Višestruka regresija i prognoza ukupne akcijske prodaje

Šunjo, Nina

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Nina Šunjo

VIŠESTRUKA REGRESIJA I PROGNOZA UKUPNE AKCIJSKE PRODAJE

Diplomski rad

Voditelj rada:
prof.dr.sc.Miljenko Marušić

Zagreb, rujan, 2014.

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. _____

Sadržaj

Uvod

U današnje vrijeme sve je više konkurentnih trgovačkih lanaca. Svi ti lanci prilagođavaju cijene svojih proizvoda potencijalnim kupcima. Da bi svaki od njih opstao na tržištu potrebno je stalno praćenje prodaje i djelovanje na temelju zapažanja. Kada je skupljeno dovoljno korisnih informacija, stvara se bolja podloga za marketing kojim se plasiraju proizvodi. Iz podataka o prodaji dobiva se uvid u želje i potrebe kupaca što se onda koristi u planiranju akcija. Postoje razne vrste potrošača, od onih koji kupuju isključivo proizvode s nižim cijenama, preko onih kojima je kvaliteta i cijena bitna, do onih koji žele kvalitetu i nije im važna cijena. Na nekim područjima Hrvatske prodaja određenog artikla je odlična, na drugima je slaba. U određeno doba godine bolje se prodaju neki artikli koji se inače loše prodaju. Mnogo je varijabli koje utječu na prodaju i zbog toga ne možemo tako jednostavno doći do zaključka kako će ubuduće ići prodaja raznih proizvoda. Zato postoje matematički modeli koji nastaju na temelju dosadašnjih podataka. Njima možemo doći do zaključka kako će utjecati povećanje ili smanjenje neke varijable na prodaju.Što bolje model predviđa buduće reakcije tržišta to je uspješnije poslovanje. Uvijek je dobro iskušavati i istraživati nove modele jer uvijek ima mjesta za napredak.

U ovom radu se obrađuje tema grupiranja podataka kao mogućeg unapređenja modela reakcije tržišta. Za takvo istraživanje potrebni su stvarni podaci. Oni su dobiveni od hrvatskog trgovačkog lanca Konzum. Podaci na kojima se radi sastoje se od 20 varijabli, a svaka varijabla ima 186 403 opažanja.

Da bi se došlo do najpovoljnijeg modela potrebno je akcijsku prodaju prikazati preko ostalih varijabli. Akcijska prodaja je zavisna varijabla, a ostale su nezavisne. Kreće se od punog modela koji sadrži sve nezavisne varijable, a zatim se smanjuje do najmanjeg adekvatnog.

Cilj je doći do modela koji će imati najmanju moguću grešku predviđanja, a birati će se između modela dobivenog grupiranjem i modela dobivenog sa cijelim skupom podataka. Ukoliko je bolji model kod grupiranih podataka, može se iskoristiti u budućim predviđanjima tvrtke.

Poglavlje 1

Grupiranje podataka

Grupiranje podataka je nenadzirana klasifikacija uzoraka (zapažanja, podatkovnih stavki, značajki) u grupe. Ono čini jedan korak u istraživačkoj analizi podataka. Analiza grupa je organizacija kolekcije uzoraka (koji su obično predstavljeni kao vektori mjerenja ili točke u višedimenzionalnom prostoru) u grupe i to na temelju sličnosti. Uzorci u valjanoj grupi su sličniji međusobno nego što su slični uzorku iz druge grupe.
Razlika između nenadzirane i nadzirane klasifikacije uzoraka je ta da kod nadzirane klasifikacije imamo skup označenih uzoraka. Problem je označiti novootkriveni, ali i neoznačeni uzorak. Označeni uzorci koje već imamo koriste se pri opisu klasa u kojima su označeni novi uzorci. Kod grupiranja, problem je grupirati danu kolekciju neoznačenih uzoraka u značajne grupe. Oznake su također vezane za grupe, ali su izvedene iz podataka.
Analizom grupa podijelimo podatke u grupe (klastere) koje su značajne, korisne, ili oboje. Ako su cilj značajne grupe, onda bi klasteri trebali prikazati prirodnu strukturu podataka. No u nekim slučajevima, analiza grupa je samo početna točka za druge svrhe, kao što je sažimanje podataka. Analiza grupa je bitna za razna područja: društvene nauke, biologiju, statistiku, prepoznavanje uzoraka, povrat informacija, strojno učenje i rudarenje podataka. Mnogo je primjena analize grupa na praktične probleme, a služi za razumijevanje ili za korisnost.

1.1 Primjene grupiranja

Klase ili značajne grupe objekata koji imaju zajedničke karakteristike, igraju važnu ulogu u tome kako netko analizira i opisuje svijet. Ljudi imaju sposobnost dijeljenja objekata u grupe i dodjeljivanja objekta nekoj grupi. U kontekstu razumijevanja podataka, klasteri su potencijalne klase, a analiza grupa je proučavanje tehnika za automatsko nalaženje klasa. Primjer klasteringa za razumijevanje je onaj koji se primjenjuje u poslovanju.
Kod poslovanja i trgovine skupljaju se velike količine podataka o trenutnim i potencijalnim mušterijama. Grupiranjem se mušterije mogu segmentirati u mali broj grupa za daljnju analizu i marketinške aktivnosti.
Analiza grupa pruža apstrakciju iz individualnih podatkovnih objekata u klastere kojima ti objekti pripadaju. Neke tehnike grupiranja okarakteriziraju klaster pomoću prototipa klastera, odnosno podatkovnog objekta koji predstavlja druge objekte u klasteru. Ti se prototipovi mogu koristiti kao temelj za mnoge analize podataka ili tehnike procesiranja podataka. U kontekstu korisnosti, klaster analiza je proučavanje tehnika nalaženja najreprezentativnijeg prototipa klastera.

Istraživanje podataka

Grupiranje se često koristi za istraživanje podataka u kojima se pokušava pronaći određena struktura. Primjeri se najprije grupiraju, te se dobivene grupe označavaju. Središta grupa se smatraju prototipnim predstavnicima grupa, a za svaku grupu može se utvrditi raspon vrijednosti karakteristika. Time se podaci mogu jednostavnije opisati, te se lakše uočavaju pravilnosti i sličnosti među podacima.

Sažimanje

Mnoge tehnike analize podataka, kao što je regresija, imaju složenost prostora $O(m^2)$ ili višu (gdje je m broj objekata), stoga nisu praktične za velike količine podataka. Umjesto da algoritam primjenimo na svim cijelom skupu podataka, moguće je primijeniti ga na smanjenom skupu podataka koji se sastoji samo od prototipova grupa. Ovisno o tipu analize, broj prototipova i preciznost kojom oni predstavljaju podatke, rezultati mogu biti usporedivi sa onima koji bi se dobili da se analizirao cijeli skup podataka. To je predobrada podataka kojoj je cilj smanjiti dimenzionalnost prostora primjera, odnosno broja značajki. Time dolazi do uštede prostora i vremena izvođenja.

Kompresija podataka

Prototipovi klastera mogu se koristiti i za kompresiju podataka. Neka je napravljena tablica koja se sastoji od prototipova za svaki klaster, svakom klasteru je dodijeljen cijeli broj

koji označava njegovu poziciju u tablici. Svaki objekt je predstavljen indeksom prototipa vezanog za njegov klaster. Ovaj oblik sažimanja zove se kvantizacija vektora.

Smanjenje dimenzionalnosti grupiranjem

Početni prostor primjera možemo prikazati matricom N*x*n. Njen i-ti redak je i-ti primjer, a stupci su značajke. Grupiramo li podatke u K<n grupa, dobili smo matricu smanjenih dimenzija, na jedan od 2 načina.
Prvi način jest da grupiramo retke matrice, odnosno primjere. Time dobijemo preslikavanje u K-dimenzionalni prostor. U tom prostoru, svakom primjeru odgovara vektor iz čijih komponenti vidimo kojoj grupi taj primjer pripada. Kod čvrstog grupiranja, jedna komponenta je jednaka jedinici dok su ostale jednake nuli. Kod mekog grupiranja više komponenti može biti različito od nule. Čvrstim grupiranjem dolazi do velikog gubitka informacija jer se mnogo primjera preslikava u istu točku K-dimenzionalnog prostora, te dobivena matrica ima manje od N redaka.
Drugi način jest da grupiramo stupce matrice, odnosno značajke primjera. U istu grupu želimo smjestiti međusobno slične značajke. Što se više vrijednosti značajki kod pojedinačnih primjera podudaraju, to su značajke sličnije. Dimenzionalnost se smanji zamjenom svih značajki koje su zajedno grupirane novom, reprezentativnom značajkom. Ta značajka naziva se centroid grupe, a pristup grupiranje značajki.

1.2 Vrste grupiranja

Postoje razne vrste grupiranja kao što su: hijerarhijsko (ugniježđeno) naspram particijskog (neugniježđeno), isključivo naspram preklapajućeg naspram neizrazitog, potpuno naspram parcijalnog.

Hijerarhijsko naspram particijskog

Particijsko grupiranje je podjela skupa podatakovnih objekata na podskupove (grupe) koji se ne preklapaju. Svaki podatkovni objekt je u točno jednom takvom podskupu.
Ako dopustimo da grupe imaju svoje podgrupe, dobiti ćemo hijerarhijsko grupiranje. Tada je skup ugniježđenih grupa organiziran kao stablo. Svaki čvor stabla je unija njegove djece, a korijen stabla je grupa koja sadrži sve objekte. Listovi stabla su često grupe sa jednim podatkovnim objektom.
Hijerarhijsko grupiranje se može ostvariti uzimanjem bilo kojeg člana tog slijeda, tj. rezanjem hijerarhijskog stabla na određenom nivou.

Isključujuće naspram preklapajućeg naspram neizrazitog

Isključujućim grupiranjem se svaki objekt dodijeli jednoj grupi. Objekt bi se u mnogo situacija mogao smjestiti u više od jedne grupe, što se može pomoću preklapajućeg grupiranja. Tim grupiranje se ističe činjenica da neki objekt može istodobno biti u više grupa. Preklapajuće grupiranje koristi često i kada je objekt između dvije ili više grupa i mogao bi biti dodijeljen bilo kojoj od tih grupa.
U neizrazitom grupiranju, svaki objekt pripada svakoj grupi sa značajnošću pripadnosti čija je vrijednost između 0 (apsolutno ne pripada) i 1 (apsolutno pripada). Grupe su tretirane kao neizraziti skupovi, a to su skupovi u kojima objekt pripada svim skupovima sa značajnosti između 0 i 1. Kod neizrazitog grupiranja često se nameće dodatno ograničenje, da je suma težina za svaki objekt jednaka 1. Slično tome, probabilističke tehnike grupiranja računaju vjerojatnost svake točke da pripada svakoj grupi, pa zbroj tih vjerojatnosti također treba imati sumu jednaku 1. Kako značajnost pripradnosti ili vjerojatnosti za svaki objekt imaju sumu 1, neizrazito ili probabilističko grupiranje se ne odnosi na prave višeklasne situacije kod kojih jedan objekt pripada u više klasa. Uglavnom se koristi da bi se izbjegla proizvoljnost dodavanja pojedinog objekta samo jednoj grupi kada ima sličnosti sa više njih. Probabilističko grupiranje je često pretvoreno u isključujuće grupiranje time što se svaki objekt dodaje grupi za koju je vjerojatnost najveća.

