SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Zorica Grmoja

PRIMJENA STATISTIČKIH METODA U OPTIMIZACIJI CIJENA U MALOPRODAJI

Diplomski rad

Voditelj rada:
prof.dr.sc. Siniša Slijepčević

Zagreb, 02.2018.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Jedno veliko HVALA mojim roditeljima koji su mi sve ovo omogućili i koji su velikim dijelom zaslužni za ovaj uspjeh. Bili ste moj oslonac kad je bilo najteže i dijelili sa mnom sreću nakon svake 'pobjede'. Zahvaljujem se i mentoru na ukazanom povjerenju i suradnji prilikom izrade ovog rada.
I svima vama koji ste cijelo ovo vrijeme bili uz mene.
Hvala vam!

Sadržaj

Uvod

Podaci su najveće blago koje neka kompanija može posjedovati, a opet jako malo njih to shvaća. Oni su ono nešto bez čega uskoro nećemo moći ostvarivati značajnije uspjehe jer upravo se u njima kriju najveće moći, a to su moći predviđanja. Pomoću podataka možemo predvidjeti ponašanje tržišta, ali i samih korisnika tog istog tržišta. Upravo će djelić toga biti tema ovog diplomskog rada, a to je određivanje elastičnosti cijena preko koje kasnije možemo kreirajući cijene utjecati na potrošaćke moći krajnjih korisnike.

Za početak, u ovom radu odlučili smo koristiti metodu linearne regresije za dobivanje željenih rezultata pa ćemo tako u Poglavlju 1 i Poglavlju 2 predstaviti tu metodu. U prvom poglavlju opisujemo jednostruku linearnu regresiju, koju u drugom poglavlju nadograđujemo na višestruku linearnu regresiju. Ovaj dio rada je teorijske prirode gdje preko iskazivanja tvrdnji pa onda i njihovog dokazivanja predstavljamo linearnu regresiju i sve vezano uz nju.

Poglavlje 3 služi nam kao kratki uvod u ekonomsku stranu cijele ove priče. Donosi nam pojam elastičnosti cijena i njene važnosti u poslovnom svijetu.

Sljedeća dva poglavlja su primjena odabrane metode na stvarne podatke jednog poduzeća, kojeg ne imenujemo zbog povjerljivosti dobivenih podataka. U Poglavlju 4 donosimo opisnu statistiku preko koje na slikovit glafički način dobivamo uvid u odnose među podacima i njihove vrijednosti. Za kraj u Poglavlju 5 provodimo linearnu regresiju na podacima i nakon toga analizu dobivenih rezultata. Nakon toga slijedi prikazi kodova korišteni za analizu i obradu dobivenih podataka.

Poglavlje 1

Jednostruka linearna regresija

Jednostruka linearna regresija je osnova na kojoj se kasnije grade složeniji regresijski modeli. Jednostrukom linearnom regresijom opisujemo vezu između podataka i to na linearan način tako da zapravo gradimo model kojem je osnovica regresijski pravac. Model izgleda ovako:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1.1}$$

Jednostavni linearni model možemo zapisati i u obliku:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{1.2}$$

gdje nam je n zapravo broj opažanja. U ovom jednostavnom regresijskom modelu β_0 i β_1 imaju jednostavne interpretacije. Kad je $x = 0$ u (1.2), $y = \beta_0$. Izraz β_0 često se naziva *konstantni član regresijske jednadžbe*. Za svaku jedinicu porasta u x, y se poveća za β_1, što se često referira kao *nagib regresijske jednadžbe*.

Važno je primijetit da su x_i-evi u ovom modelu samo brojevi, a nikako slučajne varijable. Stoga nema smisla pričati o njihovoj distribuciji. ε_i-evi su slučajne varijable kao što su i y_i-evi pošto ovise o ε_i-evima. Slučajne varijable y_i nazivaju se *opažanja*, dok $x_1, \ldots, x_n$ predstavljaju *točke projekcije* koje odgovaraju y_i-evima.

1.1 Procjena parametara β_0, β_1 i σ^2

Koristeći slučajni uzorak s n opažanja $y_1, y_2, \ldots, y_n$ i pridružene $x_1, x_2, \ldots, x_n$ možemo procijeniti parametre β_0, β_1 i σ^2. Da bismo dobili procjene $\hat{\beta}_0$ i $\hat{\beta}_1$ koristit ćemo metodu najmanjih kvadrata.

U metodi najmanjih kvadrata tražimo procjenitelje β_0 i β_1 koji minimiziraju sumu kvadrata odstupanja y_i - $\hat{y}_i$ promatranih (opažajnih) y_i-ieva od njihovih predviđenih vrijednosti $\hat{y}_i = \beta_0 + \beta_1 x_i$:

$$\begin{aligned}\mathcal{S} &= \sum_{i=1}^{n} \hat{\varepsilon_i}^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \end{aligned} \tag{1.3}$$

Primijetimo da predviđena vrijednost $\hat{y}_i$ procjenjuje $E(y_i)$, a ne y_i, tj. $\hat{\beta}_0 + \hat{\beta}_1 x_i$ procjenjuje $\beta_0 + \beta_1 x_i$, a ne $\beta_0 + \beta_1 x_i + \varepsilon_i$.

Kako je parcijalna derivacija (1.3) s obzirom na $\hat{\beta}_0$ i $\hat{\beta}_1$ jednaka

$$\frac{\partial \mathcal{S}}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \tag{1.4}$$

$$\frac{\partial \mathcal{S}}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i \tag{1.5}$$

za pronalazak vrijednosti $\hat{\beta}_0$ i $\hat{\beta}_1$ koje minimiziraju $\mathcal{S}$ u (1.3) izjednačimo (1.4) i (1.5) s nulom. Tada iz (1.4)

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \tag{1.6}$$

i zamjene $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ te $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ dobivamo

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{1.7}$$

Iz (1.5) slijedi

$$\sum_{i=1}^{n} y_i x_i - n\hat{\beta}_0 \bar{x} \tag{1.8}$$

što, kad uvrstimo (1.7) kao zamjenu za $\hat{\beta}_0$, nam daje

$$\sum_{i=1}^{n} y_i x_i - n\beta_0 \bar{x} \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0. \tag{1.9}$$

Stoga

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}. \tag{1.10}$$

Propozicija 1.1.1. *Ekvivalentan zapis za procjenitelja* $\hat{\beta}_1$ *je*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{1.11}$$

Dokaz. Kako vrijedi $\sum_{i=1}^n (y_i - \bar{y})\bar{x_1} = \bar{x_1} \sum_{i=1}^n (y_i - \bar{y}) = 0$ imamo

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})(x_{i1} - \bar{x_1}) &= \sum_{i=1}^n (y_i x_{i1} - \bar{y} x_{i1}) - \bar{x_1} \sum_{i=1}^n (y_i - \bar{y}) \\
&= \sum_{i=1}^n y_i x_{i1} - \bar{y} \sum_{i=1}^n x_{i1} \\
&= \sum_{i=1}^n y_i x_{i1} - n\bar{y}\bar{x_1}
\end{aligned}$$

uz to i

$$\begin{aligned}
\sum_{i=1}^n (x_{i1} - \bar{x_1})^2 &= \sum_{i=1}^n x_{i1}^2 - 2\bar{x_1} \sum_{i=1}^n x_{i1} + n\bar{x_1}^2 \\
&= \sum_{i=1}^n x_{i1}^2 - 2n\bar{x_1}^2 + n\bar{x_1}^2 \\
&= \sum_{i=1}^n x_{i1}^2 - n\bar{x_1}^2.
\end{aligned}$$

Stoga, koeficijent u (1.10) može se zapisati i kao

$$\hat{\beta}_1 = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \,/\, \sum_{i=1}^n (x_i - \bar{x})^2$$

□

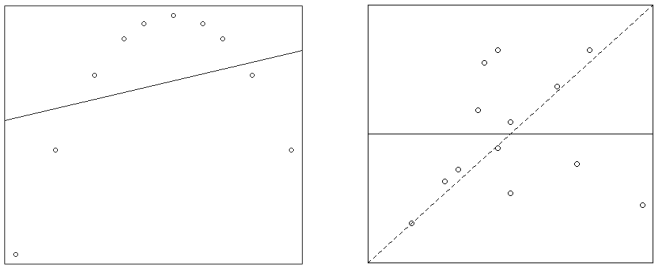
1.2 Gaus-Markovljevi uvjeti i analiza metode najmanjih kvadrata

Pitanje koje si postavljamo u ovoj odjeljku je koliko zapravo metoda najmanjih kvadrata dobro procjenjuje β-e. Zasada možemo reći da metoda najmanjih kvadrata daje

dobre procjenitelje ako su zadovoljeni određeni uvjeti (takozvani *Gauss-Markovljevi uvjeti*). Da bismo pokazali potrebu za tim uvjetima pogledajmo slučajeve navedene u [5] gdje bismo jako teško došli do dobrih procjenitelja.

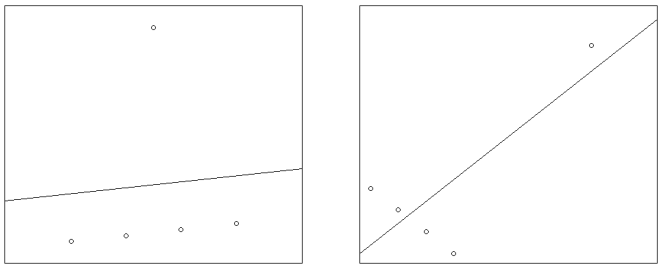
Primjer na 1.1a nam prikazuje slučaj kad ravna linija nije prigodna i kao rezultat vrlo vjerojatno nećemo dobiti dobre procjenitelje. Da bismo izbjegli takve situacije napravimo uvjet

$$E(\varepsilon_i) = 0, \forall i = 1, 2, ..., n. \tag{1.12}$$



(a) (b)

Slika 1.1: Primjer kršenja nekih Gauss-Markovljevih uvjeta



(a) (b)

Slika 1.2: Primjer outlinera i utjecajnih točki

Uvodimo oznaku $'\mathcal{I}'$ za udaljenu točku na prikazanim slikama

Ovo implicira da je očekivanje $E(y_i)$ od y_i zapravo $\beta_0+\beta_1 x_i$ u slučaju jednostavne linearne regresije kao i u slučaju višestruke linearne regresije.

Drugi tip problema koji bismo trebali spriječiti je prikazan na slici 1.1b. Ovdje pretpostavljamo da je ispravan model prikazan istočkanom linijom te da vrijedi (1.12), ali varijanca $\mathrm{Var}(\varepsilon_i)$ od ε_i povećava se s x_i. Nekolicina točaka koje su udaljene od istočkane linije mogu prouzročiti da pravac linearne regresije bude jako loš procjenitelj podataka, upravo kao što je prikazan pravac na 1.1b. Ovakav slučaj često se naziva *heteroskedastičnost* je ispravljen sljedećim uvjetom

$$\mathrm{Var}(\varepsilon_i) = E(\varepsilon_i - E(\varepsilon_i))^2 = E\left(\varepsilon_i^2\right) = \sigma^2, \forall i = 1, 2, ..., n. \tag{1.13}$$

Ponekad samo jedna točka ili jako mali skup točaka može prekršiti (1.12) i/ili (1.13). Na slici 1.2a da točka $'\mathcal{I}'$ nije prisutna regresijski pravac bi više manje prošao kroz preostale četiri točke. Međutim s točkom $'\mathcal{I}'$ regresijski pravac je prikazan na slici. Ovakav primjer će imati utjecaja na mnoge procese, no metoda najmanjih kvadrata je osobito pogođena. To je zato što razmak između kvadrata jednako udaljenih brojeva raste s porastom brojeva (npr. $10^2 - 9^2 = 19$ dok je $2^2 - 1^1 = 3$). Točke kao što je $'\mathcal{I}'$ nazivamo *outlierima* zato što su daleko od regresijskog pravca i zato što imaju jako velik utjecaj, još ih nazivamo *utjecajne točke*. Takve točke zahtijevaju značajnu pažnju upravo zato što često predstavljaju kršenje (1.12). Kad se to dogodi one ne pripadaju analizi i kao što smo već vidjeli mogu narušiti napravljenu analizu.

Još problematičnija situacija je prikazana na slici 1.2b. Slično kao u prethodnom primjeru točka $'\mathcal{I}'$ može utjecati na cjelokupni smjer pravca i što je još gore $'\mathcal{I}'$ ne bi imala veliki rezidual koji bi privlačio pažnju. Takva točka je utjecajna točka, ali nije outliner.

Zadnju vrstu potencijalnog problema možda je najlakše objasniti na ekstremnom slučaju. Kad bismo imali samo dva opažanja mogli bismo povući ravnu liniju koja bi im savršeno odgovarala, ali normalno mi bismo nerado napravili predviđanja temeljena samo na njima. Pretpostavimo da smo napravili 20 kopija svake točke podataka. Sada imamo 40 opažanja, ali sigurno nemamo koristi od njih. To je zato što su naša opažanja povezana. Zbog toga zahtijevamo da naša opažanja bude nekorelirana:

$$E(\varepsilon_i \varepsilon_j) = 0, \forall i \neq j \tag{1.14}$$

Uvjeti (1.12), (1.13) i (1.14) nazivaju se jednim imenom *Gauss-Markovljevi uvjeti* i primijetimo da oni osiguravaju da odgovarajuća predviđanja dobivena metodom najmanjih kvadrata budu dobra. Točan pojam 'dobroga' ćemo definirati kasnije, gdje će biti prikazan dokaz navedenih tvrdnji. Od sada pa nadalje pojam Gauss-Markovljevih uvjeta će nam značiti stabilnost i dobra predviđanja metode najmanjih kvadrata.

