Osječki matematički list 23 (2023), 115–131

Linearna regresija kroz primjer

Mirta Benšić*, Danijel Grahovac†, Dominik Mihalčić‡

Sažetak

U radu je prezentirana osnovna ideja linearne regresije. Glavni pojmovi objašnjeni su intuitivno. Naglasak je na konkretnom primjeru gdje je cijena automobila modelirana na temelju njegovih osnovnih karakteristika. Model je izrađen na temelju stvarnih podataka o rabljenim automobilima u Hrvatskoj.

Ključne riječi: *linearna regresija, procjena parametara, reziduali*

Linear regression through example

Abstract

The paper presents the basic idea of linear regression. The main concepts are explained intuitively. The emphasis is on a real example where the price of a car is modeled based on its basic characteristics. Real data on used cars in Croatia was used in making of this model.

Keywords: *linear regression, parameter estimation, residuals*

*Fakultet primijenjene matematike i informatike, Sveučilište J. J. Strossmayera u Osijeku, email: mirta@mathos.hr

†Fakultet primijenjene matematike i informatike, Sveučilište J. J. Strossmayera u Osijeku, email: dgrahova@mathos.hr

‡Fakultet primijenjene matematike i informatike, Sveučilište J. J. Strossmayera u Osijeku, email: dmihalci@mathos.hr

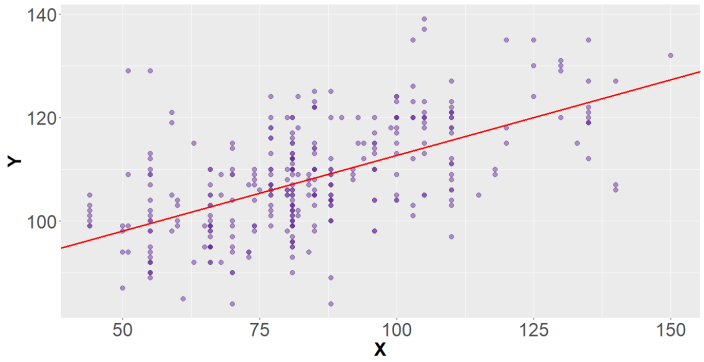
1 Uvod

Sir Francis Galton (1822.–1911.) engleski statističar i polimat

Carl Friedrich Gauss (1777.– 1855.) njemački matematičar i fizičar

Linearna regresija je statistička metoda opisivanja linearnom vezom jedne ovisne varijable pomoću jedne ili više varijabli, koje ćemo zvati **prediktori**. Naziv *regresija* dolazi od Sir Francisa Galtona koji je primijetio da je prosječna visina djece visokih roditelja bliža prosjeku cijele populacije te je tu pojavu nazvao regresija, odnosno regresija prema prosjeku. Slučaj kada imamo jedan prediktor koji opisuje ovisnu varijablu zovemo **jednostavna linearna regresija** i ona je polazišna točka za interpretaciju.

Ideja jednostavne linearne regresije je na temelju parova podataka (x_i, y_i), $i = 1, \ldots, n$, pronaći pravac $y = \alpha + \beta x$ koji te podatke najbolje opisuje u nekom smislu. Osnovni alat za to je **metoda najmanjih kvadrata** koju je još davne 1809. uveo C. F. Gauss. Pravac na slici 1 je takav da je srednje kvadratno odstupanje podataka od njega minimalno i zove se **regresijski pravac**.



Slika 1. Primjer jednostavne linearne regresije

Regresijski pravac je graf linearne funkcije koja predstavlja **model** za vezu između ovisne varijable (y) i prediktora (x). Jednom kad procijenimo jednadžbu regresijskog pravca, možemo predviđati vrijednost ovisne varijable i za one vrijednosti prediktora za koje nemamo dostupne podatke. Osim toga, koeficijent smjera α pokazuje kako se mijenja y s jediničnom promjenom x, a odsječak na y-osi β pokazuje vrijednost ovisne varijable kada je $x = 0$.

Već za više od dvije točke podataka, nemoguće je pronaći regresijski pravac koji ih savršeno opisuje, stoga će uvijek postojati greške. Pretpostavke koje greške moraju zadovoljavati su ključne da bi regresijski model bio valjan i njima treba posvetiti posebnu pažnju.

Dakle, jednostavna linearna regresija svodi se na traženje regresijskog pravca određenog dvodimenzionalnim parametrom $(\alpha, \beta) \in \mathbb{R}^2$. U općenitom modelu linearne regresije, kada imamo $k > 1$ prediktora, problem je nešto teže vizualizirati. Tada se traži **regresijska hiperravnina** određena vektorom parametara $\boldsymbol{\beta} \in \mathbb{R}^k$, no ideja ostaje ista. Najosnovnije pojmove vezane uz linearnu regresiju intuitivno ćemo objasniti u sljedećem dijelu, a nakon toga ćemo na konkretnom primjeru proći kroz osnovne karakteristike modela linearne regresije.

