Vysoká škola
ekonomická v Praze
Fakulta informatiky a statistiky

TOOLS 4F

Automatizované metody výběru podmnožiny vysvětlujících proměnných v regresním modelu a problémy s nimi spojené

Milan Bašta
Jiří Procházka
Pavel Zimmermann

Program

- Regresní model pro účely výpočtu sazeb neživotního pojištění
- Problém opakovaných testů, p-value hacking
- Simulační studie – úspěšnost automatizovaných metod

Regresní modely

- Standardní přístup je použití regresních modelů.
- zejména zobecněné lineární modely (GLM)
- Řada možností:
  - „jednorozměrné“ přirážky
  - zobecněné aditivní modely (GAM)
  - zobecněné lineární smíšené modely (GLMM)
  - neuronové sítě
  - …

Princip ekvivalence

- Rizikové/netto pojistné je stanovováno na základě principu ekvivalence.

očekávaná hodnota vybraného pojistného = očekávaná výše škody

- Pro očekávanou výši škody předpokládáme (např. kolektivní model rizika):

očekávaná výše škody =
očekávaný počet škod na smlouvu * očekávaná výše škody pokud nastane

Frekvence a severita

očekávaná výše škody =
očekávaný počet škod na smlouvu * očekávaná výše škody pokud nastane

- Zpravidla se sestavují dva regresní modely
  - model pro frekvenci,
  - model pro severitu.

Omezení

- Výsledné modely podléhají celé řadě praktických omezení.
- Ceník
  - segmenty klientů
    diskrétní faktory, dummy proměnné
  - omezený počet segmentů
  - procentuální přirážky
    multiplikativní model, logaritmická spojovací funkce

Rizikové faktory

- segmenty tvořeny kombinací úrovní rizikových faktorů
- rizikové faktory
  - diskrétní proměnné (typ vozidla, region, frekvence placení pojistného)
  - diskretizované proměnné (výkon, věk, hmotnost)
- mnohaúrovňové faktory (PSČ, tovární značka)

Konstrukce modelů

- statistická a věcná významnost faktorů
- diskretizace spojitých proměnných
- rediskretizace faktorů

Testy významnosti

- test poměrem věrohodnosti
  - celý faktor (všechny dummy proměnné)
  - pokles věrohodnosti při vyřazení faktoru
  - nárůst věrohodnosti při přidání faktoru
- Waldovy testy
  - každá úroveň faktoru (každá dummy proměnná)
  - odchylka odhadu parametru od nuly
- informační kritéria (AIC, BIC)
  - věrohodnost penalizovaná za počet parametrů

