2. zÃąpoÄĎtovÃą Ãžloha z 01RAD

Emanuel FrÃątrik

2021-12-13

1 2. zápočtová úloha z 01RAD

1.1 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
0.006	18	2.31	0	0.538	6.58	65.2	4.09	1	296	15.3	397	4.98	24.0
0.027	0	7.07	0	0.469	6.42	78.9	4.97	2	242	17.8	397	9.14	21.6
0.027	0	7.07	0	0.469	7.18	61.1	4.97	2	242	17.8	393	4.03	34.7
0.032	0	2.18	0	0.458	7.00	45.8	6.06	3	222	18.7	395	2.94	33.4
0.069	0	2.18	0	0.458	7.15	54.2	6.06	3	222	18.7	397	5.33	36.2
0.030	0	2.18	0	0.458	6.43	58.7	6.06	3	222	18.7	394	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
`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

Tabuľka 1: Deskriptívne štatistiky pre spojité (numerické) premenné

PREMENNÁ	PRIEMER	ROZPTYL	MEDIÁN	MIN	MAX	POČET NA
crim	3.614	7.40e+01	0.256	0.006	88.976	0
zn	11.364	5.44e+02	0.000	0.000	100.000	0
indus	11.137	4.71e+01	9.690	0.460	27.740	0
nox	0.555	1.30e-02	0.538	0.385	0.871	0
rm	6.285	4.94e-01	6.208	3.561	8.780	0
age	68.575	7.92e+02	77.500	2.900	100.000	0
dis	3.795	4.43e+00	3.207	1.130	12.126	0
tax	408.237	2.84e+04	330.000	187.000	711.000	0
ptratio	18.456	4.69e+00	19.050	12.600	22.000	0
black	356.674	8.33e+03	391.440	0.320	396.900	0
lstat	12.653	5.10e+01	11.360	1.730	37.970	0
medv	22.533	8.46e+01	21.200	5.000	50.000	0

1.2 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 více jak 20. Další dodatečné body mohu případně individuálně udělit za řešení mini domácích úkolů z jednotlivých hodin.

1.3 Odevzdání

Protokol ve formátu pdf (včetně příslušného Rmd souboru) odevzdejte prostřednictvím MS Teams, nejpozději do půlnoci 15. 12. 2021 (tj. za cca 3 týdny).

1.4 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`?

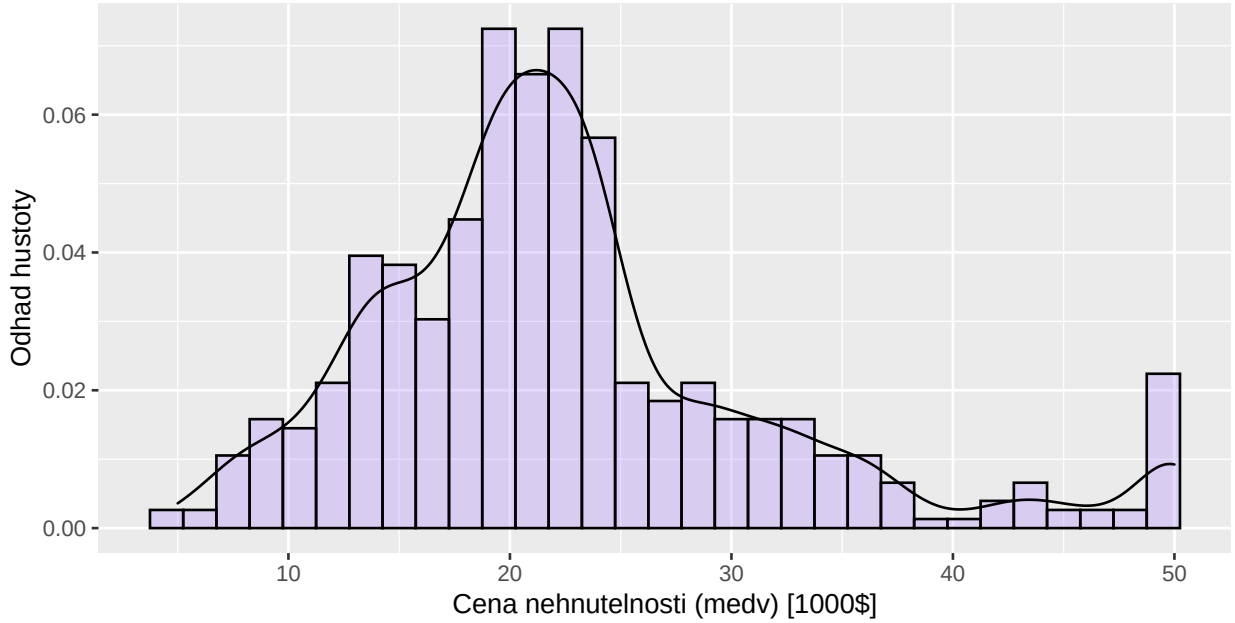
- Odpoveď 01

  Dataset `Boston` obsahuje 506 pozorovaní, popísaných štrnástimi premennými. V datasete sa nenachádzajú žiadne `NA` hodnoty. Prehľad popisných štatistík pre jednotlivé premenné sa nachádza v tabuľke 1. Premenná `chas` a premmená `rad` sú popísané diskrétnymi hodnotami a preto ich ďalej reprezentujeme faktorovými premennými. Odhad hustoty premennej `medv` je vyobrazený na obrázku 1. Ako môžme vidieť na tomto obrázku tak v okolí hodnoty `medv`=50 je náhly skok a po bližšej inšpekcii tejto premennej v sme zistili, že hodnota 50 je v datasete obsiahnutá až 16-krát. Keďže táto hodnota tvorí zároveň hraničnú hodnotu tak je možné, že vznikla zaokrúhlením všetkých vyšších cien na 50 čo spôsobilo jej zvýšenú častosť. Kvôli tomuto podozreniu je vhodné záznamy s hodnotou `medv`=50 vylúčiť.

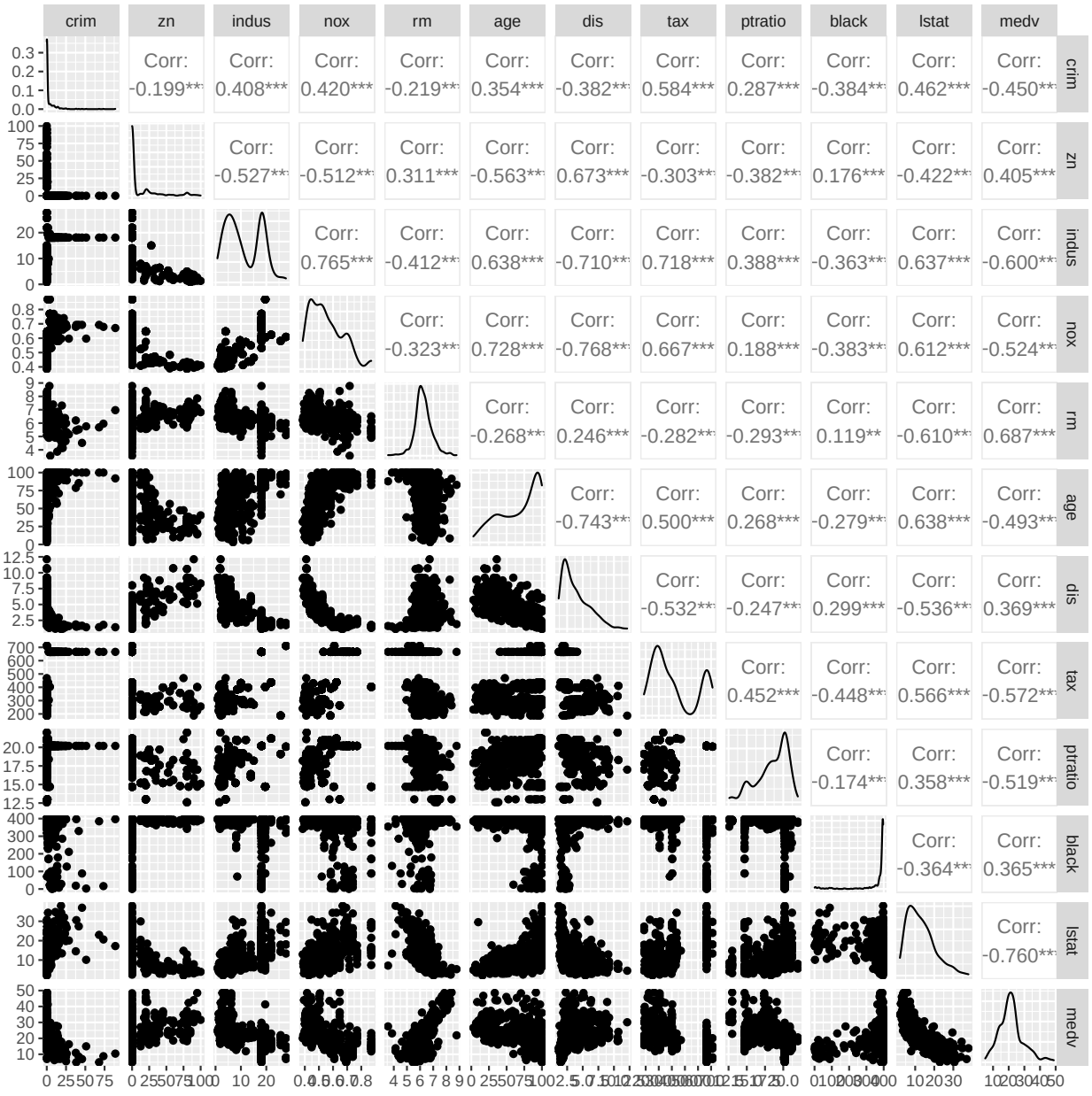
Tabuľka 2: Deskriptívne charakteristiky pre kategorické premenné