2 Linearna regresija

Pretpostavimo da neko slučajno obilježje Y želimo opisati kao funkciju slučajnih varijabli $X_1, \ldots, X_k$ uz grešku ε na način

$$Y = f(X_1, \ldots, X_k) + \varepsilon.$$

Slučajna varijabla Y zove se **ovisna varijabla**, a slučajne varijable $X_1, \ldots, X_k$ **prediktori** ili **regresori**. Prediktori se često zapisuju kao k-dimenzionalni slučajni vektor $\mathbf{X} = (X_1, \ldots, X_k)$. Uobičajeno je pretpostaviti da se među prediktorima nalazi konstantni član, pa ćemo bez smanjenja općenitosti pretpostaviti da je $X_1 = 1$. S obzirom da se ovdje bavimo linearnom regresijom, funkcija f je, naravno, linearna pa je

$$f(X_1, \ldots, X_k) = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \mathbf{X}^T \boldsymbol{\beta},$$

gdje je $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k) \in \mathbb{R}^k$ vektor parametara. Parametar β_1 zovemo **slobodni član** (eng. intercept). Funkcija f je u praktičnim problemima nepoznata pa ju treba procijeniti na osnovu podataka. Pretpostavljajući da je f linearna funkcija, sve što je ostalo nepoznato u vezi između $\mathbf{X}$ i Y je vektor parametara $\boldsymbol{\beta}$ i greška ε. Polazna jednadžba tako postaje

$$Y = \mathbf{X}^T \boldsymbol{\beta} + \varepsilon. \tag{1}$$

Jednadžbu (1) zovemo **linearni regresijski model**, a vektor parametara $\boldsymbol{\beta}$ **koeficijent linearne regresije**. U takvom modelu nužno je da greška ε bude nekorelirana s prediktorima $\mathbf{X}$, što osigurava da je očekivanje greške jednako 0, odnosno $E\varepsilon = 0$. Koeficijent $\boldsymbol{\beta}$ je takav da je

$$E(Y - \mathbf{X}^T \boldsymbol{\beta})^2 = E\varepsilon^2 = \operatorname{Var} \varepsilon$$

minimalno.

Dakle, cijeli model određen je vektorom koeficijenata $\boldsymbol{\beta}$ i ukoliko znamo distribucije od $\mathbf{X}$ i Y, možemo ga i točno izračunati (za više detalja vidjeti

[6]). Međutim, u stvarnoj situaciji radimo s podacima, koji su realizacije od $\mathbf{X}$ i Y, pa $\boldsymbol{\beta}$ možemo samo procijeniti.

Pretpostavimo da raspolažemo podacima $\mathbf{y} = (y_1, \ldots, y_n)$, koji su vrijednosti iz distribucije ovisne varijable Y, i $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$, koji su vrijednosti iz distribucije prediktora X_i, $i = 1, \ldots, k$. U primjeru iz sljedećeg poglavlja, $\mathbf{y}$ će biti cijene automobila, dok će $\mathbf{x}_i$, $i = 1, \ldots, k$, biti različite karakteristike automobila kao što su marka, snaga, starost, potrošnja i sl.

Vektor realnih brojeva $\mathbf{y}$ smatramo realizacijom slučajnog vektora $(Y_1, \ldots, Y_n)$ čije su komponente nezavisne, ali ne moraju biti jednako distribuirane. Slično, vektor $\mathbf{x}_i$ shvaćamo kao realizaciju slučajnog vektora $(X_{i1}, \ldots, X_{in})$, za $i = 1, \ldots, k$. Neka su $e_1, \ldots, e_n$ vrijednosti od ε. Kako stvarnu vrijednost od $\boldsymbol{\beta}$ ne znamo, to su nemjerljive veličine. Njih opet shvaćamo kao realizaciju slučajnog vektora grešaka $(\varepsilon_1, \ldots, \varepsilon_n)$ pri čemu je $E\varepsilon_i = 0$, $i = 1, \ldots, n$. Uobičajene pretpostavke na njegove komponente su sljedeće:

- $\varepsilon_1, \ldots, \varepsilon_n$ su nezavisne,
- ε_i je normalno distribuirana, za $i = 1, \ldots, n$,
- $\operatorname{Var} \varepsilon_i = \sigma^2$, za $i = 1, \ldots, n$.

Posljednji uvjet govori da varijanca mora biti ista za svaku grešku, odnosno homogena. Ako je zadovoljen, govorimo o **homoskedastičnom** modelu. Uočimo da je $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, za $i = 1, \ldots, n$.

Procjena za $\boldsymbol{\beta}$, koju ćemo označavati s $\hat{\boldsymbol{\beta}} \in \mathbb{R}^k$, standardno se određuje tako da suma kvadrata odstupanja opaženih vrijednosti y_i i teorijskih vrijednosti $\boldsymbol{x}_i^T \boldsymbol{\beta}$, odnosno

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$$

bude minimalna. Takav princip poznat je kao **metoda najmanjih kvadrata**. Može se pokazati da je tada [6]

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i. \tag{2}$$

Jednom kada su parametri $\hat{\beta}_i$, $i = 1, \ldots, k$ procijenjeni, možemo određivati vrijednosti ovisne varijable za proizvoljne vrijednosti prediktora. Posebno, možemo računati vrijednosti ovisne varijable i za zabilježene podatke

$$\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}, \quad i = 1, \ldots, n.$$

Vrijednosti $\hat{y}_i$ zovemo **teorijske vrijednosti**.

