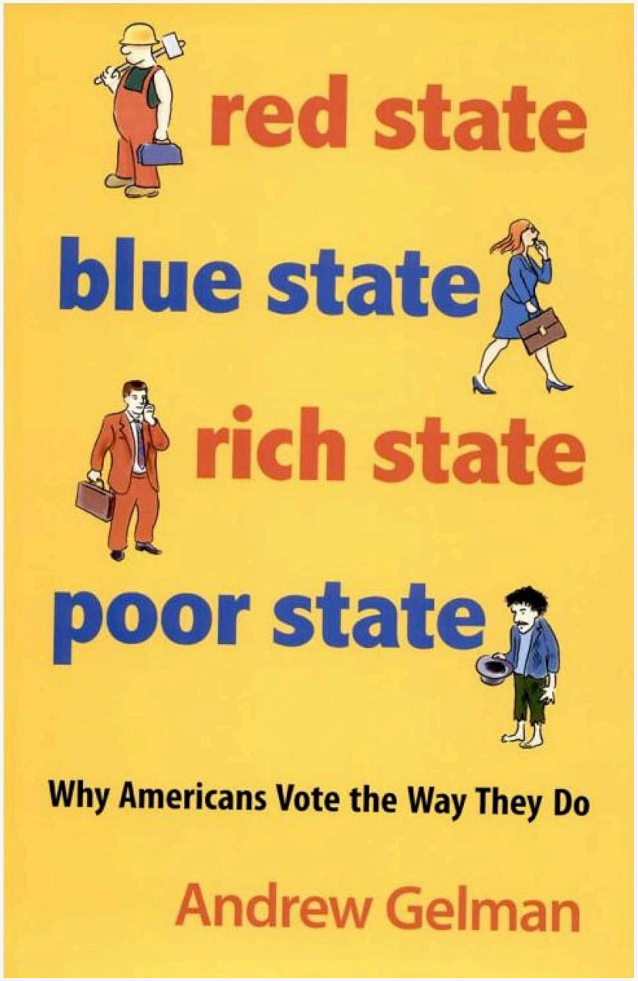
InteRmezzo!

Meziskupinové a vnitroskupinové efekty

Bohatší *státy* USA volí častěji demokraty,...

ale...

...bohatší *voliči* volí častěji repubklikány.



Meziskupinové a vnitroskupinové efekty

Častý problém:

Lidé zapomínají rozkládat mezi- a vnitro-skupinové efekty.

Extrémně časté u panelových dat!

```
stan_glmer(happiness ~ age + (1|respid))
```

Meziskupinové a vnitroskupinové efekty

Tento model splácá vnitro- a mezi-skupinový efekt do jednoho odhadu!

```
stan_glmer(happiness ~ age + wave + (1|respid))
```

Nedokážeme říct, do jaké míry s věkem roste spokojenost a do jaké míry jsou starší lidé spokojenější.

Technické poznámky

Frekventistické vs Bayesovské modely

Víceúrovňové modely jsou matematicky komplikované.

Frekventistické postupy mají problémy zohlednit standardní chybu náhodných komponent, např. `(1|region)`.

Pokud výpočetní kapacita dovolí, silně doporučuji bayesovský přístup.

Balíčky v R

Frekventistické balíčky:

- **lme4** - dobrý rozjezdový balíček
- **glmmTMB** - více modelů, efektivnější(?) implementace

Bayesovké balíčky:

- **rstanarm** - dobrý rozjezdový balíček
- **brms** - více modelů, efektivnější implementace

Otázky?

Děkuji za pozornost!