1. zÃąpoÄĎtovÃą Ãžloha z 01RAD

Emanuel FrÃątrik

2021-10-21

1 1. zápočtová úloha z 01RAD

1.1 Popis úlohy

V tomto úkolu je cílem provést předzpracování datového souboru, jeho vizualizaci a jednoduchou lineární regresní úlohu, kde budeme modelovat spotřebu automobilu v závislosti na jeho váze. K tomuto účelu poslouží datový souboru `auto_mpg_2021rad.txt`, který obsahuje 406 pozorování o 9 proměnných. Dataset byl prvně použit americkou statistickou společností v roce 1983 a lze ho též najít na UCI Machine Learning Repository, případně na kaggle.com s několika pracovními sešity.

mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
18	8	307	130	3504	12.0	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11.0	70	1	chrysler satellite
16	8	304	150	3433	12.0	70	1	chrysler rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10.0	70	1	ford galaxie 500

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 bodů. Další dodatečné body mohu případně individuálně udělit za řešení mini domácích úkolů z jednotlivých hodin.

1.3 Odevzdání

Protokol ve formátu pdf odevzdejte prostřednictvím MS Teams, nejpozději do 10. 11. 2021.

2 Předzpracování dat:

2.1 Otázka 01

Zjistěte, zdali data neobsahují chybějící hodnoty (`NA`). Pokud ano, tak rozhodněte zdali můžete příslušná pozorování z dat odstranit a proč. Které proměnné jsou kvantitativní a které kvalitativní? Jeli možno některé zařadit do obou skupin, pro kterou se rozhodnete? Které proměnné budete brát jako faktorové a proč? Spočtěte základní statistiky pro jednotlivé proměnné.

Table 1: Deskriptívne štatistiky pre spojité premenné

PREMENNÁ	PRIEMER	ROZPTYL	MEDIÁN	POČET NA
mpg	23.571	60.3641	23.0	9
displacement	194.536	10981.0718	151.0	0
horsepower	104.924	1516.7771	94.0	11
weight	2976.400	716563.9457	2811.0	0
acceleration	15.532	7.8810	15.5	0

2.2 Odpoveď 01

Dataset obsahuje celkovo 408 pozorovaní popísaných deviatimi premennými. Z toho premenné `origin` a `car_name` sú zjavne kvalitatívne premenné popisujúce isté kategórie a preto je vhodné ich popisovať faktormi. Ostatné premenné sú kvantitatívne. Nakoľko premenné `cylinders` a `model_year` popisujú každá len relatívne malé množstvo kategórii resp. diskrétnych hodnôt tak sa na ne dá taktiež pozerať ako na kvalitatívne premenné a popisovať ich ako faktorové premenné spolu s už spomenutými premennými `origin` a `car_name`. V tabuľke 1 môžme vidieť deskriptívne štatistiky pre spojité premenné. Tabuľka 2 a 3 ďalej zobrazujú zastúpenia jednotlivých kategórii v rámci zostávajúcich kategorických premenných. Taktiež vidíme, že dataset obsahuje chýbajúce dáta označené ako `NA`. Počet týchto chýbajúcich hodnôt určite nepresiahne hodnotu 20. Zahodenie týchto maximálne 20 čiastočne chybných pozorovaní je teda prijateľné nakoľko dataset obsahuje relatívne veľký počet pozorovaní.

2.3 Otázka 02

Proměnnou `mpg` nahraďte proměnnou `spotreba` kde bude místo počtu ujetých mil na galon paliva uvedena hodnota počet litrů na 100 Km. Proměnnou `cylinders` přejmenujte na `pocet_valcu`. Proměnnou `displacement` přejmenujte na `zdvihovy_objem` a převeďte z kubických palců na litry. Proměnnou `horsepower` přejmenujte na `výkon` a převeďte na kW. Proměnnou `weight` přejmenujte na `hmotnost` a převeďte z liber na kilogramy. Odstraňte proměnnou `acceleration`. Proměnnou `model.year` přejmenujte na `rok_vyroby` a upravte ji tak, aby její hodnoty popisovaly celý rok 19XX. Proměnnou `origin` přejmenujte na `puvod` a upravte ji tak, že místo 1 bude USA, místo 2 EUR a místo 3 JAP. Z proměnné `car.name` vytvořte proměnnou `vyrobce` podle prvního slova obsaženého v řetězci proměnné `car.name`.

2.4 Odpoveď 02

Náhľad na upravený dataset je zobrazený v tabuľke 4.

3 Vizualizace dat

3.1 Otázka 03

Vykreslete scatterploty pro všechny numerické proměnné. Pro proměnné `spotreba` a `hmotnost` vykreslete histogramy spolu s jádrovými odhady hustot. Pro proměnné `pocet_valcu` a `rok_vyroby` vykreslete krabicové diagramy, kde odezvou bude `spotreba`. Je z těchto grafů vidět, že některá auta mají jinou, než očekávanou spotřebu? Navrhněte úpravu těchto dvou proměných (případně úpravu datasetu) tak, aby obě proměnné `pocet_valcu` a `rok_vyroby` byly faktorové a obsahovaly právě 3 úrovně. Pro takto upravená data vykreslete místo výše zmíněných boxplotů violin ploty.

Table 2: Deskriptívne charakteristiky pre kategorické premenné

PREMENNÁ	N = 408
cylinders	
3	4 / 408 (1.0%)
4	208 / 408 (51%)
5	3 / 408 (0.7%)
6	85 / 408 (21%)
8	108 / 408 (26%)
model_year	
70	36 / 408 (8.8%)
71	30 / 408 (7.4%)
72	28 / 408 (6.9%)
73	40 / 408 (9.8%)
74	27 / 408 (6.6%)
75	30 / 408 (7.4%)
76	34 / 408 (8.3%)
77	28 / 408 (6.9%)
78	36 / 408 (8.8%)
79	29 / 408 (7.1%)
80	29 / 408 (7.1%)
81	30 / 408 (7.4%)
82	31 / 408 (7.6%)
origin	
1	256 / 408 (63%)
2	73 / 408 (18%)
3	79 / 408 (19%)

[1] n / N (%)