Potpuno naspram parcijalnog

Potpunim grupiranjem je svaki objekt dodijeljen grupi, dok to kod parcijalnog grupiranja nije slučaj. Neki objekti u skupu podataka ne moraju pripadati dobro definiranoj grupi i mogu predstavljati šum, outliere ili „nezanimljivu pozadinu“.

1.3 Vrste grupa

Da bi se podaci uspješno analizirali, grupiranjem se nalaze korisne grupe objekata. Postoji nekoliko različitih vrsta grupa koje su korisne u primjenama.

Dobro separirane

Grupa je skup objekata u kojem je svaki objekt bliži svakom drugom objektu u grupi nego što je ijednom izvan te grupe. Ponekad se koristi prag da bi se specificiralo da svi objekti u grupi moraju biti dovoljno blizu (ili slični) jedan drugome. Ovakve grupe možemo dobiti kada podaci prirodno sadrže grupe koje su dosta udaljene. Dobro separirane grupe mogu poprimiti bilo koji oblik.

Temeljene na predstavniku grupe

Grupa je skup objekata u kojima je svaki objekt bliži (ili sličniji) predstavniku grupe, nego predstavniku bilo koje druge grupe. Predstavnik grupe je često centroid, tj., prosjek svih točki u grupi. Kada centroid nije značajan, kao kada podaci imaju kategorijske atribute, predstavnik je često medoid, najreprezentativnija točka grupe. Za mnoge tipove podataka, predstavnik se može smatrati točkom koja je najbliže centru, a takve grupe teže loptastom obliku.

Temeljene na grafu

Ako su podaci prikazani kao graf čiji su čvorovi objekti, a spojnice označavaju veze među objektima, tada grupa može biti definirana kao grupa objekata koji su međusobno povezani, ali nisu povezani sa objektima izvan grupe. Primjer grupe temeljene na grafu jesu grupe temeljene na susjedstvu kod kojih su dva objekta povezana samo ako su unutar određene međusobne udaljenosti. Svaki objekt u takvoj grupi je bliži nekom drugom objektu unutar grupe nego što je ijednoj točci u drugoj grupi. Ovakva definicija grupe je korisna kada su grupe isprepletene.
Drugi tip grupe temeljene na grafu je klika, skup čvorova grafa koji su potpuno povezani. Ako dodamo veze između objekata u redoslijedu njihovih udaljenosti, grupa se formira kada skup objekata stvori kliku. Ovakve grupe također teže loptastom obliku.

Temeljene na gustoći

Grupa je gusto područje objekata koje je okruženo područjem niske gustoće. Ovakva definicija grupe se često koristi kada su grupe neregularne ili isprepletene, te kada su šum i outlieri prisutni.

Grupe zajedničkih svojstava

Grupu možemo definirati kao skup objekata koji dijele dio imovine. Ovakva definicija grupe obuhvaća sve prethodne definicije. Objekti u grupi temeljenoj na centru dijele svojstvo da su svi najbliži istom centroidu ili medoidu. No, takav pristup uključuje novu vrstu grupe.

1.4 Algoritam k-srednjih vrijednosti

Algoritam k-srednjih vrijednosti je particijska tehnika grupiranja temeljena na predstavniku, kojom se pronalazi broj grupa koje je odredio korisnik, i koje su predstavljene centroidima. Centroid je obično sredina grupe točaka.

To je najjednostavniji i najpoznatiji algoritam grupiranja. Primjeri se iz neoznačenog skupa primjera grupiraju u K čvrstih grupa. Parametar K je zadan unaprijed. Grupiranjem primjera u grupe koje su predstavljene vektorom centroida dolazi do određene pogreške koju izražava kriterijska funkcija (ciljna funkcija, mjera distorzije).
Kod jednostavnog algoritma k-srednjih vrijednosti, najprije izaberemo K početnih centroida, gdje je K unaprijed zadan željeni broj grupa. Svaka točka je zatim dodijeljena najbližem centroidu i svaki skup točaka dodijeljenih centroidu je grupa. Centroid svake grupe je ažuriran s obzirom na točke dodijeljene grupi. Dodijeljivanje ponavljamo i ažuriramo korake sve dok nijedna točka ne promijeni grupi, odnosno dok centroidi ne ostanu isti.
Za neke kombinacije funkcija neposredne blizine i tipova centroida, algoritam k-srednjih vrijednosti uvijek konvergira prema rješenju, tj., algoritam k-srednjih vrijednosti dolazi do trenutka kada se točke više ne pomiču iz grupe u grupu, pa se stoga ni centroidi ne mjenjaju.
Da bismo točku dodijelili najbližem centroidu, treba nam mjera udaljenosti koja kvantificira pojam 'najbližeg' za određene podatke koji se promatraju. Euklidska mjera udaljenosti se često koristi za podatke prikazane točkama u euklidskom prostoru. Za dani tip podataka moguće je da će postojati više prikladnih mjera udaljenosti. Mjere sličnosti koje se koriste za algoritam k-srednjih vrijednosti su relativno jednostavne jer algoritam ponavlja računanje sličnosti svake točke svakom centroidu.
Predzadnji korak algoritma jest da se ponovno izračuna centroid svake grupe. Centroidi mogu varirati, ovisno o mjeri udaljenosti za podatke i o cilju grupiranja koji je izražen ciljnom funkcijom. Ta ciljna funkcija ovisi o udaljenostima točaka jedne od druge i od centroida grupe, odnosno minimizira kvadrat udaljenosti svake točke od njenog najbližeg centroida.
Za podatke čija je mjera udaljenosti euklidska, ciljna funkcija koja mjeri kvalitetu grupiranja jest suma kvadrata pogreške (SSE). Mjerimo grešku svakog podatka, tj., njegovu euklidsku udaljenost od najbližeg centroida, zatim računamo totalnu sumu kvadriranih pogrešaka. Između dva različita skupa grupa koji su dobiveni iz dva pokretanja algoritma, biramo onaj koji ima najmanju kvadratnu grešku. To znači da centroidi grupe tog grupiranja bolje predstavljaju točke svoje grupe.
Definicija sume kvadrata pogreške:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2 \tag{1.1}$$

Dist je standardna euklidska udaljenost dvaju objekata u euklidskom prostoru, x je objekt, C_i je i-ta grupa, a c_i i-ti centroid.
Centroid koji minimizira pogrešku grupe je sredina. Sredina i-te grupe definirana je for-

mulom:

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x \tag{1.2}$$

m_i je broj objekata u i-toj grupi.
U trećem i četvrtom koraku algoritma, direktno se pokušava smanjiti ciljna funkcija. U trećem koraku se formiraju grupe tako što se točke pridruže najbližem centroidu, što minimizira SSE za dani skup centroida. U četvrtom koraku se preračunavaju centroidi za daljnje minimiziranje SSE.

Odabir početnih srednjih vrijednosti

Odabir početnih centroida je ključni korak jednostavnog algoritma k-srednjih vrijednosti. Postoji više načina za to. Jedan od načina je nasumičan odabir K primjera kao početnih vrijednosti centroida. Time se izbjegava postavljanje centroida na mjesta u prostoru primjera koja su prazna, ali se ne rješava problem ostanka u lokalnom minimumu. Problem kod ovakvog odabira predstavljaju i primjeri koji odskaču (outlieri), koji lako mogu biti dodijeljeni u zasebne grupe. Iako se možda čini da je dobro da takvi primjeri budu u zasebnim grupama to ipak nije tako jer je broj grupa K ograničen i one se trebaju poklapati sa većinskim grupama koje postoje u podacima.
Drugi način jest izračunati centroid svih primjera, a zatim toj vrijednosti dodavati manje slučajne vektore i tako dobiti K vektora. Time se rješava problem izoliranih primjera, ali ne i problem zaglavljivanja u lokalnome optimumu.
Ishod grupiranja ovisi o izboru početnih središta grupa, ali i o odabranom broju grupa, odnosno vrijednosti parametra K.

Prazne grupe

Problem kod jednostavnog algoritma k-srednjih vrijednosti jest da se mogu pojaviti prazne grupe ako nema točaka koje su prebačene u grupu tokom koraka dodjeljivanja. U tom slučaju je potrebno izabrati zamjenski centroid. Jedan način je da se izabere točka koja je najdalje od bilo kojeg trenutnog centroida. To eliminira točku koja trenutno najviše doprinosi kvadratu greške. Drugi način je da se izabere zamjenski centroid iz grupe koja ima najvišu SSE. To će razdijeliti grupu i smanjiti ukupnu SSE grupe. Ako ima više praznih grupa, proces se može ponoviti više puta.

Primjeri koji odskaču (outlieri)

Kada je korišten kriterij kvadrata pogreške, outlieri mogu prekomjerno utjecati na grupe koje su pronađene, a rezultirajući centroidi mogu biti manje reprezentativni nego što bi

inače bili i stoga će i SSE biti viša. Korisno je otkriti outliere i eliminirati ih prije. No, postoje slučajevi kada se oni ne bi trebali eliminirati. Kada je grupiranje korišteno za kompresiju podataka, svaka se točka mora grupirati. Outlieri su u nekim slučajevima, kao što je financijska analiza, najzanimljivije točke za promatranje. Ako outliere odstranjujemo prije grupiranja, izbjegavamo grupiranje točki koje se neće dobro grupirati.

Algoritam k-medoida

Kada raspolažemo samo informacijom o međusobnoj sličnosti parova primjera, ne koristimo mjeru udaljenosti, već mjeru sličnosti, ili suprotno, mjeru različitosti. Ona je izračunata između svih parova primjera. Poopćenje algoritma k-srednjih vrijednosti jest algoritam k-medoida, kod kojeg je kriterijska funkcija definirana mjerom različitosti između dva primjera. Mjera različitosti (odnosno sličnosti) općenitija je od euklidske udaljenosti i bilo koje druge mjere udaljenosti. Takva mjera ne mora ispunjavati uvjete metrike.
Kod algoritma k-medoida, predstavnici grupa su medoidi, a ne centroidi. Tipična izvedba je algoritam PAM (partitioning around medoids) kojem je veliki nedostatak visoka vremenska složenost.

Prednosti i nedostaci algoritma k-srednjih vrijednosti

Algoritam k-srednjih vrijednosti je jednostavan i može se upotrijebiti na raznim tipovima podataka. Ukupna vremenska složenost je linearna po svim parametrima što je prednost u odnosu na neke druge algoritme. Također je i dosta efikasan iako se često mora provesti više puta. Neke varijante su efikasnije i manje problematične. Algoritam k-srednjih vrijednosti ipak nije prikladan za sve tipove podataka. Ne može raditi sa grupama koje nisu loptaste ili sa grupama različitih veličina i gustoća. Također je problem kod podataka sa primjerima koji odskaču. Još jedna stvar s tim algoritmom jest da je ograničen na podatke za koje postoji definicija centra, centroida. Algoritam k-medoida nema to ograničenje, ali je skuplji.

