NSTP194 Regrese: zadání samostatně řešených úloh
(písemná část zkoušky, zadání z 21. listopadu 2011)

V tomto textu je uvedeno 21 úloh, z nich je třeba jednu si vybrat a zapsat se na evidenční list vystavený na katedrální nástěnce v prvním poschodí. Každou úlohu si může vybrat jen jeden student.

Data k úlohám pocházejí z erkových knihoven. V helpu knihoven se najde stručný popis jednotlivých proměnných, někdy i odkaz na podrobnější informace. Data standardně získáte pomocí příkazu `data(abc, package="yz")`, kde `abc` je bez uvozovek uvedené jméno datového souboru a `yz` je v uvozovkách uvedené jméno knihovny, např. `data(Animals,package="MASS")`.

Prosím, nerozšiřujte si zadanou úlohu. Datový soubor mnohdy obsahuje údaje, o nichž se v zadání nezmiňuji. Nepoužívejte je, i když byste závisle proměnnou pak dokázali vysvětlit těsnějším a snáze interpretovatelným modelem závislosti. I když je ve všech úlohách třeba ověřit splnění běžných předpokladů normálního lineárního modelu, může se stát, že se úlohu nepodaří převést na model, v němž bychom mohli předpokládat splnění všech těchto požadavků. Pak je třeba uvést i příslušné vysvětlení.

Řešení úlohy pošlete **nejpozději** do 12. hodiny dva dny před dnem zkoušky elektronickou poštou jako přílohu na adresu
`zvara@karlin.mff.cuni.cz`
Doručení Vašeho řešení potvrdím emailem. Předpokládám, že v LaTeXu či v TeXu vysázené práce pošlete ve formátu PDF nebo PS. Pokud se k mé lítosti (známku to však nesmí ovlivnit) rozhodnete pro Word, nepoužívejte, prosím, jeho nejnovější verze. Mám k disposici jen MS Word 2003, jakýsi převodník z formátu *.docx a protějšek Wordu z Open Office. Případné problémy můžeme spolu vyřešit v mých konzultačních hodinách (končí s koncem přednáškového období) nebo po vzájemné dohodě emailem i jindy. Mnohdy stačí pouhá elektronická komunikace. Počítejte však s tím, že ve zkouškovém období mám značně omezené časové možnosti.

MASS Knihovna navazuje na knihu Venables, Ripley: Modern Applied Statistics with S-Plus. Data jsou zpravidla v některém z vydání knihy také popsána. Naše úloha se však může od úlohy řešené v knize lišit.

Animals Vyšetřete závislost váhy mozku (`brain`) na hmotnosti celého těla (`body`). Použijte přitom vhodnou transformaci jedné či obou proměnných. Jsou tu nějaká odlehlá pozorování? Pokud dokážete vysvětlit, proč některá pozorování jsou „jiná“, a to včetně `Human`, neberte je při odhadování v úvahu. V takovém případě tato pozorování porovnejte s očekáváním vyplývajícím z modelu. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

Cushings Pro nejčetnější typ syndromu vyšetřete možnost předpovědi proměnné `Pregnanetriol` pomocí `Tetrahydrocortisone`. Zvažte při tom možnost použít transformace. Nakonec pro tento typ syndromu ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

Pima.tr Zhodnoťte závislost `skin` na `bmi`, ověřte splnění běžných předpokladů včetně jejich vlivu na odhad regresních koeficientů, vyjádřete se k výskytu odlehlých pozorování.

Rubber Porovnejte závislost `loss` na `hard` se závislostí `loss` na `tens` a `hard`. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu. Popište, nakolik ovlivňuje přesnost odhadu regresních koeficientů multikolinearita.

anorexia Navrhněte porovnání trojích ošetření. Zhodnoťte možnost či potřebu použít analýzu kovariance. Přihlédněte při tom k tomu, nakolik jsou splněny klasické předpoklady.

birthwt Při vyšetřování závislosti porodní hmotnosti dítěte (`bwt`) na hmotnosti matky před těhotenstvím (`lwt`) přihlédněte také k rasové příslušnosti (`race`) a kouření (`smoke`) matky. Je třeba rozlišovat hodnoty 2 a 3 proměnné `race`? Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

cabbages Vyjádřete se k závislosti obsahu vitaminu C (`VitC`) na hmotnosti zelných hlávek (`HeadWt`). Je třeba brát v úvahu také druh zelí (`Cult`) a dobu výsadby (`Date`)? Popište, jak `HeadWt` ovlivňuje obsah vitaminu C. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

cats Vyšetřete závislost hmotnosti srdce kočky (`Hwt`) na její celkové hmotnosti (`Bwt`). Je třeba přihlédnout také k pohlaví (`Sex`)? Ověřte nakolik jsou splněny běžné předpoklady normálního lineárního modelu. Bude-li třeba, pokuste se navrhnout vhodné transformace.

crabs Při vyšetřování závislosti proměnné `RW` na `BD` přihlédněte k jejich druhu (`sp`). Je třeba rozlišovat také podle pohlaví (`sex`)? Při hledání modelu zvažte také možnost transformací. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

car Knihovnu `car` (Companion to Applied Regression) vytvořil J. Fox, který je autorem mimo jiné také „klikací“ knihovny `Rcmdr`. Součástí knihovny `car` je také procedura `Anova` určená k výpočtu rozkladů součtu čtverců typu II a III.

