Metódy matematickej štatistiky

Zápočtová domáca úloha

Letný semester 2007/08

Narodené myšky

Dáta

- Načítajte dátový súbor `litters` z knižnice `DAAG` štatistického software `R` a naštudujte k nemu popis z nápovedy.
- K veličine brainwt pripočítajte vektor náhodných čísel z rozdelenia $\mathsf{N}(0, 0.00yyyymmdd)$, kde *yyyymmdd* značia cifry z vašeho dátumu narodenia v tvare rok, mesiac a deň. Takto novovytvorenú veličinu uložte ako brainwtNew a ďalej už neuvažujte veličinu brainwt.

Zadanie

1. Urobte stručný sumár dát obsahujúci základné popisné štatistiky pre trojicu veličín lsize, bodywt a brainwtNew. *(3 body)*
2. Nakreslite krabicové diagramy (boxploty) a histogramy pre tri spomínané veličiny. *(2 body)*
3. Vykreslite párový diagram (scatterplot) pre tri spomínané veličiny. *(1 bod)*
4. Spočítajte Pearsonov výberový korelačný koeficient pre každú dvojicu z našich troch veličín. *(2 body)*
5. Z čisto praktického hľadiska navrhnite a zdôvodnite, ktorá z troch veličín by mala v regresnom modeli vystupovať ako závislá (vysvetľovaná) premenná? *(1 bod)*
6. Formulujte lineárny model lineárnej regresie s interakciami, kde závislá premenná bude brainwtNew a regresory budú lsize a bodywt. Uveďte odhady parametrov tohoto modelu a ich smerodatné odchýlky. *(6 bodov)*
7. Predchádzajúci "plný" model vhodne zjednodušte. Zdôvodnite svoj postup pri redukcii plného modelu na finálny model. Uveďte odhady parametrov tohoto modelu a ich smerodatné odchýlky. *(15 bodov)*
8. Zistite koeficient mnohorozmernej korelácie a koeficient determinácie medzi veličinou brainwtNew a veličinami lsize, bodywt. *(2 body)*
9. Pomocou študentizovaných reziduí a od nich odvodených štatistík mier vplyvu pozorovania skúmajte odľahlé pozorovania pre finálny model. *(3 body)*
10. Otestujte odľahlosť prípadných podozrivých pozorovaní z predchádzajúceho bodu pomocou príslušného t-testu založenom na študentizovaných reziduách. *(2 body)*
11. Vykreslite *Cookovu* vzdialenosť pozorovaní pre finálny model. *(1 bod)*

12. Správnosť tvaru závislosti finálneho modelu graficky overte vykreslením reziduí modelu voči odhadnutým hodnotám závislej premennej. *(2 body)*

13. Konštantnosť rozptylu vo finálnom modeli graficky overte vykreslením štvorcov reziduí voči odhadnutým hodnotám závislej premennej. *(2 body)*

14. Predpoklad normality vo finálnom modeli graficky overte nakreslením *Q–Q plotu* a histogramu reziduí. *(2 body)*

15. Predpoklad homoskedasticity pre finálny model overte pomocou *Breusch-Paganovho* testu. *(2 body)*

16. Predpoklad normality pre finálny model overte pomocou *Shapiro-Wilkovho* testu. *(2 body)*

17. Napriek faktu, že z charakteru dát neočakávame autokoreláciu reziduí finálneho modelu, urobte *Durbin-Watsonov* test. *(2 body)*

18. Navrhnite kvadratický model lineárnej regresie, kde závislá premenná bude opäť brainwtNew. Je tento model vhodný? Dokumentujte číselne a vysvetlite. Dal by sa zjednodušiť? *(10 bodov)*

19. Formulujte výstižný záver, v ktorom zhrniete výsledky predchádzajúcej analýzy. Stručne diskutujte graficke overenia a testy "správnosti" predpokladov. Ak niektorá z diagnostík overovania predpokladov pre váš finálny model naznačuje ich porušenie, zvoľte iný vhodnejší finálny model a preveďte s ním opäť celú analýzu. Interpretujte model, odhady parametrov a tvar závislosti. Ako dobre model fituje dáta? Zhrňte vhodnosť finálneho modelu a jeho praktický význam. *(20 bodov)*

Bonus Čomu hovoríme multikolinearita? Ak nemôže byť multikolinearita vo vašom finálnom modeli prítomná, zdôvodnite prečo. Ak môže, skúmajte ju pomocou *VIF* (Variance Inflation Factor). Vysvetlite súvislosť medzi *VIF*, vhodným korelačným koeficientom a vhodným koeficientom determinácie. Ako sa prejavuje prípadná multikolinearita na odhadoch parametrov regresného modelu? *(5 bodov)*

Pokyny k vypracovaniu

- Vaše riešenie musí obsahovať najmä **komentár** vašich štatistických úvah. Význam všetkých štatistických veličín musí byť adekvátne **vysvetlený** a všetky výsledky i obrázky musia byť **interpretované**!
- Rozsah práce musí byť **maximálne 10 strán**. Veľkosť použitého fontu sa musí pohybovať od 10pt do 12pt.
- Komentár riešenia je vyžadovaný v **súvislých** vetách. Celý dokument musí byť (vrátane popiskov tabuliek a obrázkov) napísaný konzistentne v jednom jazyku (čeština, slovenčina alebo angličtina). Výstupný formát dokumentu musí byť `pdf`.
- Výpočty analýzy musia byť naprogramované v štatistickom software `R`. Ako vhodné knižnice môžu poslúžiť najmä `MASS`, `stats`, `DAAG`, `car`, `lattice` a `lmtest`.
- Váš **zdrojový kód** použitý k analýze a maľovaniu obrázkov musí byť dostatočne **okomentovaný** a zaslaný v kódovaní `UTF-8`.
- Aby boli vaše výsledky verifikovateľné, nastavte `set.seed(yyyymmdd)` pred generovaním náhodných čísel podľa vašeho dátumu narodenia. Tento použitý **seed** nezabudnite uviesť vo svojom riešení.

- Súbor s `R` kódom nazvite `priezvisko_meno.R`, súbor s hlavným komentárom (riešením) nazvite `priezvisko_meno.pdf`. V názvoch súborov **nepoužívajte** diakritiku! Oba súbory zabaľte do súboru nazvaného `priezvisko_meno.pripona` (prípona podľa použitého kompresného programu) a zašlite **e-mailom** cvičiacemu, od ktorého chcete zápočet. Majte na pamäti, že v dnešnej dobe odoslanie e-mailu ešte neznamená jeho prijatie adresátom. Cvičiaci v primeranej dobe potvrdí prijatie. K baleniu používajte jeden z programov: `zip`, `bzip2`, `gzip`, `tar -zcvf`.
- **Termín** odoslania úlohy je **pondelok 15. septembra 2008** (23:59 CET). Práce odoslané po tomto dátume budú mazané.
- Jedná sa o **samostatnú** prácu. Ak bude usúdené, že niektoré riešenia sa navzájom príliš podobajú, zasielatelia budú odmenení bonusom mínus 41 bodov.
- Za prácu je možné získať maximálne 100 bodov (plus 5 bonusových), pričom 80 bodov bude udelených za report, 10 bodov za komentár k `R`kovému zdrojáku a 10 bodov za celkový dojem z práce. Za úspešne riešenú zápočtovú úlohu (nutná na získanie zápočtu) sa považuje práca ohodnotená **aspoň** na **60 bodov**.

V Prahe 15. mája 2008

Kateřina Helisová — Zdeněk Hlávka — Zbyněk Pawlas — Michal Pešta — Jakub Staněk