1.3 Koeficijent determinacije jednostavne regresije

Već smo mogli primijetiti kada imamo odgovarajući model da su reziduali mali. Međutim kvaliteta odgovarajućeg modela se može mjeriti preko sume kvadrata reziduala $\sum_{i=1}^{n} \varepsilon_i^2$. Prema tome kad je $\beta_0 \neq 0$ mjera koliko je model dobar iznosi

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{1.15}$$

Mjera R^2 se zove *koeficijent determinacije* i uvijek se nalazi između 0 i 1 , što je bliže 1 to znači da mi imamo bolji model. Više o njoj možemo pronaći u [4].

Kad je $\beta_0 = 0$ mjera iznosi

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (y_i)^2}. \tag{1.16}$$

Pošto je $\sum_{i=1}^{n} y_i^2$ obično mnogo veća od $\sum_{i=1}^{n} (y_i - \bar{y})^2$ ova definicija R^2 se dosta razlikuje od definicije pod (1.15). Stoga modeli s β_0 se ne mogu uspoređivati s modela koji nemaju β_0 na temelju R^2.

1.4 Očekivanje i varijanca $\hat{\beta}_0$ i $\hat{\beta}_1$ pod Gauss-Markovljevim uvjetima

Kako $\hat{\beta}_0$ i $\hat{\beta}_1$ ovise od y_i-jevima koji su slučajne varijable, slijedi da su i $\hat{\beta}_0$ i $\hat{\beta}_1$ su slučajne varijable.

Propozicija 1.4.1. *Očekivanja i varijance slučajnih varijabli $\hat{\beta}_0$ i $\hat{\beta}_1$ dane su sa:*

$$\begin{aligned} E[\hat{\beta}_0] &= \beta_0, \;\; var[\hat{\beta}_0] = \sigma^2 \left[n^{-1} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \\ E[\hat{\beta}_1] &= \beta_1, \;\; var[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned} \tag{1.17}$$

Dokaz. Kako je $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ i

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} y_i (x_i - \bar{x}) - \bar{y} \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

iz (1.11) slijedi

$$\hat{\beta}_1 = \sum_{i=1}^{n} y_i(x_i - \bar{x}) / \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} c_i y_i$$

gdje je $c_i = (x_i - \bar{x}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$. Lako je dokazati da je

$$\sum_{i=1}^{n} c_i = 0,$$

$$\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} c_i x_i - \bar{x} \sum_{i=1}^{n} c_i = \sum_{i=1}^{n} c_i(x_i - \bar{x}) = 1$$

i

$$\sum_{i=1}^{n} c_i^2 = \Big[\sum_{i=1}^{n}(x_i - \bar{x})^2 \Big]^{-1}$$

Dakle, koristeći standardne rezultate očekivanja i varijanci linearnih kombinacija slučajnih varijabli, dobivamo

$$\begin{aligned} E(\hat{\beta}_1) &= \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i \beta_0 + \beta_1 \sum_{i=1}^{n} c_i x_i \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 = \beta_1. \end{aligned}$$

Nadalje, zato što je $\mathrm{Var}(y_i) = \mathrm{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$ imamo

$$\mathrm{Var}(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 var(y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Slično, budući da

$$E(\bar{y}) = n^{-1} \sum_{i=1}^{n} E(y_i) = n^{-1} \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \bar{x}$$

slijedi

$$E(\hat{\beta}_0) = E(\bar{y} - b_1 \bar{x}) = \beta_0 + \beta_1 \bar{x} - \bar{x} E(\hat{\beta}_1) = \beta_0 + \beta_1 \bar{x} - \bar{x} \beta_1 = \beta_0.$$

Sada možemo izraz za $\hat{\beta}_0$ napisati kao

$$\hat{\beta}_0 = n^{-1} \sum_{i=1}^{n} y_i - \bar{x} \sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n}(n^{-1} - \bar{x} c_i) y_i.$$

Odakle slijedi,

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) &= \sum_{i=1}^{n}[n^{-1}-\bar{x}c_i]^2 var(y_i)\\ &= \sigma^2\sum_{i=1}^{n}[n^{-1}-2n^{-1}\bar{x}c_i+\bar{x}^2c_i^2]\\ &= \sigma^2\Big[n^{-1}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\Big].\end{aligned}$$

Što nam upravo dovršava dokaz. □

Lema 1.4.2. *U slučaju kad je $\beta_0 = 0$ izrazi za očekivanje i varijancu su*

$$\begin{aligned}E(\hat{\beta}_1) &= \beta_1\\ \mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^{n}x_i^2}\end{aligned} \tag{1.18}$$

Dokaz. Iz (1.10) za $\beta_0 = 0$ imamo

$$\begin{aligned}\hat{\beta}_1 &= \sum_{i=1}^{n}y_ix_i/\sum_{i=1}^{n}x_i^2 = \beta_1\sum_{i=1}^{n}x_i^2/\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}\varepsilon_ix_i/\sum_{i=1}^{n}x_i^2\\ &= \beta_1+\sum_{i=1}^{n}\varepsilon_ix_i/\sum_{i=1}^{n}x_i^2.\end{aligned}$$

Odakle slijedi

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2\sum_{i=1}^{n}x_i^2/\left(\sum_{i=1}^{n}x_i^2\right)^2 = \sigma^2/\sum_{i=1}^{n}x_i^2$$

i $E(\hat{\beta}_1) = \beta_1$. □

Gore napisane formule možemo smatrati posebnim slučajem formula Teorema 2.4.1 ili se mogu dokazati raspisujući izraz iz (1.17).

Kako je očekivana vrijednost $E(\hat{\beta}_0)$ od predviđenog $\hat{\beta}_0$, $\hat{\beta}_0$ se zove nepristrani procjenitelj od β_0. Slično $\hat{\beta}_1$ se zove nepristrani procjenitelj od β_1.

Da bismo mogli koristiti varijance $\hat{\beta}_0$ i $\hat{\beta}_1$ susrećemo se s malim problemom. One ovise o σ^2, što je nepoznanica. Međutim nepristrani procjenitelj za σ^2 je, što ćemo i dokazati u višestrukoj regresiji,

$$s^2 = (n-1)^{-1}\sum_{i=1}^{n}\varepsilon_i^2 \tag{1.19}$$

i ako zamjenimo σ^2 sa s^2 u (1.17) i (1.18) dobivamo procjenitelje od $\mathrm{Var}(\hat{\beta}_0)$ i $\mathrm{Var}(\hat{\beta}_1)$. Kvadratni korijeni ovih procjenitelja nazivaju se *standardne greške* i označavaju se kao $s.e(\hat{\beta}_0)$ i $s.e(\hat{\beta}_1)$. Prema tome kad je $\beta_0 \neq 0$ imamo

$$s.e(\hat{\beta}_0) = s\big[n^{-1} + \hat{x}^2 / \sum_{i=1}^{n}(x_i - \hat{x})^2\big]^{1/2} \tag{1.20}$$

i

$$s.e(\hat{\beta}_1) = s/\big[\sum_{i=1}^{n}(x_i - \hat{x})^2\big]^{1/2}. \tag{1.21}$$

Dok za slučaj $\beta_0 = 0$ vrijedi

$$s.e(\hat{\beta}_1) = s/\big[\sum_{i=1}^{n} x_i^2\big]^{1/2} \tag{1.22}$$

gdje je $s^2 = (n-1)^{-1}\sum_{i=1}^{n}\varepsilon_i^2$.

1.5 Pouzdani intervali i test

Pretpostavimo da vrijede Gauss-Markovljevi uvjeti (1.12), (1.13) i (1.14) i dodatno pretpostavimo da su ε_i-evi normalno distribuirani. Tada su ε_i-evi nezavisno normalno distribuirani s očekivanjem 0 i varijancom σ^2. To ćemo zapisati kao $\varepsilon_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Iz toga slijedi $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Nadalje β_j-ovi kao linearna kombinacija y_i-eva su također normalno distribuirani s očekivanjem i varijancom izraženima u prethodnom odjeljku. Može se pokazati da je

$$(\hat{\beta}_j - \beta_j)/s.e.(\hat{\beta}_j) \sim t_{n-2} \tag{1.23}$$

za slučaj jednostavne linearne regresije s $\beta_0 \neq 0$ gdje je t_{n-2} Studentova t distribucija s $n-2$ stupnja slobode. Iz (1.18) možemo dobiti $(1-\alpha)\times 100$ posto pouzdani interval za β_j kao

$$\hat{\beta}_j - s.e.(\beta_j)t_{n-2,\alpha/2} < \beta_j < \hat{\beta}_j + s.e.(\beta_j)t_{n-2,\alpha/2} \tag{1.24}$$

gdje je $j = 0$ ili $j = 1$ i $t_{n-2,\alpha/2}$ označava gornju $\alpha/2$ točku t distribucije s $n-2$ stupnja slobode. Više o ovoj temi možemo pronaći na [3].

Poglavlje 2

Višestruka linearna regresija

Model linearne regresije predstavljen u prethodnom poglavlju sada će nam biti temelj na kojem ćemo graditi višestruku linearnu regresiju. Odmaknut ćemo se od jednodimenzionalnog pogleda i uvesti matrice kao temeljne elemente za analizu. Izgled modela će biti:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

gdje su $\mathbf{y}$, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$, X sve redom matrice. Logika iza toga modela će se samo nadopunjavati te će analogno vrijediti sva svojstva koja smo dokazali i u poglavlju prije. Možemo reći da ovim poglavljem objedinjujemo teorijsko znanje o linearnoj regresiji ovog rada te da rezultati ovdje dokazani su univerzalni i primjenjivi i na jednostavnije modele.

2.1 Regresijski model u matričnoj notaciji

Ovaj odjeljak započinjemo matričnim raspisom jednostruke regresije obrađene u Poglavlju 1. Prisjetimo se model je bio oblika:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \varepsilon_1 \\ &\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \varepsilon_1 \end{aligned} \tag{2.1}$$

Ako sada postavimo matrične oblike

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \\ 1 & x_{n1} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \tag{2.2}$$

lako je provjeriti da se (1.1) može zapisati i u obliku

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.3}$$

Promotrimo sada slučaj s više od jedne nezavisne varijable. Pretpostavimo da imamo k nezavisnih varijabli $x_1, x_2, \ldots, x_k$; tada je regresijski model

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{aligned} \tag{2.4}$$

Ako stavimo

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \tag{2.5}$$

model u (2.4), koji nazivamo *višestruki regresijski model*, može isto biti zapisan u formi (2.3).

Matrica X se naziva *matrica dizajna*. Kao i u jednostrukoj linearnoj regresiji β_0 se često naziva konstantni član presjecišta. Primijetimo da prvi stupac matrice X, to jest, stupac jedinica odgovara tom konstantnom članu. Ako u nekom slučaju ne želimo zadržati β_0 u našem modelu jednostavno maknemo stupac jedinca iz X. Kao što smo spomenuli u prethodnom poglavlju, zadnjih k elemenata $i$tog retka od X čine $i$tu točku dizajna modela i opažanje y_i s odgovarajućom točkom dizajna čini $i$ti slučaj ili $i$tu točku podataka.

2.2 Metoda najmanjih kvadrata

Promatramo procjenitelj $\boldsymbol{\beta}$ metodom najmanjih kvadrata u višestrukom regresijskom modelu minimizirajući

$$\begin{aligned} \mathcal{S} &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\ &= (\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'X'\mathbf{y} - \mathbf{y}'X\boldsymbol{\beta} + \boldsymbol{\beta}'X'X\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'(X'\mathbf{y}) + \boldsymbol{\beta}'(X'X)\boldsymbol{\beta} \end{aligned} \tag{2.6}$$

kako je $\mathbf{y}'X\boldsymbol{\beta}$ skalar, to je jednako $\boldsymbol{\beta}'(X'\mathbf{y})$. Kako bismo minimizirali (2.6) možemo derivirati to po svakom β_j i tu derivaciju izjednačiti s nulom. Ekvivalentno, možemo to napraviti na malo kompleksniji način koristeći diferencijaciju matrica:

$$\frac{\partial \mathcal{S}}{\partial \boldsymbol{\beta}} = -2X'\mathbf{y} + 2X'X\boldsymbol{\beta} \tag{2.7}$$

Izjednačavanjem (2.7) s nulom i zamjenom $\boldsymbol{\beta}$ s $\boldsymbol{b}$, vidimo da je procjenitelj metodom najmanjih kvadrata $\boldsymbol{b}$ od $\boldsymbol{\beta}$ dan s

$$(X'X)\boldsymbol{b} = X'\mathbf{y} \tag{2.8}$$

Da nam ovo zaista daje minimum bit će pokazano na kraju ovog odjeljka. Ukoliko $X'X$ nije singularna, (2.8) ima jedinstveno rješenje:

$$\boldsymbol{b} = (X'X)^{-1}X'\mathbf{y} \tag{2.9}$$

Definicija 2.2.1. *Kaže se da je matrica $A \in M_n(F)$ regularna ako postoji matrica $B \in M_n(F)$ takva da vrijedi $AB = BA = I$. U tom slučaju se matrica B zove multiplikativni inverz ili inverzna matrica od A i označava s A^{-1}. Matrica $A \in M_n(F)$ se naziva singularnom matricom ako nema multiplikativni inverz.*

Kada je $(X'X)$ singularna (2.8) se još uvijek može riješiti koristeći generalizirani inverz. Iz korolara u [5] dobivamo:

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} = X^{-1}\boldsymbol{y} \tag{2.10}$$

U ovom slučaj procjenitelj nije jedinstven, no iz gore spomenutog korolara slijedi da je $X(X'X)^{-1}X'$ jedinstveno pa je posljedično onda i $X\boldsymbol{b}$ jedinstven. To je jednostavno za vidjeti u slučaju kad nemamo β_0 i kad je stupac jedinica izbrisan iz X, izrazi (2.8), (2.9) i (2.10) još uvijek vrijede.