PREMENNÁ	N = 506
chas	
0	471 / 506 (93%)
1	35 / 506 (6.9%)
rad	
1	20 / 506 (4.0%)
2	24 / 506 (4.7%)
3	38 / 506 (7.5%)
4	110 / 506 (22%)
5	115 / 506 (23%)
6	26 / 506 (5.1%)
7	17 / 506 (3.4%)
8	24 / 506 (4.7%)
24	132 / 506 (26%)

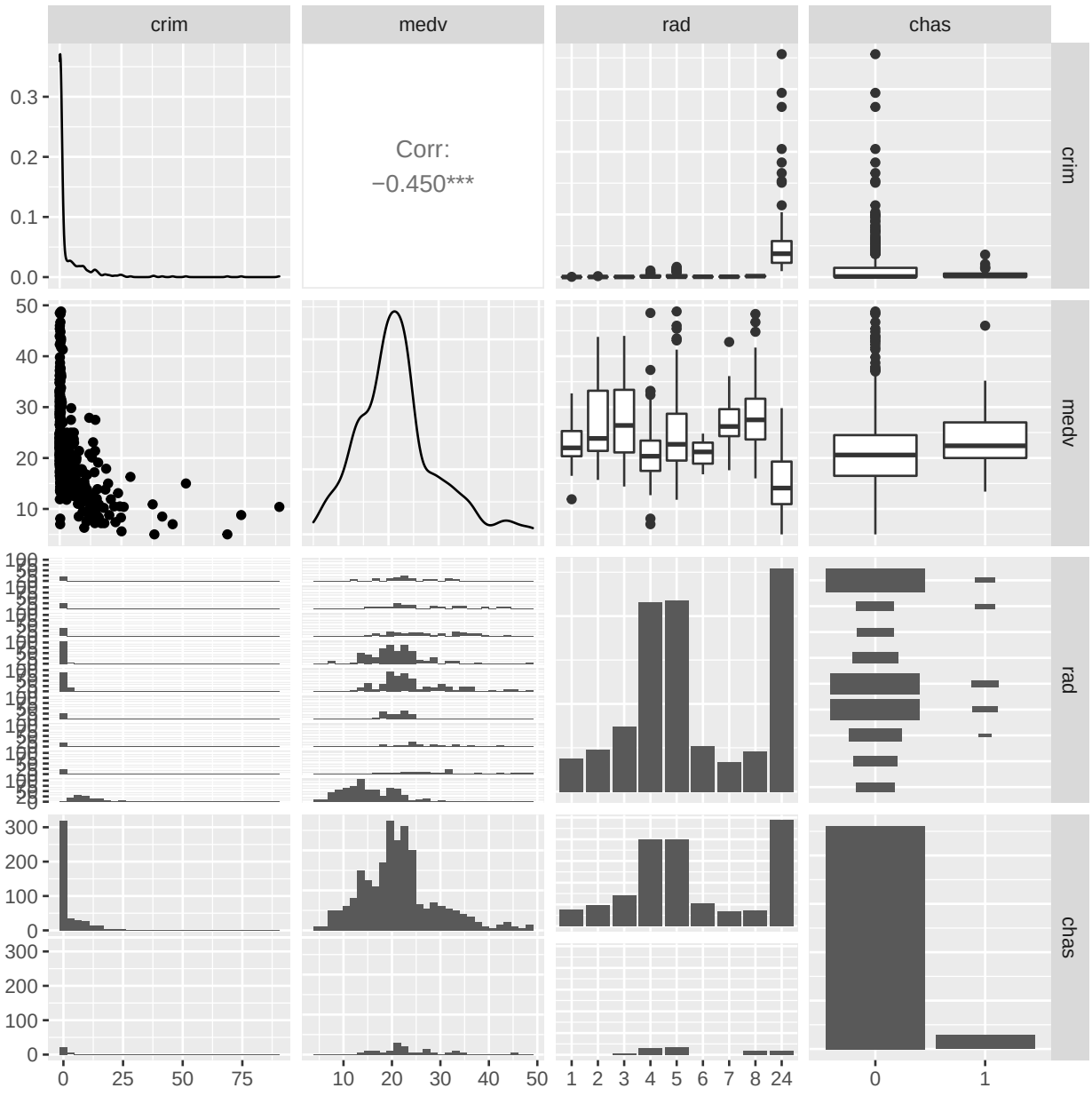
[1] n / N (%)



Obr. 1: Histogramový a jadrový odhad hustoty premennej `medv`



Obr. 2: Scatter plot medzi numerickými premennými



Obr. 3: Scatter plot medzi diskrétnymi premennými a premennými `crim` a `medv`

Tabuľka 3: Model závislosti ceny nehnuteľnosti (medv) na miere kriminality (crim)

	Lineárny model
(Intercept)	23.1147
	s.e. = 0.3443
	t-stat = 67.1432
	p-val = 0.0000
	[22.4383, 23.7911]
crim	−0.4059
	s.e. = 0.0365
	t-stat = −11.1352
	p-val = 0.0000
	[−0.4775, −0.3343]
Num.Obs.	490
R2	0.203
AIC	3305.8
BIC	3318.4
F	123.992
RMSE	7.02

1.5 Regresní model závislosti mediánu ceny nemovitosti na znečištění v okolí nemovitosti

- Otázka 2

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

- Odpoveď 2

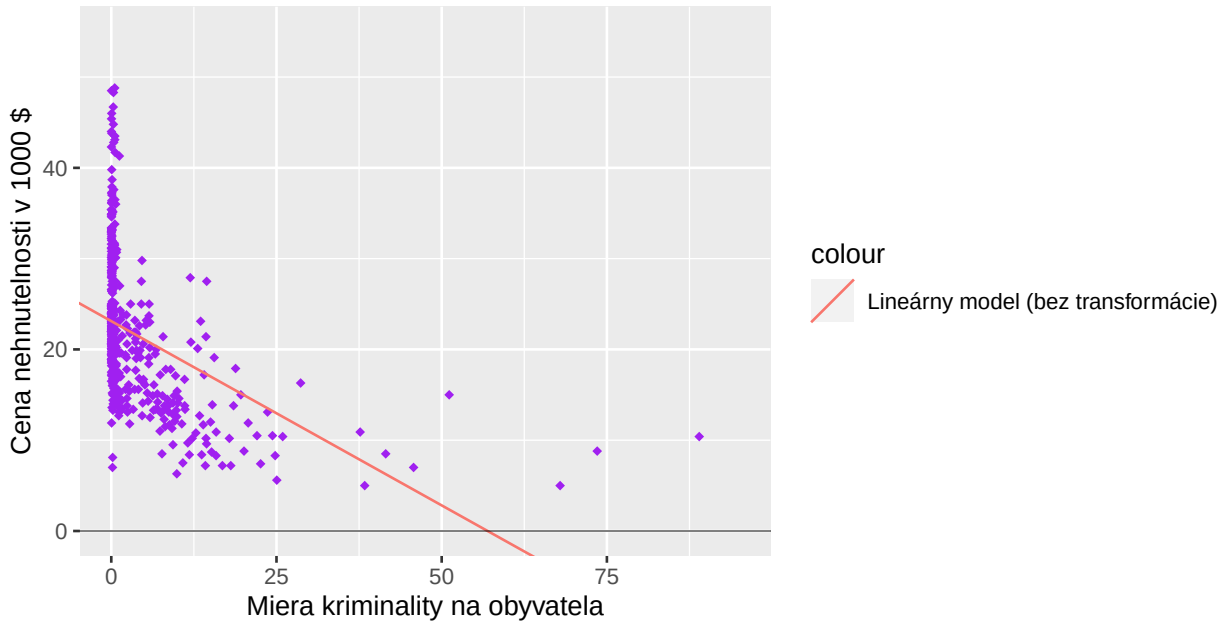
  Model závislosti ceny nehnuteľnosti (`medv`) na miere kriminality (`crim`) sa ukazuje byť štatisticky významný s p-hodnotou 8.115×10^{-26}. Podobne aj jednotlivé premenné sú štatisticky významné čo môžeme vidieť v tabuľke 3 spolu s hodnotou R^2=0.203. Na základe tohto modelu a za predpokladu splenenia všetkých predpokladov regresie môžme povedať, že pri zvýšení miery kriminality na jedného obyvateľa o jednotku sa očakávaná cena nehnuteľnosti zmenší o 405.896 $.