Table 3: Deskriptívne charakteristiky pre premennú vyrobca

PREMENNÁ	N = 389
vyrobca	
audi	7 / 389 (1.8%)
bmw	2 / 389 (0.5%)
buick	17 / 389 (4.4%)
cadillac	2 / 389 (0.5%)
chevrolet	47 / 389 (12%)
chrysler	66 / 389 (17%)
dodge	28 / 389 (7.2%)
fiat	8 / 389 (2.1%)
ford	49 / 389 (13%)
honda	13 / 389 (3.3%)
mazda	12 / 389 (3.1%)
mercedes-benz	3 / 389 (0.8%)
mercury	11 / 389 (2.8%)
nissan	24 / 389 (6.2%)
oldsmobile	10 / 389 (2.6%)
opel	4 / 389 (1.0%)
peugeot	8 / 389 (2.1%)
pontiac	16 / 389 (4.1%)
renault	3 / 389 (0.8%)
subaru	4 / 389 (1.0%)
toyota	26 / 389 (6.7%)
triumph	1 / 389 (0.3%)
volkswagen	19 / 389 (4.9%)
volvo	6 / 389 (1.5%)
VW	3 / 389 (0.8%)

[1] n / N (%)

Table 4: Náhľad na upravený dataset

spotreba	pocet_valcov	zdvihovy_objem	vykon	hmotnost	rok_vyroby	vyrobca	povod
13.06748	8	5.030840	96.94098	1589.387	1970	chevrolet	USA
15.68097	8	5.735485	123.04048	1675.116	1970	buick	USA
13.06748	8	5.211098	111.85498	1558.543	1970	chrysler	USA
14.70091	8	4.981678	111.85498	1557.182	1970	chrysler	USA
13.83615	8	4.948904	104.39798	1564.440	1970	ford	USA
15.68097	8	7.030066	147.64857	1969.044	1970	ford	USA

3.2 Odpoveď 03

Aj keď sú premenné `pocet_valcov` a `rok_vyroby` v datasete reprezentované numerickými hodnotami tak si ich dovolím z nasledujúceho scatterplotu vynechať nakoľko ako už bolo spomenuté je výhodnejšie ich vnímať ako kategorické premenné a ich vzťah voči spojitým premenným zobrazovať pomocou boxplotov. Obrázok 1 zobrazuje scatterploty medzi jednotlivými spojitými premennými. Následne obrázok 2 zobrazuje histogramový a jadrový odhad hustôt pre premenné `spotreba` a `hmotnost`. Boxploty pre premenné `pocet_valcov` a `rok_vyroby` sú zobrazené na obrázkoch 3 a 4. Z boxplotu 3 sa zdá, že autá s tromi valcami majú vyššiu spotrebu ako autá so štyrmi valcami. Toto môže byť spôsobené nedostatkom pozorovaní v kategórii áut s tromi valcami. Dataset je teda vhodné upraviť tak aby sa zlúčili kategórie obsahujúce nízky počet pozorovaní. Obrázok 5 reprezentuje violinploty pre premenné `pocet_valcov` a `rok_vyroby` po úprave datasetu navrhovaným zlúčením kategórii a to na základe tercilov.

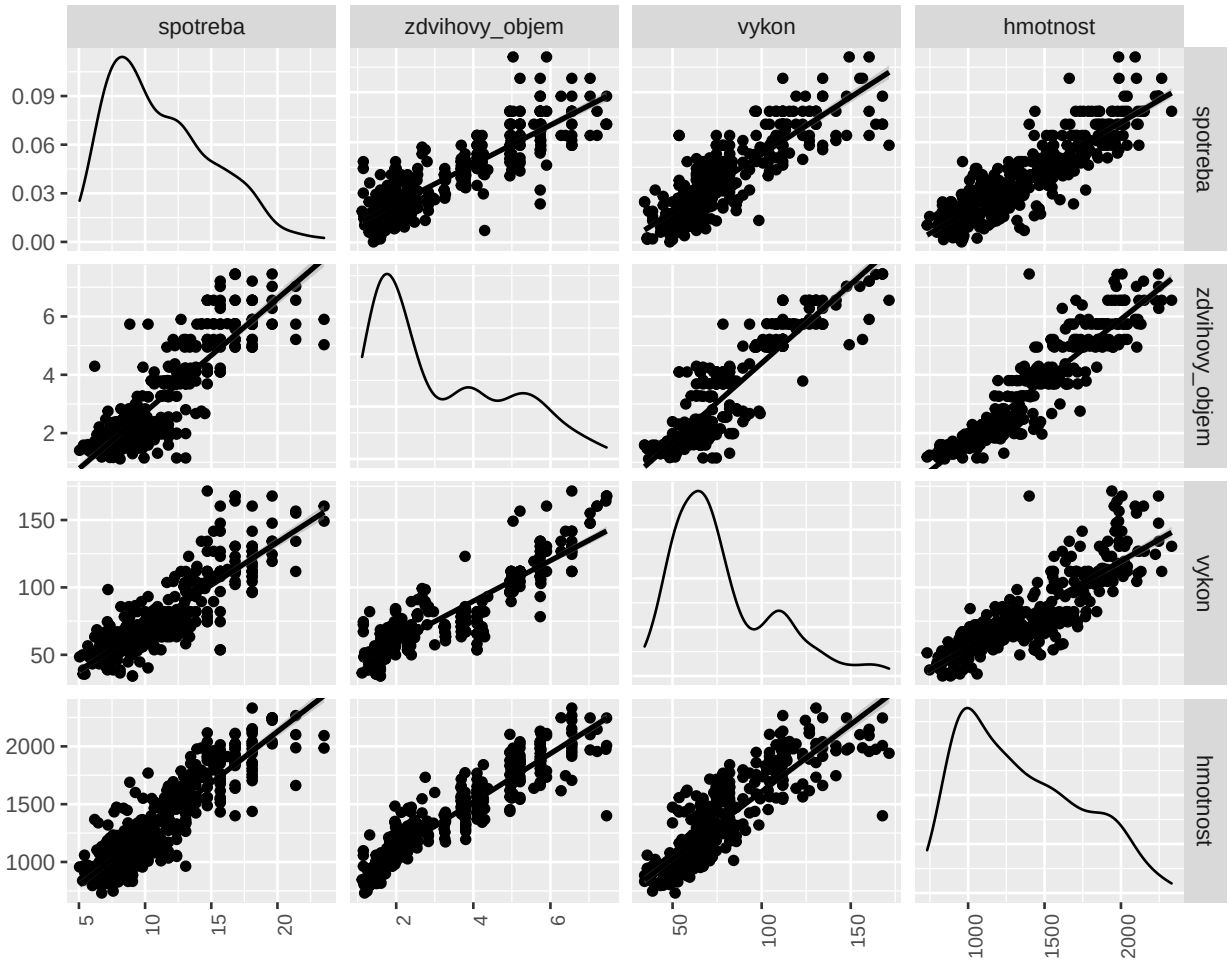

Figure 1: Scatterploty pre numerické (spojité) premenné

3.3 Otázka 04

Pro kombinace faktorizovaných proměnných `pocet_valcu`, `rok_vyroby` a `puvod` vykreslete spotřebu aut, aby bylo na obrázku vidět, jestli se liší spotřeba u aut pocházejících z různých kontinentů v závislosti na počtu válců, roku výroby a naopak. Zobrazte jen kombinace s relevantním počtem dat.

