NSTP194 Regrese: zadání samostatně řešených úloh
(písemná část zkoušky, zadání z 24. října 2012)

V tomto textu je uvedeno 42 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 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 (případně v 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), emailem nebo po vzájemné dohodě 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. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu. Nakonec vynechaná pozorování porovnejte s očekáváním vyplývajícím z výsledného modelu bez nich.

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. V obou modelech spočítejte pro průměrné hodnoty obou regresorů bodový a intervalový odhad hodnoty regresní funkce a komentujte jejich porovnání.

anorexia Navrhněte porovnání trojích ošetření (léčení anorexie). Nejprve se rozhodněte pro vhodnou závisle proměnnou: výsledná váha nebo její přírůstek? 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. Nepřehlédněte, že proměnná `race` udává hodnoty v nominálním měřítku!

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.

Angell Data se vztahují k 43 americkým městům v polovině minulého století. Města jsou rozdělena do čtyř oblastí (`region`). Je třeba rozhodnout, zda se tyto oblasti mezi sebou liší v proměnné `moral` a zda se mezi sebou oblasti liší v případě, že data adjustujeme vůči proměnným `hetero, mobility`. Je třeba také ověřit nakolik jsou v použitých modelech splněny běžné předpoklady normálního lineárního modelu.

Anscombe Vysvětlete chování proměnné `education` chováním ostatních proměnných datového souboru. Zvažte vhodnost transformace vysvětlované proměnné. Zhodnoťte nakolik výsledný model splňuje běžné předpoklady normálního lineárního modelu. Zhodnoťte vliv údajů o Aljašce na odhad regresních koeficientů,

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.

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.

Freedman Vyjádřete závislost kriminality `crime` na procentu nebělochů `nonwhite`. Zvažte potřebu vhodné transformace některé z proměnných. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu. Určete predikční 95%interval pro `nonwhite` = 3,3 %.

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 potřeby navrhněte vhodnou transformaci. Je citlivost vůči prestiži u všech typů povolání stejná?. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

Robey Je rozdíl mezi čtyřmi částmi světa (`region`) v úhrnné plodnosti žen? Je rozdíl v úhrnné plodnosti žen mezi `tfr` čtyřmi částmi světa, když nejprve vezmeme v úvahu rozšíření antikoncepce
(`contraceptors`)? Ověřte splnění předpokladů normálního lineárního modelu.

Salaries Liší se mezi sebou co do příjmu (`salary`) muži a ženy? Zůstane tato odlišnost zachována, když vezmeme v úvahu nejdřív

dobu od získání doktorátu (`yrs.since.phd`) a postavení v akademické obci (`rank`)? Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

States Vyšetřete závislost verbální složky testu SAT `SATV` na složce matematické `SATM`. Zlepší se predikce, když přihlédneme k různým regionům? Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu.

UN Vyšetřete závislost kojenecké úmrtnosti (`infant.mortality`) na HDP. Navrhněte transformace vedoucí k lineární závislosti. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu. Zjistěte, nakolik údaj o České republice odpovídá navrženému modelu.

lmtest Knihovna `lmtest` obsahuje řadu testů týkajících se normálního lineárního modelu.

Mandible Použijte pouze hodnoty splňující `age` ≤ 28. Vyšetřete závislost proměnné `length` na na `age` a nezapomeňte při tom uvažovat o vhodných transformacích či polynomech. Ověřte, nakolik jsou splněny běžné předpoklady normálního lineárního modelu. Nakreslete predikční pás kolem regresní funkce a vysvětlete jeho interpretaci. Zjistěte kolik pozorování leží mimo tento pás.

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

attenu Data popisují řadu zemětřesení v Kalifornii. My se omezíme na zeměřesení (`event`) číslo 19 a pokusíme se pomocí `dist` předpovídat hodnotu `accel`. Doporučuji nejprve najít transformaci nebo transformace, které závislost pokud možno linearizují. Dále je třeba ověřit splnění běžných předpokladů normálního lineárního modelu.

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 Po načtení dat nejprve upravte názvy sloupců matice `state.x77` tak, aby v nich nebyly mezery. Připravte pak datový soubor

příkazem `state=data.frame(state.x77)`. Vyšetřete závislost střední délky života (`LifeExp`) na dvojici charakteristik jednotlivých států: `Murder`,
`HSGrad`. Vedle obligátního ověření běžných předpokladů normálního lineárního modelu se pokuste zjistit, který ze dvou regresorů ovlivňuje dobu života více.

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 $c \cdot$ `Girth`$^2 *$ `Height`. (Zde je možno s výhodou zvolit vhodně parametr `offset`. Ověřte, nakolik 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.

cuckoos Vyšetřete závislost délky kukaččího vejce na jeho šířce. Je třeba rozlišovat jednotlivé druhy kukaček? Pokud ano, pokuste se roztřídit druhy do dvou skupin tak, abyste ztratili co nejméně informace o tvaru zkoumané závislosti.

ironslag Datový soubor obsahuje měření obsahu železa dvěma způsoby, přičemž chemická metoda je náročnější. Určete závislost proměnné `magnetic` na `chemical` a zhodnoťte nakolik jsou splněny běžné předpoklady normálního lineárního modelu. Graficky znázorněte průběh regresní funkce a pás kolem přímky spojující krajní body **predikčních** intervalů.