- Otázka 3

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

- Odpoveď 3

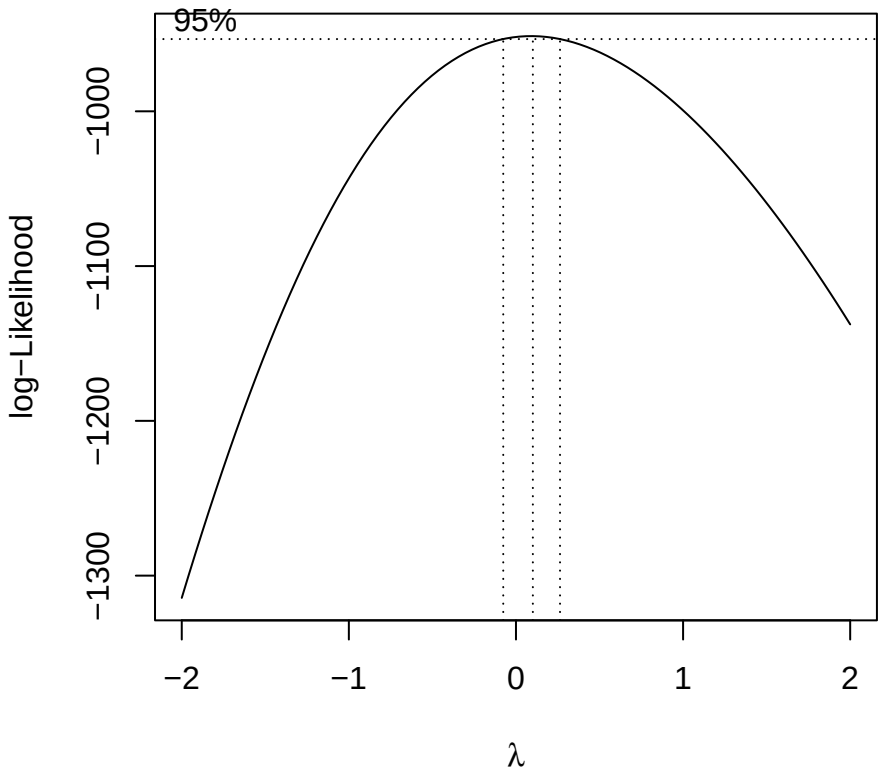
  V tabuľke 4 vidíme porovnanie jednotlivých modelov po transformáciach odozvy ako aj základný model. Použitie box-cox ako aj logaritmickej transformácie zvýšilo hodnotu R^2 štatistiky postupne o 10% a 12% oproti jednoduchému modelu bez transformácie odozvy. Avšak porovnanie R^2 štatistiky prípadne iných kritérii ako napr. `AIC` medzi týmito troma modelmi môže byť zavádzajúce nakoľko nelineárne transformácie zmenili škálu odozvy a teda aj reziduii. Takéto porovnávanie preto nemá zmysel. Oba modely po transformácii odozvy sú ako celky štatisticky významné uvažujúc hladinu 0.05. Na obrázku 5 vidíme profil log-likelihood pre odhad optimálneho parametra na boxcox transformáciu. Tento parameter bol odhadnutý a zaokrúhlený na 0.1 pričom ako je vidieť na spomínanom obrázku, 95% interval odhadu tohto parametra zahŕňa aj nulu. Výber parametra lambda na nulu je ekvivalentné logaritmickej transformácii. Výber práve logaritmickej transformácie oproti mocninnej s parametrom



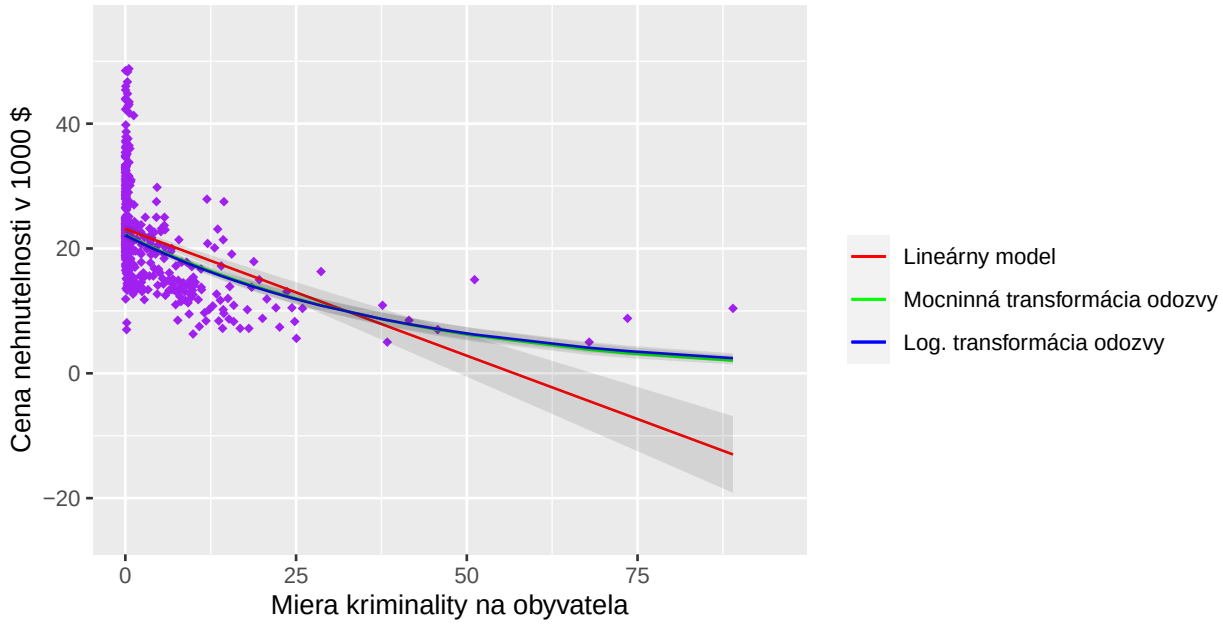
Obr. 4: Model závislosti ceny nehnuteľnosti (medv) na miere kriminality na jedného obyvateľa(crim)

Tabuľka 4: Porovnanie modelov

	Základ. model	Model po box-cox transf.	Model po log. transf.
(Intercept)	23.1147	3.6351	3.0964
	s.e. = 0.3443	s.e. = 0.0209	s.e. = 0.0155
	t-stat = 67.1432	t-stat = 174.2000	t-stat = 200.1876
	p-val = 0.0000	p-val = 0.0000	p-val = 0.0000
crim	−0.4059	−0.0326	−0.0248
	s.e. = 0.0365	s.e. = 0.0022	s.e. = 0.0016
	t-stat = −11.1352	t-stat = −14.7576	t-stat = −15.1692
	p-val = 0.0000	p-val = 0.0000	p-val = 0.0000
Num.Obs.	490	490	490
R2	0.203	0.309	0.320
AIC	3305.8	558.7	265.2
BIC	3318.4	571.3	277.8
F	123.992	217.786	230.103
RMSE	7.02	0.43	0.32



Obr. 5: Log-likelihood profil box-cox transformácie modelu



Obr. 6: Porovnanie modelov po transformácii odozvy a jednoduchého lineárneho modelu závislosti ceny nehnuteľnosti (medv) na miere kriminality na jedného obyvateľa (crim)

0.1 nám umožní jednoduchší spôsob ako vysvetliť význam premenných v našom modeli. Finálny výber jedného z týchto troch modelov by mal byť preto založený na fakte či transformácia odozvy pomohla splniť predpoklady o linearite modelu prípadne o normalite a homoskedasticite a nie na tom či zdanlivo zvýšila hodnotu R^2 štatistiky. Nakoniec na obrázku 6 môžme vidieť znázornenie priebehu modelov. Modely s transformovanou odozvou boli vypočítané na základe inverznej transformácie odozvy čo ale spôsobí, že krivky na obrázku neznázorňujú očakávanú hodnotu odozvy `medv` ale napr. v prípade logaritmickej transformácie medián odozvy `medv`.