Što se tiče slučaja jednostruke linearne regresije, definiramo reziduale e_i kao

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{2.11}$$

gdje je

$$\boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \ \hat{\boldsymbol{y}} = \begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_n} \end{pmatrix} = X\boldsymbol{b} = X(X'X)^{-1}X'\boldsymbol{y} = H\boldsymbol{y}, \tag{2.12}$$

i $H = X(X'X)^{-1}X'$. Uvedimo oznaku $M = I - H$. Tada

$$MH = (I - H)X = X - X(X'X)^{-1}X'X = X - X = 0.$$

Koristeći M i (2.3) možemo $\boldsymbol{e}$ izraziti u terminima $\boldsymbol{y}$ i $\boldsymbol{\varepsilon}$ na sljedeći način:

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{y} - H\boldsymbol{y} = M\boldsymbol{y} \\ &= MX\boldsymbol{\beta} + M\boldsymbol{\varepsilon} = M\boldsymbol{\varepsilon} \end{aligned} \tag{2.13}$$

Teorem 2.2.2. *Reziduali su ortogonalni na predviđene vrijednosti isto kao i matrica dizajna X u modelu $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$*

Dokaz. Kako vrijedi

$$X'\boldsymbol{\varepsilon} = X'M\boldsymbol{\varepsilon} = 0\boldsymbol{\varepsilon} = \boldsymbol{o}, \tag{2.14}$$

što je nul vektor, slijedi da je

$$\hat{\boldsymbol{y}}'\boldsymbol{e} = \boldsymbol{b}'X'\boldsymbol{e} = 0, \tag{2.15}$$

što dokazuje teorem. □

Iz Teorema 2.2.1 slijedi da ako je β_0 u modelu, pa je posljedično prvi stupac matrice X jednak $\mathbf{1} = (1, \ldots, 1)'$, tada vrijedi $\sum_{i=1}^{n} e_i = \mathbf{1}'\boldsymbol{e} = 0$.

Za zaključak ove odjeljka pokazat ćemo da se minimum od $\mathcal{S} = (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}$ zaista postiže u $\boldsymbol{b} = \boldsymbol{\beta}$. Primijetimo da iz Teorema 2.2.1 slijedi

$$(\boldsymbol{b} - \boldsymbol{\beta})'X'(\boldsymbol{y} - X\boldsymbol{b}) = (\boldsymbol{y} - X\boldsymbol{b})'X(\boldsymbol{b} - \boldsymbol{\beta}) = \boldsymbol{e}'X(\boldsymbol{b} - \boldsymbol{\beta}) = 0 \tag{2.16}$$

Otud

$$\begin{aligned} \mathcal{S} &= (\boldsymbol{y} - X\boldsymbol{b} + X\boldsymbol{b} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{b} + X\boldsymbol{b} - X\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - X\boldsymbol{b})'(\boldsymbol{y} - X\boldsymbol{b}) + (\boldsymbol{b} - \boldsymbol{\beta})'(X'X)(\boldsymbol{b} - \boldsymbol{\beta}). \end{aligned}$$

Oba izraza u zadnjoj liniji su kvadratne forme te su dakle pozitivne i prvi izraz ne ovisi o $\boldsymbol{\beta}$. Stoga, $\mathcal{S}$ svoj minimum postiže u $\boldsymbol{b} = \boldsymbol{\beta}$.

2.3 Gauss-Markovljevi uvjeti

Da bi procjene od $\boldsymbol{\beta}$ imale neka poželjna statistička obilježja, trebamo sljedeće pretpostavke znane pod imenom Gauss-Markovljevi (G-M) uvjeti, s kojima smo se upoznali još u Poglavlju 1:

$$E(\varepsilon_i) = 0 \tag{2.17}$$

$$E(\varepsilon_i^2) = \sigma^2 \tag{2.18}$$

$$E(\varepsilon_i \varepsilon_j) = 0, \ \forall i \neq j \tag{2.19}$$

za sve $i, j = 1, \ldots, n$. Možemo ove uvjete zapisati i u matričnoj formi kao

$$\begin{aligned} E(\boldsymbol{\varepsilon}) &= \boldsymbol{o} \\ E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') &= \sigma^2 I \end{aligned} \tag{2.20}$$

Primijetimo da je $\boldsymbol{o}$ vektor nula, odnosno nul vektor. Koristit ćemo ove uvjete u više navrata u nastavku.

Vidimo da G-M uvjeti impliciraju

$$E(\boldsymbol{y}) = X\boldsymbol{\beta} \tag{2.21}$$

kao i

$$\mathrm{Cov}(\boldsymbol{y}) = E[(\boldsymbol{y} - X\boldsymbol{\beta})(\boldsymbol{y} - X\boldsymbol{\beta})'] = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 I. \tag{2.22}$$

Također slijedi da je (pogledati (2.13))

$$E[\boldsymbol{e}\boldsymbol{e}'] = ME[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']M = \sigma^2 M \tag{2.23}$$

pošto je M idempotentna.

Definicija 2.3.1. *Matrica $A \in M_{mn}$ je idempotentna ako vrijedi $A^2 = A$.*

Upravo zato vrijedi

$$\mathrm{Var}(e_i) = \sigma^2 m_{ii} = \sigma^2[1 - h_{ii}] \tag{2.24}$$

gdje su m_{ii} i h_{ii} ijti elementi matrica M i H. Kako je varijanca nenegativna i kovarijacijska matrica barem pozitivno semidefinitna, slijedi da je $h_{ii} \leqslant 1$ i matrica M barem pozitivno semidefinitna.

Definicija 2.3.2. *Kažemo da je matrica $A \in M_{nn}$ je pozitivno semidefinitna ako je $x * Ax \geqslant 0$, za svaki $x \in \mathbb{R}^n$.*

Promotrimo sada očekivanje i varijancu procjenitelja pod G-M uvjetima. Zbog (2.20)

$$E(\boldsymbol{b}) = E[(X'X)^{-1}X'\boldsymbol{y}] = (X'X)^{-1}X'X\boldsymbol{\beta} = \boldsymbol{\beta} \tag{2.25}$$

Ako za neki parametar θ, njegov procjenitelj t ima svojstvo da je $E(t) = \theta$, tada je t *nepristrani* procjenitelj od θ. Prema tome pod G-M uvjetima, $\boldsymbol{b}$ je nepristrani procjenitelj od $\boldsymbol{\beta}$. Primijetimo da smo iskoristili samo prvi od G-M uvjeta da bismo dokazali ovo. Stoga kršenje uvjeta (2.17) i (2.18) neće voditi do pristranosti.

Teorem 2.3.3. *Pod prvim G-M uvjetom* (2.17)*, procjenitelj* $\boldsymbol{b}$ *dobiven metodom najmanjih kvadrata je nepristrani procjenitelj od* $\boldsymbol{\beta}$*. Nadalje, pod G-M uvjetima* (2.17) - (2.19) *vrijedi,*

$$\mathrm{Cov}(\boldsymbol{b}) = \sigma^2(X'X)^{-1}$$

Dokaz. Već smo utvrdili nepristranost. Označimo sada $A = (X'X)^{-1}X'$ pa je $\boldsymbol{b} = A\boldsymbol{y}$ te iz (2.22) dobivamo

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{b}) &= A\,\mathrm{Cov}(\boldsymbol{y})A' = \sigma^2 AIA' = \sigma^2 AA' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}, \end{aligned} \tag{2.26}$$

što dovršava dokaz. □

Korolar 2.3.4. *Ako je* $\mathrm{tr}[(X'X)^{-1}] \to 0$ *kad* $n \to \infty$*, tada je procjenitelj* $\boldsymbol{b}$ *konzistentan procjenitelj od* $\boldsymbol{\beta}$*.*

Dokaz. Dokaz slijedi iz činjenice kad $(X'X)^{-1} \to 0$ tada $\mathrm{Cov}(\boldsymbol{b}) \to 0$ za $n \to \infty$. □

Iz (2.13) imamo $E(\boldsymbol{e}) = E(M\boldsymbol{e}) = 0$. Stoga iz (2.24) slijedi

$$\mathrm{Cov}(\boldsymbol{e}) = E[\boldsymbol{e}\boldsymbol{e}'] = \sigma^2 MM' = \sigma^2 M. \tag{2.27}$$

Napišimo $i$tu jednadžbu u (2.4) kao

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

gdje je $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ik})$, procijenjena vrijednost od y_i može se definirati kao

$$\hat{y}_i = \boldsymbol{x}_i'\boldsymbol{b}.$$

Na isti način zaključujemo da je procijenjena vrijednost od $\boldsymbol{y}$

$$\hat{\boldsymbol{y}} = X\boldsymbol{b}.$$

Iz Teorema 2.4.1 dobivamo, koristeći rezultate kovarijanci slučajnih vektora,

$$\mathrm{Var}(\hat{y}_i) = \boldsymbol{x}_i' \,\mathrm{Cov}(\boldsymbol{b})\boldsymbol{x}_i = \sigma^2 \boldsymbol{x}_i'(X'X)^{-1}\boldsymbol{x}_i = \sigma^2 h_{ii}$$

i

$$\mathrm{Cov}(\hat{\boldsymbol{y}}) = X\,\mathrm{Cov}(\boldsymbol{b})X' = \sigma^2 X(X'X)^{-1}X' = \sigma^2 H.$$

Očito prvi od ova dva rezultat slijedi iz drugog. Isto tako slijedi i $h_{ii} \geqslant 0$ i da je matrica H barem pozitivno semi-definitna.

2.4 Procjena σ^2

Da bismo mogli upotrijebiti većinu formula iz prethodnog poglavlja trebamo σ^2, što je uglavnom nepoznanica i trebamo je procijeniti. Možemo to napraviti preko reziduala e_i. Budući da je $M = (m_{ij})$ simetrična idempotentna matrica

$$\sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} = \boldsymbol{\varepsilon}' M' M \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}' M \boldsymbol{\varepsilon} = \sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} m_{ij}\varepsilon_i\varepsilon_j. \tag{2.28}$$

Član $\sum_{i=1}^{n} e_i^2$ se često naziva *suma kvadrata reziduala* i označava se kao RSS. Iz (2.28) slijedi

$$\begin{aligned} E\left(\sum_{i=1}^{n} e_i^2\right) &= \sum_{i=1}^{n} m_{ii}E(\varepsilon_i^2) + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} m_{ij}E(\varepsilon_i\varepsilon_j) \\ &= \sigma^2 \sum_{i=1}^{n} m_{ii} = \sigma^2 trM = (n-k-1)\sigma^2 \end{aligned}$$

kad je konstantan član u regresijskoj jednadžbi prisutan i imamo k nezavisnih varijabli, pošto tada vrijedi $\mathrm{tr}M = \mathrm{tr}I_n$ - $\mathrm{tr}H = n - k - 1$. Prema tome, ako stavimo

$$s^2 = \sum_{i=1}^{n} e_i^2/(n-k-1) \tag{2.29}$$

vidimo da je s^2 nepristrani procjenitelj za σ^2. Kada je konstantni član regresije odsutan i imamo k nezavisnih varijabli

$$s^2 = \sum_{i=1}^{n} e_i^2/(n-k) \tag{2.30}$$

je nepristrani procjenitelj za σ^2. Djelitelj $n-k$ u zadnjoj formuli i $n-k-1$ u (2.29) su stupnjevi slobode.

Također $s^2 \to \sigma^2$ po vjerojatnosti kada $n \to \infty$, tj. s^2 je konzistentan procjenitelj od σ^2. Stoga, kada je σ^2 nepoznat, konzistentan i nepristran procjenitelj za $\mathrm{Cov}(\boldsymbol{b})$ je dan s

$$\widehat{\mathrm{Cov}(\boldsymbol{b})} = s^2(X'X)^{-1} = G = (g_{ij}). \tag{2.31}$$

Matrica

$$(g_{ij}/g_{ii}^{1/2}g_{jj}^{1/2}) \tag{2.32}$$

je procjena korelacijske matrice od $\boldsymbol{b}$. Gore navedene rezultate možemo sumirati u sljedećem teoremu:

Teorem 2.4.1. *Na temelju G-M uvjeta, s^2 je nepristran i konzistentan procjenitelj od σ^2 i $s^2(X'X)^{-1}$ je nepristran i konzistentan procjenitelj od* $\mathrm{Cov}(\boldsymbol{b})$.

Reziduali i suma kvadrata reziduala imaju vrlo važnu ulogu u regresijskoj analizi. Kao što smo upravo vidjeli, suma kvadrata reziduala kad je podijeljena s $n-k-1$ daje nepristrani procjenitelj od σ^2. Zapravo, pod pretpostavkom normalnosti opažanja ovo je najbolji nepristrani procjenitelj u smislu da ima uniformno najmanju varijancu među svim nepristranim procjeniteljima koji su kvadratne funkcije $y_i$eva (tj. procjenitelji forme $\boldsymbol{y}'A\boldsymbol{y}$, gdje je A simetrična matrica; primijetimo da je $s^2 = \boldsymbol{y}'M\boldsymbol{y}/[n-k-1]$).

Reziduali su inače korišteni i za prepoznavanje prisutnosti outliera i utjecajnih točaka, provjeru normalnosti podataka, prepoznavanje adekvatnosti modela, itd. Ukratko, koristimo ih da bismo utvrdili je li opravdano pretpostaviti da su zadovoljeni Gauss-Markovljevi uvjeti. Očito ih možemo koristiti i za određivanje kvalitete regresijskog modela. U sljedećem odjeljku definirat ćemo mjeru koja određuje kvalitetu nekog modela. No, prvo ćemo iskazati sljedeći teorem koji nam pokazuje vezu između sume kvadrata reziduala, ukupne sume kvadrata $\sum_{i=1}^n y_i^2$ i predviđene sume kvadrata $\sum_{i=1}^n \hat{y}_i{}^2$.