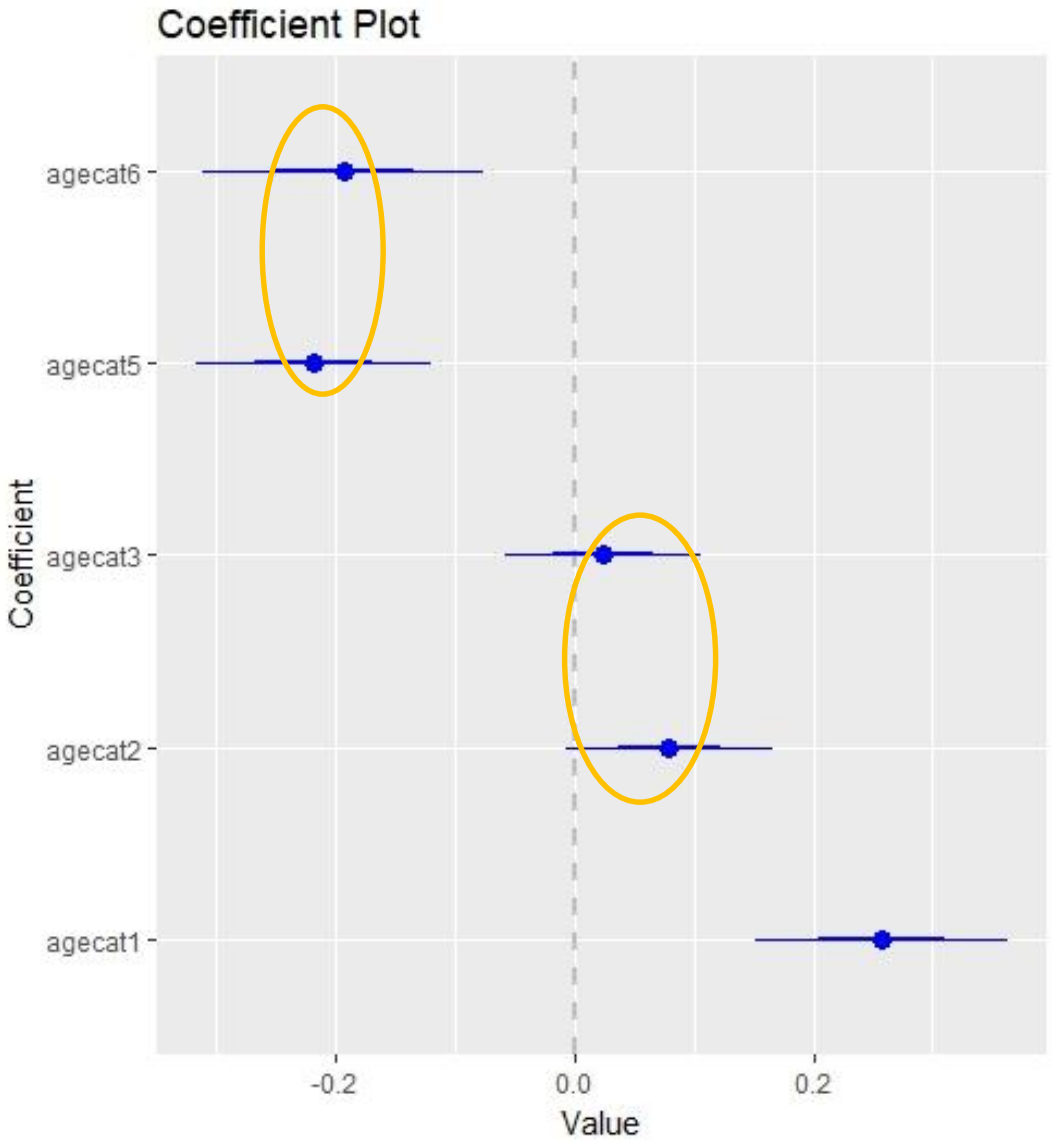
Diskretizace spojitých proměnných

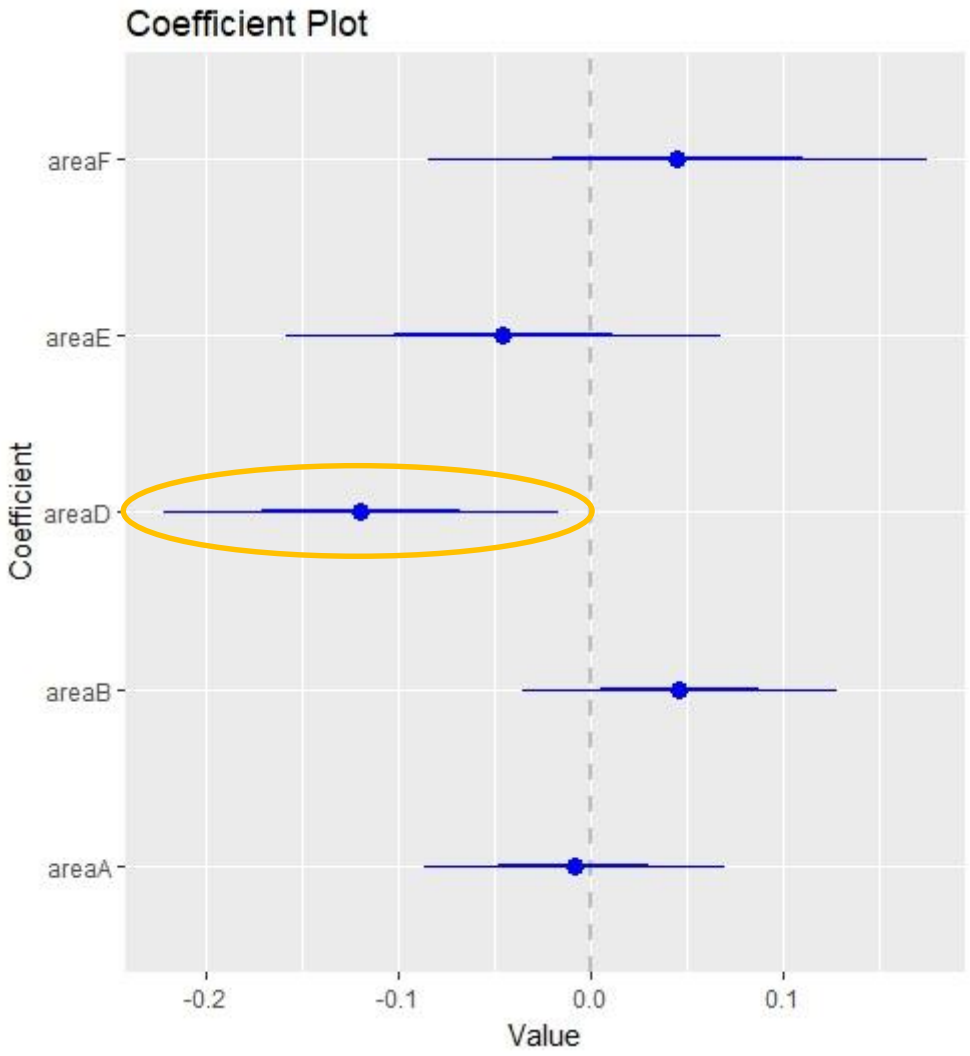
- expertní (restriktivní omezení na počet segmentů)
- ekvidistantní dle hodnot – některé málo zastoupené
- dle kvantilů – podobné hodnoty v různých kategoriích
- complete linkage (hierarchické shlukování)
- Fisherův algoritmus (natural breaks) – iterativní minimalizace vnitroskupinového součtu čtverců

Rediskretizace

důvody

- málo zastoupené
- redukce počtu segmentů
- **statistická významnost?**


Coefficient Plot
agecat6
agecat5
agecat3
agecat2
agecat1
Coefficient
-0.2
0.0
0.2
Value