- Otázka 4

  Z předchozího modelu vyčtěte procentuální navýšení/pokles ceny nemovitostí při změně míry znečištění o jeden stupeň (odpověď typu: Střední cena nemovitosti v lokalitách okolo Bostonu, liších se o počet kriminálních deliktů na 1000 obyvatel daného města, klesá/roste zhruba o XX% na každou 1 jednotku nárůstu kriminálních deliktů.

- Odpoveď 4

  Očakávaná cena nehnuteľností sa pri náraste miery kriminality na 1000 obyvateľov o jednotku podľa log-transformovaného modelu zníži o 0.002%.

- Otázka 5

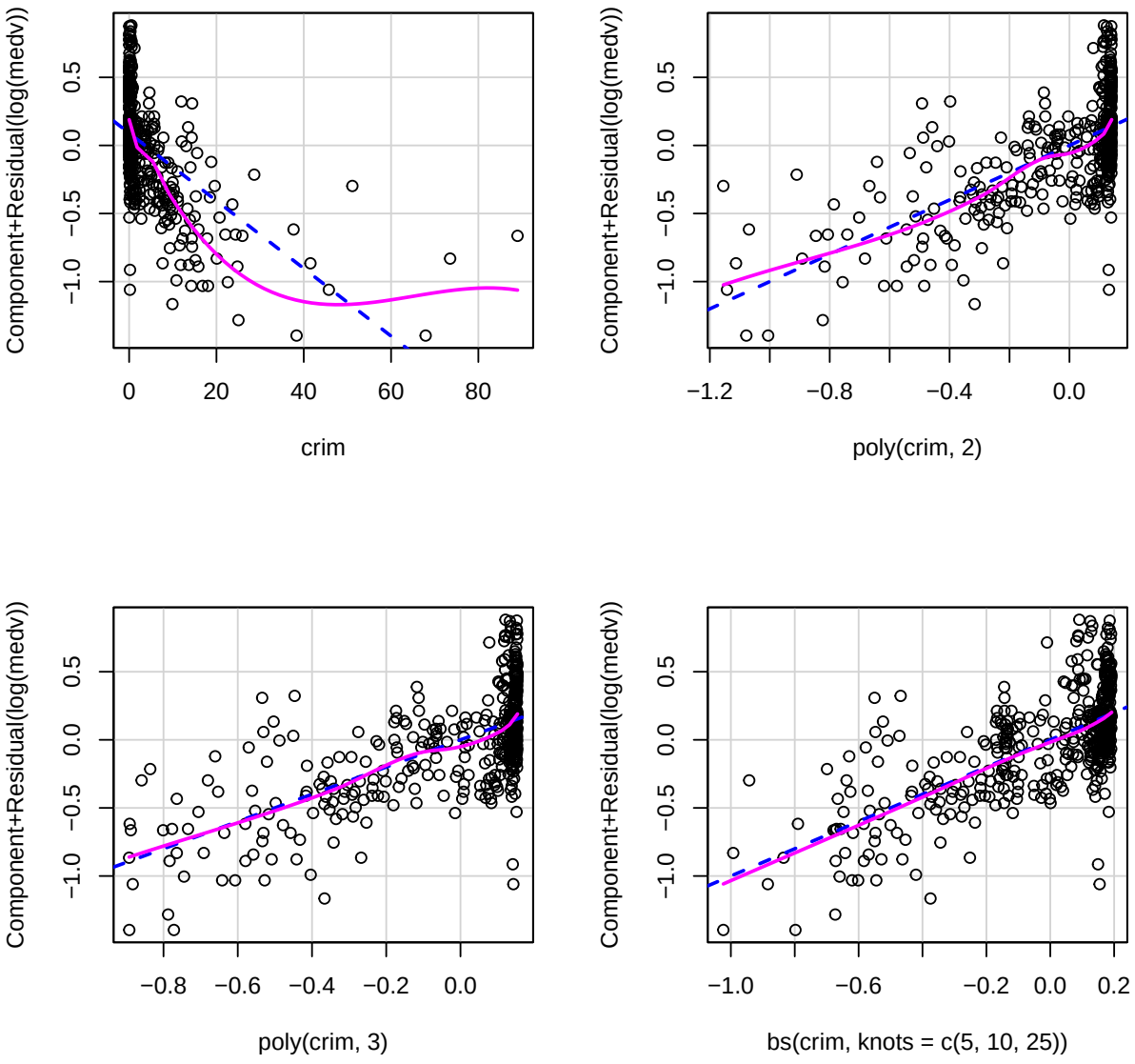
  Zachovejte logaritmickou trasformaci odezvy a zkuste transformovat i nezávislou proměnnou `crim`. Vyzkoušejte například po částech konstantní transformaci, splines a polynomiální transformace (kvadratickou a kubickou). Zkuste využít informací získaných například z `crPlots(model)`. Lze některé z těchto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

- Odpoveď 5

  Obrázok 7 zobrazuje čiastočné reziduálne grafy pre modely bez transformovaného regresoru `crim`, s kvadratickou a kubickou transformáciou a nakoniec s použitím splinov. Z týchto štyroch grafov sa dá vypozorovať, že kvadratická, kubická transformácia pomohli zlepšiť linearitu modelu. Model s použitím

Tabuľka 5: Porovnanie modelov

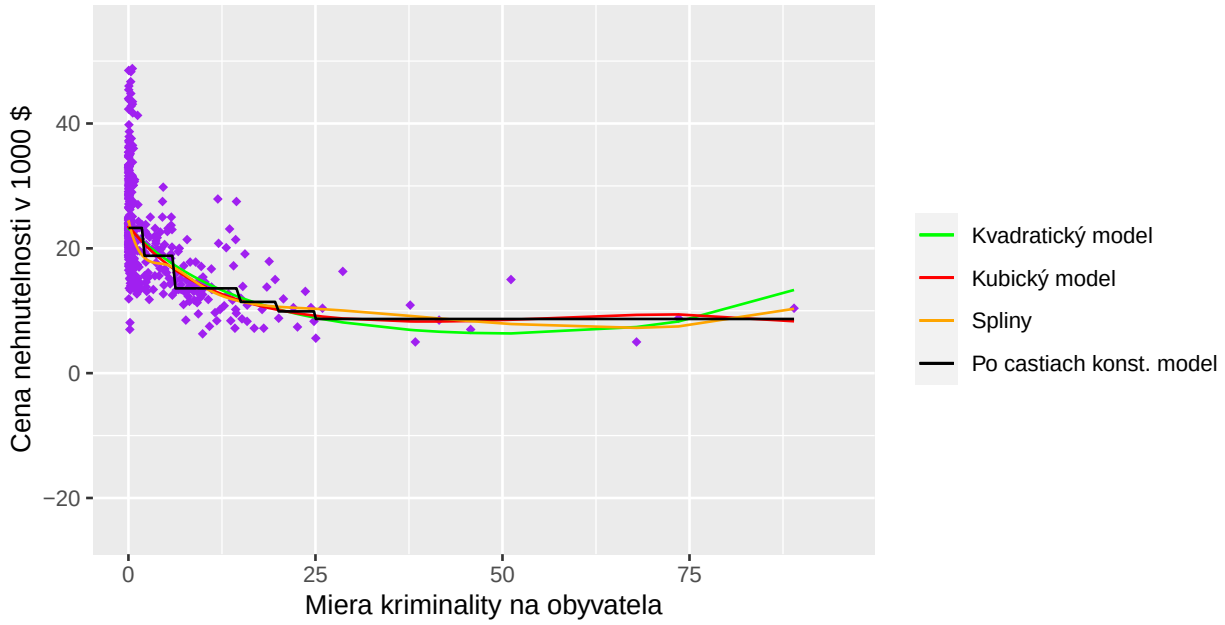
	Kvadratický	Kubický	Spliny
(Intercept)	3.0059	3.0059	3.1982
	s.e. = 0.0130	s.e. = 0.0129	s.e. = 0.0193
	p-val = 0.0000	p-val = 0.0000	p-val = 0.0000
poly(crim, 2)1	−4.7917		
	s.e. = 0.2887		
	p-val = 0.0000		
poly(crim, 2)2	2.8446		
	s.e. = 0.2887		
	p-val = 0.0000		
poly(crim, 3)1		−4.7917	
		s.e. = 0.2862	
		p-val = 0.0000	
poly(crim, 3)2		2.8446	
		s.e. = 0.2862	
		p-val = 0.0000	
poly(crim, 3)3		−0.8932	
		s.e. = 0.2862	
		p-val = 0.0019	
bs(crim, knots = c(5, 10, 25))1			−0.3712
			s.e. = 0.0890
			p-val = 0.0000
bs(crim, knots = c(5, 10, 25))2			−0.2551
			s.e. = 0.0759
			p-val = 0.0008
bs(crim, knots = c(5, 10, 25))3			−0.8304
			s.e. = 0.0955
			p-val = 0.0000
bs(crim, knots = c(5, 10, 25))4			−0.8517
			s.e. = 0.3258
			p-val = 0.0092
bs(crim, knots = c(5, 10, 25))5			−1.5922
			s.e. = 0.4748
			p-val = 0.0009
bs(crim, knots = c(5, 10, 25))6			−0.8651
			s.e. = 0.2760
			p-val = 0.0018
Num.Obs.	490	490	490
R2	0.433	0.444	0.459
AIC	178.2	170.4	163.7
BIC	194.9	191.4	197.2
F	186.225	129.626	68.236
RMSE	0.29	0.29	0.28