Ne možemo očekivati da prediktori savršeno opišu svaku vrijednost ovisne varijable. Tako će se teorijske vrijednosti obično razlikovati od onoga što je zabilježeno u podacima. Te razlike nazivamo **reziduali** i računamo kao

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}, \quad i = 1, \ldots, n.$$

Za razliku od grešaka $e_1, \ldots, e_n$, koje su nemjerljive veličine, reziduale možemo izračunati iz podataka. Međutim, oni nam daju samo informaciju o greškama modela te ih od njih treba razlikovati. Jednom kada se napravi procjena parametara, pretpostavke koje se odnose na slučajni vektor grešaka treba provjeriti na dobivenim rezidualima. Ta provjera ključni je korak regresijskog modeliranja koji osigurava valjanost dobivenog modela.

Osim (2), postoje i brojni drugi načini kako procijeniti nepoznate parametre. Međutim, metoda najmanjih kvadrata zaista je najbolji pristup ukoliko su zadovoljene neke pretpostavke. Postoje brojne generalizacije prethodnog modela te različiti skupovi pretpostavki i metode procjene. Više detalja o tome i regresijskim modelima može se vidjeti u [1], [6] i [7].

3 Primjer

Kreirat ćemo i analizirati linearni regresijski model koji opisuje cijenu rabljenog automobila na temelju osnovnih karakteristika, a to su: marka, starost, vrsta motora, vrsta mjenjača, snaga, prijeđeni kilometri, potrošnja goriva i emisija CO_2. Iz populacije, koju čine svi automobili koji se trenutno prodaju u Hrvatskoj, napravili smo bazu sačinjenu od 349 automobila pronađenih na internetskom oglasniku. Obuhvaćeno je 15 najprodavanijih marki automobila u Hrvatskoj starosti do 10 godina u cjenovnom rasponu do 150 000 kuna. Najprije ćemo ukratko opisati svaku od varijabli.

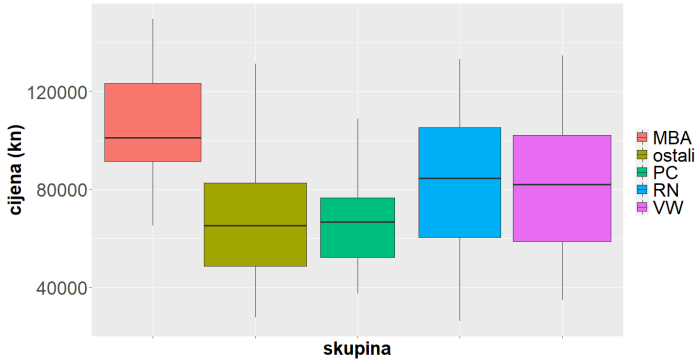
3.1 Varijable

S obzirom da baza sadrži 15 različitih marki vozila, kategorizirat ćemo ih na sljedeći način:

- VW skupina (Volkswagen, Seat, Škoda),
- Peugeot-Citroën skupina,
- Renault skupina (Renault, Dacia, Nissan),
- MBA skupina (Mercedes, BMW, Audi),

- ostali (Ford, Kia, Opel, Toyota).

Najmanji udio čine Peugeot i Citröen (13.18 %), a potom slijede Renault (16.05 %), MBA (22.35 %) i VW (23.21 %). Preostalih 25.21 % čine Ford, Kia, Opel i Toyota. Na slici 2 prikazani su kutijasti dijagrami cijene u odnosu na pet skupina marki. Kutijasti dijagram prikazuje pet karakterističnih točaka skupa podataka: minimum, donji kvartil, medijan, gornji kvartil i maksimum (vidi [2] za više detalja). Možemo vidjeti da su najjeftinija vozila iz Peugeot-Citroën skupine, dok među najskuplje ubrajamo Mercedes, BMW i Audi. Marka će očekivano imati velik efekt na cijenu.



Slika 2. Kutijasti dijagrami cijene u odnosu na kategoriju automobila

Prosječni automobil u Hrvatskoj star je 15 godina, što nas prema starosti voznog parka svrstava ispod europskog prosjeka. Međutim, većina automobila u bazi stara je između 5 i 8 godina pa možemo reći da sadržava relativno mlade automobile.

U bazi prevladavaju automobili s dizelskim motorom (85.1 %), dok preostalih 14.9 % čine benzinci. Hibridni i električni automobili nisu obuhvaćeni jer su još uvijek rijetka pojava na hrvatskim cestama. Automobili s ručnim mjenjačem čine 82.52 % uzorka, a ostatak čine automatici. Prosječna cijena automatika iznosi 100 356.30 kn što je znatno više od prosječne cijene automobila s ručnim mjenjačem (77 208.28 kn).

Za preostale četiri numeričke varijable (snagu, prijeđene kilometre, potrošnju goriva i emisiju CO_2) navodimo deskriptivnu statistiku. Snaga je izražena u kW, potrošnja u l/100 km te emisija u g/km.