1.5 Hijerarhijsko grupiranje

Tehnike hijerarhijskog grupiranja su druge po važnosti metoda grupiranja. Time se dobije hijerarhija grupa koja se može prikazati dendrogramom[1].
Dva su pristupa generiranju hijerarhijskog grupiranja: aglomerativno i divizivno. Kod aglomerativnog[2] grupiranja počinje se sa grupama koje sadrže po jedan primjer. Nakon

[1] Dendrogram je stablo kod kojega listovi odgovaraju primjerima, a vodoravne linije odgovaraju povezivanjima na određenoj udaljenosti

[2] Aglomeracija je cjelina dobivena nakupljanjem

toga se grupe spajaju dok se svi primjeri ne stope u jednu grupu.
Divizivno grupiranje kreće od jedne grupe koju postepeno razdjeljuje.
Za hijerarhijsko grupiranje potrebne su funkcije udaljenosti ili mjere sličnosti. Njima se pronalaze grupe primjera koji su najmanje udaljeni. Funkcija udaljenosti je funkcija

$$d : XxX \to \mathbb{R}$$

za koju vrijede ova svojstva:

1. $d(x^a, x^b) \geqslant 0$
2. $d(x^a, x^b) = 0$ ako i samo ako $x^a = x^b$
3. $d(x^a, x^b) = d(x^b, x^a)$
4. $d(x^a, x^b) + d(x^b, x^c) \geqslant d(x^a, x^c)$

Najčešće korištena metrika jest Minkowskijeva udaljenost:

$$d(x^a, x^b) = \left(\sum_{j=1}^{n} \left(x_j^a - x_j^b \right)^p \right)^{1/p} \tag{1.3}$$

Za p=1 to je L1-udaljenost (Manhattan distance), a za p=2 euklidska udaljenost.
Mahalanobisova udaljenost je poopćenje euklidske udaljenosti:

$$d(x^a, x^b) = \left(x^a - x^b \right)^T \Sigma^{-1} \left(x^a - x^b \right) \tag{1.4}$$

Σ je kovarijacijska matrica. .

Mjera sličnosti i njoj komplementarna mjera različitosti su općenitiji pojmovi od udaljenosti. To nisu metrike, već funkcije $s : XxX \to [0, 1]$ sa svojstvima:

1. $s(x, x) = 1$
2. $0 \leqslant s\left(x^a, x^b\right) \leqslant 1$
3. $s(x^a, x^b) = s(x^b, x^a)$

Mjera udaljenosti i mjera sličnosti mogu se preslikati jedna u drugu monotono padajućom funkcijom.
Pretvorba sličnosti u udaljenost je nešto teža zbog uvjeta nejednakosti trokuta. Udaljenost i sličnost se mogu uglavnom jednako koristiti, no u nekim slučajevima je sličnost fleksibilnija.

Mnoge aglomerativne tehnike grupiranja počinju sa pojedinačnim točkama kao grupama, zatim se sukcesivno[3] spajaju dvije grupe dok ne ostane samo jedna.

Jednostavni algoritam hijerarhijskog aglomerativnog grupiranja:

Računaj matricu udaljenosti, ako je potrebno
Ponavljaj:
Spoji dvije najbliže grupe
Ažuriraj matricu udaljenosti za usporedbu udaljenosti između novih grupa i originalnih grupa
Dok ne ostane samo jedna grupa

Najbitnija operacija navedenog algoritma jest računanje udaljenosti između dvije grupe. Definicija udaljenosti grupa razlikuje razne aglomerativne hijerarhijske tehnike grupiranja. Aglomerativne hijerarhijske tehnike grupiranja kao što su MIN, MAX i prosjek grupe, dolaze od grafički temeljenog pogleda na grupe. MIN je udaljenost između dvije najbliže točke koje su u različitim grupama, odnosno najkraći brid između dva čvora u različitim podskupovima čvorova (grupiranje jednostruke povezanosti). MAX je udaljenost između dvije najudaljenije točke u različitim grupama, odnosno najdulji brid između dva čvora u različitim podskupovima čvorova (grupiranje potpunom povezanošću). Sa te dvije udaljenosti rezultati se ne razlikuju puno, osim ako grupe nisu prirodno dobro odvojene.
U teoriji grafova, stapanje dviju grupa odgovara uvođenju brida između odgovarajućih primjera u tim dvjema grupama. Kod jednostrukog povezivanja to su dva najbliža primjera iz svake grupe. Bridovi se uvode između primjera različitih grupa. Ako je zadani broj grupa jednak 1, algoritam generira minimalno razapinjuće stablo[4].
Kod potpunog povezivanja, stapanje dviju grupa podrazumijeva uvođenje bridova između svih parova primjera, pa algoritam generira potpuno povezani graf.
Jednostruko i potpuno povezivanje su dosta osjetljivi na šum. Zato postoji grupiranje temeljem prosječne povezanosti. Kod te vrste grupiranja, za udaljenost grupa uzimaju se prosječne udaljenosti parova prosječne duljine bridova i to svih parova točaka iz različitih grupa.
Ako umjesto grafičkog pogleda na grupe uzmemo prototipni, u kojem svaku grupu predstavlja njen centroid, udaljenost grupa uobičajeno je definirati kao udaljenost između centroida grupa. Druga metoda je Wardova, također pretpostavlja da su grupe predstavljene centroidima, ali mjeri udaljenost između dvije grupe kao povećanje u SSE što rezultira stapanjem dviju grupa. Wardovom metodom se pokušava minimizirati suma kvadrata udaljenosti točaka od centroida njihove grupe, što je slučaj i kod algoritma k-srednjih vrijednosti.

[3] u slijedu

[4] stablo s putem između svaka dva brida kod kojeg je ukupan zbroj težina bridova

Prostorna složenost hijerarhijskog aglomerativnog algoritma je $O(N^2)$, dok je vremenska složenost $O(N^3)$. Kubna složenost je nedostatak algoritma, posebno jer su složenosti algoritma k-srednjih vrijednosti linearne. Moguće je implementirati algoritam tako da se umjesto matrica udaljenosti koriste neke druge strukture podataka čime se složenost može smanjiti na $O(N^2 \log N)$, no već kvadratna prostorna složenost može biti veliki problem.

1.6 Probabilistički pristup grupiranju

Kod probabilističkog pristupa postoji određena vjerojatnost pripadnosti primjera grupama. Jedan primjer može pripadati u više grupa, stoga su granice između njih meke. Za uzorke koji se trebaju grupirati smatra se da pripadaju različitim distribucijama. Zato je cilj probabilističkog grupiranja pronaći parametre svake od distribucija grupa i ako je to moguće, njihov broj. U tu svrhu koristi se miješani model što je linearna kombinacija razdioba svake od grupa. Obično se smatra da komponente mješavine imaju Gaussovu razdiobu čije parametre onda želimo naći.
Ako imamo K grupa na koje želimo podijeliti podatke, miješana gustoća (*mixture density*) je linearna kombinacija K funkcija gustoća vjerojatnosti:

$$p\left(x\right) = \sum_{k=1}^{K} \pi_k p\left(x \mid \theta_k\right) \tag{1.5}$$

Pritom su $p\left(x \mid \theta_k\right)$ komponente mješavine (*mixture components*), svaka s parametrima θ_k. Koeficijenti mješavine su π_k (*mixture coefficients*). Mora vrijediti $\sum_{k=1}^{K} \pi_k = 1$, te pošto je $p\left(x \mid \theta_k\right) \geqslant 0$ i $p\left(x\right) \geqslant 0$, mora vrijediti $0 \leqslant \pi_k \leqslant 1$. Koeficijenti π_k mogu se tumačiti kao vjerojatnosti. Marginalnu gustoću $p\left(x\right)$ možemo izraziti kao

$$p(x) = \sum_{k=1}^{K} P(c_k) p(x \mid c_k) \tag{1.6}$$

gdje je $\pi_k = P(c_k)$ prethodno nepoznata vjerojatnost odabira komponente k, dok je $p(x \mid c_k)$ gustoća od x uz odabranu komponentu k.
Bayesovim pravilom možemo dobiti aposteriorne vjerojatnosti $P(c_k \mid x)$:

$$P(c_k \mid x) = \frac{P(c_k)p(x \mid c_k)}{p(x)} = \frac{P(c_k)p(x \mid c_k)}{\sum_j P(c_j)p(x \mid c_j)} = \frac{\pi_k p(x \mid \theta_k)}{\sum_j \pi_j p(x \mid \theta_j)} = \tag{1.7}$$

koje nazivamo odgovornost, ona iskazuje kolika je vjerojatnost da primjer x pripada k-toj komponenti.
Parametre modela $\theta = \{P(c_k), \theta_k\}_{k=1}^{K}$ možemo procijeniti metodom najveće izglednosti, ali

je procjena složenija jer ne znamo koji primjer pripada kojoj komponenti. Funkcija log-izglednosti jednaka je:

$$\ln L(\theta \mid D) = \ln \prod_{i=1}^{N} p(x^{(i)}) = \ln \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k p(x^{(i)} \mid \theta_k) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k p(x^{(i)} \mid \theta_k) \tag{1.8}$$

$D = \{x_1, ..., x_N\}$ je skup za učenje sa neoznačenim primjerima. Za maksimizaciju ovog izraza koristi se metoda maksimizacije očekivanja.

Algoritam maksimizacije očekivanja

Algoritam maksimizacije očekivanja ili EM-algoritam je iterativan optimizacijski postupak za rješavanje problema najveće izglednosti kod modela s latentnim varijablama. To su slučajne varijable čije realizacije ne opažamo izravno, već zaključujemo na temelju drugih, opaženih varijabli. Te varijable se uvode ili samo kao sredstvo apstrakcije ili da modelira nešto stvarno, ali nedostupno. Tada latentnu varijablu nazivamo skrivena varijabla.
Cilj algoritma je naći parametre θ koji maksimiziraju log-izglednost $\ln L(\theta \mid D)$ gdje X predstavlja podatke. Model $p(X \mid \theta)$ proširujemo skupom latentnih varijabli Z te radimo sa zajedničkom gustoćom $p(X, Z \mid \theta)$ čijom marginalizacijom možemo dobiti marginalnu gustoću $p(X \mid \theta)$:

$$p(X \mid \theta) = \sum_{Z} p(X, Z \mid \theta) \tag{1.9}$$

Skup X, Z je potpun, a X je nepotpun skup podataka. Također, $\ln L(\theta \mid X, Z)$ je potpuna log-izglednost, a $\ln L(\theta \mid X)$ nepotpuna log-izglednost.

$$\ln L(\theta \mid D) = \ln p(X \mid \theta) = \ln \sum_{Z} p(X, Z \mid \theta) \tag{1.10}$$

$$\ln L(\theta \mid X, Z) = \ln p(X, Z \mid \theta) \tag{1.11}$$

U prvome slučaju optimizacija je teška, dok u drugome djeluje izravno na gustoću i lagano se izvodi.
Kod algoritma maksimizacije očekivanja ne radimo izravno s potpunom log-izglednošću već s očekivanjem potpune log-izglednosti, $E\left[\ln L\left(\theta \mid X, Z\right)\right]$. Ideja je iterativno podešavati parametre θ u svrhu maksimiziranja očekivanja. Time dolazi i do povećanja nepotpune log-izglednosti što je ustvari potrebno.
Maksimizacija očekivanja $E\left[\ln L\left(\theta \mid X, Z\right)\right]$ postiže se provođenjem E-koraka i M-koraka. U E-koraku se računa očekivanje potpune log-izglednosti uz fiksirane trenutne vrijednosti parametara $\theta^{(t)}$. Označimo to očekivanje sa:

$$Q(\theta \mid \theta^{(t)}) = E_{Z|X,\theta^{(t)}}\left[\ln L\left(\theta \mid X, Z\right)\right] = E_{Z|X,\theta^{(t)}}\left[\ln p(X, Z \mid \theta)\right] = \sum_{Z} P(Z \mid X, \theta^{(t)}) \ln p(X, Z \mid \theta) \tag{1.12}$$

Vjerojatnost $P(Z \mid X, \theta^{(t)})$ je aposteriorna vjerojatnost latentne varijable uz trenutne vrijednosti parametara. U izrazu su nam slobodni jedino parametri θ, a to su upravo parametri koje moramo optimizirati. U M-koraku, koraku maksimizacije, odabiremo nove parametre $\theta^{(t+1)}$:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)}) \tag{1.13}$$

Slijed algoritma: Krenemo od nekih početno odabranih parametara $\theta^{(0)}$, zatim se E-korak i M-korak izmjenjuju sve dok $\theta^{(t)}$ ne konvergira. Konvergencija je zagarantirana jer se u svakoj iteraciji povećava očekivanje izglednosti. Važno je odabrati dobre početne vrijednosti parametara jer algoritam ne nalazi nužno globalni optimum. Problem se rješava tako da se početne vrijednosti dobiju algoritmom k-srednjih vrijednosti, odnosno grupiranje se u prvih nekoliko koraka provodi algoritmom k-srednjih vrijednosti, te se tako dobivene vrijednosti koriste kao početne vrijednosti EM-algoritma.