Teorem 2.4.2. *Neka je $\bar{y} = n^{-1}\sum_{i=1}^n y_i$. Tada*

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i{}^2 = \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right) - \left(\sum_{i=1}^n \hat{y}_i{}^2 - n\bar{y}^2\right).$$

Dokaz. Budući da, iz Teorema 2.2.2, $\hat{\boldsymbol{y}}'\boldsymbol{e} = \sum_{i=1}^{n} e_i\hat{y}_i = 0$,

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}\hat{y}_i{}^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i \\ &= \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}\hat{y}_i{}^2 + 2\sum_{i=1}^{n} e_i\hat{y}_i \\ &= \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}\hat{y}_i{}^2\end{aligned}$$

Drugi dio teorema je očigledan. □

Korolar 2.4.3. *Ako u modelu postoji konstantan član β_0, tada vrijedi*

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Dokaz. Budući da u modelu imamo konstantan član, po Teoremu 2.2.1 vrijedi $\mathbf{1}'\boldsymbol{e} = \boldsymbol{e}'\mathbf{1} = \sum_{i=1}^{n} e_i = 0$. Odakle slijedi $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{y}_i$. Stoga u ovom slučaju je očekivanje opažanja jednako očekivanju predviđenih vrijednosti. Pa po Teoremu 2.5.2 imamo

$$\sum_{i=1}^{n} e_i^2 = \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) - \left(\sum_{i=1}^{n}\hat{y}_i{}^2 - n\bar{y}^2\right) = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

što dokazuje tvrdnju korolara. □

2.5 Koeficijent determinacije

Koeficijent determinacije R^2, koji smo predstavili u Odjeljku 1.3 i zadali ga s (1.15) u slučaju postojanja konstantnog člana te s (1.16) u slučaju nepostojanja istoga, je primjenjiv i na višestruku regresiju. Drugi korijen od R^2 je uzajamna ovisnost između y_i-eva i $\hat{y}_i$-eva, to jest u slučaju regresijskog modela s konstantnim članom,

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2(\hat{y}_i - \bar{y})^2\right]^{1/2}} \tag{2.33}$$

pošto, kao što smo vidjeli u Korolaru 2.5.3, vrijedi $n^{-1}\sum_{i=1}^{n} y_i = n^{-1}\sum_{i=1}^{n} \hat{y}_i$.

Da bismo vidjeli da je kvadrat (2.33) zaista (1.15), primijetimo da zbog (2.15) , $\sum_{i=1}^{n} e_i\hat{y}_i = 0$,

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y}) &= \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})(\hat{y}_i-\bar{y})\\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\\ &= \sum_{i=1}^{n} e_i\hat{y}_i-\bar{y}\sum_{i=1}^{n} e_i+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\\ &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.\end{aligned}$$

Dakle,

$$R=\Big[\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\Big]^{1/2}/\Big[\sum_{i=1}^{n}(y_i-\bar{y})^2\Big]^{1/2}$$

te zbog Korolara 2.5.3 dobivamo

$$R^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \tag{2.34}$$

Iz (2.34), koristeći ponovno Korolar 2.5.3, slijedi da R^2 poprima vrijednosti između 0 i 1 te jer je R nenegativan vrijedi $0 \leqslant R \leqslant 1$. Dokazano je da je R^2 uzajamna višestruka ovisnost između zavisne varijable y i nezavisnih varijabli $x_1, \ldots, x_k$.

U analizama se još koristi i *prilagođeni* R^2 u oznaci R_a^2 koji je dan s

$$R_a^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-k-1)}{\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)} = 1-s^2/\Big[\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)\Big]$$

R_a^2 se prilagođava veličini uzorka, budući da se često mala veličina uzorka povezuje sa sklonošću da povećava R^2. No, R_a^2 može poprimiti negativne vrijednosti.

Alternativno, s^2 se isto može koristiti kao mjera kvalitete modela, manje vrijednosti s^2 znače bolje odgovaranje modela podacima. "Gruba" praktična primjena s^2 proizlazi iz činjenice da kada je broj opažanja n velik, $4s$ je aproksimativna širina 95 postotnog pouzdanog intervala za nadolazeća opažanja. Ako smo primarno zainteresirani za predviđanja, ovo nam pruža odličnu oznaku kvalitete modela.

U slučaju kad regresijska jednadžba nema konstantni član, točnije kad je $\beta_0 = 0$, tada definiramo kvadrat uzajamne ovisnosti između y_i-eva i $\hat{y}_i$-eva kao

$$R^2 = \Big[\sum_{i=1}^{n} y_i \hat{y}_i\Big]^2 / \Big[(\sum_{i=1}^{n} y_i^2)(\sum_{i=1}^{n} \hat{y}_i{}^2)\Big]. \tag{2.35}$$

Budući da je

$$\begin{aligned}
\sum_{i=1}^{n} y_i \hat{y}_i &= \boldsymbol{y}'\hat{\boldsymbol{y}} = \boldsymbol{y}' X\boldsymbol{b} = \boldsymbol{y}' X(X'X)^{-1}X'\boldsymbol{y} \\
&= \boldsymbol{y}' X(X'X)^{-1}(X'X)(X'X)^{-1}X'\boldsymbol{y} = \boldsymbol{b}'X'X\boldsymbol{b} \\
&= \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} \\
&= = \sum_{i=1}^{n} \hat{y}_i{}^2
\end{aligned}$$

i iz Teorema 2.5.2 imamo

$$\begin{aligned}
\sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n} \hat{y}_i{}^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} \hat{y}_i{}^2
\end{aligned}$$

pa (2.35) postaje

$$\left(\sum_{i=1}^{n} \hat{y}_i{}^2\right) / \left(\sum_{i=1}^{n} y_i^2\right) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} y_i^2}.$$

Primijetimo da ovdje također vrijedi $0 \leqslant R \leqslant 1$ i $0 \leqslant R^2 \leqslant 1$.

2.6 Gauss-Markovljev teorem

U većini slučajeva prilikom korištenja regresije zanimaju nas procjene nekih linearnih funkcija $\boldsymbol{L\beta}$ ili $\boldsymbol{l}'\boldsymbol{\beta}$ od $\boldsymbol{\beta}$, gdje je $\boldsymbol{l}$ vektor i $\boldsymbol{L}$ matrica. Procjene ove vrste uključuju procijenjene vrijednosti $\hat{y}_i$, procjenu $\hat{y}_0$ budućeg opažanja, $\hat{\boldsymbol{y}}$ pa čak i sam $\boldsymbol{b}$. Prvo ćemo promotriti $\boldsymbol{l}'\boldsymbol{\beta}$, općenitije funkcije vektora ćemo naknadno razmatrati.

Iako može biti nekoliko mogućih procjenitelja, mi ćemo se ograničiti na linearne procjenitelje, tj. na procjenitelje koji su linearne funkcije $y_1, \ldots, y_n$, recimo $\boldsymbol{c}'\boldsymbol{y}$. Uz to ćemo zahtijevati da te linearne funkcije budu nepristrani procjenitelji od $\boldsymbol{l}'\boldsymbol{\beta}$ i

pretpostavit ćemo da takav nepristran linearni procjenitelj za $\boldsymbol{l}'\boldsymbol{\beta}$ postoji. U tom slučaju $\boldsymbol{l}'\boldsymbol{\beta}$ se naziva procjenljivim.

U teoremu što slijedi pokazat ćemo da među svim linearnim nepristranim procjeniteljima, procjenitelj dobiven metodom najmanjih kvadrata $\boldsymbol{l}'\boldsymbol{\beta} = \boldsymbol{l}(X'X)^{-1}X'\boldsymbol{y}$, koji je također linearna funkcija od $y_1, \ldots, y_n$ i za koji je već pokazano u (2.25) da je nepristran, ima najmanju varijancu. To jest, $\mathrm{Var}(\boldsymbol{l}'\boldsymbol{\beta}) \leqslant \mathrm{Var}(\boldsymbol{c}'\boldsymbol{y})$ za sve $\boldsymbol{c}$ takve da je $E(\boldsymbol{c}'\boldsymbol{y}) = \boldsymbol{l}'\boldsymbol{\beta}$. Takav procjenitelj se naziva najbolji nepristran linearni procjenitelj (BLUE, tj. *best linear unbiased estimator*).

Teorem 2.6.1. ***(Gauss-Markovljev)*** *Neka je $\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y}$ te $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Tada pod G-M uvjetima, procjenitelj $\boldsymbol{l}'\boldsymbol{b}$ procjenljive funkcije $\boldsymbol{l}'\boldsymbol{\beta}$ je BLUE.*

Dokaz. Neka je $\boldsymbol{c}'\boldsymbol{y}$ proizvoljan nepristran linearan procjenitelj od $\boldsymbol{l}'\boldsymbol{\beta}$. Kako je $\boldsymbol{c}'\boldsymbol{y}$ nepristran procjenitelj od $\boldsymbol{l}'\boldsymbol{\beta}$ vrijedi $\boldsymbol{l}'\boldsymbol{\beta} = E(\boldsymbol{c}'\boldsymbol{y}) = \boldsymbol{c}'X\boldsymbol{\beta}$ za sve $\boldsymbol{\beta}$ i zato je

$$\boldsymbol{c}'X = \boldsymbol{l}'. \tag{2.36}$$

Sada imamo,

$$\mathrm{Var}(\boldsymbol{c}'\boldsymbol{y}) = \boldsymbol{c}'\,\mathrm{Cov}(\boldsymbol{y})\boldsymbol{c} = \boldsymbol{c}'(\sigma^2 I)\boldsymbol{c} = \sigma^2\boldsymbol{c}'\boldsymbol{c},$$

i

$$\mathrm{Var}(\boldsymbol{l}'\boldsymbol{b}) = \boldsymbol{l}'\,\mathrm{Cov}(\boldsymbol{b})\boldsymbol{l} = \sigma^2\boldsymbol{l}'(X'X)^{-1}\boldsymbol{l} = \sigma^2\boldsymbol{c}'X(X'X)^{-1}X'\boldsymbol{c},$$

iz tvrdnji (2.26) i (2.36). Stoga

$$\begin{aligned}\mathrm{Var}(\boldsymbol{c}'\boldsymbol{y}) - \mathrm{Var}(\boldsymbol{l}'\boldsymbol{b}) &= \sigma^2[\boldsymbol{c}'\boldsymbol{c} - \boldsymbol{c}'X(X'X)^{-1}X'\boldsymbol{c}] \\ &= \sigma^2\boldsymbol{c}'[I - X(X'X)^{-1}X']\boldsymbol{c} \geqslant 0,\end{aligned}$$

jer je $I - X(X'X)^{-1}X' = M$ pozitivna semi-definitna matrica, što dokazuje tvrdnju teorema. □

Teorem 2.6.2. *Pod G-M uvjetima, procjenitelj $L\boldsymbol{b}$ procjenljive funkcije $\boldsymbol{l}'\boldsymbol{\beta}$ je BLUE u smislu da je*

$$\mathrm{Cov}(C\boldsymbol{y}) - \mathrm{Cov}(L\boldsymbol{b})$$

pozitivno semi-definitno, gdje je L proizvoljna matrica i $C\boldsymbol{y}$ jedan od nepristranih linearnih procjenitelja od $L\boldsymbol{\beta}$

Detaljnije o Gauss-Markovljevo teormu možemo naći na [1]. Nadalje, ovaj teorem implicira da ako želimo procijeniti nekoliko povezanih linearnih funkcija od β_j, ne možemo to napraviti bolje nego metodom najmanjih kvadrata.

Dokaz. Kao u dokazu prethodnog teorema, nepristranost $C\boldsymbol{y}$ povlači $L\boldsymbol{\beta} = CE(\boldsymbol{y}) = CX\boldsymbol{\beta}$ za sve $\boldsymbol{\beta}$, međutim $L = CX$ i pošto je $\mathrm{Cov}(C\boldsymbol{y}) = \sigma^2 CC'$ i

$$\mathrm{Cov}(L\boldsymbol{b}) = \sigma^2 L(X'X)^{-1}L' = \sigma^2 CX(X'X)^{-1}X'C',$$

slijedi

$$\mathrm{Cov}(C\boldsymbol{y}) - \mathrm{Cov}(L\boldsymbol{b}) = \sigma^2 C[I - X(X'X)^{-1}X']C',$$

što je pozitivno semi-definitno, jer je matrica $[I - X(X'X)^{-1}X'] = M$ najmanje pozitivno semi-definitna. □

Ako Gauss-Markovljevim uvjetima pridodamo još i pretpostavku da su y_i-evi normalno distribuirani, tada se može pokazati da $\boldsymbol{l'b}$ ima najmanju varijancu unutar cijele klase nepristranih procjenitelja i da s^2 ima najmanju varijancu među svim nepristranim procjeniteljima od σ^2.

Poglavlje 3

Elastičnost cijena

3.1 Uvod

Kao mali uvod u ovo poglavlje predstavit ćemo pojam elastičnosti cijena, način na koji se ona primjenjuje isto kao i njen značaj. Za početak uvodimo kraticu PED (eng. Price Elasticity Demand) koja predstavlja cjenovnu elastičnost potražnje i definira se kao promjena u potražnji za zadanu promjenu cijene. Njome mjerimo potrošačku osjetljivost na promjene u cijeni za dobivenu uslugu ili proizvod i kreće se od visoke osjetljivosti što onda predstavlja elastičnost do niske osjetljivosti što je neelastičnost. Zapisano u obliku forme imamo,

$$E = -\frac{(\mu_{P1} - \mu_{P0})/\mu_{P0}}{(P1 - P0)P0}, \tag{3.1}$$

gdje je

E = PED

$P0$ = početna cijena

$P1$ = nova cijena

μ_{P0} = potražnja pri početnoj cijeni

μ_{P1} = potražnja pri novoj cijeni.