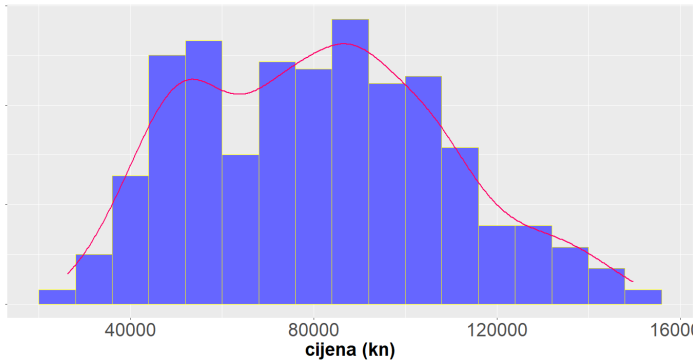
Tablica 1. Deskriptivna statistika preostalih numeričkih varijabli

	Min.	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Max.
kW	44	70	81	84.41	96	150
km	26 238	109 000	143 000	141 549	175 300	284 000
l/100 km	3.5	4.5	5	5.02	5.5	6.5
g/km	84	100	107	108.2	115	139

Ovisna varijabla je cijena automobila. Izražena je u kunama, a kreće se između 26 204 kn i 149 705 kn s prosjekom 81 254 kn. Gornja granica od 150 000 kn odabrana je s razlogom jer je pretpostavka da bi si osoba za iznose veće od toga radije priuštila novi nego rabljeni automobil. Empirijska raspodjela cijena promatranih automobila ilustrirana je histogramom na slici 3.

Tablica 2. Deskriptivna statistika cijene automobila (u kunama)

Min.	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Max.
26 204	57 591	81 894	81 254	101 101	149 705

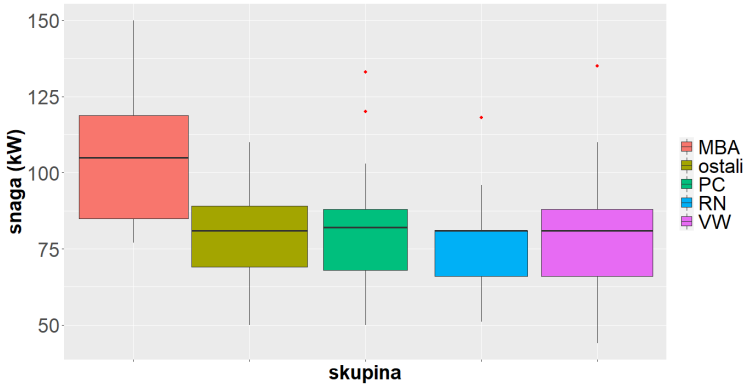


Slika 3. Histogram cijene automobila

Varijable koje ćemo koristiti kao prediktore nisu ni u kojem smislu neza-

visne. Primjerice, za očekivati je da postoji rastuća veza između snage te potrošnje i emisije CO_2. Naime, jači automobili u pravilu imaju veću potrošnju goriva pa posljedično i proizvode veću količinu CO_2. Pearsonov koeficijent korelacije između snage i potrošnje iznosi 0.52, a između snage i emisije 0.59, što potvrđuje navedene veze.

Još jedan očiti primjer je veza između starosti i prijeđenih kilometara. Pearsonov koeficijent korelacije za te dvije varijable iznosi 0.58. Značajna je i veza između snage i marke automobila. Štoviše, pokazuje se da postoje razlike u distribucijama snage među 5 skupina marki. To je ilustrirano kutijastim dijagramom na slici 4, a može se potvrditi provođenjem *Kruskal-Wallis* testa koji daje p-vrijednost vrlo blisku 0 (vidjeti [8] za više detalja o ovom statističkom testu). Toj vezi trebat će posvetiti posebnu pozornost kada se bude analizirala *multikolinearnost* u okviru modela.



Slika 4. Kutijasti dijagrami snage u odnosu skupinu

3.2 Transformacija varijabli

Procjena parametara modela u praksi nije naročito teška. Primjerice, u programskom jeziku R [9] dobiva se pomoću funkcije `lm()` kojoj proslijedimo ovisnu varijablu, zatim simbol "$\sim$" i oznake za prediktore odvojene znakom "+". No, da bi model bio valjan, puno stvari treba uzeti u obzir.

Prilikom izrade modela za cijenu, odmah u startu dvije stvari trebalo je korigirati. Ranije smo spomenuli da je homoskedastičnost grešaka poželjno svojstvo. Međutim, greške su nemjerljive veličine pa ga provjeravamo pomoću reziduala. Linearni model za cijenu nije bio homoskedastičan, stoga ga je trebalo promijeniti tako da bude. Jedan od načina za to je transformacija ovisne varijable Y pomoću *Box-Cox* procedure. Ona daje

parametar $\lambda \in \mathbb{R}$ koji tada sugerira sljedeću transformaciju

$$\begin{cases} Y^{\lambda}, & \text{ako je } \lambda \neq 0, \\ \log Y, & \text{ako je } \lambda = 0. \end{cases}$$

Dobiveni λ iznosio je $0.4848 \approx 0.5$ što je sugeriralo transformaciju ovisne varijable drugim korijenom. Dakle, u nastavku umjesto originalnih podataka o cijeni, modeliramo drugi korijen iz cijene, što nam osigurava homoskedastičnost. Za procjene dobivene modelom to ne predstavlja problem jer samo trebamo kvadrirati procjene cijene koje dobijemo iz modela. Osim toga, korijen je rastuća funkcija pa veća cijena znači i veći korijen cijene, i obratno.