Obr. 7: Čiastočné reziduálne grafy

Tabuľka 6:

	Po častiach konštantný model
(Intercept)	3.1470
	s.e. = 0.0154
	p-val = 0.0000
crim(2,6]	−0.2128
	s.e. = 0.0419
	p-val = 0.0000
crim(6,15]	−0.5380
	s.e. = 0.0401
	p-val = 0.0000
crim(15,20]	−0.7117
	s.e. = 0.0841
	p-val = 0.0000
crim(20,25]	−0.8549
	s.e. = 0.1094
	p-val = 0.0000
crim(25,90]	−0.9857
	s.e. = 0.0877
	p-val = 0.0000
Num.Obs.	490
R2	0.446
AIC	173.4
BIC	202.7
F	77.843
RMSE	0.28



Obr. 8: Porovnanie modelov po transformácii a jednoduchého lineárneho modelu závislosti ceny nehnuteľnosti (medv) na miere kriminality na jedného obyvateľa (crim)

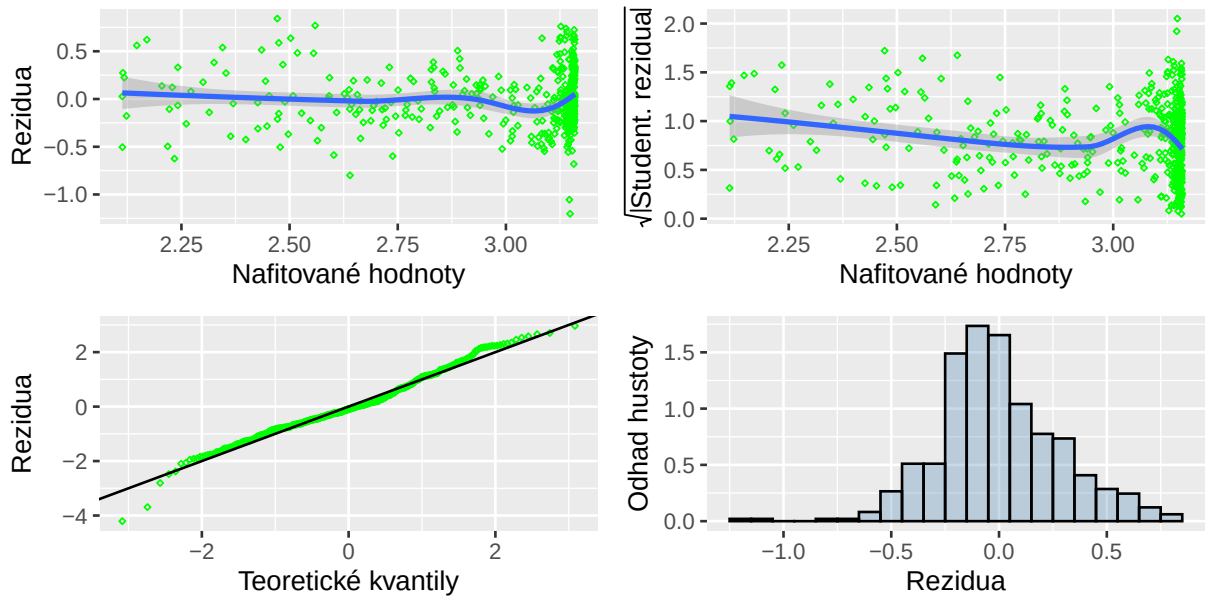
Tabuľka 7: Výsledky F-testu pre kvadratický a kubický model.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
487	40.6				
486	39.8	1	0.798	9.74	0.002

Tabuľka 8: Výsledok Breusch-Paganovho testu heteroskedasticity reziduii kubického modelu

Test	statistic	p-value
Breusch-Pagan	6.04	0.11

splinov vylepšil linearitu zdanlivo najlepšie. Dané transformácie boli teda vhodné. Nakoľko všetky modely používajú logaritmicky transformovanú odozvu tak je možné ich zrovnávať napr. pomocou R^2 štatistiky a tak získať náhľad na ich kvalitu. Porovnávať F-testom je možné len vnorené modely a síce v tomto prípade kvadratický a kubický model. V ostatných prípadoch je vhodné použiť na porovnanie modelov informačné kritéria ako napr. `AIC` alebo `BIC` prípadne porovnávať pomocou `RMSE` a R^2. Porovanie kvadratického a kubického modelu F-testom je zobrazené v tabuľke 7. Tento F-test poukazuje na štatisticky významný rozdiel medzi kvadratickým a kubickým modelom na hladine významnosti 0.05 a preto spomedzi týchto modelov je vhodné vybrať ten lepší a síce kubický model. Tabuľky 5 a 6 sumarizujú všetky modely. Na základe štatistík R^2, `AIC` ale aj `RMSE` sa ukazuje za najlepšie popísujúci model, model využívajúci spliny. Na druhú stranu `BIC` kritérium preferuje oproti modelu so splinami jednoduchší, kubický model. Po častiach konštantný model je taktiež štatisticky významný a dosahuje porovnateľné hodnoty R^2, `AIC` a `RMSE` s ostatnými modelmi (viz. tab. 5 a 6). V jednorozmernom prípade máme možnosť vizuálne porovnať kvalitu fitu modelu. Takáto vizualizácia je zobrazená na obrázku 8 opäť získaného po spätnej transformácii odozvy `medv`. Vizuálne sú všetky štyri modely porovnateľne dobré čo je v zhode s väčšinou spomínaných štatistík.



Obr. 9: Diagnostika reziduii kubického modelu

- Otázka 6

Tabuľka 9: Výsledok testov normality reziduii kubického modelu

Test	statistic	p-value
Shapiro-Wilk	0.983	0
Lilliefors	0.074	0

Tabuľka 10: Porovnanie viacrozmerných modelov + kubického modelu. (PRESS je vypočítaná po spätnej transformácii)

	df	BIC	AIC	R2_adj	RMSE	PRESS
model_my	8	-186	-220	0.75	0.192	7248
model_max	22	-206	-298	0.79	0.174	5999
model_backward_F	20	-217	-301	0.79	0.174	5963
model_backward_AIC	19	-221	-301	0.79	0.174	5910
model_forward_BIC	10	-233	-275	0.78	0.181	6372
model_both_BIC	9	-237	-275	0.78	0.181	6419
model_log_3	5	191	170	0.44	0.286	21216
model_corrected_AIC	21	-291	-379	0.82	0.161	5159

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.)