Jedan od benefita PED-a je da omogući tvrtkama poboljšanje strategija za određivanje cijena kroz bolje razumijevanje elastičnosti cijena ciljanog im tržišta. Na primjeru iz članka [2] možemo vidjeti jednostavan primjenu elastičnosti cijena. Neki računovođa je tijekom godina stekao vjerne korisnike, no zbog boljeg budućeg poslovanja odluči

se na privlačenje novih korisnika i to tako da im da popust na svoje usluge. Zna da tim potezom može završiti u negativnom profitu i zato se odluči na povećanje cijena lojalnim korisnicima, koje bi bilo minimalno i zbog kojeg većina korisnika ne bi promijenila njega kao računovođu. Drugim riječima, on zapravo cjenovnu strategiju bazira na elastičnosti cijene njegovog tržišta. Za nove korisnike koji su više cjenovno osjetljivi on snižava cijenu, dok postojećim korisnicama koji su zapravo manje cjenovno osjetljivi, povećava dosadašnju cijenu. Ovakvom strategijom cijena on planira kroz duži period povećati broj korisnika isto kao i iznos profita. Ovo je dobar primjer da vidimo na koliko se faktora treba obratiti pažnju prilikom osmišljavanja cijena i sveukupne poslovne strategije, i poslužio nam je kao uvid za korištenje PED-a.

3.2 Model potražnje

U statističkim modelima, korištenim za ovakve svrhe, imamo dva osnovna tipa podatkovnih varijabli. Varijabla odziva se dobiva kao rezultat modela, dok je prediktorska varijabla ulazni parametar za model. Za dobivenih set prediktorskih varijabli, model nam daje varijable odziva. Stvaran odziv u podacima ima dva dijela. Signal predstavlja predvidljivo ponašanje dok je šum slučajno ponašanje. Cilj modela je odvojiti signal od šuma. Također, pokušava se dobiti funkcija koja bi primijenjena na prediktorske varijable dala varijablu odziva koja bi predstavljala signal u stvarnom odzivu.

Kod modeliranja cjenovne elastičnosti potražnje prvo trebamo jasno definirati stvarni odziv. Povrh toga, korisno je klasificirati različite tipove prediktora koja bi se mogli pojaviti u podacima. Klasificiranje podataka za model elastičnosti poprilično je zahtjevno.

Modeliranje elastičnosti cijena formira model u kojem je stvarni odziv individualno korisničko prihvaćanje ili odbijanje navedene ponude. Prediktori u ovakvim modelima mogu biti klasificirani na osnovi cijene, odnosno oni koji jesu ili nisu povezani s cijenom. Postavljena elastičnost je izvedena iz koeficijenata koji odgovaraju prediktorskim varijablama iz modela. Primijetimo da bi bilo korisno dodatno klasificirati cjenovne faktore kao vanjske (utjecaj cijena na konkurente) i unutarnje (utjecaj cijena na aktivnosti tvrtke).

U sljedećem odjeljku bavit ćemo se strukturom modela, gdje ćemo sa zadanim stvarnim odzivom i prediktorskim varijablama moći izračunati varijablu odziva i pripadajuću elastičnost. Bit ćemo fokusirani na distribuciju pogrešaka stvarnog odziva, funkcionalni oblik koji povezuje stvari odziv i prediktore i strukturu prediktora. Osim toga, vidjet ćemo i načine provjere kvalitete modela.

Distribucija pogrešaka

Za razliku od mnogo kompliciranijih modela, odabir distribucije pogreški za model elastičnosti cijena je vrlo jednostavan. Budući da stvaran odziv može poprimiti dvije vrijednosti *da/ne*, binomna distribucija se čini kao najprikladniji odabir.

Korisno bi bilo prije svega shvatiti što to znači imati binomnu distribuciju. Distribucija pogrešaka je definirana u empirijskim uvjetima isto kao i funkcionalni oblik. Empirijski, stvarni odziv može biti samo 0 ili 1, što znači da je odabir distribucije očito dobar. Funkcionalni oblik distribucije pogrešaka povezuje varijancu stvarnog odziva s predviđenim odzivom. Veza varijance i očekivanja u binomnoj distribuciji daje nam sljedeće:

$$\text{varijanca} = \text{očekivanje } (1 - \text{očekivanje}).$$

Varijanca rezultata predstavlja nam stupanj neodređenosti povezan s predviđanjima rezultata. U troškovnom modelu je uobičajeno pretpostaviti ako imamo veće predviđene vrijednosti da ćemo imati veću neodređenost, odnosno varijabilnost procjenitelja. U modelu elastičnosti imamo drukčiju situaciju, naime i visoke i niske predviđene vrijednosti imaju manju varijabilnost dok prosječne predviđene vrijednosti imaju veću varijabilnost. To je svojstvo naslijeđeno od binomnih modela. Ako sva zapažanja uvijek odbiju ponudu (stvarni odaziv je nizak, tj. nula je) tada modelari mogu sa sigurnošću tvrditi da će buduće ponude također biti odbijene. Analogno, ako sva zapažanja prihvate ponudu (stvarni odaziv je visok, tj. jedan je) onda s istom sigurnošću mogu tvrditi da će sljedeća ponuda biti prihvaćena. No, ako imamo srednji odaziv, oko 50% tada ne možemo sa sigurnošću tvrditi vezu s procijenjenim vrijednostima. Sad kad imamo odabranu i definiranu distribuciju, sljedeći korak u strukturi je link funkcija.

Link funkcija

Link funkcija je funkcionalni oblik koji povezuje varijablu odziva s prediktorskim varijablama. Dva su glavna zahtjeva koja trebamo poštivati prilikom odabira link funkcije za model elastičnosti. Prvi je taj da nam predviđeni odaziv vjerojatnosti prihvaćanja ili odbijanja cjenovne ponude uvijek bude između 0 i 1. Zadaća link funkcije je da transformira koeficijente dobivene iz strukturnog dizajna u rezultate koji su u skladu sa zakonom vjerojatnosti. Kombinacija koeficijenata iz strukturnog dizajna modela često se naziva linearni procjenitelj. Drugi zahtjev je da se predviđeni odaziv približi prihvaćanju ili odbijanju, ali da ih zapravo nikad ne dohvati. Kad bismo dopustili predviđenim vrijednostima da u potpunosti dođu do čisto prihvaćanja ili odbijanja to bi dovelo do previše sigurnosti u predviđanjima.

Za razliku od distribucije pogrešaka, u slučaju link funkcija imamo više kandidata koji zadovoljavaju gore navedene uvjete. Dvije najčešće korištene u praksi su:

1. Logistička funkcija
$$\mu = f(x) = \frac{1}{1 + 1/exp(x)} \tag{3.2}$$

2. Normalna funkcija
$$\mu = f(x) = \Phi(x) \tag{3.3}$$

gdje je x linearni procjenitelj i Φ označava normalnu distribuciju.

Prilikom odabira između različitih link funkcija, bilo bi dobro i korisno provesti test valjanosti koristeći koncept elastičnosti. No ne možemo elastičnost individualno promatrati , nego je trebamo izvesti iz modela za predviđeni odaziv. Postoje mnogi načini definiranja elastičnosti, ali mi ćemo se fokusirati na sljedeća dva:

1. Klasična elastičnost, je postotak promjene u potražnji u odnosu na postotak promjene u cijeni. Kao što je definirano u uvodu, za zadanog korisnika:
$$E = -\frac{(\mu_{P1} - \mu_{P0})/\mu_{P0}}{(P1 - P0)P0}. \tag{3.4}$$
Pretpostavimo da je $P1 > P0$, tada je početna očekivanja potražnja fiksirana i elastičnost u tom slučaju ima linearan odnos s potražnjom. Primijetimo da je μ_{Pi} očekivana vjerojatnost uspjeha povezana s cijenom P_i.

2. Elastičnost linearnog procjenitelja, možemo definirati kao promjenu u linearnom procjenitelju uslijed postotka promjene u cijeni. Specijalno, za zadanog korisnika:
$$E = -\frac{\left(\beta_0 + \alpha \times \frac{P1}{P0}\right) - \left(\beta_0 + \alpha \times \frac{P0}{P0}\right)}{(P1 - P0)/P0}. \tag{3.5}$$
Pretpostavljajući da je linearan procjenitelj koji nema nelinearne komponente i jedan cjenovni faktor:
$$E = -\alpha \tag{3.6}$$
gdje je α koeficijent povezan s cijenom.
Primijetimo da u ovom jednostavnom slučaju, E ne mijenja s potražnjom. Prema tome, ako potražnja poraste mi očekujemo da će E ostati konstantan. Ova veza se održava bez obzira na broj parametara koji nisu cjenovni u strukturnom modelu.

Sad kad imamo odabranu distribuciju i link funkciju, sljedeći korak je određivanje strukture dizajna prediktora.

Struktura modela

Gradnja strukture modela je ravnoteža između *overfiitng*-a i *underfitting*-a. Kada je struktura modela prekompleksna, tada je veća vjerojatnost over-fittinga podataka. To općenito dođe do izražaja u gubitku snage predviđanja. Međutim, kad je struktura prejednostavna, velika je vjerojatnost under-fitting podataka, što je isto izraženo prilikom gubitka snage predviđanja. Strukturni dizajn modela se često naziva linearni procjenitelj. On odražava kombinaciju koeficijenata koji su primijenjeni na određen opažanje. Primijetimo da modelari zapravo zadaju strukturni dizajn, link funkciju i distribuciju pogrešaka. Sam model nakon toga računa skup parametara koji su temeljeni na tim pretpostavkama.

Linearni procjenitelj sadrži niz različitih objekata koji predstavljaju različite prediktore. Kao dodatak kategorizaciji prediktora na one faktore koji su cjenovni i one koji nisu, oni mogu biti klasificirani kao kategorički ili kao kontinuirani elementi. U slučaju da je prediktor kategorički, parametri se računaju za svaki nivo u prediktoru. Kad imamo kontinuirani prediktor, tada se parametri baziraju na obliku funkcije koja je predstavljena.

Prisjetimo da je glavni cilj kreiranja modela potražnje projiciranje očekivanje cjenovne osjetljivosti na aktivne korisnike, to jest elastičnost. Iz definicija iskazanih u prvom odjeljku vidimo da je elastičnost zapravo funkcija nagiba povezana s cjenovnim faktorima iz modela. Kod kreiranja modela često tražimo interakciju između cjenovnih faktora i onih koji to nisu. U okvirima linearnog procjenitelja, analitičari bi mogli napraviti strukturni model u kojem bi svi cjenovni faktori bili u interakciji s necjenovnim faktorima. Takav model ima jako velike šanse za over-fitting podataka i postoji vjerojatnost da bi elastičnost izvedena iz tog modela mogla biti negativna.

No proizvoljno dodavanje interakcije nije dobra ideja. To bi moglo voditi do over-fittinga koji bi oslabio prediktivnu snagu modela. Generalno bilo kakvo dodavanje kompleksnost u model treba provjeriti i potvrditi nekim od testova, bilo to statistički ili preko testa konzistentnosti. Ako nam struktura ne daje usklađene rezultate, tada ona ne bi trebala biti uključena u model. Tome se može stati na kraj raznim tehničkim rješenjima preko kojih je moguće graditi kompleksnije modele s uključeno interakcijom bez negativnog odraza na elastičnost. Nakon što ustvrdimo da imamo dobar strukturni model prelazimo na zadnji korak, a to je provjera valjanosti.

Validacija

Modele predstavljene u prethodnim odjeljcima sada trebamo validirati koristeći bezvremenske podatke za provjeru njihovih performansi. Dosada su u poglavljima te provjere bile povezane s link funkcijama, dok ćemo se sada bazirati na dvije dodatne vrste provjera povezane sa skupom valjanosti podataka. Taj skup podataka može biti

ili slučajan uzorak uzet iz originalnih podataka ili bezvremenski podaci. Vrijednosti bezvremenskog skupa podataka je ta da oni odražavaju ispravnu mjeru prediktivne točnosti. Njihova mana je mogućnost značajne promjene u poslovanju koja nije nastala kao posljedica povijesnih podataka. Zato manjak prediktivne moći nije greška modela nego činjenica da se vrijeme mijenja. Sličnu tvrdnju možemo iskazati i o slučajnom uzorku, on je dobra mjera koja objašnjava snagu modela, ali omogućava slab uvid u buduće ponašanje modela. Na kraju modelari odlučuju o varijacijama ovih dvaju modela.

Modeliranje potražnje i izvođenje elastičnosti pruža nam slikovito razumijevanje korisničke reakcije na predstavljenu cijenu. To razumijevanje je ključno u postavljanju ispravne mjere za rizik.

Za kraj, kod izrađivanja modela ključno je shvaćanje da bez obzira na strategiju najbolji model je uvijek mješavina tehničke stručnosti s poslovnom sposobnošću.