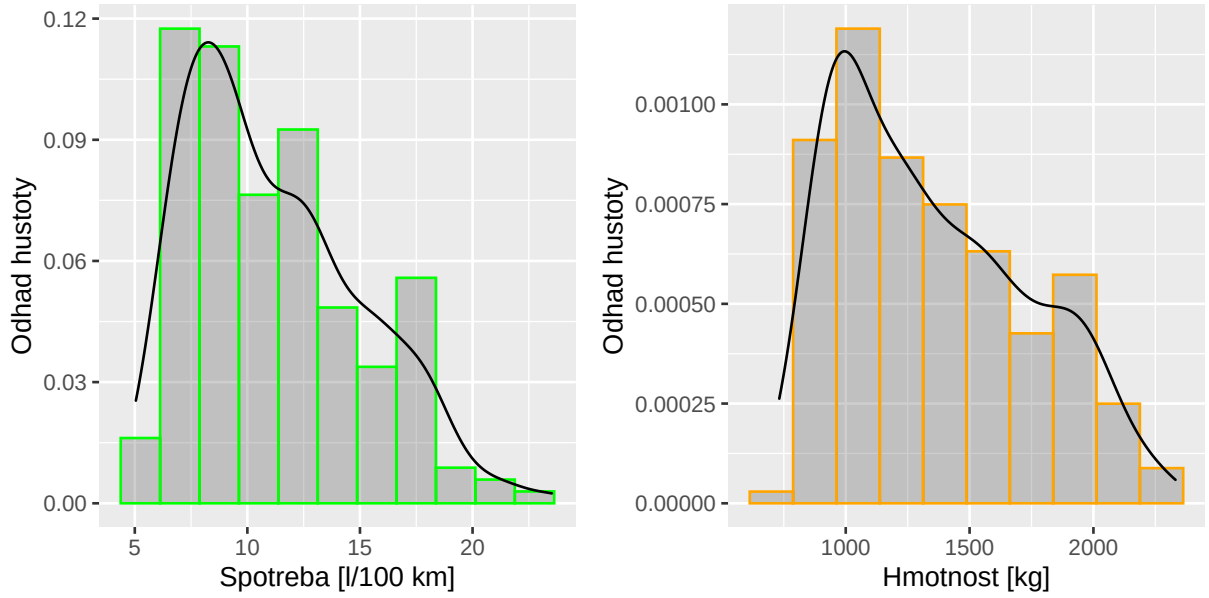

Figure 2: Histogramový a jadrový odhad hustoty premenných spotreba a hmotnost

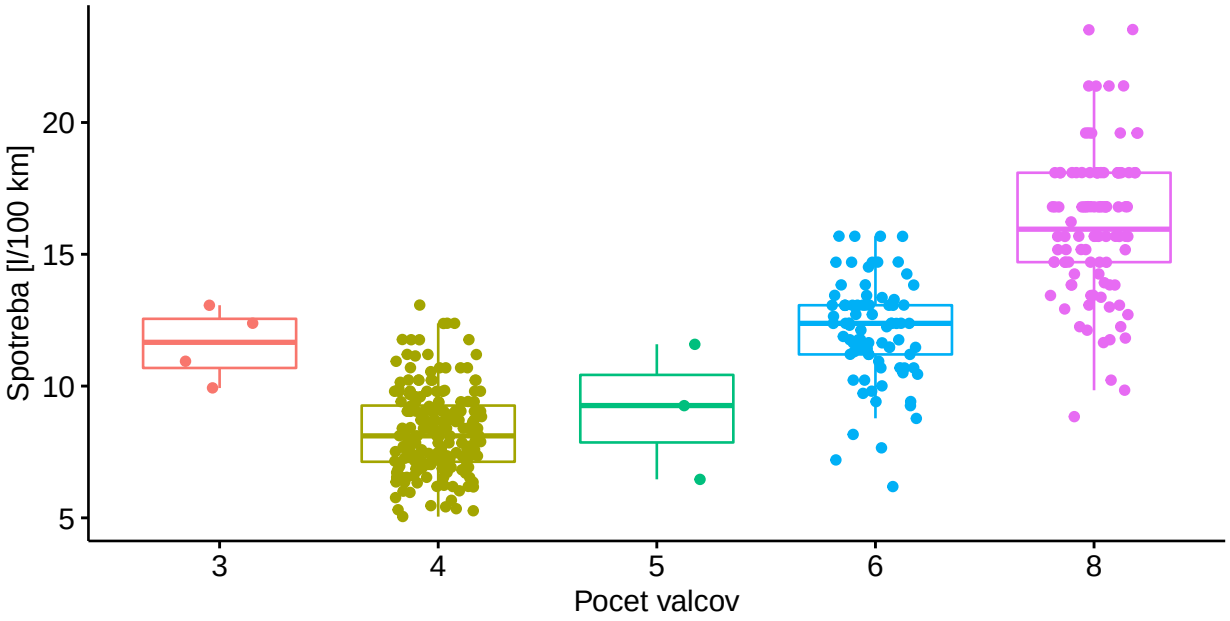

Figure 3: Boxplot pre premennú pocet_valcov

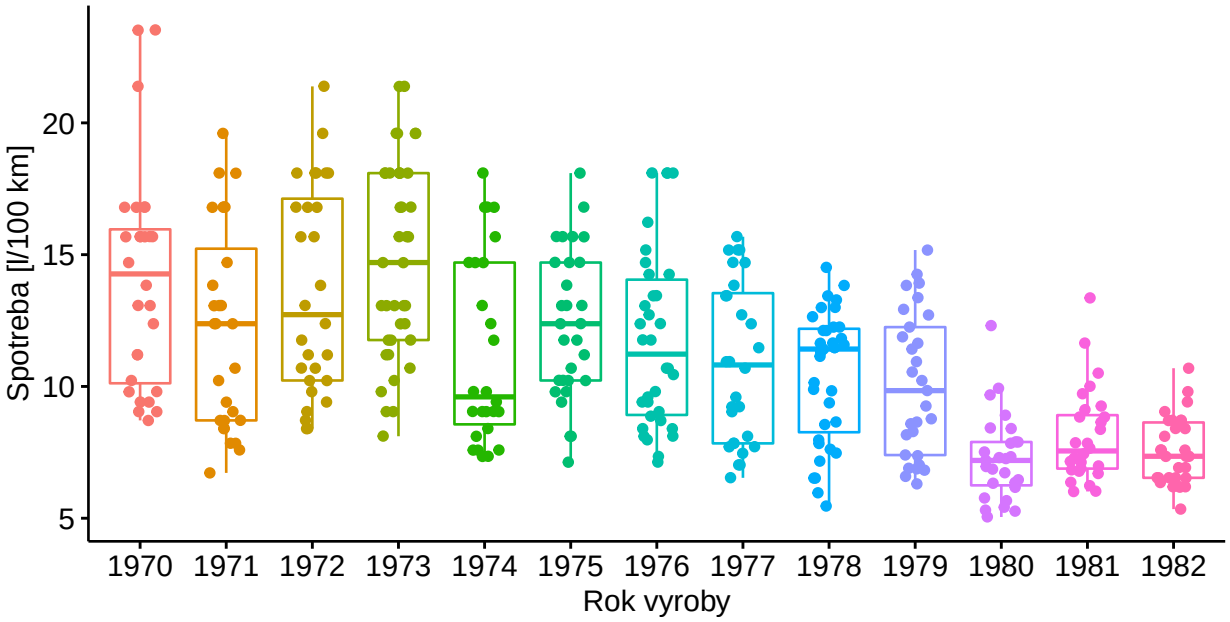

Figure 4: Boxploty pre premennú rok_vyroby

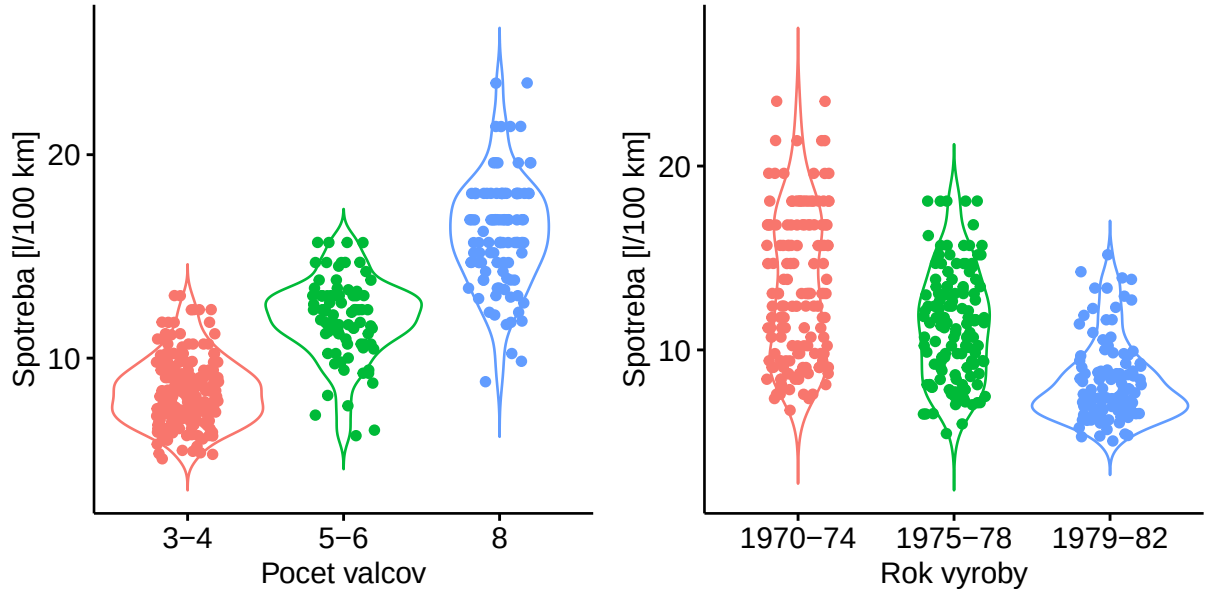