1.7 Provjera grupa

U nadziranoj klasifikaciji, provjera dobivenih modela klasifikacije je sastavni dio razvoja modela klasifikacije. Zbog prirode grupiranja, provjera grupa nije dobro razvijena ili dio analize čije je korištenje uobičajeno. Provjera grupa (validacija grupa) je bitan dio analize grupa.

Analiza grupa je često izvedena kao dio istraživačke analize podataka, stoga izgleda da je nepotrebna provjera nečeg što bi trebao biti informativan proces. Ipak, analiza grupa bi trebala biti dio svake analize grupa, najviše zato što će gotovo svaki algoritam grupiranja naći grupe unutar skupa podataka, čak i kada skup podataka nema prirodnu strukturu grupa.

Jedan od glavnih problema kod grupiranja je odabir broja grupa, odnosno parametra K, što je potrebno odrediti unaprijed. U idealnom slučaju, broj grupa odgovara broju „prirodnih“ grupa u skupu podataka. Odabir optimalnog parametra K znači odabir optimalne složenosti modela.

Poteškoća kod validacije grupa jest to što primjeri nisu označeni, pa nije moguće koristiti metodu unakrsne provjere. Postoje razni načini na koje se može napraviti provjera grupa: Broj grupa je poznat u nekim primjenama, ili se pak određuje primjenom neke od tehnika redukcije dimenzionalnosti. Zatim se podaci prikazuju dvodimenzionalno te se pokuša pronaći prirodan broj grupa.

Treći način jest da broj grupa povećavamo sve dok kriterijska funkcija ne padne ispod vrijednosti koju smo odredili kao maksimalno dozvoljenu.

Nekada je moguće ručno provjeriti rezultat grupiranja i zaključiti ima li unaprijed zadan parametar K smisla. Označimo manji podskup primjera, zajedno grupiramo označene i neoznačene, zatim provjeru radimo na označenim primjerima. Za mjeru pogreške može se koristiti Randov indeks (*rand index*), mjera normalizirane uzajamne informacije (NMI) ili

mjera F_1. Mjera F_1 predstavlja vanjski kriterij, a kriterij koji se koristi za grupiranje predstavlja unutarnji kriterij. Pretpostavka je da su unutarnji i vanjski kriterij dobro usklađeni. Mjera F_1 je harmonijska sredina preciznosti i odziva. Randov indeks mjeri sličnost dvije grupe i povezan je sa točnošću. Ima vrijednosti između 0 i 1. Što je vrijednost bliža jedinici to su grupe sličnije. NMI je mjera koja dopušta zamjenu kvalitete grupiranja sa brojem grupa. Ako gledamo grafički ovisnost kriterijske funkcije o parametru K, tražimo koljeno krivulje. Porastom vrijednosti parametra K vrijednost kriterijske funkcije pada. Kada K postane dovoljno velik, počinju se stvarati prirodne grupe. Kod hijerarhijskog grupiranja možemo prikazati ovisnost broja grupa o udaljenosti, te na mjestima gdje broj stagnira opet imamo prirodne grupe koje su u podacima. Još jedan način je minimiziranje kriterija koji kombinira kriterijsku funkciju i složenost modela. Opći oblik kriterija jest:

$$K^* = \arg\min_K (J(K) + \lambda K) \tag{1.14}$$

$J(K)$ je vrijednost kriterijske funkcije za model s K grupa, a λ težinski faktor. Što je veća vrijednost faktora λ, naginje se rješenju s manjim brojem grupa. Za $\lambda = 0$ optimalan broj grupa je $K^* = N$. Parametar λ biramo ili na temelju iskustva sa grupiranjem sličnih skupova ili korištenjem nekog teorijskog kriterija, kao što je Akaikeov informacijski kriterij (AIC):

$$AIC = \arg\min_K (-2 \ln L(K) + 2k) \tag{1.15}$$

$-\ln L(K)$ predstavlja negativnu log-izglednost podataka za K grupa, k predstavlja broj parametara modela s K grupa.

1.8 Grupiranje u R-u

Za testiranje podataka i predviđanje akcijske prodaje, koristiti će se programski jezik R. U skladu s time, koristiti će se i za grupiranje podataka pomoću kojeg ćemo dobiti bolje rezultate i kvalitetniju analizu. U R-u je dostupno više unaprijeđenih algoritama za grupiranje podataka, a nalaze se u sklopu biblioteke **cluster**. Ona sadrži veliki izbor metoda za grupiranje podataka, te mogućnosti usporedbi starih metoda sa novima da bi se došlo do saznanja je li novija metoda naprednija ili ne.
Algoritam k-srednjih vrijednosti koji je jedna od najstarijih metoda analize grupa, u R-u je dostupan preko **kmeans** funkcije. Za algoritam k-medoida, konkretno, PAM (partitioning around medoids) imamo funkciju **pam**. Niti kod algoritma k-srednjih vrijednosti, niti kod PAM-a, nemamo garanciju da će struktura koja je otkrivena sa malim brojem grupa biti održana i kada povećamo broj grupa.
Za aglomerativno hijerarhijsko grupiranje u R-u imamo funkciju **hclust**. Također imamo i funkciju **agnes** koja koristi istu tehniku kao **hclust**, no s korištenjem manje kratica kod

ažuriranja matrice udaljenosti. Kada se za računanje udaljenosti koristi metoda srednje vrijednosti, **hclust** uzima samo dva zapažanja ili grupe koje su se nedavno spojile pri ažuriranju matrice udaljenosti. Funkcija **agnes** računa te udaljenosti kao prosjek svih udaljenosti između svih zapažanja u dvije grupe. Te dvije tehnike se prilično slažu kada se za ažuriranje koristi metoda maksimuma ili minimuma. Do primjetnih razlika može doći kada se koristi prosječna udaljenost ili Wardova metoda.
Da bi se dobio uvid o mogućim grupama, dendrogram je glavni grafički alat kod hijerarhijskog grupiranja. Nakon što se napravi analiza grupa pomoću jedne od funkcija **hclust** i **agnes**, dendrogram ćemo vidjeti pomoću funkcije **plot**.
Za računanje matrice udaljenosti postoje dvije funkcije koje se mogu koristiti, **dist** i **daisy**. Funkcija **dist** je u svakoj verziji R-a, a **daisy** je dio **cluster** biblioteke i ima više pogodnosti. U slučaju da su podaci prikazani u različitim mjernim jedinicama, htjeli bismo ih standardizirati prije analize. Funkcija **daisy** to radi automatski, no ne daje potpunu kontrolu. Ako želimo određenu standardizaciju, onda koristimo funkciju **scale**. U tu funkciju unesemo matricu ili strukturu koju želimo standardizirati, te dva opcionalna vektora, **center** i **scale**. Prvi je vektor vrijednosti, jedan za svaki stupac koji želimo standardizirati, a drugi je sličan prvom, ali se koristi da bi podijelio vrijednosti u svakom stupcu. Za **center** možemo uzeti vektor srednjih vrijednosti, a za **scale** vektor standardnih devijacija. Ti se vektori mogu napraviti funkcijom **apply** koja obavlja istu operaciju na svakom retku ili stupcu matrice.
U slučaju da nam matrica udaljenosti treba za nešto izvan analize grupa, potrebno ju je konvertirati u regularnu matricu koristeći funkciju **as.matrix**. To je potrebno jer je u R-u matrica udaljenosti donjetrokutasta.
Funkcija **pam** nudi dodatne dijagnostičke informacije o rješenju grupiranja, a nalazi i zapažanja iz originalnih podataka koja su najbliža centru grupe.

Poglavlje 2

Regresija

Grupiranje je metoda preliminarne analize podataka nakon koje slijedi modeliranje. Jedna od metoda modeliranja koje se koriste u upravljanju odnosom s kupcima i klijentima jest regresijska analiza. Ona je ujedno i najčešće korištena u ekonometriji. Metodama regresijske analize doznajemo kako jedna varijabla ovisi o jednoj ili više drugih varijabli, dakle postoji li uopće povezanost među njima, koja je jakost te veze, te može li se varijabla koju analiziramo prognozirati pomoću opaženih vrijednosti drugih varijabli. Da bismo to mogli, prvo definiramo koja je varijabla zavisna, a koja nezavisna. Regresijskom analizom na matematički način utvrđujemo utjecaj zavisne varijable na nezavisne. Regresijski model kojim je to izraženo jest funkcija kojom se ta zavisnost predstavlja.
Opći oblik modela regresije je:

$$y = f(x_1, x_2, ..., x_k) + \varepsilon$$

ili

$$y = f(x_1, x_2, ..., x_k) * \varepsilon$$

Varijabla y je zavisna varijabla (regresand), a varijable $x_1, x_2, ..., x_k$ su nezavisne (regresorske).
Iz općeg oblika modela regresije vidimo da zavisna varijabla y može biti prikazana kao linearna ili nelinearna funkcija nezavisnih varijabli. Vrijednost zavisne varijable se predviđa, a za predviđanje se koriste nezavisne varijable čije su vrijednosti poznate.
Deterministički dio modela jest funkcija $f(x_1, x_2, ..., x_k)$ kojom je matematički izražena zavisnost varijable y od varijabli $x_1, x_2, ..., x_k$. Varijabla ε je slučajna promjenjiva varijabla koja predstavlja grešku relacije, odnosno odstupanje od zavisnosti, te je stohastički dio modela.
Modele regresije možemo podijeliti na jednostruku i višestruku regresiju, ovisno o broju nezavisnih varijabli koje su unutar modela. Također ih možemo podijeliti na linearne i nelinearne, ovisno o obliku matematičke funkcije kojom je model opisan.

Kod jednostavne regresije, model ima jednu zavisnu i jednu nezavisnu varijablu, dok kod višestruke ima jednu zavisnu i više nezavisnih varijabli.
Nelinearni modeli se koriste kada povezanost pojava ne možemo prikazati na linearan način. Najčešće korišteni modeli su oni koji se transformacijom mogu prevesti u modele linearne regresije, kao na primjer eksponencijalni i logaritamski model.
Općenito najkorištenije metode su linearna i logistička regresija.
Višestruku regresiju koristimo kada imamo veliki broj zapažanja. Broj slučajeva mora biti značajno veći od broja prediktornih varijabli koje koristimo.