U modelu ćemo koristiti numeričke i kategorijalne varijable. Dodavanje numeričkih varijabli u model je prilično jednostavno. Pripadni koeficijent govori kako se prosječno mijenja ovisna varijabla uslijed jediničnog povećanja prediktora, uz ostale uvjete nepromijenjene (lat. *ceteris paribus*). S druge strane, dodavanje kategorijalne varijable u model zahtijeva nešto veću pažnju. Kategorijalnu varijablu s dvije kategorije najprije je potrebno transformirati u 0-1 (dihotomnu) varijablu. Primjerice, varijabla *motor* ima dvije kategorije: *dizel*, koju ćemo označiti s 0, i *benzin*, koju ćemo označiti s 1. Kategorija označena s 0 naziva se **bazna** i ključna je u interpretaciji koeficijenata. U ovom primjeru, koeficijent uz takvu varijablu se interpretira kao prosječna razlika u korijenu iz cijene između dizelaša i benzinaca. Ako je taj koeficijent negativan, to znači da su u prosjeku dizelaši skuplji od benzinaca. U slučaju kada imamo $m > 1$ kategorija, potrebno je kreirati $m - 1$ dihotomnih varijabli te (proizvoljno) identificirati baznu kategoriju.

Marku automobila iz baze, koja ima 15 kategorija, smo reducirali na 5 podjednakih skupina: Peugeot-Citröen (PC), Renault (RN), Volkswagen (VW), MBA i ostali. Korištenjem takve varijable u modelu, dobili bismo 4 koeficijenta: $\beta_{marka:PC}$, $\beta_{marka:RN}$, $\beta_{marka:VW}$ i $\beta_{marka:ostali}$ (kategorija MBA uzeta je kao *bazna*). S obzirom da u modelu imamo još 7 prediktora, ideja je dodatno smanjiti broj kategorija pa u tu svrhu testiramo hipotezu

$$\mathcal{H}_0 : \ \beta_{marka:ostali} = \beta_{marka:PC} \ \ \& \ \ \beta_{marka:ostali} = \beta_{marka:RN}.$$

Korištenjem *Waldovog testa* (vidi [5]) dobivamo p-vrijednost $0.1651 > 0.05$ što sugerira „spajanje" 3 kategorije (PC, RN i ostali) u jednu koju ćemo nazvati *preostali*.

Sada kada je marka automobila podijeljena u tri kategorije, trebat ćemo dvije dihotomne varijable u modelu. Prvu ćemo označiti s *marka:preostali*, koja ima vrijednost 1 ako je marka iz skupine preostalih i 0 inače, a drugu s *marka:VW*, koja se definira analogno. Dakle, za baznu kategoriju bit će

ponovno uzeti automobili čija je marka iz skupine MBA. Interpretacija koeficijenata je slična kao i u slučaju dodavanja varijable koja ima dvije kategorije. Primjerice, koeficijent uz *marka:VW* govori za koliko se prosječno razlikuje korijen iz cijene takvih automobila u odnosu na baznu skupinu (MBA).

Dakle, ovisna varijabla u modelu bit će korijen iz cijene, a marka automobila podijeljena u 3 skupine:

- VW (Volkswagen, Seat, Škoda),
- MBA (Mercedes, BMW, Audi),
- preostali (Peugeot, Citroën, Renault, Dacia, Nissan, Ford, Kia, Opel, Toyota).

Osim nje imamo još 7 prediktora.

3.3 Procjena parametara modela

Nakon prethodno opisanih transformacija cijene i marke automobila, možemo procijeniti parametre modela koristeći R. Rezultati procjene navedeni su u tablici 3.

Tablica 3. Tablični prikaz modela

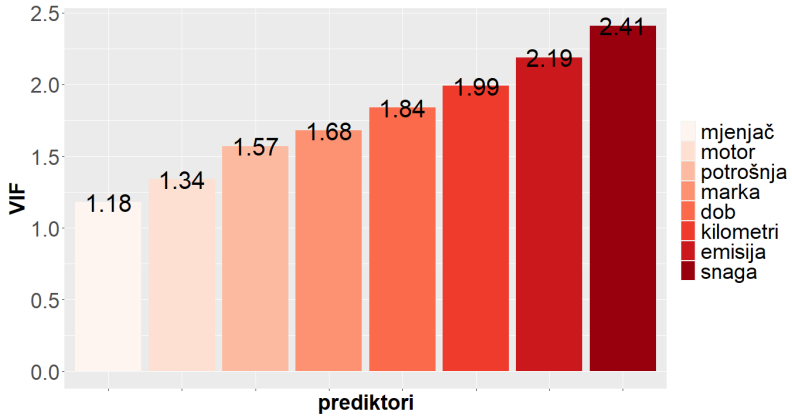
Varijabla	Koeficijent	Standardna greška	p
slobodni član	308.67	18.25	≈ 0
marka:preostali	-96.30	15.10	$6.0{\cdot}10^{-10}$
marka:VW	-94.50	16.01	$8.8{\cdot}10^{-9}$
motor:benzin	-18.19	3.66	$1.0{\cdot}10^{-6}$
dob	-13.22	0.70	≈ 0
mjenjač:automatski	9.26	3.14	$3.4{\cdot}10^{-3}$
potrošnja	7.26	2.17	$9.3{\cdot}10^{-4}$
emisija	0.56	0.16	$4.8{\cdot}10^{-4}$
snaga	0.35	0.13	$7.3{\cdot}10^{-3}$
kilometri	$-1.70{\cdot}10^{-4}$	$3.00{\cdot}10^{-5}$	$3.9{\cdot}10^{-9}$
marka:preostali*snaga	0.62	0.17	$2.4{\cdot}10^{-4}$
marka:VW*snaga	0.48	0.16	$2.5{\cdot}10^{-3}$

U prvom stupcu nalaze se varijable korištene u modelu. Od njih su četiri

numeričke i tri kategorijalne. Zadnja dva retka odnose se na interakciju između marke i snage, a nju ćemo komentirati nešto kasnije.