Burt Data pocházejí ze studie sledující vývoj dvojčat, z nichž jen jedno bylo vychováváno biologickými rodiči. Pokuste se předpovědět IQ dvojčete vychovávaného pěstouny (`IQfoster`) pomocí IQ dvojčete vychovávaného biologickými rodiči (`IQbio`) s přihlédnutím k sociálnímu postavení pěstounů (`class`). Nestačilo by rozlišovat jen kategorii `low` a druhé dvě kategorie spojit? Projeví se případně zjištěný vliv sociálního postavení, když budeme hodnotit pouze rozdíl IQ mezi dvojčaty? Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

Chirot Na základě historického materiálu byly jednotlivé územní jednotky v Rumunsku zhodnoceny s ohledem na selskou rebelii. Pokuste se předpovědět sílu rebelie (`intensity`) pomocí podílu obdělávané ploch (`commerce`) a gramotnosti (`tradition`). Zvažte případnou neaditivitu jejich vlivu. Porovnejte výsledný model s modelem, v němž použijete centrované nezávisle proměnné. Pokuste se názorně vysvětlit svoje zjištění. Pomůckou může být tabulka predikcí pro několik kombinací nezávisle proměnných (např. pro jejich první, druhé a třetí kvartily). Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

Prestige Vyjádřete závislost příjmu (`income`) na odpovídající prestiži povolání (`prestige`) a typu povolání (`type`). Podle potřeba navrhněte vhodnou transformaci. Je citlivost vůči prestiži u všech typů povolání stejná?. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

Davis Rozhodněte o závislosti váhy, jak ji udává sama vyšetřovaná osoba (`repwt`) na objektivně zjištěné váze (`weight`). Souvisí hodnota `repwt` s pohlavím osoby? Opravte případně zjištěné hrubé chyby v datech. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

datasets Knihovna `datasets` patří k těm, které jsou standardně nahrávány při spuštění erka. Obsahuje více než 80 datových souborů.

mtcars Navrhněte vhodnou funkci spotřeby paliva (`mpg`) tak, aby byla co nejlépe vysvětlitelná obsahem válců motoru (`disp`), případně funkcí obsahu válců, přičemž je možno přihlédnout také k počtu válců (`cyl`). Další údaje neberte v úvahu. Zhodnoťte, nakolik jsou splněny běžné požadavky na normální lineární model.

state Na základě dat obsažených v matici (`state.x77`) vyšetřete závislost střední délky života (`Exp.Life`) na dvojici charakteristik jednotlivých států: (`Murder`), (`HS.Grad`). Vedle obligátního ověření běžných předpokladů normálního lineárního modelu se pokuste zjistit, který ze dvou regresorů ovlivňuje dobu života více. Data zpřístupníte příkazem `attach(data.frame(state.x77))`.

trees K disposici jsou informace o množství dřeva (`Volume`) získaného z poraženého stromu, jeho výčetní tloušťka (`Girth`) a výška (`Height`). Navrhněte předpověď množství dřeva na základě těchto údajů a porovnejte svůj návrh s předpovědí pomocí vztahu $c \cdot \texttt{Girth}^2 * \texttt{Height}$. (Zde je možno s výhodou zvolit vhodně parametr `offset`. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

medExpenses Vyšetřete závislost rodinných medicínských výdajů (vztažených na osobu – `expenses`) na počtu členů domácnosti (`family size`). Vystačíme s lineární závislostí na počtu členů? Porovnejte délku 90% predikčních intervalů pro tříčlennou a pětičlennou domácnost. Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.

DAAG Knihovna obsahuje data ze cvičení a příkladů knihy Maindonald, J.H. and Braun, W.J. (2003, 2007) nazvané Data Analysis and Graphics Using R. Kromě toho jsou zde různé diagnostické funkce.

ais Popište závislost váhy (`wt`) na výšce (`ht`), když vezmete v úvahu také pohlaví (`sex`) a druh provozovaného sportu (`sport`). Pokuste se navrhnout vhodné transformace jedné či obou spojitých veličin. Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.

leaftemp Vyšetřete závislost rozdílu teplot listu a vzduchu (`tempDiff`) na veličině `wapPress`, přihlédněte k úrovni kysličníku uhličitého (`CO2level`). Lze výsledné přímky považovat za rovnoběžné? Nezapomeňte ověřit, zda jsou splněny běžné předpoklady normálního lineárního modelu.

litters Vyšetřete závislost hmotnosti mozku myši (`brainwt`) na její celkové hmotnosti (`bodywt`) s přihlédnutím k velikosti vrhu (`lsize`). Rozhodněte, zda lze závislost na velikosti vrhu považovat za lineární funkci této velikosti. Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.

faraway Knihovna doprovází zajímavý text o lineárních modelech umístěný na `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`.

mammalsleep Vyjádřete závislost celkové denní doby spánku (`sleep`) na délce březosti (`gestation`), přičemž přihlédnete k proměnné `danger`. Stačí veličinu `danger` použít jako číselnou proměnnou nebo je třeba ji uvažovat jako faktor? Je citlivost doby spánku na změny délky březosti pro všechny hodnoty proměnné `danger`. Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.