Figure 5: Violinloty pre upravené premenné pocet_valcov a rok_vyroby

3.4 Odpoveď 04

Na obrázku 6 je zobrazená závislosť spotreby na pôvode, počte valcov a roku výroby auta. Zobrazené sú len kombinácie s viac ako 5 pozorovaniami.

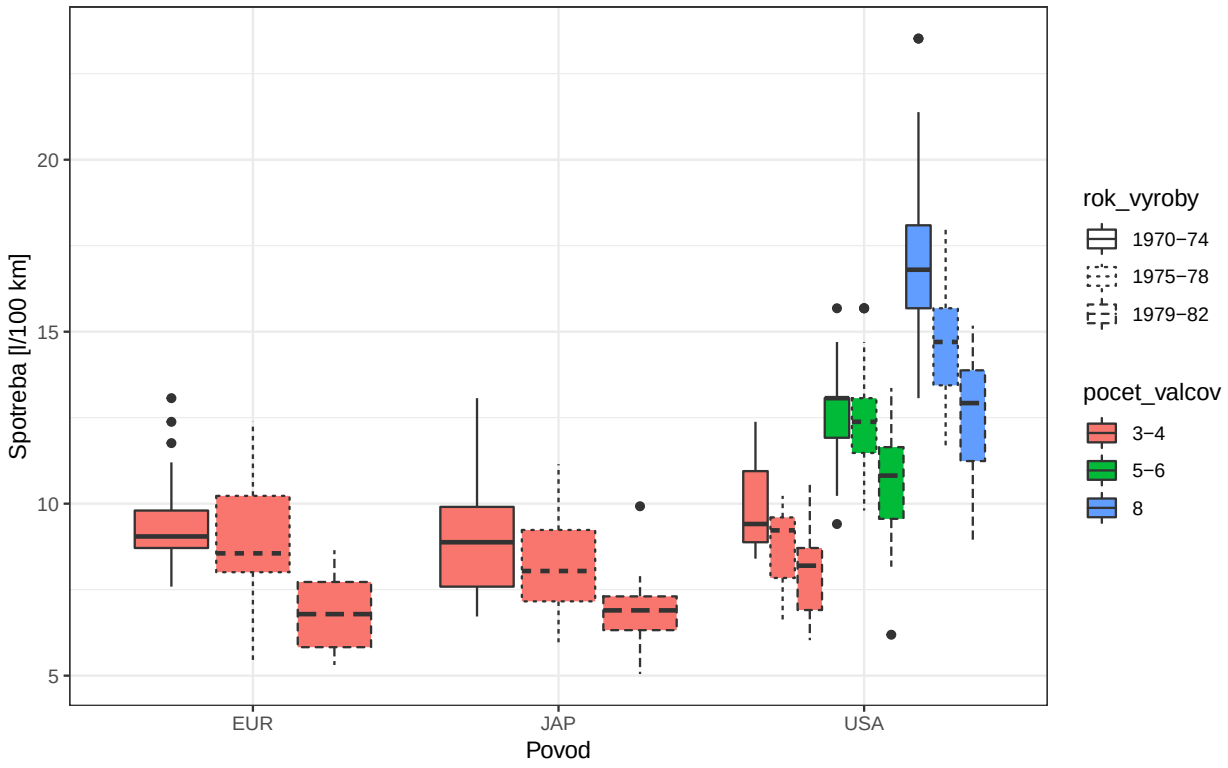

Figure 6: Závislosť spotreby na pôvode, počte valcov a roku výroby auta.