Pomoću standardne greške možemo kreirati 95 %-tni pouzdani interval za procijenjeni koeficijent. Manja standardna greška daje uži pouzdani interval pa je time procjena za koeficijent preciznija. Ako pouzdani interval ne sadrži 0, koeficijent je statistički značajno različit od 0 ili, kraće, statistički značajan. To odgovara činjenici da je p-vrijednost u zadnjem stupcu manja od 0.05, što je upravo naš slučaj.

Prije same interpretacije, potrebno je provjeriti postoji li u modelu **multikolinearnost**. Ona podrazumijeva veliku koreliranost između prediktora, a njezina pojava narušava kvalitetu modela na više načina: otežava odabir prediktora za model, koeficijenti imaju veliku standardnu grešku i dovodi u pitanje interpretabilnost modela. Najčešće korištena mjera multikolinearnosti su VIF-ovi (eng. *variance inflation factor*). VIF se izračunava za pojedini prediktor te ako je veći od 5, model moramo praviti ispočetka. Na sreću, multikolinearnost nam ne predstavlja problem jer su svi VIF-ovi manji od 2.5, a najveći iznosi 2.41 i odnosi se na snagu (slika 5). Dodavanjem interakcijskog člana snage i marke u model osigurana je jasnija interpretacija efekta snage na cijenu automobila.



Slika 5. Faktori inflacije varijance

Da bismo mogli interpretirati koeficijente, trebaju nam njihovi *marginalni efekti*. Oni govore kako promjena u prediktoru utječe na ovisnu varijablu. U slučaju kada u modelu nema interakcija, oni su jednaki procijenjenim koeficijentima, no kod nas su nešto drugačiji za one varijable koje imaju značajnu interakciju, a dani su u tablici 4.

Tablica 4. Marginalni efekti

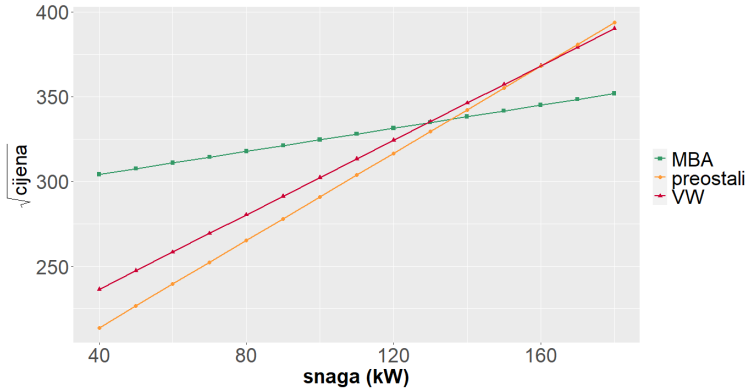
Varijabla	Marginalni efekt
marka:preostali	-56.04
marka:VW	-42.40
motor:benzin	-18.19
dob	-13.22
mjenač:automatski	9.26
potrošnja	7.26
emisija	0.56
snaga	0.75
kilometri	$-1.70 \cdot 10^{-4}$

Konačno, interpretacija koeficijenata u modelu je sljedeća:

- ako je marka iz skupine VW, to uzrokuje prosječno smanjenje korijena iz cijene za 42.40 u odnosu na skupinu MBA,
- ako je marka iz skupine preostalih, to uzrokuje prosječno smanjenje korijena iz cijene za 56.04 u odnosu na skupinu MBA,
- ako je automobil benzinac, to uzrokuje prosječno smanjenje korijena iz cijene za 18.19 u odnosu na dizelaša, a ako je s automatskim mjenjačem, prosječno povećanje korijena iz cijene za 9.26 u odnosu na automobil s ručnim mjenjačem
- za svaku godinu starosti, korijen iz cijene automobila prosječno se smanji za 13.22,
- jedinično povećanje potrošnje uzrokuje povećanje korijena iz cijene za 7.26, a jedinično povećanje emisije CO_2 uzrokuje prosječno povećanje od 0.55,
- za svaki dodani kW snage, korijen iz cijene se prosječno poveća za 0.75,
- za svakih prijeđenih 100 000 km, korijen iz cijene automobila prosječno se smanji za 17.

Potrebno je još razjasniti ulogu interakcijskog člana marke i snage. S obzirom da marka ima tri kategorije (dvije nisu bazne), imamo dva interakcijska koeficijenta, a interpretiraju se na sljedeći način:

- ukoliko je marka iz skupine preostalih, jedinično povećanje snage prosječno povećava korijen iz cijene za 0.6173,
- ukoliko je marka iz skupine VW, jedinično povećanje snage prosječno povećava korijen iz cijene za 0.4769.



Slika 6. Interakcija snage i marke

To je jasnije prikazano na slici 6. Poanta je u tome da se utjecaj snage na cijenu mijenja ovisno o marki automobila. Primjerice, automobili iz skupine MBA (Mercedes, BMW, Audi) smatraju se luksuznijima pa dodani kW snage njihovu cijenu ne povećava previše. S druge strane, ako se recimo radi o Peugeotu (skupina preostali) koji se smatra automobilom srednje klase, dodani kW će mu znatno povećati cijenu.

