Zápočtová úloha z 01REAN, 7. 12. 2017

Popis datového souboru Boston Housing

Datový soubor `Boston` je obsažen v balíku `MASS` a lze použít rovnou po načtení příslušné knihovny.

```
library(MASS)
fix(Boston)
head(Boston)
? Boston
```

Obsahuje celkem 506 záznamů z obcí v předměstí města Boston, MA, USA a data pocházejí ze studie v roce 1978.
Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5, 81–102.

Základní charakteristiky ohledně jednotlivých proměnných získáte pomocí funkcí `str(Boston)` a `summary(Boston)`.

Data celkem obsahují 14 proměnných, přičemž naším cílem je prozkoumat vliv 13 z nich na ceny nemovitostí `medv`. Přičemž anglický popis jednotlivých proměnných (sloupců) je následující:

Feature	Description
`crim`	per capita crime rate by town
`zn`	proportion of residential land zoned for lots over 25,000 sq.ft
`indus`	proportion of non-retail business acres per town
`chas`	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
`nox`	nitrogen oxides concentration (parts per 10 million)
`rm`	average number of rooms per dwelling
`age`	proportion of owner-occupied units built prior to 1940
`dis`	weighted mean of distances to five Boston employment centres
`rad`	index of accessibility to radial highways
`tax`	full-value property-tax rate per $10,000
`ptratio`	pupil-teacher ratio by town
`black`	$1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
`lstat`	lower status of the population (percent)
`medv`	median value of owner-occupied homes in $1000s

Pro načtení a nainstalování potřebných balíků můžete použít tento kód:

```
load.libraries <- c('data.table','car','MASS','ggplot2','ISLR','graphics','effects','lattice')
install.lib <- load.libraries[!load.libraries %in% installed.packages()]
for(libs in install.lib) install.packages(libs, dependences = TRUE)
sapply(load.libraries, require, character = TRUE)
```

Požadavky k vypracování a odevzdání

Zápočtovou úlohu vypracujte a odevzdejte samostatně. V případě konzultace a spolupráce s kolegy, uveďte u dané otázky s kým jste na daném řešení spolupracovali. Vypracované řešení odešlete e-mailem na adresu jiri.franc@fjfi.cvut.cz. Vždy ve zprávě uveďte číslo otázky a také příslušný výpis z funkce (například `summary(model)`), nebo obrázek, aby bylo jasné na základě čeho je odpověď formulována. Dostačující formát, je dobře okomentovaný R kód, ale samozřejmě hezký a upravený protokol v LaTeXu je vítán.
Odevzdání zprávy a získání zápočtu je podmínkou pro získání zkoušky.

Zadání

Vypracujte následující body zadání a zodpovězte příslušné otázky:

Průzkumová a grafická část:

Q01: Zjistěte, zdali data neobsahují chybějící hodnoty (`NA`), pokud ano tak příslušná pozorování z dat odstraňte. Ověřte rozměry datového souboru a shrňte základní popisné charakteristiky všech proměnných.

Q02: Vykreslete histogram a odhad hustoty pro odezvu `medv`.

Q03: Pro proměnné `crim, nox, rm, lstat, ptratio, dis` vykreslete `scatterplot` - závislost dané proměnné na odezvě a proložte body jak lineárním odhadem tak vyhlazenou křivkou (`lines(lowess(X,Y))`).

Q04: Pro proměnné `chas` a `rad` a jejich vztah k odezvě vykreslete krabicové diagramy (boxploty). Proměnnou `rad` transformujte tak, aby obsahovala pouze dvě úrovně (levely) a vykreslete opět krabicový diagram.

Q05: Navrhněte další zobrazení datového souboru. Proveďte ho a popište jeho účel.