3.5 Otázka 05

Pro auta výrobce Chrysler vykreslete závislost spotřeby na váze automobilu, kde jednotlivé události označíte barvou podle počtu válců a velikost bodů v grafu bude odpovídat objemu motoru.

3.6 Odpoveď 05

Požadovaná závislosť spotreby áut značky Chrysler v závislosti na váhe a počte valcov je vyobrazená na obrázku 7.

4 Jednoduchý lineární model

4.1 Otázka 06

Sestavte jednoduchý regresní model (s i bez interceptu), kde vysvětlovaná proměnná bude spotřeba automobilu. Spočtěte pro oba modely R^2 a F statistiky, co nám o modelech říkají. Vyberte jeden z nich a zdůvodněte proč ho preferujete. Na základě zvoleného modelu zjistěte, zdali spotřeba automobilu závisí na hmotnosti

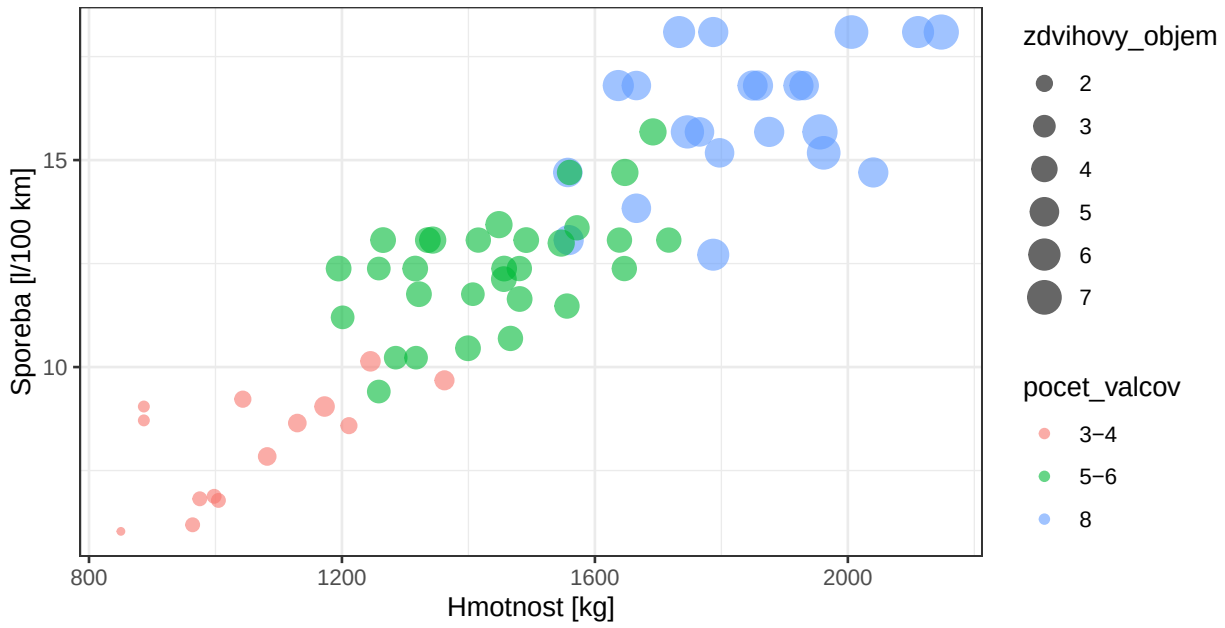

Figure 7: Závislosť spotreby áut značky Chrysler na hmotnosti v závislosti na zdvihovom objeme a počte valcov

automobilu. Pokud ano, o kolik se změní očekávaná spotřeba automobilu pokud se jeho hmotnost zvýší o 1000kg?

4.2 Odpoveď 06

Model bez interceptu dosahuje hodnotu štatistiky R^2=0.9770402 a model s interceptom hodnotu 0.7849673. Model s interceptom dosiahol hodnotu F štatistiky rovnú 1412.7262376 a model bez interceptu hodnotu 1.6511128×10^4. Tieto dve štatistiky ale nie sú vhodné na porovnanie daných modelov keďže model bez interceptu bude mať vždy štatistiku R^2 väčšiu ako model s interceptom a rovnaké tvrdenie platí pre F štatistiku. Namiesto toho by som zvolil zrovnanie napr. pomocou MSE. Model bez interceptu dosiahol hodnotu $MSE = 3.2318258$ a model s interceptom hodnotu 3.1979685. Vzhľadom k tejto metrike sa oba modely takmer nelíšia. Aj keď model s interceptom predpovedá pre očakávanú hodnotu spotreby pre autá s hmotnosťou blízko nuly záporné hodnoty tak by som preferoval práve tento model pretože ponechaním interceptu v modeli zabezpečíme dodátočnú voľnosť a teda možnú lepšiu lineárnu aproximáciu na intervale [731.6443799, 2331.464422]. Zhrnutie a porovnanie modelov je zobrazené v tabuľke 5. Model s interceptom je ako celok štatisticky významný na hladine 0.05 s p-hodnotou $3.1600719 \times 10^{-131}$ a podobne aj oba koeficienty sú významné na hladine 0.05. Model teda za predpokladu splenia predpokladov rozumne lineárne aproximuje dáta a popisuje vzťah medzi spotrebou a hmotnosťou. Podľa modelu s interceptom sa očakávaná spotreba automobilu pri zmene o 1000 kg zmení o 8.8633965 l/100 km.