- Odpoveď 6

  Za najvhodnejší model považujem kubický model nakoľko dosahuje relatívne vysoké hodnoty R^2 resp. nízke hodnoty `RMSE`, `AIC` a najnižšiu hodnotu kritéria `BIC`, ktoré, je známe tým, že viac penalizuje nadbytočné parametre modelu. Podľa ostatných metrík je lepším modelom, model so splinami, ktorý má ale dvakrát viac parametrov čo môže viesť k overfitu. Podľa môjho názoru prichádza do úvahy aj po častiach konštantný model ale ako bolo spomenuté vyberám kubický model. Obrázok 9 zobrazuje diagnostiku rezidui kubického modelu. Na spodných dvoch podobrázkoch vidíme, že model nemá zásadný problém s predpokladom normality čo ale nie je v zhode s výsledkami štatistických testov, ktoré detekovali významnú deviáciu od normality na hladine 0.05 viď. tabuľka 9). Čo sa týka homoskedasticity rezidui tak Breasch-Paganov test sa ukázal byť štatisticky nevýznamný a preto nezamietame nulovú hypotézu o homoskedasticite rezidui (viz tabuľka 8). Toto je v zhode s oboma hornými podobrázkami v 9. Nevidno na nich, žiaden trend a rezidua sú rozmiestnené viac-menej náhodne okolo modrých kriviek ktoré sú približne vodorovné. Model taktiež nemá problém s predpokladom linearity čo usudzujem podľa ľavého-dolného podobrázka v 7. Reziduá sú tam rovnomerne rozmiestnené okolo modrej prerušovanej priamky resp. táto priamka je v približnej zhode s ružovou lokálne regresnou krivkou.

1.6 Vícerozměrný regresní model

- Otázka 7

Tabuľka 11: Výsledky F-testu pre prvú skupinu vnorených viacrozmerných modelov

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	Model
483	17.7					model_my
472	14.4	11	3.364	10.04	0.000	model_backward_AIC
471	14.3	1	0.045	1.47	0.226	model_backward_F
469	14.3	2	0.040	0.66	0.517	model_max

Tabuľka 12: Výsledky F-testu pre druhú skupinu vnorených modelov.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	Model
482	15.8					model_both_BIC
481	15.7	1	0.065	2.14	0.144	model_forward_BIC
472	14.4	9	1.361	4.97	0.000	model_backward_AIC
471	14.3	1	0.045	1.47	0.226	model_backward_F
469	14.3	2	0.040	0.66	0.517	model_max

Zkonstruujte 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 `crim` a `medv`, pokud existuje, vysvětlit pomocí jiných proměnných? Tj, že například oblasti s vyšší kriminalitou jsou v blízkosti dálnic, je tam větší znečištění atd.?

- Odpoveď 7

  Na konštrukciu optimálneho modelu môžme preskúmať metriky modelov založených na každej podmnožine regresorov. Toto je ale zdĺhavé. Ďalšou možnosťou je použiť doprednú alebo postupnú regresiu alebo spätnú elimináciu na základe F-testov alebo informačných kritérii. Taktiež môžme využiť našu expertnú znalosť a vybrať podľa nás optimálnu množinu regresorov a na jej základe vystavať model. Takto vzniklo celkovo 10 modelov plus model obsahujúci všetky regresory. V tabuľke 10 sú zobrazené všetky odlišné modely porovnané podľa metrík `AIC`, `BIC`, `RMSE`, R^2 a nakoniec podľa `PRESS` štatistiky, ktorá bola vypočítaná po spätnej exponenciálnej transformácii predikovaných hodnôt kvôli tomu aby sa táto štatistika dala použiť na porovnanie predikčnej kvality všetkých doteraz uvažovaných modelov. Výsledný model bol vybratý na základe F-testu medzi vnorenými modelmi. Tabuľka 11 prezentuje výsledok F-testu medzi prvou množinou vnorených modelov. Z tejto tabuľky môžme vyčítať, že medzi modelom, v ktorom boli vybraté podľa mňa najrelevantnejšie regresory a modelom získaným pomocou spätnej eliminácie s kritériom AIC je štatisticky významný rozdiel na hladine 0.05 ale na druhú stranu sa tento model štatisticky významne nelíši od modelu so všetkými premennými resp. s modelom získaného spätnou elimináciou podľa F-testu. Teda je z tejto množiny najvhodnejšie vybrať model získaný pomocou `AIC`. Z tabuľky 12 plynie rovnaký záver a síce, že najvhodnejšie je vybrať model získaný spätnou elimináciou pomocou kritéria `AIC`. Treba dodať, že postupná regresia ale aj dopredná regresia pomocou `AIC` vybrali rovnaké regresory. Čo sa týka vysvetlenia vzťahu medzi `medv` a `crim` tak je možné, že napr. oblasti s vyšším percentuálnym zastúpením nevzdelaných obyvateľov (`lstat`) budú vykazovať vyššiu mieru kriminality. Takýto záver naznačuje relatívne vysoká korelácia resp. trend v scatterplote medzi `crim` a `lstat` (viď. obrázok 2). Na podobrázkoch v prvom riadku vidíme korelácie `crim` so všetkými spojitými premennými. Na druhú stranu je možno vhodnejšie preskúmať scatterploty nakoľko korelačný koeficient môže byť blízky nule a zároveň premenné môžu vykazovať istý spoločný trend. Odhliadnúc teda od korelácie tak zo scatterplotov v prvom stĺpci obrázku 2 je možné, že ešte premenné `age` a `dis` vykazujú vzťah s premennou `crim`, pričom o príčinnom vzťahu medzi nimi silno pochybujem. Z boxplotov v 3 sa zdá, že by mohla existovať slabá závislosť medzi kriminalitou a prístupnosťou k diaľnici. Opäť ale môže platiť, že táto zdanlivú závislosť môže byť iba náhodná a nie príčinná.

- Otázka 8

  Použijte ve výsledném modelu krimalitu (proměnnou `crim`) a porovnejte jak se změnil vliv kriminality na medián ceny nemovitostí oproti jednoduchému regresnímu modelu s log transformovanou odezvou (viz otázka 4). Jaké je snížení průměrné ceny nemovitostí při vzrůstu kriminality o jednu jednotku na 1000 obyvatel? Pokud proměnnou `crim` v modelu nemáte tak ji pro tuto otázku do modelu přiřaďte.

- Odpoveď 8

  Vo vybratom modely na základe `AIC` je premenná `crim` stále štatisticky významná. Jej vplyv na odozvu oproti jednoduchému modelu sa ale vo viacrozmernom modely znížil 2.438 krát. V novom viacrozmernom

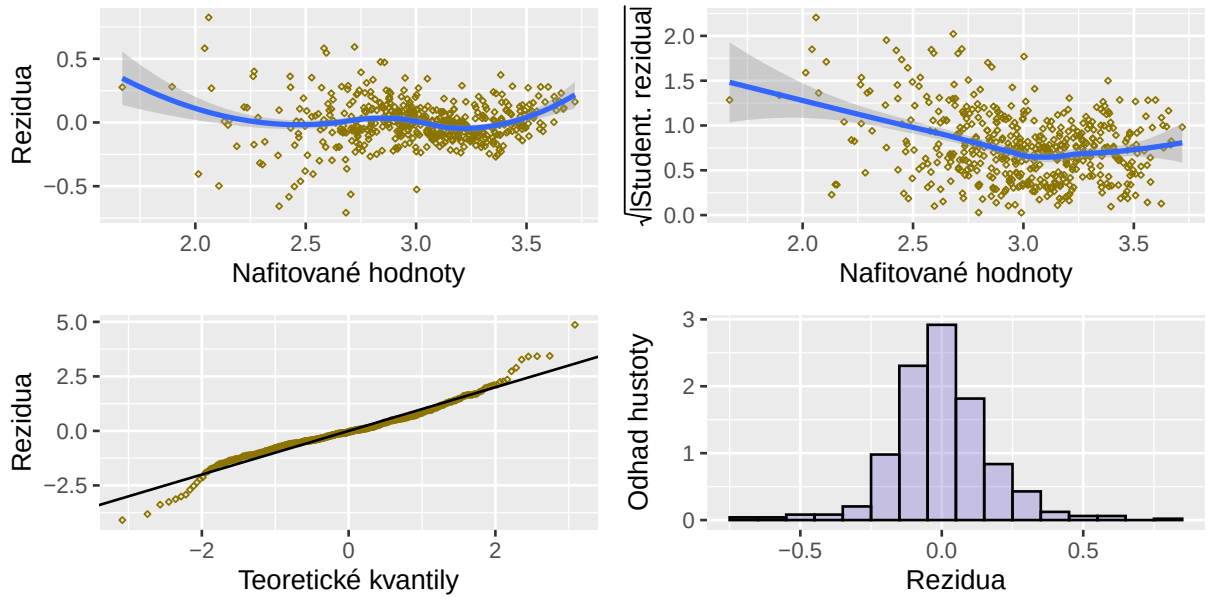
Tabuľka 13: Výsledok Breusch-Paganovho testu heteroskedasticity reziduii vo vybratom viacrozmernom modely

Test	statistic	p-value
Breusch-Pagan	89	0

Tabuľka 14: Výsledok testov normality reziduii vo vybratom viacrozmernom modely

Test	statistic	p-value
Shapiro-Wilk	0.960	0
Lilliefors	0.065	0

modely sa očakávaná cena nehnuteľností pri náraste miery kriminality na 1000 obyvateľov o jednotku (ceteris paribus) zníži o 0.001%.



