

M9750 Robustní a neparametrické statistické metody

cvičení 6 - robustní odhady vícerozměrného parametru

1. Zkoumejte chování vícerozměrného výběrového průměru a Jamesova - Steinova odhadu na základě náhodného výběru z p -rozměrného normálního rozdělení o rozsahu $n = 30$ se střední hodnotou $\boldsymbol{\theta} = (1, \dots, 1)'$ a varianční maticí $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p$. Pro jednoduchost volte $p = 3$ a $\sigma^2 = 1$.
 - (a) Pro daný náhodný výběr spočítejte oba příslušné odhady a spočítejte jejich empirickou střední čtvercovou chybu.
 - (b) Celý postup opakujte 10 000 krát. Na základě těchto simulací odhadněte střední čtvercovou chybu obou odhadů (průměr empirických středních čtvercových chyb).
 - (c) Výsledky z bodu (b) porovnejte. Výsledek porovnejte i s teoretickou hodnotou střední čtvercové chyby pro výběrový průměr.
2. Uvažujte data `heptathlon` z knihovny `HSAUR` (proměnné `hurdles` a `shot`).
 - (a) Data si nejprve vykreslete a tipněte si, která budou mít největší a která nejmenší hloubku.
 - (b) Ke každému pozorování do grafu přidejte jeho hloubku.
 - (c) Nakreslete graf kontur Tukeyho hloubky.
 - (d) Najděte oblast s největší hloubkou a Tukeyho medián.
 - (e) Přidejte jej do grafu, stejně tak i marginální a L_1 medián. Výsledky porovnejte.
3. Pracujte se stejnými daty.
 - (a) Proveďte jednorozměrnou exploratorní analýzu proměnných `hurdles` a `shot`. Obsahují proměnné nějaké odlehlé hodnoty?
 - (b) Proveďte mnohorozměrnou exploratorní analýzu. Vykreslete si bagplot. Obsahují data nějaká odlehlá pozorování?
4. Pracujte se stejnými daty.
 - (a) Pomocí klasické lineární regrese modelujte závislost výsledku vrhu koulí na výsledku z běhu na 100 metrů překážek.
 - (b) Odhadnutou regresní přímku přidejte do grafu.
 - (c) Posuďte vhodnost zvoleného modelu. Obsahují data nějaká vlivná nebo odlehlá pozorování?
 - (d) Odhadněte regresní koeficienty pomocí 50% regresního kvantilu a opět odhadnutou regresní přímku přidejte do grafu.
 - (e) A konečně, uvažujte model bez odlehlého pozorování a proveďte odhad metodou nejmenších čtverců na těchto datech; odhadnutou regresní přímku přidejte do grafu.
 - (f) Všechny tři odhady porovnejte. Který z nich nejlépe popisuje sledovanou závislost?

5. Uvažujte data `CYGOB1` z knihovny `HSAUR` (logaritmus povrchové teploty `logst` a logaritmus světelné intenzity `logli` 47 hvězd).
- (a) Data si nejprve vykreslete a poté pomocí lineární regrese modelujte závislost logaritmu světelné intenzity na logaritmu povrchové teploty.
 - (b) Odhadnutou regresní přímku přidejte do grafu.
 - (c) Posuďte vhodnost zvoleného modelu. Obsahují data nějaká vlivná nebo odlehlá pozorování?
 - (d) Odhadněte regresní koeficienty pomocí 50% regresního kvantilu a opět odhadnutou regresní přímku přidejte do grafu.
 - (e) A konečně, uvažujte model bez odlehlých pozorování a proveďte odhad metodou nejmenších čtverců na těchto datech; odhadnutou regresní přímku přidejte do grafu.
 - (f) Všechny tři odhady porovnejte. Který z nich nejlépe popisuje sledovanou závislost?

Funkce, které by se mohly hodit: `apply`, `points`, `abline`, funkce `mvrnorm` z knihovny `MASS`, funkce `depth`, `isodepth`, `med` z knihovny `depth`, funkce `bagplot` z knihovny `aplpack` a funkce `rq` z knihovny `quantreg`.