2.1 Višestruka linearna regresija

Ako zavisna varijabla y ovisi o više nezavisnih varijabli, a želimo koristiti linearni model u regresiji, model višestruke linearne regresije je:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon = \beta^T x + \varepsilon \tag{2.1}$$

gdje je $\beta = (\beta_0, ..., \beta_k)^T$ vektor stupac nepoznatih parametara, a $x = (x_1, ..., x_k)^T$, vektor stupac nezavisnih varijabli.
U slučaju da analizu želimo napraviti na temelju n opažanja, za svako opažanje imamo linearnu jednadžbu, dakle imamo n linearnih jednadžbi:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{aligned} \tag{2.2}$$

Matrična notacija je pogodnija, stoga stavljamo da je matrica X jednaka:

$$\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Broj stupaca matrice X je jednak broju značajki plus 1, odnosno $m = k + 1$. Matrica X je dimenzije mxn i to je matrica vrijednosti nezavisnih varijabli u n opažanja.
Vektor β je definiran. Sada definirajmo preostale:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

gdje je y vektor stupac opaženih vrijednosti zavisne varijable, a ε vektor stupac slučajnih varijabli.
Model sada možemo zapisati u matričnom obliku:

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

Procjena parametara

Regresijskom analizom procjenjujemo nepoznate parametre (vektor β) i nepoznate varijance σ^2 slučajnih varijabli ε_i. Za svako opažanje, postoji distribucija vjerojatnosti slučajne varijable ε_i i zavisne varijable y_i.
Pretpostavke modela višestruke linearne regresije su da postoji linearna veza između zavisne varijable i nezavisnih varijabli koje smo odabrali, te da su su međusobno nezavisne i normalno su distribuirane s istim parametrima, očekivana vrijednost im je nula, a varijanca σ^2 . Iz toga slijedi da slučajan vektor ε ima n-dimenzionalnu normalnu distribuciju:

$$\varepsilon \sim N(0, \sigma^2 I) \tag{2.4}$$

Još jedna pretpostavka jest da su x_i međusobno nezavisni vektori, te da je X matrica punog ranga.
Ako uvrstimo vektore koje smo dobili procjenom u 2.3, imamo:

$$y = X\hat{\beta} + \hat{\varepsilon} \tag{2.5}$$

$$\hat{\varepsilon} = y - X\hat{\beta} \tag{2.6}$$

pri čemu je $\hat{\varepsilon}$ procjena varijable ε , a nazivamo ga vektor rezidualnih odstupanja.
Za procjenu vektora β koristimo metodu najmanjih kvadrata koja je najčešće korištena. Želja je da dobijemo procjenu koja minimizira sumu kvadrata rezidualnih odstupanja:

$$\min_{\hat{\beta}}(\hat{\varepsilon}^T\hat{\varepsilon}) = \min_{\hat{\beta}}((y - X\hat{\beta})^T(y - X\hat{\beta})) \tag{2.7}$$

Kako su prve parcijalne derivacije jednake nuli u točki u kojoj funkcija dostiže minimum, minimizacija se svodi na rješavanje sustava jednadžbi:

$$\frac{\partial(\hat{\varepsilon}^T\hat{\varepsilon})}{\partial\hat{\beta}_j} = 0 \qquad j = 0, 1, ..., k \tag{2.8}$$

Odakle dobijemo:

$$X^T y = X^T X\hat{\beta} \tag{2.9}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.10}$$

Ako su regresorske varijable nezavisne, matrica $X^T X$ je invertibilna. Iz 2.9 slijedi:

$$X^T (y - X\hat{\beta}) = X^T \hat{\varepsilon} = 0 \tag{2.11}$$

$$\hat{y} = X\hat{\beta} = X(XTX)^{-1} X^T y = hy \tag{2.12}$$

Tada vrijedi:

$$\hat{y} \sim N(y, \sigma^2 h) \tag{2.13}$$

gdje je $h = X(XTX)^{-1}X^T$ matrica koja je simetrična i idempotentna, odnosno $H^2 = H, H^T = H$.
Procjenitelj vektora parametara dobiven metodom najmanjih kvadrata je najbolji linearni nepristrani procjenitelj u slučaju da su ispunjene polazne pretpostavke o modelu. Također, zadovoljava ova svojstva:

- $\mathbb{E}(\hat{\beta}) = \beta$
- $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \equiv \Sigma$
- $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$ (Svaka komponenta ima distribuciju $\hat{\beta} \sim N(\beta_j, \sigma^2 s_{jj})$, gdje je s_{jj} dijagonalni element matrice $(X^T X)^{-1}$.)
- Približan 1-α interval pouzdanosti za β_j je $\hat{\beta}_j \pm t_{\frac{\alpha}{2}} \hat{se}(\hat{\beta}_j)$ gdje je $\hat{se}(\hat{\beta}_j)$ drugi korijen j-tog dijagonalnog elementa matrice $\sigma^2 (X^T X)^{-1}$ i to su standardne pogreške.

Standardne pogreške se mogu tumačiti kao odstupanje procijenjenih vrijednosti od stvarne vrijednosti parametra. Za omjer:

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\hat{se}(\hat{\beta}_j)} \tag{2.14}$$

možemo pokazati da ima t-distribuciju sa $(n - k - 1)$ stupnjeva slobode. Procjena pouzdanog intervala parametra β_j je interval koji uz zadanu vjerojatnost uključuje i stvarnu vrijednost parametra. Interval procjene za omjer t_j jest:

$$P\left\{-t_{\alpha/2} < t_j < t_{\alpha/2}\right\} = 1 - \alpha \tag{2.15}$$

Koeficijent $t_{\alpha/2}$ je koeficijent poudarnosti i odgovarajuća vrijednost t-distribucije s (n-k-1) stupnjeva slobode.
Do procjene intervala pouzdanosti za parametar β_j dođemo tako da umjesto t_j uvrstimo omjer koji predstavlja.

Reziduali su $\hat{\varepsilon} = y - \hat{y}$ i procjenjuju vrijednosti slučajne varijable regresijskog modela (ε), stoga je suma kvadrata reziduala jednaka $RSS = \hat{\varepsilon}^T\hat{\varepsilon} = \sum_{i=1}^{n} \hat{\varepsilon}_i{}^2$, a varijanca reziduala je procijenjena sa:

$$\hat{\sigma}^2 = \frac{RSS}{n-k-1} \tag{2.16}$$

Ako izračunamo pozitivni drugi korijen procijenjene varijance, dobiti ćemo procjenu standardne devijacije regresije. Analiza varijance je potrebna za testiranje regresijskog modela, a jednadžba analize varijance jest:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.17}$$

Ili:

$$TSS = PSS + RSS \tag{2.18}$$

TSS=Total Sum of Squares
PSS=Regression Sum of Squares
RSS=Residual Sum of Squares

Ukupna varijanca od y (TSS) rastavlja se na sumu kvadrata odstupanja regresijskih vrijednosti od prosjeka (PSS), i sumu kvadrata odstupanja regresijskih od opaženih vrijednosti (RSS).
Nezavisne procjene komponenti varijance dobijemo dijeljenjem zbrojeva kvadrata s odgovarajućim stupnjevima slobode. Kod sume kvadrata reziduala, stupnjevi slobode su $n-k-1$ i njenim dijeljenjem stupnjevima slobode smo dobili u 2.16 procijenjenu varijancu regresije.
Kod zbroja kvadrata regresije broj stupnjeva slobode je k. Omjer sredine kvadrata regresije i sredine kvadrata reziduala ima F-distribuciju. F-omjer jest:

$$\frac{PSS/k}{RSS/(n-k-1)} \tag{2.19}$$

Ako je vrijednost F-omjera velika, model koji smo odabrali je statistički značajan. U suprotnom, moramo izabrati drugi model, jer je testirani model statistički neprihvatljiv.
U slučaju da je odabrani model dobar, računamo koeficijent determinacije:

$$R^2 = \frac{PSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{(n-k-1)\hat{\sigma^2}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.20}$$

R^2 je koeficijent determinacije čija je vrijednost između 0 i 1. Što je vrijednost koeficijenta determinacije bliže jedinici, to je model reprezentativniji, odnosno to podaci su više u skladu s odabranim modelom.

Multiplikativni model

Multiplikativni model je jedan od modela koji se svode na linearan zbog lakše procjene. Do jednostavnijeg modela se dođe logaritmiranjem postojećeg.
Multiplikativni model:

$$y = e^{\beta_0} x_1^{\beta_1} x_2^{\beta_2} ... x_k^{\beta_k} \tag{2.21}$$

Logaritmirani model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{2.22}$$

gdje je k broj prediktora, odnosno nezavisnih varijabli.
U ovom radu ćemo koristiti multiplikativan model kod procjene parametara zbog prilagođavanja podataka normalnoj distribuciji.

Pojednostavljivanje višestruke regresije

Da bi se pojednostavnio postojeći model višestruke regresije, od skupa prediktora se biraju najbolji, odnosno najkorisniji. Nakon odabira, dobije se skup prediktora koji su najviše potrebni i bitni u modelu. Mnogi modeli višestruke regresije sadrže varijable čije t-statistike imaju neznačajne p-vrijednosti. Te varijable ne predstavljaju statistički značajnu sposobnost predviđanja u odnosu na ostale prediktore i mogu se izbaciti iz modela.
U slučaju da postoji razlog za isticanje nekog prediktora prije nego što su podaci prikupljeni, kažemo da varijabla ima prediktivnu vrijednost ako dostigne statističku značajnost. No, ako se u obzir uzima više prediktora koji nisu posebno istaknuti prije sakupljanja podataka, te ih promatramo kao da je svaki od njih jedini prediktor koji uzimamo u obzir, neki će dostići statističku značajnost samo zbog slučajnosti.
U bilo kojem uzorku, neke varijable će biti bolji prediktori od drugih. Ako mnogo varijabli predviđa odaziv jednako dobro, to ne znači da su one uistinu bolje, već tako izgledaju zbog slučajnosti. Zbog toga nam je potrebna procedura kojom možemo razlikovati varijable koje su uistinu bolji prediktori od onih koje tako samo izgledaju. Procedura koja to postiže ne postoji, ali jedna od procedura koje su se približile cilju jest **stepwise** procedura.