3.4 Analiza kvalitete modela

Najčešće korištena mjera kvalitete linearnog regresijskog modela je R^2 koji se definira na sljedeći način

$$R^2 = 1 - \frac{(n-1)\sum_{i=1}^{n} \hat{e}_i^2}{(n-k)\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Ovdje je n broj podataka, k broj prediktora, a $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Interpretira se kao *udio varijabilnosti podataka opisan modelom* pa je poželjno da je što veći. U našem slučaju on iznosi 0.8234 što sugerira da je ovaj model pouzdan i koristan za predikciju.

Posebnu pažnju potrebno je posvetiti rezidualima koje smo dosad već nekoliko puta spomenuli. Njihove vrijednosti kreću se između -58 164 i

63 972 s medijanom -0.668 (tablica 5). S obzirom da među prediktorima imamo konstantni član, prosjek je izostavljen jer je on u takvim modelima uvijek 0.

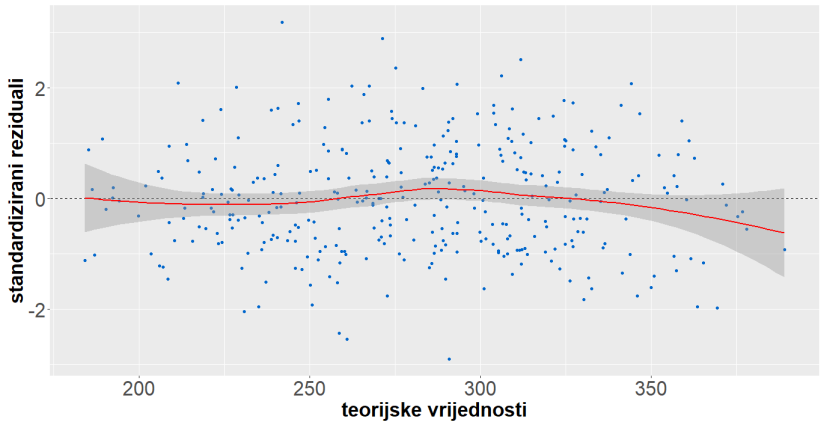
Tablica 5. Deskriptivna statistika reziduala

Min.	Donji kvartil	Medijan	Gornji kvartil	Max.
-58 164	-15 413	-0.668	13 568	63 972

Provođenje *ncv* testa [3] daje p-vrijednost 0.458, što svjedoči o homoskedastičnosti, no to je osigurano transformacijom ovisne varijable drugim korijenom prije samog kreiranja modela. Na slici 7 prikazani su *standardizirani* reziduali $\hat{r}_i$, $i = 1, \ldots, n$. Oni se računaju kao

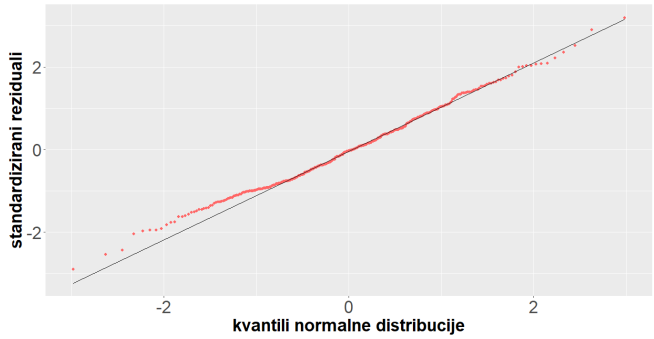
$$\hat{r}_i = \frac{\hat{e}_i}{\sqrt{\frac{1}{n-k}\sum_{i=1}^{n} \hat{e}_i{}^2 (1 - h_{ii})}},$$

gdje je h_{ii} dijagonalni element *hat*-matrice (vidi [6]). Ako su po apsolutnoj vrijednosti veći od 3, podatak od kojega dolaze smatramo *stršećim*, o čemu će više riječi biti u nastavku. Homoskedastičnost se očituje u njihovoj ujednačenoj raspršenosti, a crvena linija, koja bi trebala biti što bliže 0, aproksimira njihov prosjek.



Slika 7. Standardizirani reziduali

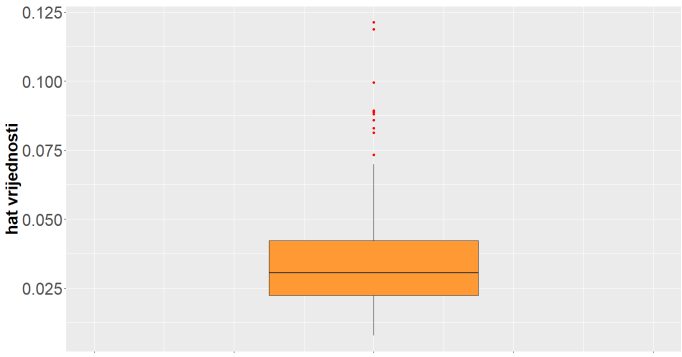
Poželjna je i normalnost standardiziranih reziduala koja se da naslutiti iz QQ-grafa na slici 8. Provođenje *Shapiro-Wilk* [10] testa o normalnosti, zaključujemo da hipotezu o normalnoj distribuiranosti ne odbacujemo uz p-vrijednost 0.133.