4.3 Otázka 07

Sestavte obdobný model jako v předchozí otázce, ale pouze na základě dat výrobce Chrysler. Liší se tento model od předchozího? Jaký model vykazuje silnější linearní vztah mezi hmotností a spotřebou a proč? O kolik roste spotřeba s rostoucí hmotností pro vozy Chrysler rychleji než pro libovolný automobil? Spočtěte 95% konfidenční intervaly pro regresní koeficienty popisující sklon regresnní přímky v obou modelech a zjistěte, zdali se protínají? Co z toho můžeme vyvozovat? Na základě těchto modelů zjistěte o kolik procent bude mít

Table 5: Porovnanie modelov s a bez interceptu

	Model bez interceptu	Model s interceptom
hmotnost	0.0084	0.0089
	s.e. = 0.0001	s.e. = 0.0002
	t-stat = 128.4956	t-stat = 37.5863
	p-val = 0.0000	p-val = 0.0000
(Intercept)		−0.7475
		s.e. = 0.3307
		t-stat = −2.2600
		p-val = 0.0244
Num.Obs.	389	389
R2	0.977	0.785
R2 Adj.	0.977	0.784
F		1412.726

Table 6: Porovnanie modelov s a bez interceptu pre auta značky Chrysler

	Model bez interceptu	Model s interceptom
hmotnost	0.0085	0.0090
	s.e. = 0.0001	s.e. = 0.0005
	t-stat = 74.3061	t-stat = 16.4144
	p-val = 0.0000	p-val = 0.0000
(Intercept)		−0.6838
		s.e. = 0.8345
		t-stat = −0.8195
		p-val = 0.4156
Num.Obs.	66	66
R2	0.988	0.808
R2 Adj.	0.988	0.805
F		269.434

automobil značky Chrysler a hmotnosti 1,5 tuny vyšší očekávanou spotřebu než průměrný automobil o stejné hmotnosti.

4.4 Odpoveď 07

Nový model popisujúci len autá značky Chrysler sa líši od predchádzajúceho modelu čo vidíme na mierne odličných odhadoch regresných koeficientov a taktiež vykazuje silnejší lineárny vzťah čo môžeme usudzovať z hodnoty štatistiky R^2=0.8080577 ktorá je pre model s interceptom rovná druhej mocnine korelačného koeficientu kvantifikjúceho silu lineárneho vťahu. Pre vozidlá značky rastie spotreba s hmotnosťou 1.0138312 krát rýchlejšie ako pre ľubovoľné vozidlá. Model popisujúci všetky autá má postupne intervaly spoľahlivosti pre intercept a koeficient lineárneho členu [-1.3977294, -0.0972131] a [0.0083998, 0.009327]. A model popisujúci autá značky Chrysler má nasledujúce intervaly spoľahlivosti pre intercept [-2.3508234, 0.983196] a lineárny člen [0.0078923, 0.0100796]. Z toho vidíme, že interaly spoľahlivosti pre koeficienty modelu popisujúceho všetky autá sú vnorené do prislúchajúcich intervalov pre model popisujúci autá značky Chrysler a teda regresné koeficienty modelu popisujúci všetký autá sú presnejšie odhadnuté. Automobil značky Chrysler s hmotnosťou 1.5 t bude mať spotrebu o 1.9728371 % väčšiu spotrebu než priemerný automobil s hmotnosťou 1.5 t.

4.5 Otázka 08

Vykreslete scatterplot hmotností automobilů a jejich spotřeby. Do tohoto grafu vykreslete regresní přímku modelu s interceptem i bez. Sestrojte navíc lineární model, kde budete uvažovat, že spotřeba závisí na kvadrátu hmotnosti. Příslušnou křivku popisující odhady středních hodnot z tohoto modelu přidejte do obrázku k oboum předchozím modelům. Pro účely predikce spotřeby automobilů, na základě jakých statistik byste mezi těmito modely vybírali, nebo byste se rozhodovali na základě něčeho jiného a proč?

4.6 Odpoveď 08

Modely sú vidieť na obrázku 8. Pre účely predikcie by som modely porovnával podľa MSE aby bolo možné zrovnávať aj model bez interceptu ale v prípade jednoduchého modelu s jedným prediktorom by som vybral na základe grafu. Zdá sa, že závislosť by mohla byť kvadratická ale podľa môjho názoru lineárny model popisuje dané dáta veľmi dobre. Ako som už popisoval v odpovedi 06 tak by som vybral konkrétne lineárny model s interceptom.

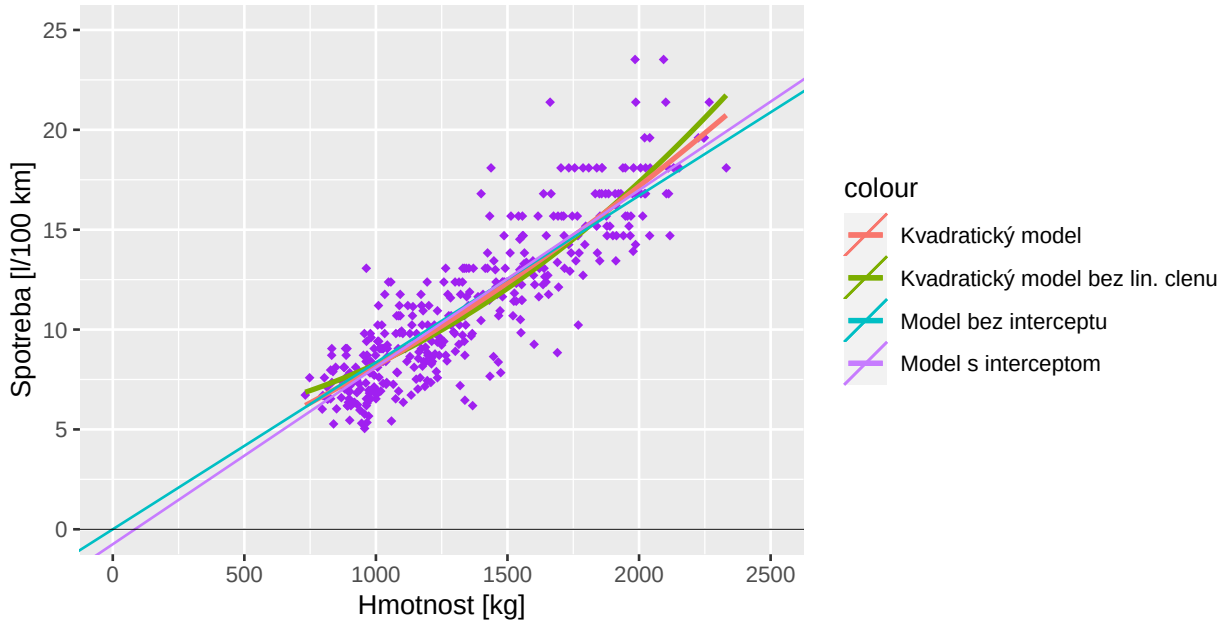

Figure 8: Vizuálne porovnanie modelov popisujúcich závislosť spotreby na hmotnosti

4.7 Otázka 09

Pro vámi vybraný finální lineární model popisující vztah mezi hmotností a spotřebou automobilu ověřte předpoklady pro použití metody nejmenších čtverců. Každý předpoklad zmiňte a uveďte jak byste ho validovali pomocí reziduí.