2.2 Stepwise procedura

Stepwise procedure su: odabir unaprijed, eliminacija unatrag, te stepwise regresija. Te procedure se sastoje od toga da dodaju ili izbacuju jednu po jednu varijablu sve dok nije zadovoljeno neko pravilo zaustavljanja.
Metoda odabirom unaprijed započinje s praznim modelom. Zatim se u model ubacuje varijabla koja ima najmanju p-vrijednost kada je ona jedini prediktor. U svakom sljedećem

koraku dodaju se varijable sa najmanjom p-vrijednosti u odnosu na ostale prediktore koji su već u jednadžbi. Dodaju se jedna po jedna, sve dok je p-vrijednost dovoljno mala. Uobičajena dovoljno mala vrijednost je manja od 0.05 ili 0.10.
Eliminacija unatrag započinje sa svim prediktorima u modelu. Iz modela se izbacuje varijabla sa najvećom p-vrijednosti te se zatim model prilagođava. U svakom sljedećem koraku iz modela se izbacuje varijabla sa najvećom p-vrijednosti sve dok ne dobijemo skup varijabli od kojih svaka ima p-vrijednost manju od neke vrijednosti, 0.05 ili 0.10.
Stepwise regresija je slična odabiru unaprijed, sa razlikom da se varijable izbacuju iz modela ako postanu beznačajne kada se dodaju drugi prediktori.
Prednost eliminacije unatrag jest da postoji mogućnost da skup varijabli ima znatnu predvidljivost iako ju nijedan drugi podskup varijabli nema. Ta predvidljivost se može primjetiti jer metoda eliminacije unatrag počinje sa svim varijablama u modelu. Metodom odabira unaprijed i stepwise regresijom ne možemo prepoznati takve varijable jer same po sebi ne predviđaju dobro, stoga ne ulaze u model. Zbog toga ne možemo primjetiti njihovo zajedničko ponašanje.
Procjene bilo čega vezanog za predviđanje mogu uputiti u krivom smjeru jer odabrane varijable samo izgledaju kao dobri prediktori. Važnost koeficijenata regresije često izgleda veća nego što jest. Tada je t-statistika veća, a standardna pogreška je manja nego što bi bila kada bi se promatrala ponavljana studija, te su intervali pouzdanosti preuski, a pojedinačne p-vrijednosti premale dok je R^2 i prilagođeni R^2 prevelik. Ukupan F omjer je prevelik, a njegova p-vrijednost je premala, što je slučaj i sa standardnom pogreškom procjene.
Kada varijable imaju visok koeficijent korelacije, one koje su u modelu su tamo zbog slučajnosti i mogu se promijeniti ubacivanjem jednog ili više zapažanja. Problem kod višestruke regresije se pojavljuje ako se koeficijenti i p-vrijednosti mogu promijeniti, a intervali pouzdanosti koeficijenata regresije ostanu isti. Kod stepwise procedura mana je što se značajke modela mogu drastično mijenjati zbog ubacivanja i izbacivanja varijabli. Također, kod stepwise procedura može doći do problema kada nedostaju neki podaci. Po proceduri se mora izostaviti zapažanje koje nema u sebi nikakav potencijalni prediktor, a moguće je da će neka od tih zapažanja imati prediktore u konačnom modelu koji su bitni.
Zbog toga što se stepwise regresijom pretražuje velik prostor mogućih modela, često dolazi do pretjerane prilagodbe podataka. Model koji dobijemo tada odgovara puno bolje uzorku na kojem radimo nego što će odgovarati novim podacima koji nisu u uzorku. Da bi se to izbjeglo, potrebno je naći dovoljno strogi kriterij za dodavanje ili izbacivanje varijable.

Poglavlje 3

Provedba istraživanja

U ovom poglavlju prelazimo na rad sa stvarnim podacima koji su prikupljeni u hrvatskom vodećem trgovačkom lancu Konzum. Da bi neka trgovina što uspješnije poslovala, potrebno je konstantno prikupljati podatke o dosadašnjoj prodaji. Time se dolazi do uvida o prodaji pojedinih artikala, te se može predvidjeti što bi moglo u budućnosti donijeti zaradu za nastavak poslovanja, ali i koje bi cijene i artikli mogli zadovoljiti kupce. Potrebe kupaca variraju i nužno je pratiti te trendove. Tako je olakšano predviđanje akcijske prodaje jer prikupljena zapažanja daju odgovore na mnoga pitanja u vezi prodaje, te se time pospješuje rad trgovine. Konzumovih 700 prodavaonica smješteno je po cijeloj Republici Hrvatskoj i kao što je slučaj sa svim velikim lancima koji su rasprostranjeni na više različitih područja, prikupljanje, analiza, modeliranje podataka je važan faktor za daljnji i uspješan rad. Oni artikli koji se dobro prodaju na jednom području, možda se uopće ne prodaju na nekom drugom području, i ta nam informacija govori o potrebama kupaca. Ukoliko ima dovoljno takvih informacija i dobro su iskorištene, moguće je smanjiti gubitke poslovanja, povećati dobit i dalje se razvijati tokom vremena.
Jedan dio razvoja jest analiza, prilagodba i modeliranje podataka koji su dostupni. Što je veće poklapanje stvarne i predviđene akcijske prodaje, to je model bolji i uspjeh veći.

3.1 Modeliranje

Za provedbu istraživanja dobiveni su podaci iz Konzuma. Iz podataka su odmah uklonjena opažanja kod kojih nijedan artikl nije bio prodan. Iz njihovog sažetka zaključujemo da se sastoje od 186 403 opažanja za svaku od varijabli, kojih ima 20. Za modeliranje nećemo koristiti neke od varijabli, kao što su identifikator prodavaonice i identifikator artikla koji imaju mnogo kategorija, te datum početka i kraja akcije. Neke od ostalih varijabli, koje su kategorijske, pretvorili smo u identifikatorske, odnosno *dummy* varijable. To su varijable:

regija (7 kategorija), odjel (prehrana i neprehrana) i format prodavaonice (MALI, MAXI, SUPER).
Umjesto da izaberemo neku identifikatorsku varijablu po kojoj bismo podijelili podatke u grupe na kojima bismo radili modele, odlučili smo se za *clustering* podataka. Pomoću *k-means* algoritma podijelili smo podatke na K=5 grupa. Za varijable po kojima ćemo grupirati podatke uzeli smo akcijsku cijenu, relativnu promjenu cijene i akcijski udio u ukupnom prometu. Sve su vezane za akcije i nezavisne su.
Nakon što su utvrđene grupe, moguće je prijeći na istraživanje. Cilj je predvidjeti jednu varijablu na temelju utjecaja više varijabli iz preostalog skupa. Za metodu istraživanja izabrali smo višestruku regresiju.
U ovom slučaju, zavisna varijabla je akcijski promet, što je stvarni broj prodanih artikala tokom akcije. Nezavisnih varijabli kojima se predviđa akcijski promet ima 18. One sadrže razne informacije, kao što je promet artikla izvan akcije, način oglašavanja, koliko je trajala akcija, u kojoj regiji je ostvaren određeni promet, o kojoj se vrsti proizvoda radi, itd.
Kada smo se odlučili za metodu istraživanja i na kakvim podskupovima će se vršiti, ono može početi. Višestruku regresiju ćemo primijeniti na cijelom skupu podataka vezanih za varijable koje koristimo, te za svaku od pet grupa dobivenih *k-means* algoritmom.
Za početak, moramo provjeriti je li zavisna varijabla normalno distribuirana. Ukoliko nije, koristimo se nekom transformacijom varijabli, koju sami izaberemo. Kada dobijemo normalno distribuirane podatke, možemo razne testove vršiti. Novi skup podataka je dobijen tako što su iz početnog izbačene neke varijable iz daljnjeg razmatranja, neke su zamijenjene transformiranim varijablama. Neke od njih su logaritmirane, a krajnji skup ima 18 nezavisnih varijabli.
Kada dođemo do krajnjih modela, moramo izračunati pogreške predviđanja da bismo vidjeli je li dobro predviđaju akcijsku prodaju. Za to ćemo koristiti WMAPE (*Weighted Mean Absolute Percentage Error*), mjeru koja dodaje težinu u izračun da ne bi došlo do iskrivljenosti razultata zbog malih volumena.

$$WMAPE = \frac{\sum_{i=1}^{n} |\frac{\hat{y}_i - y_i}{y_i}| y_i}{\sum_{i=1}^{n} y_i} \tag{3.1}$$

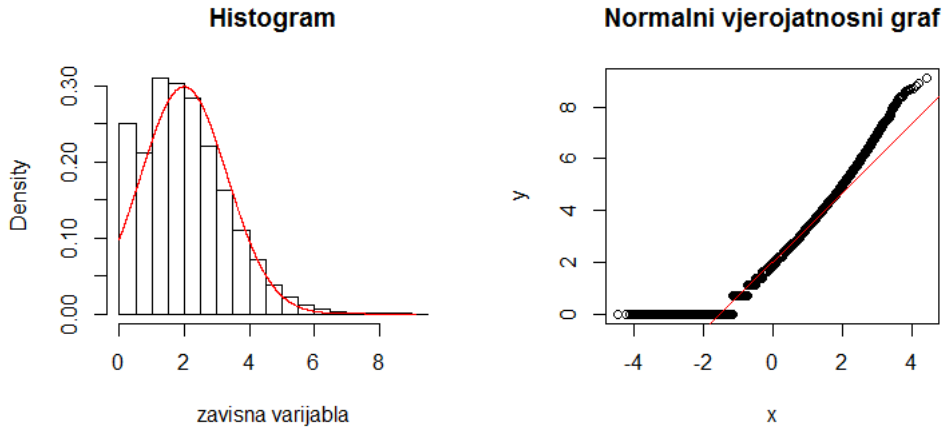
gdje je n broj opažanja.

3.2 Procjena na cijelom skupu

Prije nego što krenemo na modeliranje i procjenu, trebamo se upoznati sa svojstvima zavisne varijable. Da bi se mogle testirati razne hipoteze, potrebna je normalna distribuiranost varijable. Histogram i normalni vjerojatnosni graf su pokazali da zavisna varijabla ne prati normalnu distribuciju koliko bismo htjeli. Stoga slijedi transformacija. Nakon

što je zavisna varijabla logaritmirana, vidimo da je došlo do poboljšanja (slika 3.1). Od ostalih varijabli one koje nisu *dummy* također transformiramo da bismo mogli nastaviti sa procjenom. Zbog potrebe za transformacijama u procjeni se koristi log-linearni model. Za stvarne podatke koristimo multiplikativan model koji logaritmiranjem postaje linearan. Na novom modelu ćemo koristiti višestruku linearnu regresiju.

Da ne bi došlo do problema u računanju, varijable koje logaritmiramo, a imaju nule,



Slika 3.1: Histogram i normalni vjerojatnosni graf zavisne varijable

povećavamo za 1. Neke varijable koje nisu potrebne za predviđanje su izbačene iz daljnjeg razmatranja, a neke su zamijenjene novima s kojima je lakše raditi.

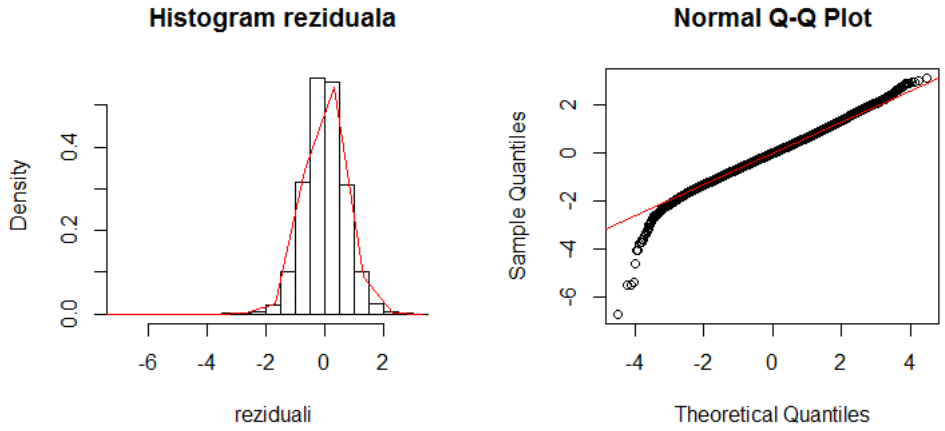
Konačan skup na kojem se radi analiza ima 19 varijabli, svaka varijabla ima 186 403 opažanja. Podijelimo ga na skup za učenje i skup za validaciju, a kriterij je datum početka akcije koji smo sami odredili da ćemo uzeti za granicu. Podaci prije tog datuma su oni na kojima dolazimo do najmanjeg adekvatnog modela, a na podacima nakon toga testiramo model.

Počinjemo od punog linearnog modela u kojem imamo jednu zavisnu varijablu, a prediktori su sve ostale varijable. Sažetkom dobijemo uvid u značajnost pojedine nezavisne varijable u modelu i dolazimo do zaključka da samo jedna varijabla ima jako malu značajnost, stoga je izbacujemo modela. U novom modelu su sve varijable jako značajne stoga smo došli do najmanjeg adekvatnog modela. Način na koji smo došli do njega jest procedura eliminacije unatrag (*stepwise backward*).

