

2. zápočtová úloha z 01RAD

Jiří Franc

2020-11-24

2. zápočtová úloha z 01RAD

Popis úlohy

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

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

Obsahuje celkem 506 záznamů z obcí v předměstí města Boston, MA, USA a data pocházejí ze studie v roce 1978. Viz 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 cenu nemovitostí `medv`. Přičemž anglický popis jednotlivých proměnných (sloupců) je následující:

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

Podmínky a body

Úkol i protokol vypracujte samostatně. Pokud na řešení nějaké úlohy budete přesto s někým spolupracovat, radit se, nezapomeňte to u odpovědi na danou otázku uvést. Tato zápočtová úloha obsahuje 10 otázek po 1 bodu. Celkem za 3 zápočtové úlohy bude možné získat 30 bodů, přičemž pro získání zápočtu je potřeba 20. Další dodatečné body mohou případně individuálně udělit za řešení mini domácích úkolů z jednotlivých hodin.

Odevzdání

Protokol ve formátu pdf (včetně příslušného Rmd souboru) odevzdejte prostřednictvím MS Teams, nejpozději do půlnoci 14. 12. 2020 (tj. za cca 3 týdny).

Průzkumová a grafická část::

- Otázka 01

Zjistěte, zdali data neobsahují chybějící hodnoty, ověřte rozměry datového souboru a shrňte základní popisné charakteristiky všech proměnných. Vykreslete histogram a odhad hustoty pro odezvu `medv`. Medián ceny nemovitostí je spojitá proměnná, zkontrolujte tabulku četností jednotlivých hodnot. Diskutujte zdali některé hodnoty nejsou způsobeny zaokrouhlením, useknutím a podobně. Měření která považujete z tohoto pohledu za nedůvěryhodná odstraňte. Co to znamená z pohledu modelu odezvy `medv`?

Regresní model závislosti mediánu ceny nemovitosti na znečištění v okolí nemovitosti

- Otázka 2

Sestavte jednoduchý regresní model a na jeho základech zjistěte zdali znečištění `nox` v okolí ovlivňuje cenu nemovitostí určených k bydlení. Pokud ano, o kolik je cena nemovitostí nižší v závislosti na míře znečištění?

- Otázka 3

Vyzkoušejte model s mocninou a logaritmickou transformací odezvy. Pro výběr mocninové transformace vykreslete optimální log-věrohodnostní profil u Box-Coxovy transformace a porovnejte navrženou transformaci s provedenou logaritmickou.

- Otázka 4

Z předchozího modelu vyčtěte procentuální navýšení/pokles ceny nemovitostí při změně míry znečištění o jeden stupeň (odpověď typu: Střední cena nemovitostí v lokalitách okolo Bostonu, liších se koncentrací NO_x , klesá/roste zhruba o XX% na každou 1 jednotku nárůstu koncentrace NO_x %).

- Otázka 5

Zachovejte logaritmickou transformaci odezvy a zkuste transformovat i nezávislou proměnnou `nox`. Vyzkoušejte například po částech konstantní transformaci, splines a polynomiální transformace (kvadratickou a kubickou). Zkuste využít informací získaných například z `crPlots(model)`. Lze některé z těchto modelů testovat mezi sebou F-testem? Pokud ano, proveďte a diskutujte.

- Otázka 6

Vyberte jeden z předešlých modelů, zdůvodněte jeho výběr a validujte ho pomocí příslušných testů hypotéz na rezidua (normalita, homoscedasticita, ...) a pomocí příslušných obrázků (QQplot, residua vs. fitted, atd.)

Vícerozměrný regresní model

- Otázka 7

Zkonstruuje lineární model s logaritmicky transformovanou odezvou `medv` a zkuste najít vztah mezi cenou a dalšími nezávislými proměnnými, které máte k dispozici (stačí aditivní model bez interakcí). Na základě kritérií jako jsou AIC, BIC, R^2 , F, atd. vyberte podle vás nejvhodnější model. Lze vztah mezi `nox` a `medv`, pokud existuje, vysvětlit pomocí jiných proměnných? Tj, že například ve znečištěných oblastech bydlí chudší lidé a je tam větší kriminalita atd.?

- Otázka 8

Použijte ve výsledném modelu znečištění (proměnnou `nox`) a porovnejte jak se změnil vliv znečištění na medián ceny nemovitostí oproti jednoduchému regresnímu modelu s log transformovanou odezvou (viz otázka 4). Jaké je snížení průměrné ceny nemovitostí při vzrůstu znečištění o jednu jednotku? Pokud proměnnou `nox` v modelu nemáte tak ji pro tuto otázku do modelu přiřaďte.

- Otázka 9

Prezentujte váš výsledný model pro predikci `medv`, diskutujte výsledné parametry R^2 , σ , F a porovnejte je s jednoduchým lin. modelem z otázky 6. Jak se změnila a dala se tato změna očekávat? Validujte model (jak graficky, tak pomocí příslušných testů hypotéz).

- Otázka 10

Na základě vašeho modelu odpovězte, zdali si myslíte, že pokud bychom dokázali zlepšit kvalitu ovzduší v dané lokalitě, vedlo by to ke zvýšení cen nemovitostí určených k bydlení v dané lokalitě?