4.8 Odpoveď 09

Medzi predpoklady lineárnej regresie patrí predpoklad lineárneho vzťahu medzi vysvetľovanou premennou a vysvetľujúcimi premennými. Ďalším predpokladom je, že rezidua tvoria náhodný výber z normálneho rozdelenia s nulovou strednou hodnotou a rovnakým rozptylom. Posledným predpokladom je nezávislosť

pozorovaní. Deviácie od linearity by som overoval pomocou grafu 9 kde by som pozoroval či v grafe nie je pozorovateľný nejaký trend. V rovnakom grafe by som overoval predpoklad homoskedasticity a to tak, že by som pozoroval či sa so zmenou nafitovaných hodnôt spotreby mení šírka oblaku dátovych bodov. Zmena šírky naznačuje heteroskedasticitu. Nezávislosť sa taktiež dá overovať z rovnakého grafu pričom v ideálnom prípade budú dáta ležať náhodne rozptýlené okolo osy x. Normalitu rezidui môžeme skontrolovať pomocou qqplotov alebo histogramov rezidui. Dáta na obrázku 9 vyzerajú byť náhodne rozptýlené okolo osy x a nepozoroujem v nich nejaký význačný trend. Šírka daného oblaku dát vyzerá byť zhruba rovnaká vzhľadom k zmene nafitovaných hodnôt. Na obrázku 11 vidíme hisotgramový odhad hustoty rezidui. Podľa tohto grafu to vyzer, že reziduá tvoria skutočne náhodný výber zo štandardného normálneho rozdelenia. Podobný záver plynie aj z pozorovania qqplotu rezidui na obrázku 10. Predpoklady metódy najmenších štvorcov boli teda splnené.

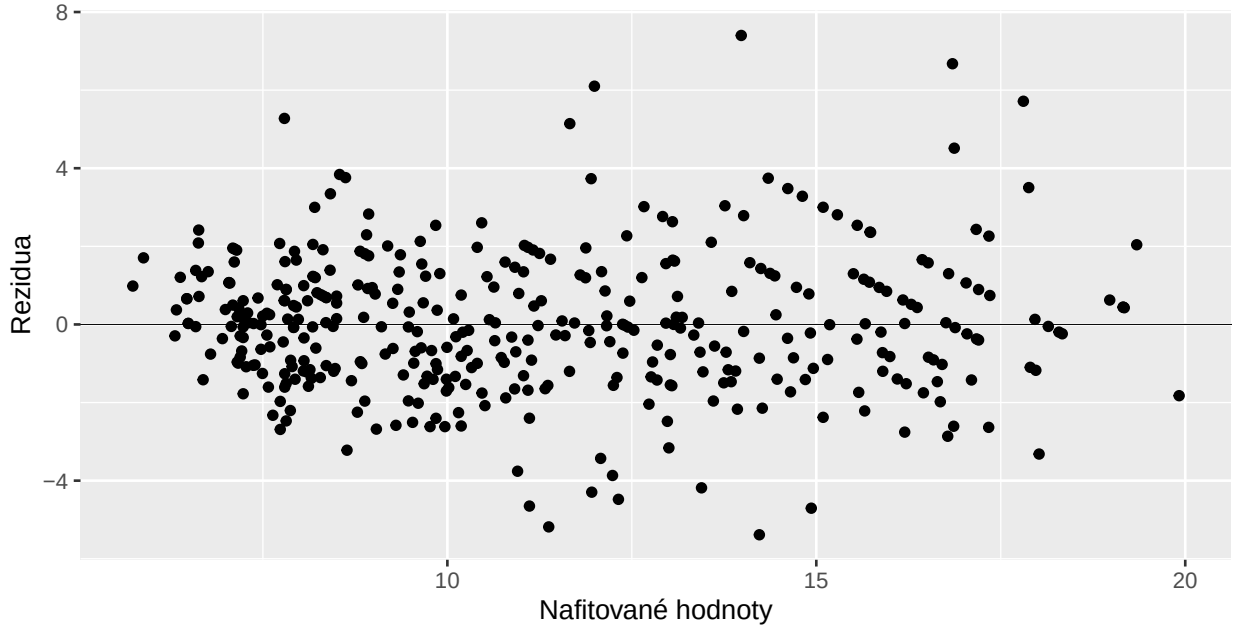

Figure 9: Graf rezidui vzhľadom k nafitovaným hodnotám spotreby

4.9 Otázka 10

Přidejte k vysvětlující proměné `hmotnost`, i proměnnou `puvod`. Navrhněte aditivní lineární model (případně 3 modely pro každý region zvlášť), ve scatterplotu vykreslete 3 skupiny různými barvami a data proložte třemi odpovídajícími regresními přímkami. Uvažujeme 3 auta o hmotnosti 2 tuny zastupující jednotlivé regiony původu. Sestrojte 90% konfidenční intervaly okolo očekávaných spotřeb a na jejich základě rozhodněte, zdali a jak se očekávané spotřeby budou lišit. Je to porovnávání správné? Zdůvoněte.

4.10 Odpoveď 10

Na obrázku 12 vidíme závislosť spotreby na hmotnosti vzhľadom k pôvodu auta. Podľa modelu pre autá z USA bude mať 2 tonové auto z USA spotrebu s prevdepodobnosťou 90% v intervale [16.8339412, 17.4500121] l/100 km. Podobne 2 tonové auto z Japonska bude spotrebu s pravdepodobnosťou 90% v intervale [12.5948108, 17.1436032] l/100 km a auto z Európy [12.5725901, 15.5604959] l/100 km. Porovnávanie týchto troch predikcii nie je vhodné nakoľko autá s hmotnosťou okolo dvoch ton majú zástupcov len medzi autami z USA. Preto