Poglavlje 4

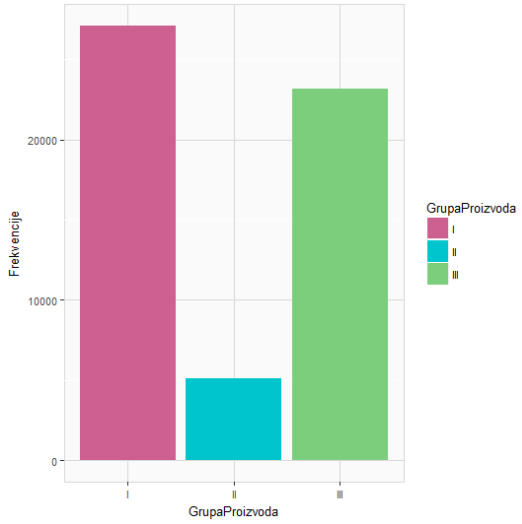
Opisna statistika

U ovom poglavlju bavit ćemo se slikovitim prikazom podataka i njihovom opisnom analizom. Analize su rađene u programskom jeziku *R*-u i za njihovu izradu je korišten paket *ggplot2*, o kojem se može više pročitati u [6]. Kao što je napisano u uvodu ovaj diplomski rad je temeljen na stvarnim prodajnim podacima. Podaci su dobiveni suradnjom fakulteta s jednom od vodećih prehrambenih industrija u državi. Naime oni su dostavili podatke vezane za određene grupe njihovih proizvoda koji su se skupljali u periodu od 3 godine, točnije od 2013. do 2015. godine.

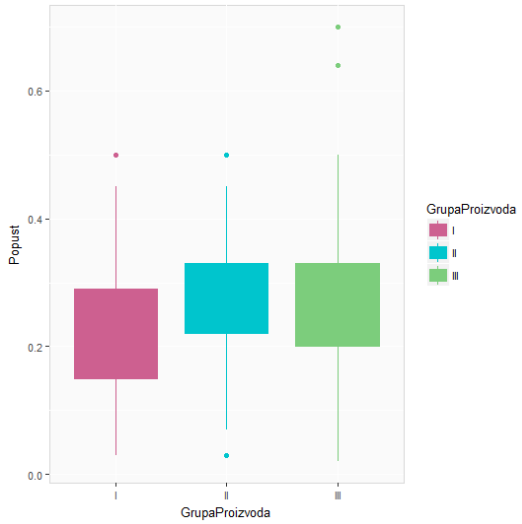
Za početak ćemo opisati varijable iz skupa podataka i predstaviti njihove vrijednosti. Kao što smo rekli podatke dijelimo na 3 grupe proizvoda, na koje ćemo se kroz ovaj rad referirati kao I, II ili III grupu proizvoda. Te grupe detaljnije u sebi sadržavaju proizvode koji su jedinstveno određeni preko njihovog ID-ja. Zapravo će nam proizvod biti osnovna točka na kojoj će se bazirati naša analiza. Za svaki od proizvoda imamo podatke o popustu koji je u određenom vremenskom periodu vrijedio, prodanu količinu i naravno akcijsku cijenu proizvoda. To su sve kontinuirane numeričke varijable, dok su mjesec, godina, grupa proizvoda i tjedan zapravo kategoričke. U nastavku možemo vidjeti dio podataka iz tablice:

ProizvodID	Mjesec	Godina	Tjedan	GrupaProizvoda	ProdanaKoličina	Popust	AkcijskaCijena
33048	4	2013	16	II	51	0.03	0.79
33049	3	2013	10	II	41	0.03	0.86
33049	4	2013	16	II	39	0.03	1.11
33048	3	2013	10	II	58	0.03	1.11
34365	11	2013	47	III	7	0.03	1.11
36110	2	2013	6	III	26	0.20	3.99

Tablica 4.1: Prikaz dijela podataka



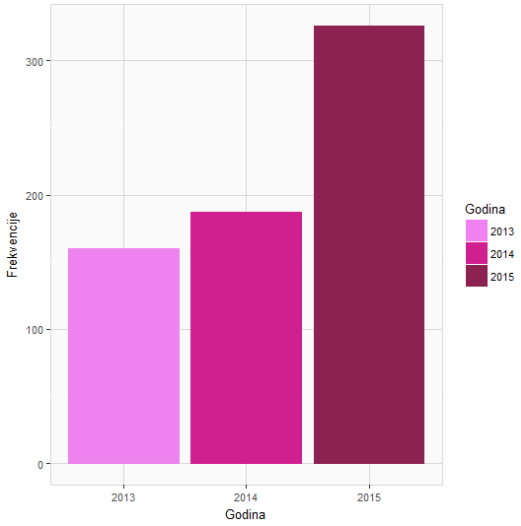
Slika 4.1: Prikaz frekvencija po grupama proizvoda



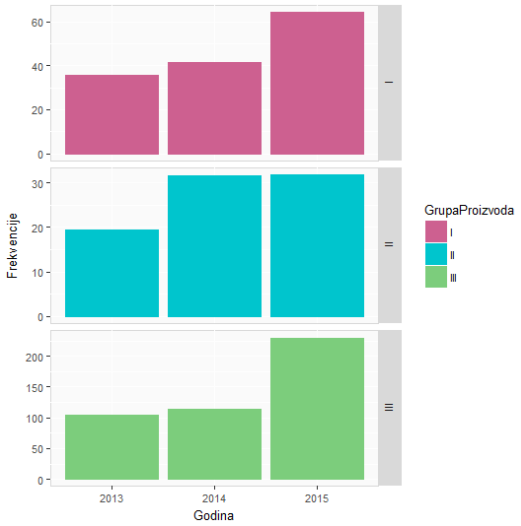
Slika 4.2: Boxplot dijagram različitih grupa proizvoda u odnosu na popuste

Počnimo sada s opisnom statistikom. Na slici 4.1 možemo vidjeti ukupni prikaz prodanih proizvoda podijeljenih u grupe. Zaključujemo da nemamo ravnomjernu prodaju jer jedna grupa proizvoda značajnije odstupa od ostale dvije, točnije GrupaProizvoda I i GrupaProizvoda III su prodane u više od 20000 komada dok je GrupaProizvoda II ostala na otprilike 5000 komada.

Pogledajmo sada boxplot dijagram na slici 4.2 koji nam prikazuje svojstva po



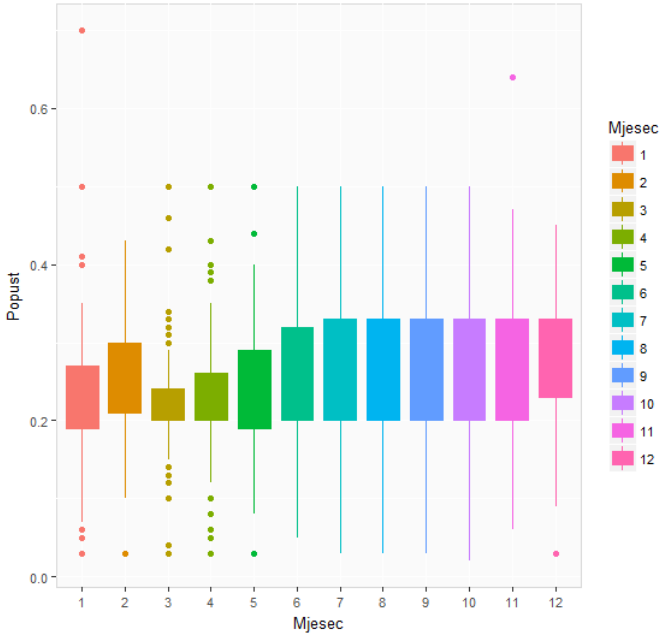
Slika 4.3: Prikaz popusta po godinama



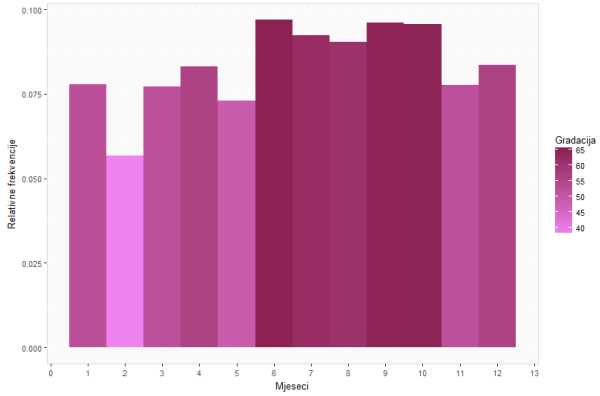
Slika 4.4: Prikaz popusta po godinama za različite grupe proizvoda

grupama proizvoda u odnosu na popuste. Znamo da nam boxplot dijagrami služe za prikaz karakteristične petorke uzorka i kod nas vidimo da nema baš prevelike razlike između izgleda dijagrama za različite grupe proizvoda. Vidimo još i da svaka od grupa ima par outlinera što je prihvatljivo s obzirom na veličinu uzorka.

Sljedeći graf na slici 4.3 prikazuje odnose između popusta i godina. Kako godine prolaze vidimo da su popusti sve više zastupljeniji. Najveći broj popusta za proma-



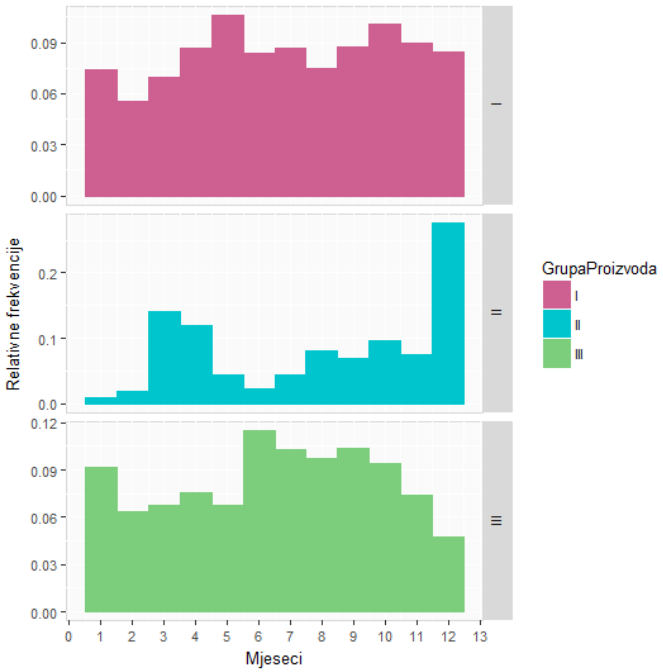
Slika 4.5: Boxplot dijagram različitih mjeseci u odnosu na popuste



Slika 4.6: Histogram popusta svih proizvoda po mjesecima

trane proizvode je bio u 2015. godini. Na slici možemo primijetiti jako velik porast od početka praćenja prodaje, porast od čak 50 posto.

Na idućoj slici 4.4 prikazani su isti odnosi kao i na prethodnoj 4.3, no ne za sve proizvode ukupno već razdijeljeno po grupama proizvoda. Vidimo da se rast u popustima kroz godine nastavio i u svakoj od GrupaProizvoda I, II, III. Primjećujemo razlike u rastu među grupama, no možemo zaključiti da je 2015. godina u svakoj od njih bila najpopunjenija popustima. Vidimo i da se frekvencije među grupama razli-



Slika 4.7: Histogram popusta različitih grupa proizvoda po mjesecima

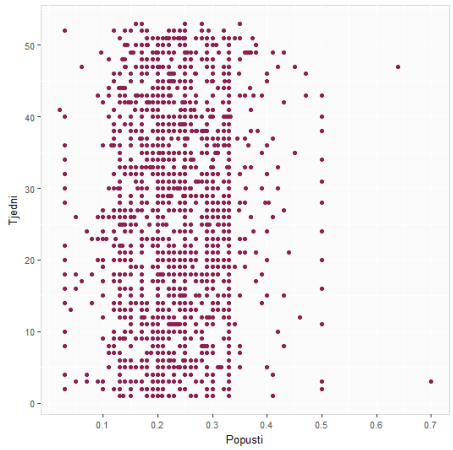
kuju. GrupaProizvoda I i GrupaProizvoda II imaju puno manje frekvencije popusta u odnosu na GrupuProizvoda III.

Slika 4.5 donosi nam novi boxplot dijagram i to ovog puta vezu mjeseci i popusta. Slično kao i prethodni boxplot dijagram donosi nam usporedbu karakterističnih petorki uzoraka podijeljenih po mjesecima u odnosu s popustima.

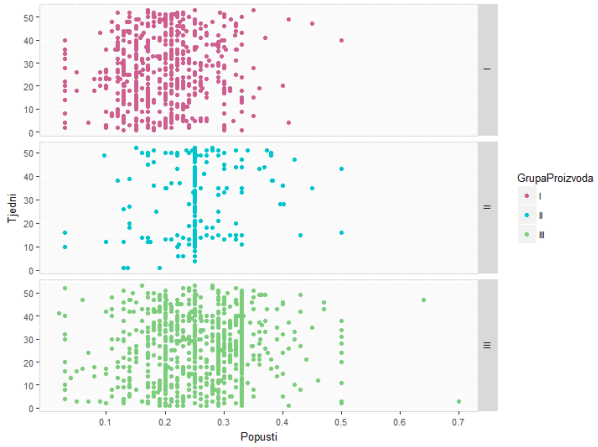
Na slici 4.6 je prikazan histogram popusta po mjesecima i vidimo da je koncentracija popusta u jednoj godini najveća u periodu od lipnja do listopada. Najmanji popusti su zabilježeni u veljači. Općenito su popusti zastupljeniji u toplijim ljetnim mjesecima, dok u zimskom periodu vidimo blagi pad.

Sljedeća slika 4.7 je histogram koji proizvode iz 4.6 dijeli na GrupeProizvoda I, II, III. Najznačajniju razliku uočavamo na grupi proizvoda II koja tijekom cijele godine ne bilježi veće zastupljenosti popusta i onda u prosincu dolazi do naglog skoka. Do izražaja dolaze i ožujak i travanj, no ne toliko značajno kao prosinac što je najbolje vidljivo na slici.