Vrijednosti R^2 i prilagođenog R^2 su jednake i iznose 0.67, što znači da se 67% varijance zavisne varijable može objasniti odabranim modelom. Vrijednosti su bile iste i kod punog modela, stoga nema razlike u prilagodbi modela podacima.

Pogreška modela je izračunata pomoću mjere WMAPE i iznosi 60.4%. Da bismo to

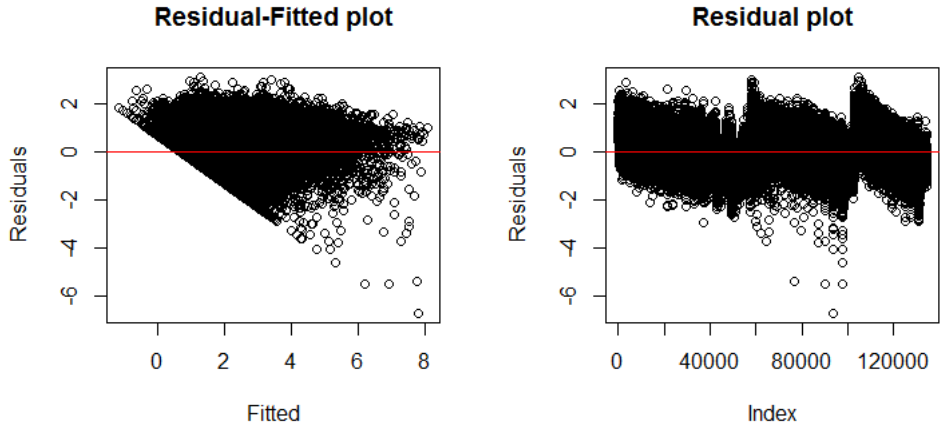
izračunali, model smo testirali na validacijskom skupu pomoću funkcije *predict()* koja za parametre prima odabrani model i validacijski skup podataka. Time dobijemo predviđene vrijednosti zavisne varijable za taj skup, nakon čega moramo na njih primjeniti funkcije inverzne funkcijama kojima smo transformirali zavisnu varijablu. Za linearnu regresiju moraju biti zadovoljene i pretpostavke normalne distribucije pogreške, nezavisnosti prediktora, nezavisnost pogrešaka i konstantnost varijance. Prediktori su nezavisni jer je matrica punog modela također i punog ranga. Iz slika 3.2 i 3.2 možemo zaključiti da su pogreške normalno distribuirane, ali da razlike u varijancama ima.



Slika 3.2: Histogram i normalni vjerojatnosni graf reziduala

3.3 Procjena na grupama

Ono što prethodi procjeni na grupama jest samo grupiranje podataka. To smo napravili pomoću *k-means* algoritma implementiranog u funkciji R-a koja prima podatke i broj koji predstavlja broj grupa na koje želimo podijeliti podatke. Za grupiranje se koristi samo nekoliko varijabli koje opisuju zavisnu varijablu. To su ujedno i varijable koje su direktno vezane za zavisnu.
U ovom slučaju smo uzeli tri varijable po kojima ćemo grupirati i koje će te grupe predstavljati. Kako varijable nisu u istoj skali, potrebno ih je transformirati. Ovdje su najprije normirane, povećane za 1 i zatim logaritmirane. Time izbjegavamo probleme sa računanjima i možemo doći do boljih rezultata. Još nešto što poboljšava rezultat jest čišćenje podataka

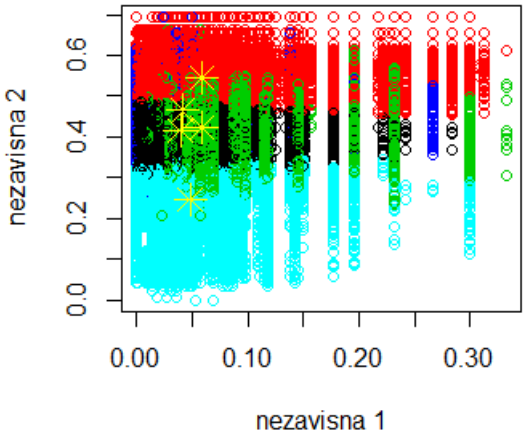


Slika 3.3: Grafovi raspršenosti

od outliera.
Nakon što je primjenjen algoritam, dobiveno je pet grupa (slika 3.3). U rezultatima se mogu vidjeti centri svih grupa po svakoj varijabli koji su njihovi predstavnici. Također smo dobili omjer sume kvadrata među grupama i cjelokupne sume kvadrata koji iznosi 66.5% što upućuje na to da bi za veći broj grupa mogli dobiti bolji omjer i time bolji model.

Odredili smo datum do kojeg ćemo uzeti podatke za modeliranje, a nakon toga je skup za validaciju koji je manji. Na njemu ćemo testirati model koji dobijemo. Imamo grupe, stoga možemo prijeći na regresiju svake od njih pet. Ponovno kao i na cijelom skupu podataka, počinjemo s multiplikativnim modelom, od kojeg prelazimo na linearni pomoću logaritmiranja.
Na slici 3.3 možemo primjetiti da za neke grupe postoje odstupanja distribucije zavisne varijable od normalne, ali je došlo do poboljšanja u odnosu na početne podatke i uglavnom su normalno distribuirane.

Sada za svaku grupu tražimo najmanji adekvatan model. Počinjemo od punog modela, a zatim izbacujemo varijable, jednu po jednu, koje nisu značajne za model. Grupe su različite, stoga se razlikuju i varijable koje su izbačene. Broj prediktora varira od 15 do 17. Ujedno i R^2 i prilagođeni R^2 variraju od grupe do grupe, ali se ne mijenjaju pri prelasku u manji model.
Kada je dobiven najmanji adekvatan model za svaku grupu, potrebno je doći do pogrešaka. Dakle, procijenjene β koeficijente moramo primijeniti na skupu za validaciju. U R-u je to pojednostavljeno funkcijom *predict()* kojom dobijemo predviđene vrijednosti zavisne va-



Slika 3.4: Podjela podataka po grupama i njihovi centri

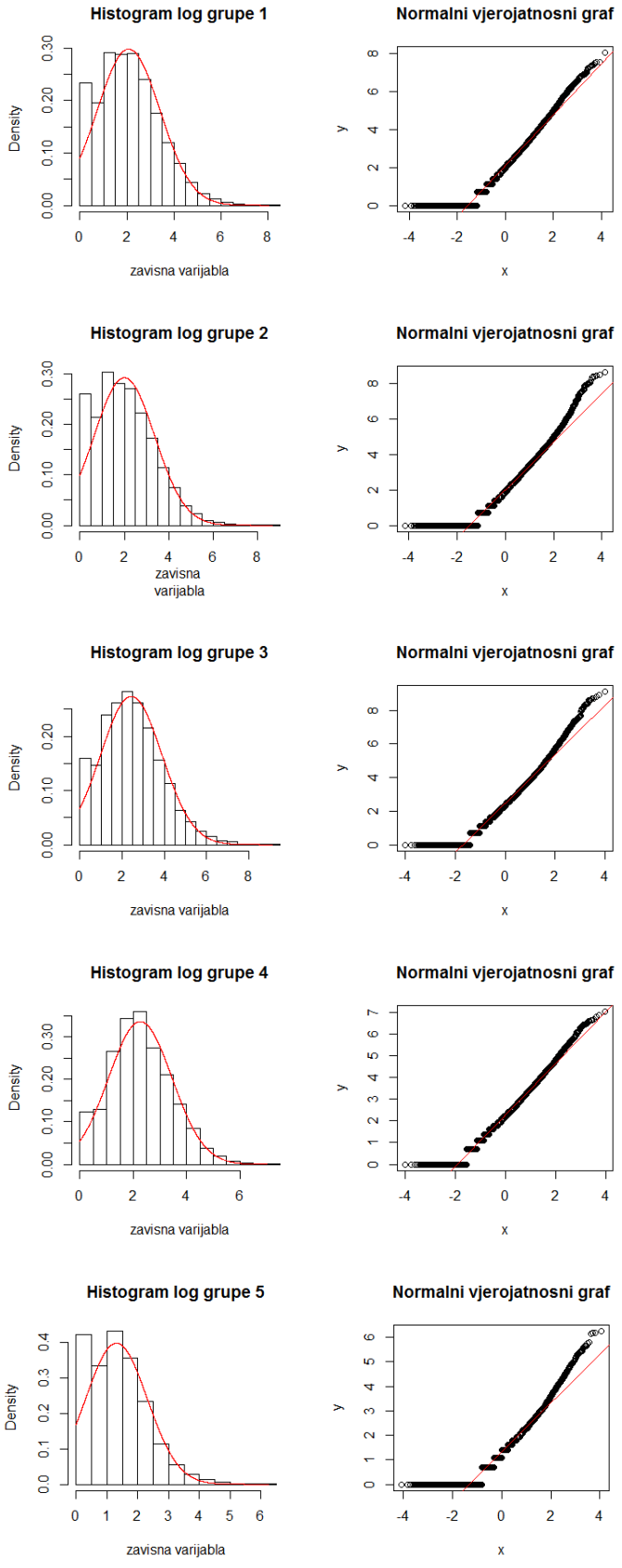
rijable za taj skup. Nakon toga predviđene vrijednosti vraćamo u početnu skalu zavisne varijable kao kod procjene na cijelom skupu. Nove predviđene vrijednosti zavisne varijable za sve grupe spojimo u jedan vektor i izračunamo pogreške na cijelom validacijskom skupu. Pogrešku ponovno računamo pomoću mjere WMAPE, a iznosi 64.6%.
Na slici 3.3 su histogrami i normalni vjerojatnosni grafovi reziduala iz kojih vidimo da dobro prate normalnu distribuciju.

Iz grafova raspršenosti (slika 3.3) vidimo da kao i na cijelom skupu tako i na grupama ima razlike u varijancama. Ima sličnosti u načinu raspršenosti, ali postoji vidljiva razlika između grupa.