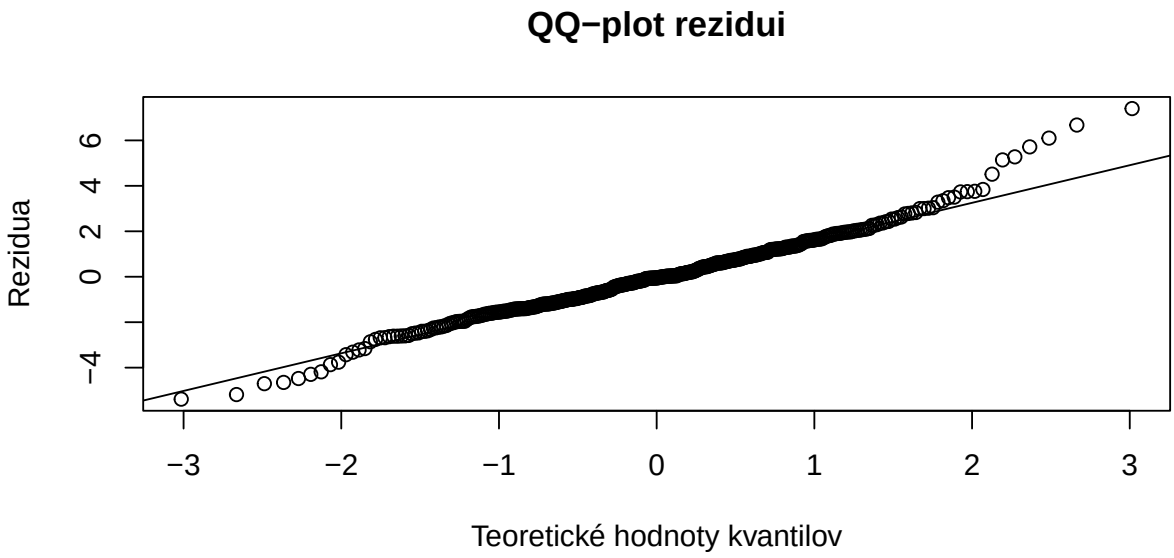

Figure 10: QQplot rezidui modelu s interceptom

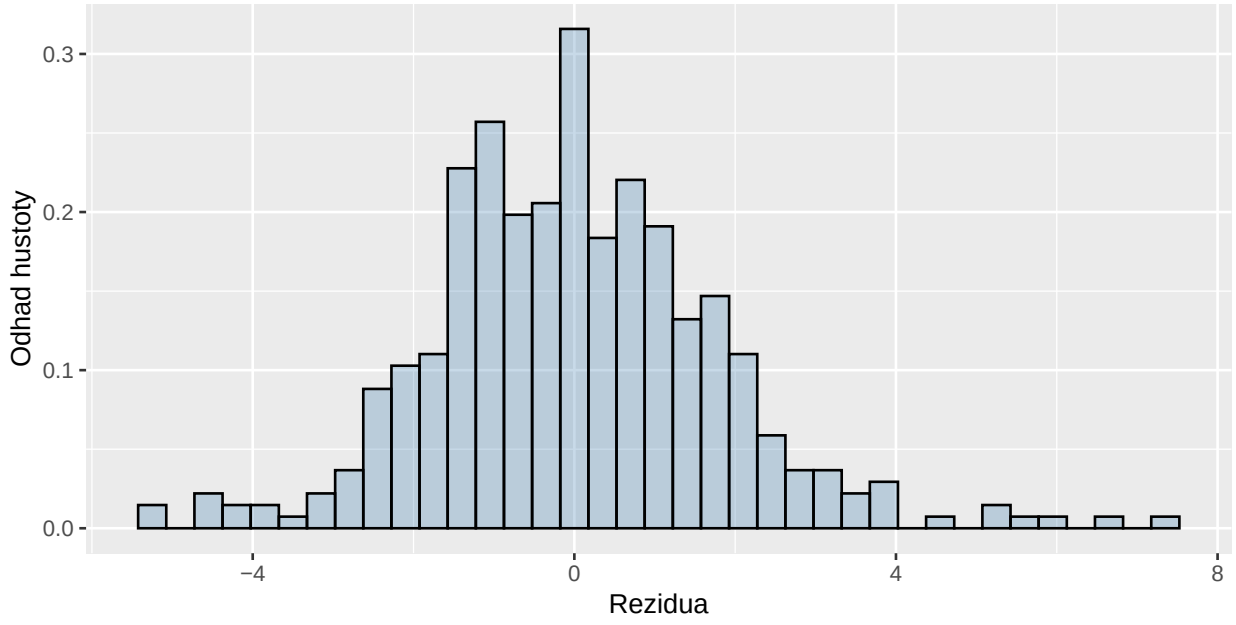

Figure 11: Histogramový odhad hustoty rezidui

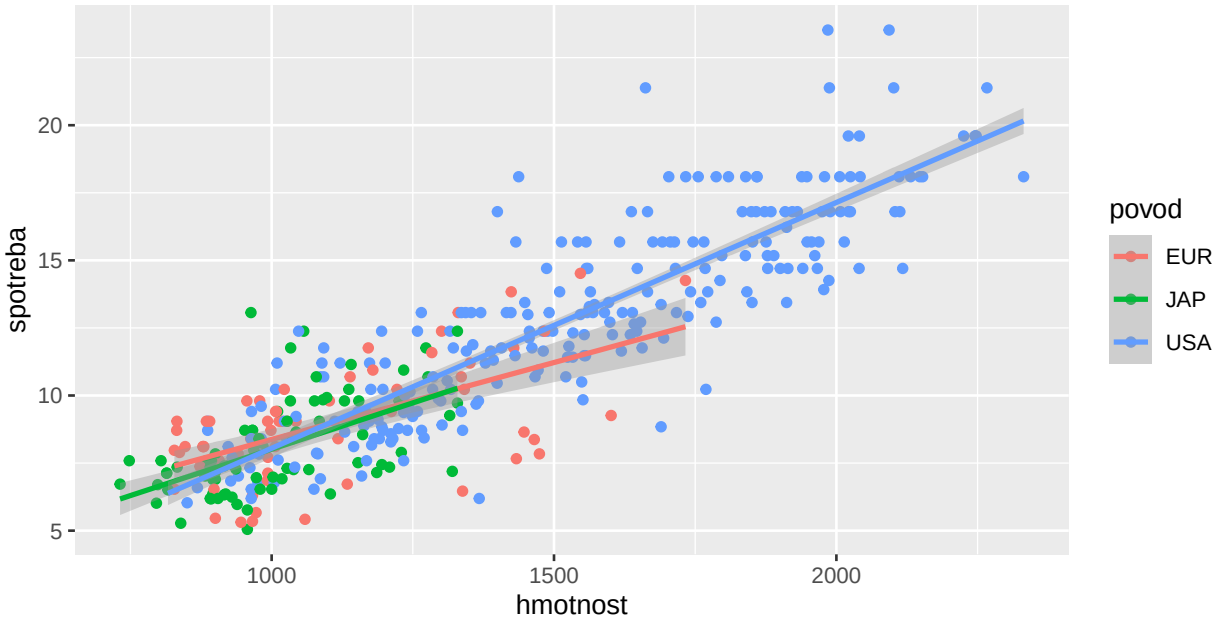

Figure 12: Závislosť spotreby na hmotnosti auta vzhľadom na jeho pôvod

nevieme či by nafitované lineárne modely pre japonské a európske autá dobre popisovali aj nové dáta mimo interval v ktorom dáta máme.