leaftemp Vyšetřete závislost rozdílu teplot listu a vzduchu (`tempDiff`) na veličině `vapPress`, 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? Porovnejte délku 90% predikčních intervalů pro tříčlennou a pětičlennou domácnost. 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`.

cathedral Porovnejte výšky katedrál `x` ve dvou skupinách podle proměnné `style`. Bude rozdíl průkazný, když přihlédnete k délkám staveb `y`? Graficky znázorněte pásy spolehlivosti regresních přímek zvlášť pro každou kategorii katedrál.

diabetes Pkuste se vysvětlit závislost obvodu pasu (`waist`) na výšce (`height`) a váze (`weight`). Je třeba brát v úvahu také pohlaví osoby?. Zvažte potřebu transformací. Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.

chicago Datový soubor obsahuje informace o jednotlivých částech města. Pokuste se vysvětlit počet požárů na 100 bytů (`fire`) pomocí podílu minorit (`race`) a podílu domů postavených před rokem 1939 (`age`). Zvažte potřebu transformace některých proměnných. Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu. Která z nezávisle proměnných více ovlivňuje počet požárů?

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

nepali nejprve náhodně vyberte z datového souboru 100 dětí starších 10 let. Vyjádřete váhu dětí (`wt`) jako funkci jejich věku (`age`) s přihlédnutím k pohlaví (`sex`). Je třeba k oběma těmto faktorům přihlížet? Ověřte, zda jsou splněny běžné předpoklady normálního lineárního modelu.

pima Data obsahují údaje o Indiánech kmene Pima. Zdá se, že numerické proměnné mají nulu jako chybějící pozorování, proto u těchto proměnných nahraďte nuly symbolem NA pro chybějící hodnoty. Vysvětlete hodnotu BMI jako funkci tloušťky trojhlavého svalu. Podle potřeby zvolte vhodnou transformaci, případně odstraňte odlehlé pozorování. Pro výsledný model ověřte splnění běžných předpokladů normálního lineárního modelu.

pipeline Data obsahují dvojí měření hloubky defektů: v polních podmínkách (`Field`) a v laboratoři (`Lab`). Dále je informace o dávce, v níž bylo měření provedeno (`Batch`). Informaci o dávce považujte za nenáhodnou a nezpůsobující stochastickou závislost mezi pozorováními. Pokuste se vyjádřit polní měření jako funkci měření laboratorního, zvolte při tom vhodné transformace, které pokud možno zachovají linearitu a stabilizují rozptyl. Pro výsledný model ověřte splnění běžných předpokladů normálního lineárního modelu.

psid Omezte se jen na pozorování z roku 1968 (proč?). Vyšetřete závislost příjmu () na věku v roce 1968 (`age`) nebo na délce školní docházky (`educ`) s případným přihlédnutím k pohlaví (`sex`). Výsledný model graficky znázorněte a ověřte splnění běžných předpokladů normálního lineárního modelu.

rats Najděte vhodný tvar závislosti doby přežívání (`time`) na použitém jedu (`poison`) a na zvoleném ošetření (`treat`). Pro výsledný model ověřte splnění běžných předpokladů normálního lineárního modelu.

sat Zjistěte, jak výdaje na jednoho studenta (`expend`) ovlivní výsledek matematické části testu SAT (`math`). Změní se tato závislost, když přihlédneme k tomu, jaká část studentů (`takers`) testy SAT vlastně skládá? Nezapomeňta na možnost hledání vhodné transformace. Interpretujte svoje zjištění. Pro výsledný model ověřte splnění běžných předpokladů normálního lineárního modelu.

twins Data udávají hodnoty IQ jednovaječných dvojčat, z nichž jedno bylo vychováváno biologickými rodiči (`Biological`) a jedno pěstouny (`Foster`). Pokuste se vysvětlit IQ dvojčete vychovávaného pěstouny pomocí IQ dvojčete u biologických rodičů. Je možno přihlédnout také k sociálnímu postavení pěstounů (`Social`). Jak dopadne hypotéza, že koeficient u proměnné (`Biological`) je roven jedné? Pro výsledný model ověřte splnění běžných předpokladů normálního lineárního modelu.

Zájemci o úlohy z regrese

Animals (MASS)	Cushings (MASS)	Pima.tr (MASS)
Rubber (MASS)	anorexia (MASS)	birthwt (MASS)
cabbages (MASS)	cats (MASS)	crabs (MASS)
Angell (car)	Anscombe (car)	Brt (car)
Davis (car)	Freedman (car)	Chirot (car)
Prestige (car)	Robey (car)	Salaries (car)
States (car)	UN (car)	Mandible (lmtest)
attenu (datasets)	mtcars (datasets)	state (datasets)
trees (datasets)	ais (DAAG)	cuckoos (DAAG)
ironslag (DAAG)	leaftemp (DAAG)	litters (DAAG)
medExpenses (DAAG)	cathedral (faraway)	diabetes (faraway)
chicago (faraway)	mammalsleep (faraway)	nepali (faraway)
pima (faraway)	pipeline (faraway)	psid (faraway)
rats (faraway)	sat (faraway)	twins (faraway)

Řešení pošlete na adresu `zvara@karlin.mff.cuni.cz` nejpozději dva dny před dnem zkoušky, do 12:00 hodin.

24. října 2012, Karel Zvára