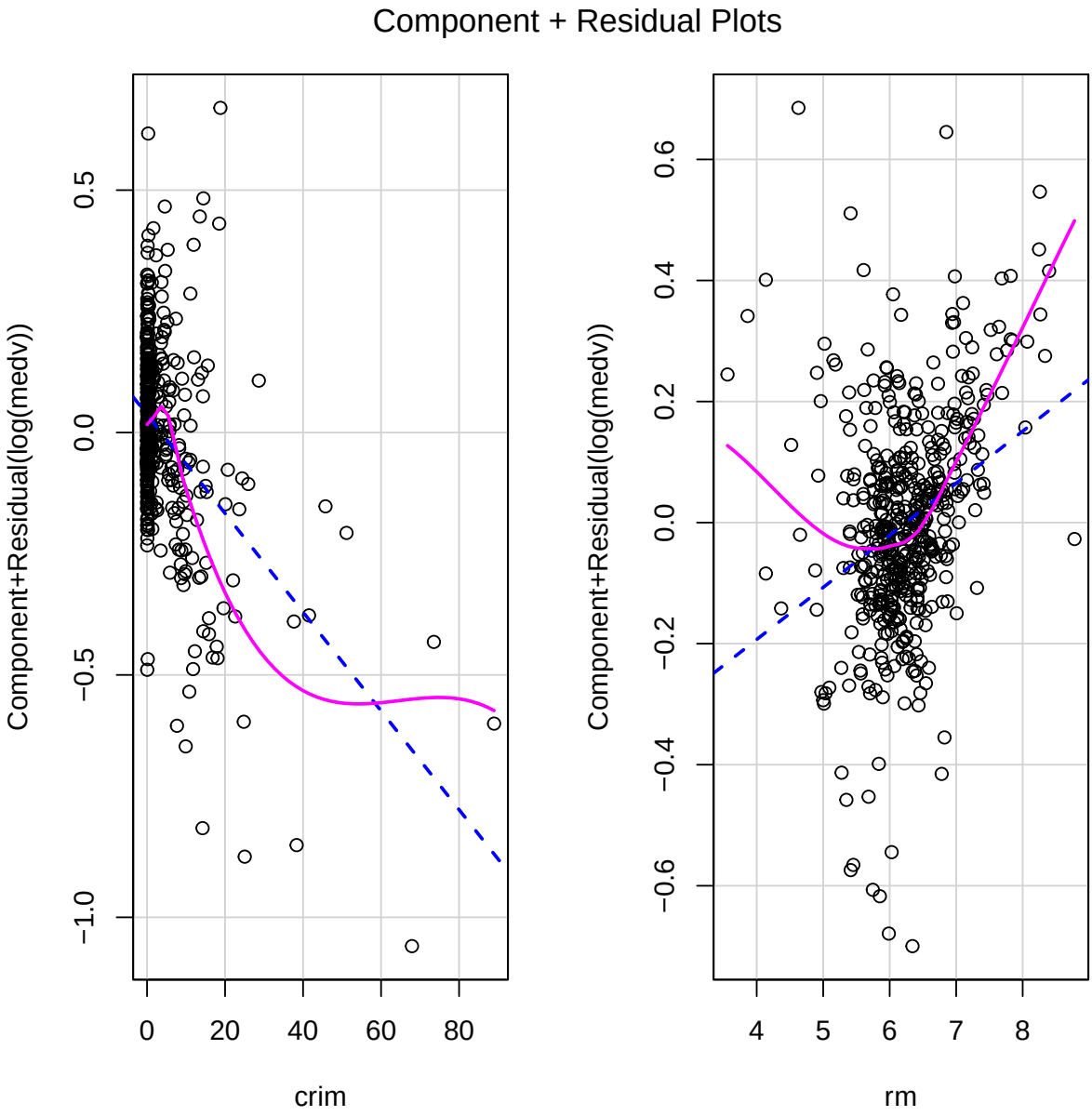
Obr. 10: Diagnostika reziduii v multirozmernom modely

- 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 změnily a dala se tato změna očekávat? Validujte model (jak graficky, tak pomocí příslušných testů hypotéz).

- Odpoveď 9

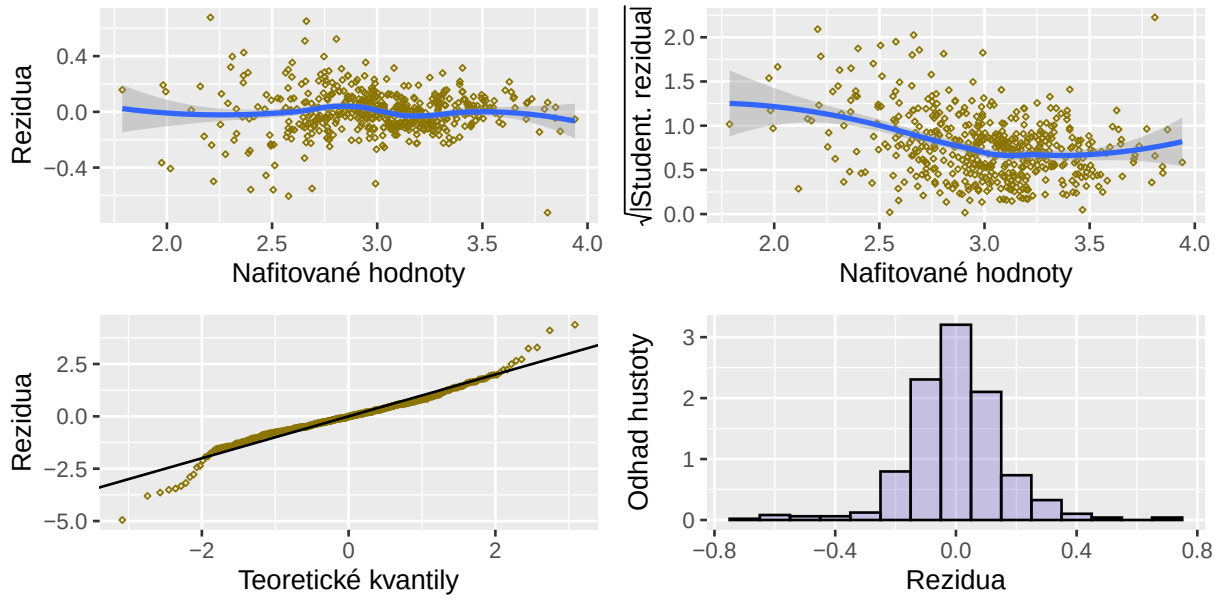
  Obrázok 10 zobrazuje diagnostiku reziduii viacrozmerného modelu získaného pomocou spätnej eliminácie na základe `AIC` kritéria. Na spodných dvoch podobrázkoch vidíme, že model nemá zásadný problém s predpokladom normality čo ale nie je v zhode s výsledkami štatistických testov, ktoré detekovali významnú deviáciu od normality na hladine 0.05 (viď tabuľka 14). Podobne aj homoskedasticita reziduii bola zamietnutá podľa Breasch-Paganovho testu (viď tabuľka 13). Toto je v zhode s oboma hornými podobrázkami v 10 kde vidíme, že rozptyl reziduii sa zmenšuje so zväčšujúcou sa hodnotou nafitovanej hodnoty. Rezidua tiež vykazujú mierny nelineárny trend čo môže naznačovať problém s predpokladom



Obr. 11: Čiastočné reziduálne grafy viacrozmerného modelu

Tabuľka 15:

	Multirozmerný model podľa AIC	Viacrozmerný model s transformovanými prediktormi ‘crim‘ a ‘rm‘
(Intercept)	3.8770	4.3174
	p-val = 0.0000	p-val = 0.0000
crim	−0.0102	
	p-val = 0.0000	
zn	0.0013	0.0010
	p-val = 0.0176	p-val = 0.0435
nox	−0.6242	−0.6553
	p-val = 0.0000	p-val = 0.0000
rm	0.0860	
	p-val = 0.0000	
dis	−0.0432	−0.0381
	p-val = 0.0000	p-val = 0.0000
rad2	0.1041	0.0776
	p-val = 0.0599	p-val = 0.1281
rad3	0.1744	0.1560
	p-val = 0.0006	p-val = 0.0009
rad4	0.1079	0.0996
	p-val = 0.0166	p-val = 0.0165
rad5	0.1368	0.1366
	p-val = 0.0029	p-val = 0.0013
rad6	0.0968	0.0993
	p-val = 0.0771	p-val = 0.0494
rad7	0.2065	0.1756
	p-val = 0.0005	p-val = 0.0014
rad8	0.2161	0.1603
	p-val = 0.0001	p-val = 0.0023
rad24	0.3187	0.3862
	p-val = 0.0000	p-val = 0.0000
tax	−0.0005	−0.0005
	p-val = 0.0001	p-val = 0.0000
ptratio	−0.0350	−0.0313
	p-val = 0.0000	p-val = 0.0000
black	0.0004	0.0003
	p-val = 0.0002	p-val = 0.0036
lstat	−0.0250	−0.0234
	p-val = 0.0000	p-val = 0.0000
poly(crim, 2)1		−2.7929
		p-val = 0.0000
poly(crim, 2)2		1.0542
		p-val = 0.0000
poly(rm, 2)1		1.2328
		p-val = 0.0000
poly(rm, 2)2		1.4940
		p-val = 0.0000
Num.Obs.	490	490
R2	0.800	0.831
AIC	−301.0	−379.2
BIC	−221.3	−291.1
F		121.214
RMSE	0.17	0.16



Obr. 12: Diagnostika reziduii vo viacrozmernom modely po transformácii `crim` a `rm`

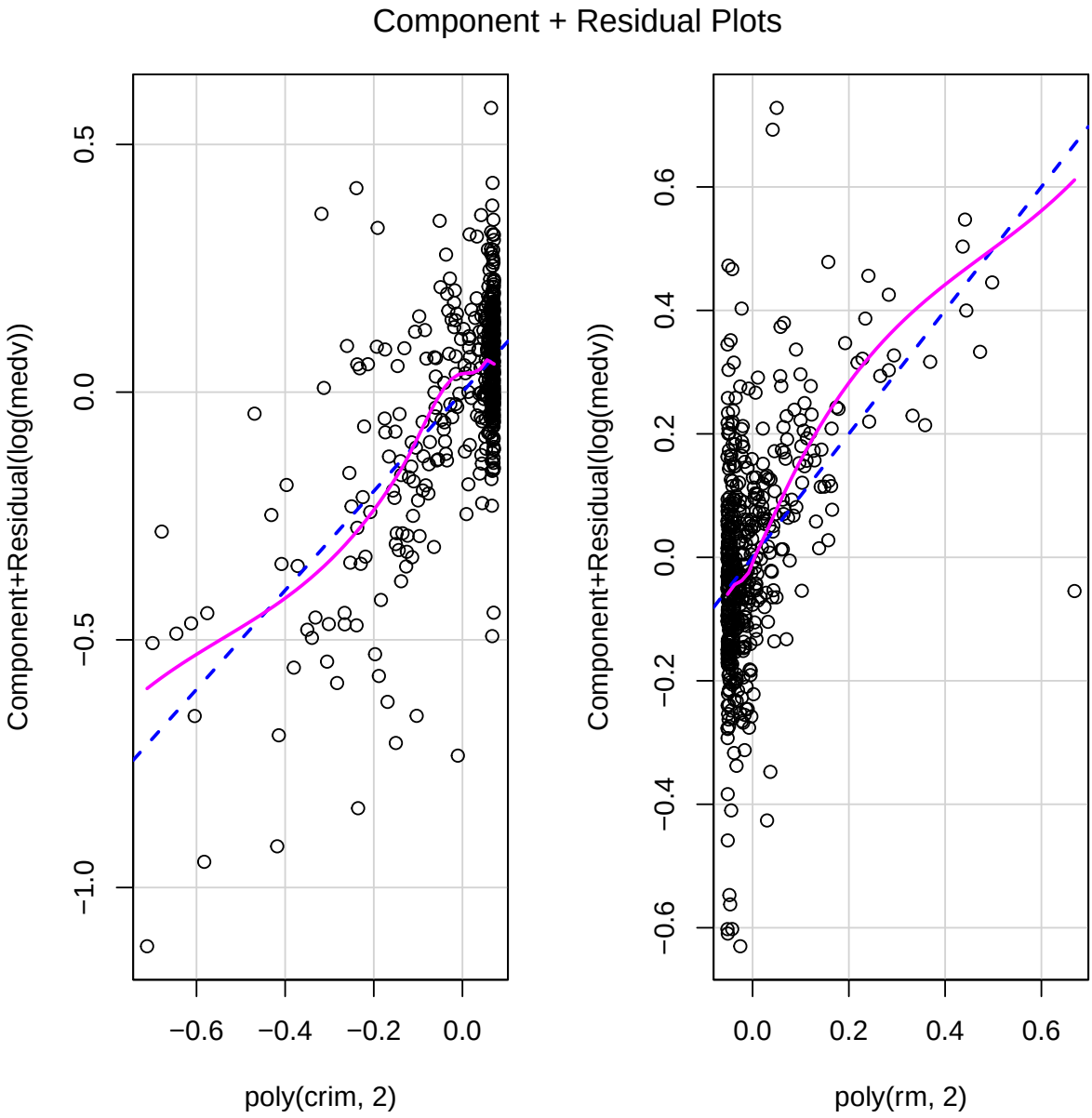
linearity. Problém s linearitou v premennej `crim` a `rm` naznačuje aj čiastočný reziduálny graf 11. Preto navrhujeme transformovať tieto premenné napr. kvadraticky. Porovnanie vybraného modelu a modelu po transformácii premenných `crim` a `rm` je v tabuľke 15. Obrázky 11 a 12 naznačujú, že transformácia `crim` a `rm` pomohla s linearitou. Mierna deviácia od homoskedasticity stále pretrváva. V tabuľke 10 vidíme porovnanie všetkých viacrozmerných modelov, vrátane modelu po transformácii `crim` a `rm` ako aj jednoduchého kubického modelu (z otázky 6). Ako sa dalo očakávať tak pridaním nových regresorov bola do modelu vnesená nová informácia čo zlepšilo kvalitu modelu oproti jednoduchému kubickému modelu (model_log_3) vzhľadom k všetkým uvádzaným metrikám. Taktiež vidíme, že kvadratická transformácia `crim` a `rm` priniesla ďalšie zlepšenie všetkých metrík oproti všetkým modelom. Čo sa týka predikcie tak je asi najrelevantnejšie merať `PRESS` štatistiku ktorá bola ako už bolo spomenuté vypočítaná po spätnej exponenciálnej transformácii čo síce nebolo nutné keďže všetky modely v tabuľke majú log. transformovanú odozvu ale prípadne to umožní formálne porovnanie s modelmy bez tranformácie odozvy. Vybraný viacrozmerný model (model_backward_AIC) dosiahol spomedzi ostatných viacrozmerných modelov najlepšie hodnotu tejto `PRESS` štatistiky pričom použitie transformácie `crim` a `rm` prinieslo ešte ďalšie zlepšenie. Na predikciu by som teda vybral model označený ako (model_corrected_AIC) (viď tabuľka 15).

- Otázka 10

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

- Odpoveď 10

  Na základe nášho modelu by bola odpoveď zjavná ak by platila nezávislosť medzi všetkými prediktormi resp. by pri zmene kriminality ostali všetky ostatné premenné konštantné. To v praxi ale nevieme zaistiť ani zistiť pretože ako bolo popísané v odpovedi 7 tak premenná `crim` môže byť vo vzťahu s inou premennou v našom modely a tak jej zmena môže spôsobiť zmenu tejto na nej závislej premennej a teda výsledná zmena ceny by bola ovplyvnená oboma týmito zmenami kriminality ako aj na nej závislom regresore. Taktiež sa teoreticky môže stať, že kriminalita ovplyvňuje neznámu premennú, ktorú v modely ani nemáme ale táto neznáma premenná tiež ovplyvňuje cenu nehnuteľnosti. Z nášho



Obr. 13: Čiastočné reziduálne grafy viacrozmerného modelu po transformácii `crim` a `rm`

modelu teda plynie iba to, že za predokladu nezávislosti iných premenných na kriminalite by sa cena nehnuteľnosti mala zvýšiť pri znížení kriminality (viď odpoveď 8).