Slika 8. QQ-plot standardiziranih reziduala

Na kraju analize modela, potrebno je komentirati stršeće vrijednosti (eng. *outliers*). To su podaci u bazi koji značajno odstupaju od ostalih, a imaju mogućnost pokvariti regresijski model. Stoga ih je potrebno detektirati i odlučiti treba li ih ostaviti u modelu ili ne.

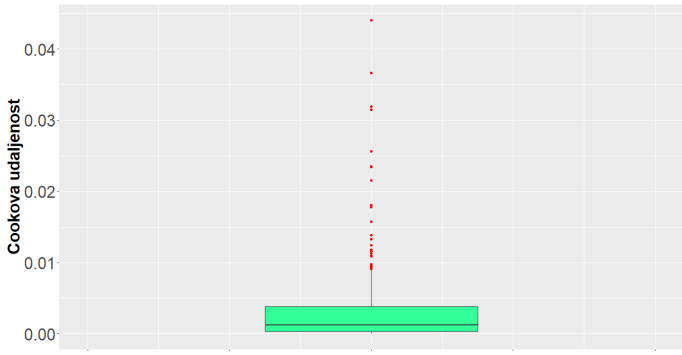
Prilikom detektiranja stršećih vrijednosti, najčešće se koriste *hat*-vrijednosti i *Cookova udaljenost*. Hat-vrijednost nekog podatka govori koliko je on u određenom smislu udaljen od prosjeka svih podataka. Ako je ona veća od dozvoljene granice, reći ćemo da pripadni podatak ima velik utjecaj (eng. *leverage*). Kutijasti dijagram hat-vrijednosti dan je na slici 9.



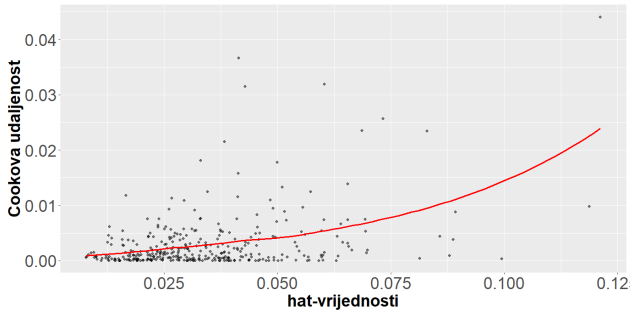
Slika 9. Hat-vrijednosti

Za dozvoljenu granicu uzet je broj $\frac{4k}{n} = 0.09169$, no može se uzeti $\frac{2k}{n}$, $\frac{3k}{n}$ i sl. (vidjeti poglavlje 4 u [4] za više detalja). Kod tri podatka, hat-vrijednosti prelaze tu granicu. Konkretno, riječ je o sljedećim automobilima: *Seat Leon* (86 123 kn), *Seat Leon* (106 269 kn), *Škoda Roomster* (34 283 kn).

Cookova udaljenost za neki podatak pokazuje za koliko bi se cijeli model promijenio ako bismo taj podatak izbacili. Kutijasti dijagram njihovih vrijednosti dan je na slici 10. Maksimalna Cookova udaljenost u našem slučaju iznosi 0.044, a postiže se na 265. podatku (već spomenuti *Seat Leon* po cijeni od 106 269 kn). Dozvoljena granica za Cookovu udaljenost podatka iznosi 0.9288 pa prema tom kriteriju, niti jedan podatak nije detektiran kao stršeća vrijednost.



Slika 10. Cookova udaljenost



Slika 11. Cookova udaljenost i hat-vrijednosti

Na slici 11 prikazane su hat-vrijednosti i Cookova udaljenost za sve podatke. Ustanovili smo da podatak s rednim brojem 265. ima najveću hat-

vrijednost, a ujedno i najveću Cookovu udaljenost. To ne znači nužno da taj podataka treba izbaciti iz modela, već samo da treba dodatno istražiti zbog čega se on izdvaja od drugih. Međutim, ispostavilo se da je model jednako kvalitetan s tim podatkom, kao i bez njega pa smo ga odlučili ostaviti u modelu.

Literatura

[1] A. Basilevsky, *Statistical Factor Analysis and Related Models: Theory and Applications*, Wiley-Interscience, New York, 1994.

[2] M. Benšić, N. Šuvak, *Uvod u vjerojatnost i statistiku*, Sveučilište J. J. Strossmayera u Osijeku, Odjel za matematiku, Osijek, 2022.

[3] T. S. Breusch, A. R. Pagan, *A simple test for heteroscedasticity and random coefficient variation*, Econometrica, **47**(5) (1979), 1287–1294.

[4] S. Chatterjee, A. S. Hadi, *Sensitivity analysis in linear regression*, John Wiley & Sons, New York, 1988.

[5] N. R. Draper, H. Smith, *Applied Regression Analysis*, John Wiley & Sons, 1998.

[6] B. Hansen, *Econometrics*, Princeton University Press, 2022.

[7] F. E. Harrell, *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*, Springer, 2001.

[8] M. Hollander, D. A. Wolfe, *Nonparametric Statistical Methods*, John Wiley & Sons, 1973.

[9] *R - project*, URL: https://www.r-project.org

[10] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, Chapman and Hall/CRC, 2003.