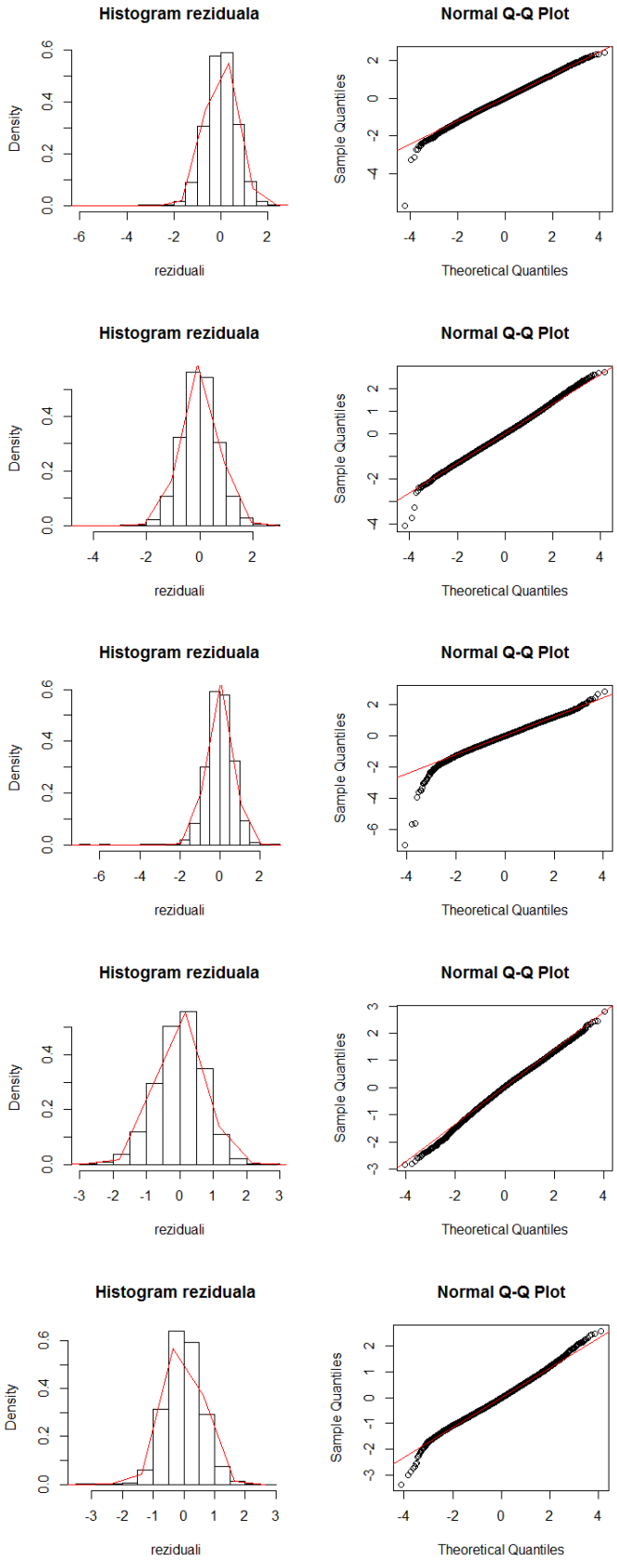
	Grupa 1	Grupa 2	Grupa 3	Grupa 4	Grupa 5	Svi podaci
Veličina	56925	45561	26967	24841	32048	186 403
R2	0.7097	0.6571	0.7679	0.5554	0.4813	0.6518
AR2	0.7096	0.6749	0.7677	0.555	0.4809	0.6517
WMAPE	0.6398811					0.6041209

Tablica 3.1: Značajke modela

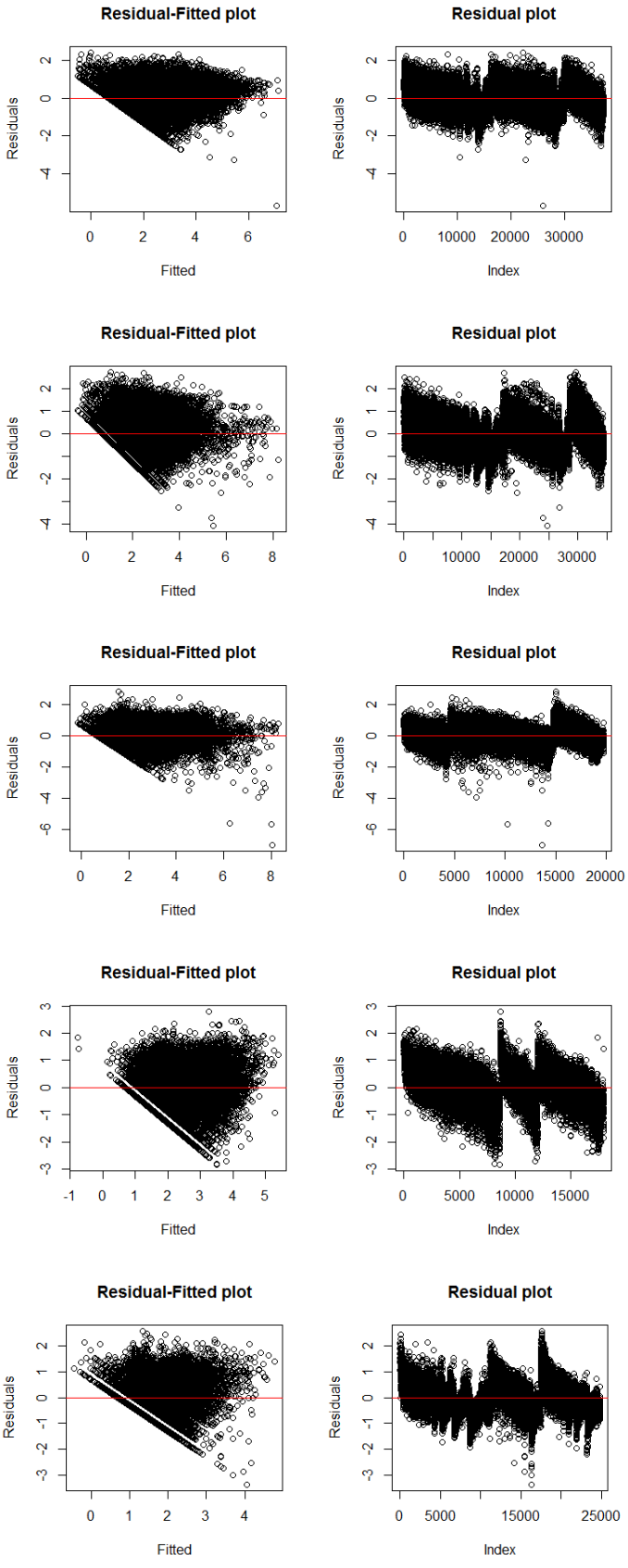
Iz tablice 3.3 vidimo da su za modele grupa R^2 i prilagođeni R^2 variraju. Za neke grupe su lošiji nego kod modela sa svim podacima, dok su za neke bolji. WMAPE je veći kod grupiranih podataka, stoga je to i lošiji model.



Slika 3.5: Histogram i normalni vjerojatnosni graf zavisne varijable



Slika 3.6: Histogram i normalni vjerojatnosni graf reziduala



Slika 3.7: Grafovi raspršenosti

Poglavlje 4

Zaključak

U ovom radu teorijski je obrađena tema grupiranja podataka i višestruke linearne regresije, nakon čega je provedeno istraživanje na stvarnim podacima dobivenih od suradnika iz Konzuma. Istraživanjem smo htjeli doznati je li predviđanje akcijske prodaje moguće i poboljšano kada vršimo grupiranje podataka u odnosu na predviđanje na cijelom skupu. Kada bi se uspostavilo da jest bolje, novim modelom bi bilo moguće povećati uspješnost prodaje.

Tokom istraživanja vršena je linearna regresija na cijelom skupu, te na grupiranim podacima. Zavisna varijabla čije vrijednosti želimo predvidjeti jest akcijski promet. Na cijelom skupu smo počeli sa punim modelom i doznali da je od 18 nezavisnih varijabli, 17 njih značajno. Grupe smo dobili *k-means* algoritmom i to za $k = 5$ čime smo dobili pet grupa. Za to smo koristili tri nezavisne varijable vezane za akcijsku prodaju. Nakon toga smo radili regresiju na svakoj od tih pet grupa opet počevši od punog modela da bismo došli do najmanjeg adekvatnog modela. Broj nezavisnih varijabli u krajnjim modelima varira od 15 do 17.

Pomoću predviđenih vrijednosti zavisne varijable izračunali smo pogreške modela. Za mjeru smo koristili WMAPE koja iznosi 64% za grupirane podatke i 60% za negrupirane podatke.

Cilj je bio provjeriti bi li grupiranje podataka smanjilo pogrešku predviđanja i dobili smo negativan odgovor. Greška predviđanja je veća kod grupiranog skupa i model je nesigurniji. To ne znači da bismo grupiranje trebali odbaciti kao opciju, već ga iskušati sa većim brojem grupa ili drugačijim modelima. Ukoliko se dođe do boljih rezultata, dosad korišteni model se zamijeni boljim. Novi model osigurava bolju upoznatost ureda marketinga sa prilikama i potrebama kupaca. Na temelju toga se formira i marketinška kampanja. Dobra procjena i kampanja za rezultat imaju i dobru prodaju.

Bibliografija

[1] http://statmethods.net/.

[2] http://www.gardenersown.co.uk/Education/Lectures/R/regression.htm.

[3] http://hr.wikipedia.org/wiki/.

[4] http://cran.r-project.org/.

[5] Bojana Dalbelo Bašić i Jan Šnajder, *Bilješka 5: Grupiranje podataka*, Strojno učenje, predavanja, FER, Zagreb (2011).

[6] Kenneth Benoit, *Linear regression models with logarithmic transformations*, London School of Economics, London (2011).

[7] Dominique M. Hanssens i Leonard J. Parsons, *Handbooks in Operations Research and Management Science*, Elsevier, 1993.

[8] Dominique M. Hanssens, Leonard J. Parsons i Randall L. Schultz, *Market response models: Econometric and Time Series Analysis*, Kluwer Academic Pub, 2003.

[9] Richard M. Heiberger i Burt Holland, *Statistical Analysis and Data Display*, Springer, 2004.

[10] Miljenko Huzak, *Vjerojatnost i matematička statistika*, predavanja, PMF-MO, Zagreb (2006).

[11] Anil K. Jain, M. Narasimha Murty i Patrick J. Flynn, *Data clustering: a review*, ACM computing surveys (CSUR) (1999).

[12] Vipin Kumar, Michael Steinbach i Pang Ning Tan, *Cluster analysis: basic concepts and algorithms*, 2006.

[13] Yanchang Zhao, *R and Data Mining: Examples and Case Studies*, Elsevier, 2012.

Sažetak

Cilj ovog istraživanja bio je upoznati se s koristima grupiranja podataka u razvitku matematičkog modela za predviđanje akcijske prodaje, te razvitak tog modela. Za grupiranje podataka korišten je *k-means* algoritam, a metodom višestruke linearne regresije vršene su procjene. Regresijom je izrađen model uz korištenje stvarnih podataka dobivenih od trgovačkog lanca Konzum.
Rezultati su pokazali da grupiranje dobivenih podataka na pet grupa ne smanjuje grešku predviđanja, čak je povećava. Mijenjanjem raznih parametara modeliranja te korištenjem drugih vrsta podataka možda bi se i došlo do boljih rezultata, stoga je potrebno daljnje istraživanje u svrhu poboljšanja postojećeg modela.

Summary

The aim of this study was to become familiar with the benefits of grouping data in the development of a mathematical model to predict the action of sales, and the development of the model. K-means algorithm was used to group the data, and the method of multiple linear regression was used to perform evaluation. Regression model was created using actual data provided by the retail chain Konzum. The results showed that clustering data into five groups does not reduce the prediction error, but increases. By changing various parameters of modeling and the use of other types of data might achieve better results, so further research is needed to improve the existing model.

Životopis

Rođena sam 11. travnja 1988. godine u Dubrovniku. U gradu Korčuli, na istoimenom otoku, završila sam Osnovnu školu Petra Kanavelića. Nakon toga sam se upisala u Srednju školu Petra Šegedina, smjer Opća gimnazija, također u gradu Korčuli. Završila sam srednju školu 2006. godine, nakon čega sam u rujnu iste godine upisala preddiplomski studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Preddiplomski studij sam završila 2010. godine. Tada sam upisala diplomski studij Matematička statistika na istom fakultetu.
Tokom osnovne i srednje škole sam stekla znanje engleskog i njemačkog jezika, rada u Microsoft Office i programskom jeziku Pascal. Za vrijeme fakultetskog obrazovanja proširila se lista poznavanja programskih jezika. Koristila sam: C++, Python, SAS, Matlab, Mathematica, LaTex, SQL i R.