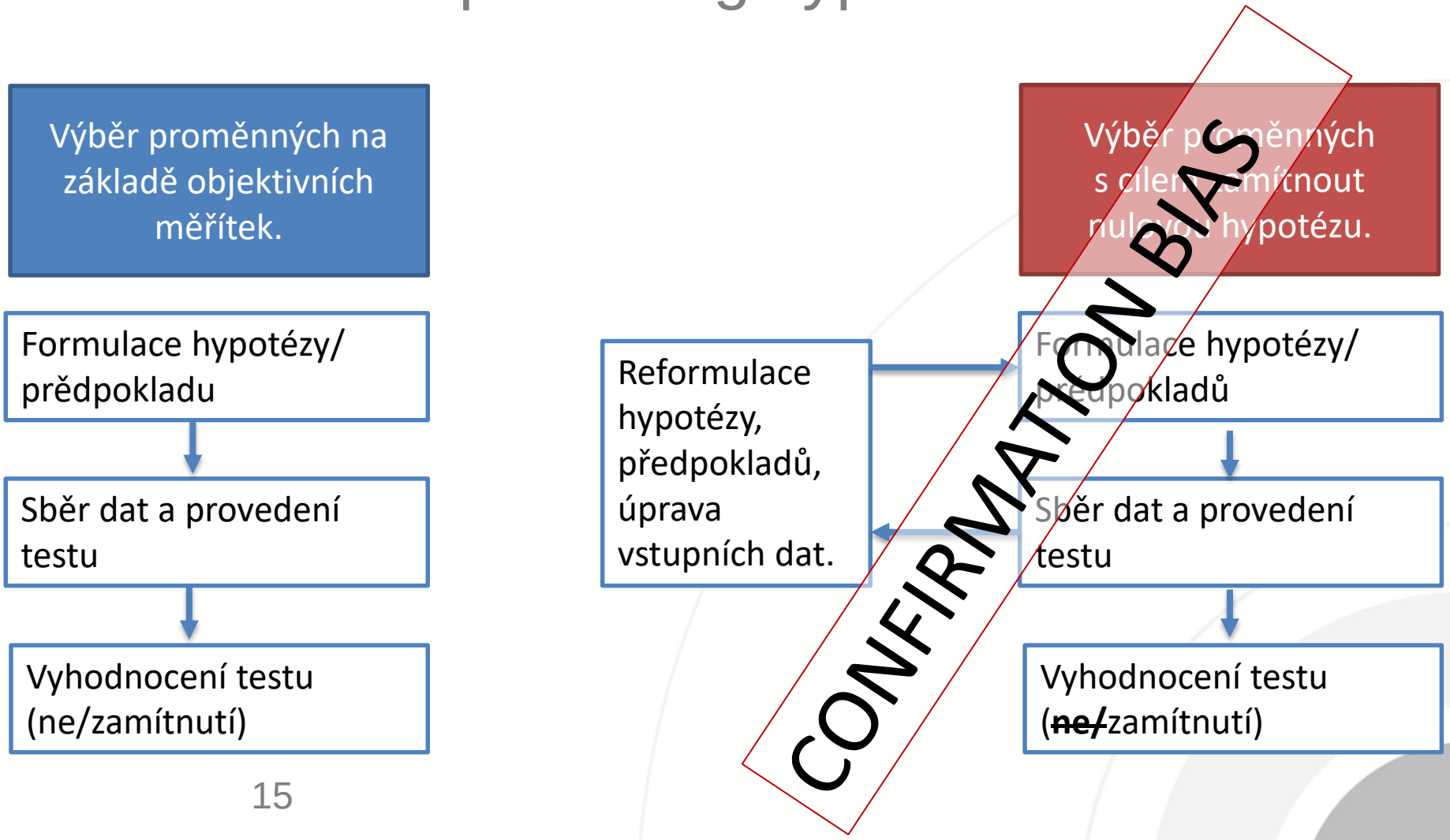

Coefficient Plot
areaF
areaE
areaD
areaB
areaA
Coefficient
-0.2
-0.1
0.0
0.1
Value

Co je to p-hacking?

(*data-dredging, snooping, fishing, significance-chasing, double-dipping*)

- Postup, jehož cílem není získání objektivních závěrů pomocí statistické indukce.
- Jediným cílem p-hackingu je docílit (téměř všemi prostředky) zamítnutí nulové hypotézy.
- Je možné se p-hackingu dopouštět vědomě i nevědomě.

Jak p-hacking vypadá?



Nejčastější metody p-hackingu

- Ukončíme sběr dat v momentu, kdy p-hodnota < 0.05.
- Omezíme sesbíraná data tak, aby p-hodnota < 0.05.
- Testujeme velké množství hypotéz, prezentujeme pouze ty, kde p-hodnota < 0.05.
- Vybíráme pouze ty vysvětlující proměnné u kterých je p-hodnota < 0.05.
- Transformujeme data tak, aby p-hodnota < 0.05.

Otázky ohledně p-hackingu?

- Jaká jsou hlavní nebezpečí p-hackingu?
- Je p-hacking vždycky špatně?
- Hraje v kontextu p-hackingu nějakou roli síla testu?
- Je vůbec p-hodnota míra designovaná k vytváření induktivních úsudků?

Kontakty

Pavel Zimmermann	zimmermann@tools4f.com zimmerp@vse.cz