

NSTP194 Regrese: zadání samostatně řešených úloh (písemná část zkoušky, zadání ze 4. prosince 2008)

V tomto textu je uvedeno 30 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")`.

I když ve všech úlohách je třeba ověřit splnění běžných předpokladů normálního lineárního modelu, může se stát, že se úlohu nepodaří převést na model, v němž bychom mohli splnění těchto požadavků předpokládat. Pak je třeba uvést i příslušné vysvětlení.

Řešení úlohy napsané podobným způsobem jak má vypadat zápočtová práce pošlete **nejpozději** do 12. hodiny dva dny před dnem zkoušky elektronickou poštou jako přílohu na adresu `Karel.Zvara@mff.cuni.cz`. Doručení Vašeho řešení potvrdím emailem. Předpokládám, že v LaTeXu či 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 dispozici jen MS Word 2003, případně protějšek z Open Office. Případné problémy můžeme spolu vyřešit v mých konzultačních hodinách (končí s koncem přednáškového období) nebo po vzájemné dohodě i jindy. Mnohdy stačí pouhá elektronická komunikace. Počítejte však s tím, že ve zkuškovém období mám dost 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? Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

Cushings Pro nejčtenější 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.

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 vitamínu C (`VitC`) na hmotnosti zelných hlávek (`HeadWt`). Je třeba brát v úvahu také druh zelí (`Cult`) a dobu výsadby (`Date`)? 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 vychovááno 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.

Chiot Na základě historického materiálu byly jednotlivé územní jednotky v Rumunsku zhodnoceny s ohledem na selské povstání. Pokuste se předpovědět sílu rebelie (**intensity**) pomocí podílu obdělávaných 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.

Davis Rozhodněte o závislosti váhy, jak ji udává vyšetřovaná osoba (**repwt**) na skutečně zjištěné váze (**weight**). Souvisí (**repwt**) s pohlavím osoby? Opravte případně zjištěné hrubé chyby v datech. Podobně se vyjádřete k případné závislosti u výšky. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu. Při problémech s normalitou můžete ověřit p-hodnoty pomocí permutací hodnot regresorů.

Prestige Vyjádřete závislost prestiže povolání (**prestige**) na odpovídajícím příjmu (**income**) a typu povolání (**type**). Zjistěte, u kterého typu je citlivost vůči příjmu nejmenší. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

States Pokuste se vysvětlit **SATV** (verbální složka testu SAT) jako funkci proměnných **pay**, **dollars**. Lze považovat vliv dvou zmíněných regresorů za aditivní? Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

UN Použijte jen řádky bez chybějících pozorování. Navrhněte vhodnou transformaci pro vyjádření závislosti novorozenecké úmrtnosti

(`infant.mortality`) na velikosti HDP (`gdp`). Pokuste se prokázat, že novorozenecká úmrtnost je v ČR prokazatelně menší, než by odpovídalo jejímu hrubému domácímu produktu. Ve výsledném modelu tolerujte problém ve špičatosti rozdělení náhodné složky a ověřte ostatní běžné předpoklady normálního lineárního modelu.

lmtest Knihovna `lmtest` je zaměřena především na nejrůznější postupy regresní diagnostiky.

Mandible Použijte jen údaje splňující `age <= 28`. Vysvětlete závislost velikosti dolní čelisti (`length`) na stáří plodu (`age`). Zvažte potřebu transformace závisle proměnné, u nezávisle proměnné lze použít polynom druhého stupně. 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.

attenu V datovém souboru najděte událost (`event`), která má měření z největšího počtu stanic. Ověřte, že pak máte z každé stanice nejvýše jediné měření. Navrhněte vhodnou transformaci pro vyjádření závislosti (`accel`) na vzdálenosti zemětřesení od stanice (`dist`). Snažte se při tom, aby byly splněny běžné předpoklady normálního lineárního modelu.

morley Rozhodněte, zda v klasickém pokusu učení rychlosti světla je rozdíl mezi jednotlivými pokusy resp. zda v rámci pokusu záleží na pořadí jednotlivých měření. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

mtcars Navrhněte vhodnou funkci spotřeby paliva (`mpg`) tak, aby byla co nejlépe vysvětlitelná obsahem válců motoru (`disp`), přičemž je možno přihlédnout také k počtu válců (`cyl`). Další údaje neberte v úvahu. Přihlédněte přitom k tomu, 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 dispozici 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 $c \cdot \text{Girth}^2 * \text{Height}$. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

warpbreaks Data (snad?) udávají, jak často na jednotku délky vyrobené přize dojde k jejímu přetržení (**breaks**) v závislosti na kvalitě vlny (**wool**) a zvoleném napětí (**tension**). Rozhodněte, zda na těchto faktorech opravdu závisí, zvolte při tom vhodnou transformaci proměnné (**tension**). 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.

medExpenses Vyšetřete závislost rodinných medicínských výdajů (vztažených na osobu – **expenses**) na počtu členů domácnosti (**family size**). Vystačíme s lineární závislostí na počtu členů? Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.

toycars V jednoduchém pokusu byla opakovaně zjišťována vzdálenost (**distance**), v jaké se zastaví autíčko, které bylo spuštěno z rampy se sklonem (**angle**). Při modelování závislosti této vzdálenosti na úhlu rampy rozlišujte také tři použití autíčka (**car**). Je mezi autíčky průkazný rozdíl? Pokud ano, popište jej. 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>.

cahedral Vyšetřete závislost výšky hlavní lodi katedrály (**x**) na její délce (**y**). Vezměte v úvahu rozlišení románských a gotických kostelů uvedené v proměnné (**style**). Zvažte přitom možnost transformace. Je rozdíl mezi styly průkazný? Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.

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? Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.

nepali Nejprve z datového souboru náhodně vyberte 100 dětí starších 10 let. Vyjádřete váhu dětí (**wt**) jako funkci jejich věku (**age**) s přihlédnutím k pohlaví (**sex**) a ke gramotnosti matky (**lit**). Je třeba k těmto dvěma faktorům přihlížet? Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.