Regresní model závislosti mediánu ceny nemovitosti na míře kriminality:

Q06: Sestavte jednoduchý regresní model a na jeho základech zjistěte zdali kriminalita v okolí ovlivňuje cenu nemovitostí určených k bydlení. Pokud ano, o kolik je cena nemovitostí nižší v závislosti na míře kriminality? Ověřte

předpoklady pro použití lineárního modelu (validujte např. symetrii a normalitu residui) a diskutujte výstup.

Q07: Vyzkoušejte model s logaritmickou transformací odezvy. Vykreslete optimální log-věrohodnostní profil u Box-Coxovy transformace a porovnejte navrženou transformaci s provedenou logaritmickou.

Q08: Z předchozího modelu vyčtěte procentuální navýšení/pokles ceny nemovitostí při změně míry kriminality o jeden stupeň (odpověď typu: cena nemovitosti v průměru klesne o ???% při nárůstu míry kriminality o 1 jednotku). Využijte znalosti, že

$$\begin{aligned} \log(Y) &= \beta_0 + \beta_1 X + \varepsilon \\ Y &= e^{\beta_0+\beta_1 X} e^{\varepsilon} \\ \mathbb{E}\left[Y|X=x\right] &= e^{\beta_0+\beta_1 x}\mathbb{E}\left[e^{\varepsilon}\right] \\ e^{\beta_1} &= \frac{\mathbb{E}\left[Y|X=x+1\right]}{\mathbb{E}\left[Y|X=x\right]} \end{aligned}$$

Q09: Zkuste transformovat proměnnou `crim`. Vyzkoušejte například po částech konstantní transformaci, lineární transformaci, splines a polynomiální transformaci (kvadratickou a kubickou). Zkuste využít informací získaných například z `crPlots(model)`.

Q10: Vykreslete scatterplot predikovaných cen nemovitostí na základě vybraného modelu, proložte skrze data odhadnutou regresní přímku a vykreslete efekty pomocí `plot(allEffects(model))`. Validujte výsledný model pomocí příslušných testů na rezidua a pomocí příslušných obrázků (QQplot, residua vs. fitted, atd.)

Vícerozměrný regresní model:

Q11: Medián ceny nemovitostí je spojitá proměnná, vypište tabulku četností jednotlivých hodnot. Diskutujte zdali některé hodnoty nejsou způsobeny zaokrouhlením, useknutím a podobně. Měření která považujete z tohoto pohledu za nedůvěryhodná, případně za outliery odstraňte.

Q12: Zkonstruujte lineární model s logaritmicky transformovanou odezvou `medv` a všemi nezávislými proměnnými, které máte k dispozici. Na základě kritérií jako jsou `AIC`, `BIC`, R^2, `F`, `atd`. Vyberte nejvhodnější model. Ten validujte a okomentujte jeho výběr.

Q13: Zkoumejte případnou multikolinearitu. Spočtěte korelace mezi jednotlivými proměnnými, porovnejte s vaším výběrem a pomocí VIF a dalších nástrojů validujte váš výběr.

Q14: Pokud ve vašem výsledném modelu máte zahrnutou kriminalitu (proměnnou `crim`) porovnejte jak se změnil vliv kriminality na medián ceny nemovitostí. Jaké je snížení průměrné ceny nemovitostí při vzrůstu kriminality o jednu jednotku? Pokud `crim` v modelu nemáte tak ji pro tuto otázku do modelu přiřaďte.

Q15: Prezentujte váš výsledný model pro predikci `medv`, diskutujte výsledné parametry R^2 a σ tohoto modelu. Validujte model (jak graficky, tak pomocí příslušných testů hypotéz).

Kam dál?

Q16: Diskutujte jak by šlo případně zlepšit predikci, jaké transformace jednotlivých proměnných by mohli pomoci. Převedli byste některé spojité proměnné na diskrétní (na faktory)? Jaké další kroky byste při analýze navrhli?

Q17: Myslíte, že pokud bychom cíleně dokázali potlačit kriminalitu v daném městě, vedlo by to ke zvýšení cen nemovitostí určených k bydlení v dané lokalitě?