Na slici 4.8 vidimo drukčiji grafički prikaz od dosad viđenih koji nam donosi odnos između grupiranih popusta i tjedana. Uočimo da se većina popusta nalazi do neke granice 0,35 i da to vrijedi za sve tjedne. Neki tjedni imaju i više izražene popuste, vidimo da neki čak i do 0,7.



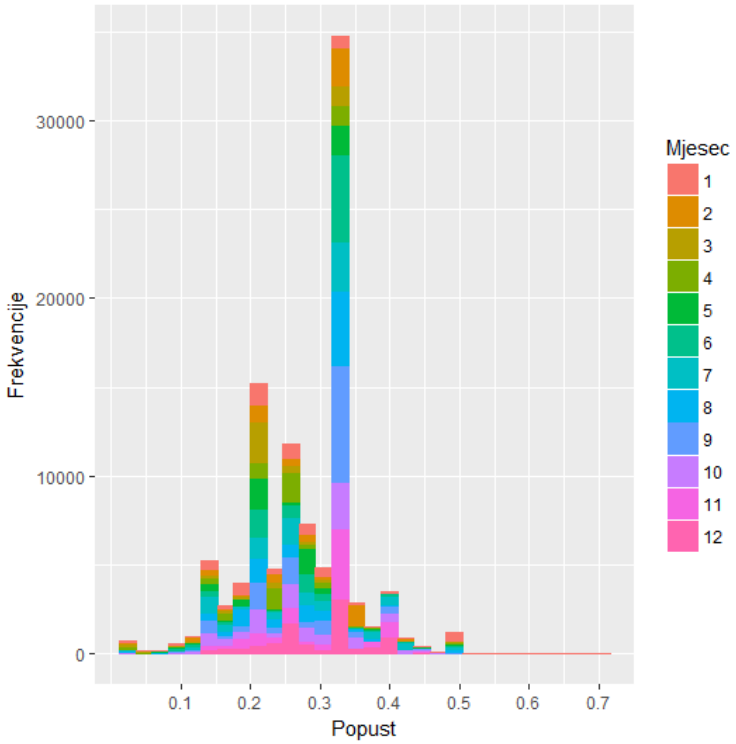
Slika 4.8: Prikaz popusta svih proizvoda po tjednima



Slika 4.9: Prikaz popusta različitih grupa proizvoda po tjednima

Slika 4.9 je nadogradnja prethodne 4.8 slike. Na njoj možemo vidjeti usporedbu među grupama proizvoda I, II i III na temelju prije opisanih uvjeta. Vidimo da su popusti najizražajniji upravo za grupu proizvoda III te da za tu grupu imamo i najveće vrijednosti popusta. Nakon nje slijedi grupa proizvoda II koja nema možda količinski puno popusta, no prosjek popusta je veći od prosjeka grupe proizvoda I, što se jasno i vidi iz prikaza. Ovo je možda jedna od boljih slika gdje se vidi omjer podataka po grupama proizvoda i možemo zaključiti da u našem setu podataka imamo najviše prodajnih podataka o grupi proizvoda III, pa slijedi grupa proizvoda I i za kraj ostaje grupa proizvoda II.

I za kraj donosimo graf 4.10 koji nam slikovito predočava prodane količine akcij-



Slika 4.10: Prikaz prodanih proizvoda po popustima za različite mjesece

skih proizvoda iz našeg seta podataka po grupiranim popustima za svaki od mjeseci u godini.

S ovim završavamo poglavlje opisne statistike u kojem smo na slikovit način pokušali približiti odnose među varijablama i podacima. No ovo nam je služilo samo kao uvod u podatke i zapravo na tome se ništa dalje ne može graditi niti se neke odluke mogu na tome temeljiti. Za to nam je potrebna daljnja analiza podataka koju ćemo odraditi u sjedećem poglavlju.

Poglavlje 5

Analiza rezultata linearne regresije

Nakon što smo u prethodnom poglavlju na slikovit način prikazali odnose među varijablama sada nam slijedi primjena linearne regresije na naše podatke u svrhu dobivanja što boljih željenih rezultata. Cilj nam je doći do modela koji najbolje opisuje podatke i pomoću njega odrediti međusoban odnos popusta i prodane količine akcijskih proizvoda. Krajnji rezultat će nam odrediti cjenovnu elastičnost proizvoda, što je zapravo i željeni rezultat našeg rada.

Analiza podatka je rađena kao i opisna statistika u programskom jeziku *R* koji već ima ugrađene funkcije za potrebe linearne regresije. U tablici 5.1 donosimo usporedbu tri modela linearne regresije. Razlika među modelima je u varijabli odziva koju smo modelirali na različite načine. Varijabla odziva u našem slučaju je Akcij-

```
Calls:
z: lm(formula=AkcijskaCijena~., data=data)
z1: lm(formula=log(AkcijskaCijena)~., data=data)
z2: lm(formula=yjPower(AkcijskaCijena,lambda = 0)~., data=data)

=============================================================================
                              z              z1              z2
-----------------------------------------------------------------------------
  (Intercept)             306.444          42.558          33.013
                         (302.173)        (25.392)        (21.317)
  ProizvodID               -0.000'*'        -0.000'***'     -0.000'***'
                            (0.000)          (0.000)         (0.000)
  Mjesec                     0.265           -0.004          -0.003
                            (0.401)          (0.034)         (0.028)
  Godina                    -0.142           -0.020          -0.015
                            (0.150)          (0.013)         (0.011)
  Tjedan                    -0.063            0.001           0.001
                            (0.092)          (0.008)         (0.007)
  GrupaProizvoda: II      -12.373'***'     -1.070'***'     -0.961'***'
                            (0.425)          (0.036)         (0.030)
  GrupaProizvoda: III     -12.999'***'     -1.220'***'     -1.084'***'
                            (0.299)          (0.025)         (0.021)
  ProdanaKolicina            0.000'***'       0.000'***'      0.000'***'
                            (0.000)          (0.000)         (0.000)
  Popust                    -4.029'*'        -1.360'***'     -1.073'***'
                            (1.581)          (0.133)         (0.112)
-----------------------------------------------------------------------------
  R-squared                  0.500            0.576           0.598
  adj. R-squared             0.499            0.574           0.596
  sigma                      6.161            0.518           0.435
  F                        341.170          462.353         506.010
  p                          0.000            0.000           0.000
  Log-likelihood         -8849.234        -2075.861       -1597.453
  Deviance              103477.841          730.703         514.999
  AIC                    17718.467         4171.722        3214.907
  BIC                    17777.606         4230.861        3274.046
  N                       2735             2735            2735
=============================================================================
```

Tablica 5.1: Usporedba modela linearne regresije

skaCijena jer želimo vidjeti ponašanje i ovisnost te varijable o ostalim varijablama iz skupa podataka. U prvom modelu z varijabla je ostavljena u izvornom obliku i nije modificirana. Model $z1$ ima transformacije i to tako da smo spomenutu varijablu odziva logaritmirali, dok smo u modelu $z2$ primijenili Yeo-Johnson transformaciju. Iz tablice iščitavamo da je $z2$ najprihvatljiviji model upravo zbog najvećeg utjecaja ostalih varijabli na traženu varijablu odziva. Nadalje, za taj model imamo i najveći R^2 koji nam služi za procjenu kvalitete nekog modela.

Od sada pa nadalje analiziramo model $z2$ za koji smo ustanovili da nam pruža najprihvatljivije rezultate. U tablici 5.1 možemo vidjeti i da neke od varijabli kao što su Mjesec, Godina, Tjedan nemaju utjecaj na AkcijskuCijenu pa ćemo zato napraviti detaljniju analizu odabranog modela. Poslužit ćemo se AIC kriterijem (eng. *Akaike Information Criterion*) za dobivanje modela koji će biti ravnoteža između veličine i pristajanja podacima.

$$AIC = -2(log - likelihood) + 2(p + 1)$$

gdje za linearne modele vrijedi

$$-2(log - likelihood) = nlog(SSE/n) = nlog\left(\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}\right)$$

Koristit ćemo funkciju *step* koja generira model s najmanjom vrijednošću AIC kriterija. Kao rezultat dobivamo model prikazan u tablici 5.2.

Promotrimo sada mjere prikazne u toj tablici. Za početak vidimo da su nam od početnih varijabli u modelu ostale samo one značajne ProizvodID, GrupaProizvoda, ProdanaKoličina i Popust. Njihov utjecaj možemo vidjeti po jako niskim p-vrijednostima u ovom modelu.

U tablici vidimo nekoliko vrsta mjera koje ćemo sada pobliže objasniti. Kolona *Estimate*, odnosno procijenjena vrijednost prikazuje nam nagib pravca određenog s varijablom odziva i promatranom varijablom. Dakle povećanje za jedan u varijabli odziva znači povećanje za vrijednost iskazanu u ovoj koloni promatrane varijable. Na primjeru Popusta i AkcijskeCijene vidimo da ukoliko se AkcijskaCijena poveća za 1 da će se Popust smanjiti za 1.080, to jest nagib pravca će biti negativan i iznosit će -1.080. Kod ove mjere vidimo da najveći utjecaj na AkcijskuCijenu imaju Popust, GrupaProizvodaIII i slobodan član, oni su jedini s vrijednostima koeficijenta reda nula.

Sljedeća kolona *Std.Error* standardna devijacija mjeri prosječne vrijednosti odstupanja procijenjenih vrijednosti varijabli od stvarne prosječne vrijednosti varijable. U našem slučaju te vrijednosti su jako male u odnosu na procijenjene iznose što je jako dobro i znači da nemamo previše varijacija unutar varijabli.

```
Call:
lm(formula = yjPower(AkcijskaCijena, lambda = 0) ~ ProizvodID +
    GrupaProizvoda + ProdanaKolicina + Popust, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30448 -0.28652 -0.00599  0.25264  2.93649

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.196e+00  2.965e-02 107.815  < 2e-16 '***'
ProizvodID        -3.057e-07  5.112e-08  -5.979 2.53e-09 '***'
GrupaProizvodaII  -9.597e-01  2.983e-02 -32.168  < 2e-16 '***'
GrupaProizvodaIII -1.084e+00  2.102e-02 -51.560  < 2e-16 '***'
ProdanaKolicina    1.658e-23  2.945e-24   5.632 1.96e-08 '***'
Popust            -1.080e+00  1.110e-01  -9.733  < 2e-16 '***'
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4346 on 2729 degrees of freedom
Multiple R-squared:  0.5973,    Adjusted R-squared:  0.5966
F-statistic: 809.5 on 5 and 2729 DF,  p-value: < 2.2e-16
```

Tablica 5.2: Prikaz odabranog modela i njegovih mjera

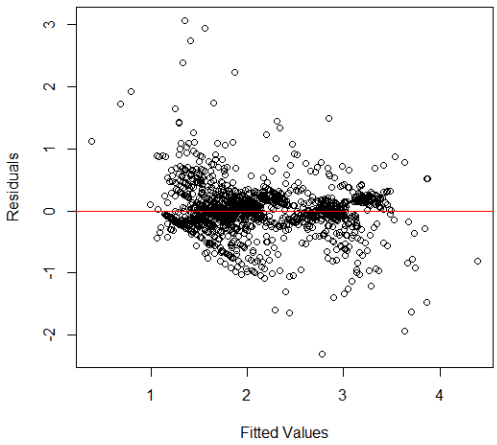
Zatim gledamo *t value* t statistiku koji nam predstavlja mjeru za udaljenost standardne devijacije promatrane varijable od nule. Preko nje dolazimo do iduće kolone $Pr(> |t|)$ p-vrijednosti koja predstavlja vjerojatnost da testna statistika poprimi vrijednosti koju su uz pretpostavku da je nulta hipoteza istinita, manje ili jednako vjerojatne od opažene vrijednosti testne statistike. Za svaku od promatranih varijabli p-vrijednost je jako mala pa možemo zaključiti da na nivou značajnosti od 5% odbacujemo nultu hipotezu o nepostojanosti veze između promatrane i varijable odziva.

Dakle došli smo do zaključaka da na razini značajnosti od 5% postoji veza između AkcijkeCijene i ostalih varijabli. Točnije naš model se pokazao kao dobar procjenitelj podataka. Sad nam ostaje samo još testirati uvjete za korištenje tog modela.

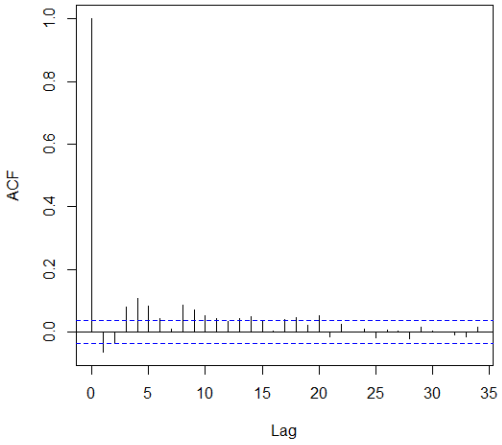
Da bismo mogli koristiti rezultate odabranog modela linearne regresije trebaju nam biti zadovoljene sljedeće pretpostavke:

1. linearan odnos između varijable odziva i promatranih varijabli poticaja
2. nezavisnost grešaka
3. homogenost grešaka
4. normalna distribuiranost grešaka

Pogledamo prvo *residual-fit* graf na slici 5.1 koji nam prikazuje odnos između reziduala i predviđenih vrijednosti kao i homogenost grešaka. Uočimo da su po-



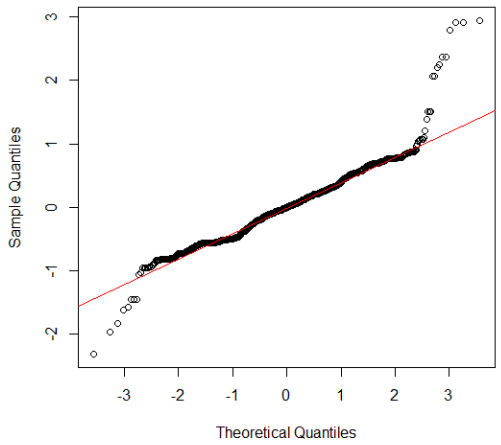
Slika 5.1: Prikaz reziduala i predviđenih vrijednosti



Slika 5.2: Grafički prikaz autokorelacijske funckije

daci većinom ravnomjerno raspršeni oko apscise, osim nekolicine outlinera koji nam narušavaju sklad podataka. No za potrebne naše analize možemo pretpostaviti homogenost grešaka.

Na slici 5.2 vidimo grafički prikaz vrijednosti autokorelacijske funkcije s pripadnom 95%-pouzdanom prugom. Primjećujemo da su reziduali nekorelirani, što povlači i njihovu nezavisnost. Dakle većina uvjeta za korištenje odabranog modela linearne



Slika 5.3: Prikaz normalnog vjerojatnosnog grafa

regresije je zadovoljna, ostaje nam još samo provjeriti normalnost.

Slika 5.3 donosi nam normalni vjerojatnosni graf na kojem možemo vidjeti da podaci većim dijelom odgovaraju normalnom pravcu. Vidimo da imamo mala odstupanja na krajevima koja nam impliciraju da se tu radi o greškama koje imaju distribuciju teškog repa što bi moglo stvarati probleme. No u našem slučaju rezultati prikazani na ovom grafu će biti dovoljni.

Nakon što smo ustvrdili da vrijede preduvjeti potrebni za korištenje linearne regresije, preostaje nam sumirati zaključke. Cilj ovog rada bio je ispitati elastičnost cijena i pronaći utjecaj dostupnih varijabli na cijenu pomoću kojih bismo kasnije istu tu cijenu i kreirali. U terminima našeg skupa podataka to znači ispitati ovisnost AkcijskeCijene o ostalim varijablama i pokušati doći do nekih značajnijih veza među njima. Odlučili smo se na model generiran linearnom regresijom za koji se koeficijenti prikazani u 5.2. Tu možemo vidjeti da je p-vrijednost jako mala ($p < 2.2e-16$) što nam znači da veze među varijablama postoje. Pripadajući R^2 i prilagođeni R^2 imaju redom vrijednosti 0.5973 i 0.5966 pa zaključujemo da odabrani model jako dobro opisuje dobivene podatke. Preko procijenjenih vrijednosti koje smo komentirali prije u tekstu dolazimo do finalnog dijela ovog rada, odnosno formule modela po kojoj se dolazi do optimalne vrijednosti AkcijskeCijene.

```
AkcijskaCijena = 3.196e+00 - 3.057e-07ProizvodID
        - 9.597e-01GrupaProizvodaII - 1.084e+00GrupaProizvodaIII
        + 1.658e-23ProdanaKolicina - 1.080e+00Popust
```

Za kraj možemo reći da se linearna regresija pokazala kao dovoljna metoda za procjenu elastičnosti cijena, no definitivno tu ima prostora za napredak i poboljšanja. Za detaljniju analizu i poželjnu optimizaciju cijena smatram da bi neki drugi modeli bili puno prihvatljiviji izbor za analizu dobivenih podataka.

Poglavlje 6

Dodatak

U dodatku ćemo predstaviti isječke kodova koji su korišteni za prikaz podataka i statističku analizu u ovom radu. Kronološki će biti poredani tako da se može pratiti tijek njihovog izvršavanja.

Učitavanje podataka

Učitavanje podataka i potrebnih paketa za crtanje u *R*-u:

```
library(ggplot2)
library(memisc)
podaci<-read.csv("Ispravni podaci.csv",sep=";", stringsAsFactors=FALSE)
```

Nakon učitavanja bilo je potrebno napraviti normalizaciju prodajnih podataka. Odlučili smo se na normalizaciju podataka na mjesečnoj bazi:

```
pi<-aggregate(podaci$Koli\v{c}ina,by=list(podaci$ProizvodID,podaci$Mjesec,
        podaci$Godina,podaci$Tjedan,podaci$GrupaProizvoda), FUN=sum)

p<-aggregate(podaci$Koli\v{c}ina,by=list(podaci$ProizvodID,podaci$Godina,
        podaci$Tjedan,podaci$GrupaProizvoda), FUN=sum)

p[,5]<-p[,5]/12

pv<-aggregate(podaci$Popust,by=list(podaci$ProizvodID,podaci$Mjesec,
        podaci$Godina,podaci$Tjedan,podaci$GrupaProizvoda), FUN=mean)

pac<-aggregate(podaci$AkcijskaCijena,by=list(podaci$ProizvodID,
        podaci$Mjesec,podaci$Godina,podaci$Tjedan,podaci$GrupaProizvoda),
        FUN=mean)

pi<-pi[order(pi[,1],pi[,2],pi[,3],pi[,4],pi[,5]),]
```

```
pv<-pv[order(pv[,1],pv[,2],pv[,3],pv[,4],pv[,5]),]
pac<-pac[order(pac[,1],pac[,2],pac[,3],pac[,4],pac[,5]),]

data <- data.frame(pi, pv[,6], pac[,6])
```

Opisna statistika

Slika 4.1: Prikaz frekvencija po grupama proizvoda

```
ggplot(data)+
geom_bar(aes(GrupaProizvoda,weight=Koli\v{c}ina,fill=GrupaProizvoda))+
scale_color_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
scale_fill_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
xlab("GrupaProizvoda") + ylab("Frekvencije") + labs(fill="GrupaProizvoda") +
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
panel.grid.major = element_line(colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Slika 4.2: Boxplot dijagram različitih grupa proizvoda u odnosu na popuste

```
ggplot(data,
aes(x=GrupaProizvoda,y=Popust,color=GrupaProizvoda,fill=GrupaProizvoda))+
geom_boxplot() +
labs(x="GrupaProizvoda", y="Popust") +
scale_color_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
scale_fill_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Slika 4.3: Prikaz popusta po godinama

```
ggplot(data)+
geom_bar(aes(Godina, weight=Popust,fill=Godina,color=Godina))+
scale_color_manual(values=c("violet", "violetred", "violetred4"))+
scale_fill_manual(values=c("violet", "violetred", "violetred4"))+
xlab("Godina") + ylab("Frekvencije") + labs(fill="Godina")+
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
panel.grid.major = element_line(colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Slika 4.4: Prikaz popusta po godinama za različite grupe proizvoda

```
ggplot(data)+
geom_bar(aes(Godina,weight=Popust,color=GrupaProizvoda,fill=GrupaProizvoda))
+labs(x="Godina", y="Frekvencije")+
facet_grid(GrupaProizvoda~., scales = "free")+
scale_color_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
scale_fill_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Slika 4.5: Boxplot dijagram različitih mjeseci u odnosu na popuste

```
ggplot(data,aes(x=Mjesec,y=Popust,color=Mjesec,fill=Mjesec))+
geom_boxplot()+
labs(x="Mjesec", y="Popust")+
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Slika 4.6: Histogram popusta svih proizvoda po mjesecima

```
ggplot(data)+
geom_histogram(aes(Mjesec,..density..,weight=Popust,fill=..count..),
binwidth=1)+xlab("Mjeseci") + ylab("Relativne frekvencije")+
scale_x_continuous(breaks=0:13)+
scale_fill_gradient("Gradacija",low = "violet",high = "violetred4")+
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Slika 4.7: Histogram popusta različitih grupa proizvoda po mjesecima

```
ggplot(data)+
geom_histogram(aes(Mjesec,..density..,weight = Popust,color=GrupaProizvoda,
fill=GrupaProizvoda),binwidth=1)+
labs(x="Mjeseci",y="Relativne frekvencije",fill="GrupaProizvoda")+
scale_x_continuous(breaks=0:13) +
facet_grid(GrupaProizvoda~., scales = "free") +
scale_color_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
scale_fill_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
legend.text = element_text(size = 8),    text=element_text(size=10))
```

Slika 4.8: Prikaz popusta svih proizvoda po tjednima

```
qplot(Popust, Tjedan, data=data, color="violetred4")+
labs(x="Popusti", y="Tjedni") +
scale_x_continuous(breaks=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1))+
scale_y_continuous(breaks=c(0,10,20,30,40,50,60))+
scale_color_manual(values=c("violetred4"))+
theme(legend.position="none",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Slika 4.9: Prikaz popusta različitih grupa proizvoda po tjednima

```
qplot(Popust,Tjedan, data=data,fill=GrupaProizvoda,color=GrupaProizvoda)+
labs(x="Popusti",y="Tjedni")+
scale_x_continuous(breaks=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1))+
scale_y_continuous(breaks=c(0,10,20,30,40,50,60))+
facet_grid(GrupaProizvoda~., scales = "free")+
scale_color_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
scale_fill_manual(values=c("hotpink3", "turquoise3", "palegreen3"))+
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Slika 4.10: Prikaz prodanih proizvoda po popustima za različite mjesece

```
ggplot(data)+
geom_bar(aes(Popust,weight=Koli\v{c}ina,color=Mjesec,fill=Mjesec))+
labs(x="Popust",y="Frekvencije")+
scale_x_continuous(breaks=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1))+
facet_grid(Mjesec~., scales = "free")+
theme(legend.position="right",
panel.background = element_rect(fill = "gray98", colour = "gray85"),
legend.text = element_text(size = 8),
text=element_text(size=10))
```

Linearna regresija

Tablica 5.1: Usporedba modela linearne regresije

```
z <- lm(formula = AkcijskaCijena~. , data=data)
z1 <- lm(formula = log(AkcijskaCijena)~. , data=data)
z2 <- lm(formula = yjPower(AkcijskaCijena, lambda=0)~. , data=data)
mtable(z,z1,z2)
```

Tablica 5.2: Prikaz odabranog modela i njegovih mjera

```
zn<-step(z2)
summary(zn)
```

Slika 5.1: Grafički prikaz reziduala i predviđenih vrijednosti

```
plot(zn2$fit,zn2$res,xlab="Fitted Values",ylab="Residuals")
abline(0,0,lty=2)
abline(mean(zn2$res),0,col="red")
```

Slika 5.2 Grafički prikaz autokorelacijske funkcije

```
acf(zn$res, main="")
```

Slika 5.3: Grafički prikaz normalnog vjerojatnosnog grafa

```
qqnorm(zn$res)
qqline(zn$res,col="red")
```

Bibliografija

[1] Friedman, Jerome, Trevor Hastie i Robert Tibshirani: *The elements of statistical learning*, svezak 1. Springer series in statistics New York, 2001.

[2] Guven, Serhat i Michael McPhail: *Beyond the Cost Model: Understanding Price Elasticity and Its Applications*. U *Casualty Actuarial Society E-Forum, Spring 2013*. Citeseer, 2013.

[3] Keener, Robert W: *Theoretical statistics: Topics for a core course*. Springer, 2011.

[4] Rencher, Alvin C i G Bruce Schaalje: *Linear models in statistics*. John Wiley & Sons, 2008.

[5] Sen, Ashish i Muni Srivastava: *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 2012.

[6] Wickham, Hadley: *ggplot2: elegant graphics for data analysis*. Springer, 2016.

Sažetak

Cilj ovog rada je pronaći odgovarajući statistički model za procjenu elastičnosti cijena pomoću kojeg bismo mogli predviđati buduća ponašanja. Koristili smo metodu linearne regresije opisanu u prva dva poglavlja ovog rada. Kako su dolazni podaci bili raznovrsni prije početka obrade i analize proveli smo normalizaciju tih podataka i to na temelju mjeseče prodaje. Dobivene normalizirane podatke smo zatim metodom linearne regresije procijenili te još jednom transformirali za dobivanje boljih rezultata. Na konačno oblikovane i procijenjene podatke primijenili smo provjeru Gauss-Markovljevih uvjeta, koji su ispali zadovoljeni pa je korištenje linearne regresije kao modela postalo opravdano. Na kraju je predstavljen model za procjenu cijena u maloprodaju po kojem bi se ubuduće mogle te iste cijene kreirati. Zaključak nam donosi osvrt na cjelokupnu analizu i prijedlog za odabir prilagođenijih metoda za detaljniju optimizaciju cijena.

Summary

The main goal of this thesis is to find adequate statistical model for estimating price elasticity which could be used to anticipate future pricing behaviour. For that purpose we will be using linear regression described in first two chapters. Since we had raw sales data, we normalized them on monthly bases. Using theoretical knowledge we defined model and its coefficients based on linear regression method. Next thing was checking assumptions known as Gauss-Markov conditions. Satisfied assumptions implied validity of chosen model. Main formula from our model is presented at the end and it could be used to predict future prices in sale. Conclusin of this thesis brings us overview and analytic point of view for choosing linear regression as our method.

Životopis

Rođena sam 11. lipnja 1993. godine u Splitu. Do odlaska na fakultet živjela sam i odrastala u Trogiru gdje sam i završila Osnovnu školu Majstora Radovana. Godine 2012. završavam Opću gimnaziju Ivana Lucića u Trogiru i upisujem Prirodoslovno-matematički fakultet u Zagrebu, smjer Matematika. Završavam preddiplomski dio studija 2016. godine i stječem titulu univ.bacc.math te se iste godine upisujem na Diplomski studij Matematičke statistike na Prirodoslovno-matematičkom fakultetu u Zagrebu.

Od 2014. do 2017. godina bila sam aktivna članica studentske udruge eSTUDENT koja promiče aktivnost studenata i njihovu suradnju s poduzećima. Od ožujka 2017. godine radim u Zagrebačkoj banci u odjelu Razvoj poslovne inteligencije.