SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Martina Škrnjug

PRIMJENA STATISTIČKIH METODA NA ANALIZU UTJECAJA DEMOGRAFSKIH VARIJABLI U MALOPRODAJI

Diplomski rad

Voditelj rada:
prof.dr.sc.Siniša Slijepčević

Zagreb, veljača 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. _____

Hvala mojoj obitelji i prijateljima na potpori i ohrabrenjima.

Najveće hvala mojoj mami, tati, seki i mom Ferdi na beskrajnom strpljenju, ljubavi i podršci koju mi neumorno pružaju. Bez vas ništa ne bi vrijedilo.

Sadržaj

Uvod

U današnje je vrijeme normalna i očekivana snažna konkurencija u svim oblicima poslovanja. Da bi poduzeća uspjevala ostvariti rast i razvijati se nije dovoljno samo prodavati proizvode, već je nužno praćenje tržišta i prilagođavanje njegovom kretanju. Matematika ima bitnu ulogu u tom procesu. Zahvaljujući raznim mogućnostima praćenja potrošača, moguće je zbilježiti njegovu potražnju. Prikupljeni podaci kasnije se mogu iskoristiti kako bi se implementirao model koji opisuje potrošaćevu potražnju u ovisnosti o raznim čimbenicima. U ovom radu probat ćemo opisati ovisnost godišnje prodaje određenog proizvoda o demografskim čimbenicima. U tu svrhu poslužit će nam model višestruke linearne regresije koji ćemo prethodno teorijski obraditi.

Prvo poglavlje započinjemo oblikovanjem modela jednostavne linearne regresije na što kasnije nastavljamo definiranje višestruke linearne regresije. Predstavljena je procjena parametara modela te njena geometrijska interpretacija, kao i bitni rezultati vezani uz linearnu regersiju.

Drugo poglavlje opisuje metode odabira podskupa varijabli višestrukog modela linearne regresije. Metode koje su opisane su odabir najboljeg podskupa, stepwise unaprijed i unatrag te stagewise unaprijed regresija.

U trećem poglavlju pozornost je posvećena metodama sažimanja linearnih regresijskih modela i to ridge i lasso regresiji.

Zadnje poglavlje sadrži primjenu linearne regresije na demografskim podacima. Pri tome je provedena opisna statistika, izrada modela, provjera ispravnosti njegova korištenja te analiza dobivenih rezultata.

Na kraju rada nalazi se dodatak s pripadnim kodom koji je korišten za analizu podataka i modeliranje, a napravljen je u programskom jeziku R.

Poglavlje 1

Modeli linearne regresije

Svrha linearne regresije je izrada modela kojim opisujemo podatke. Regresijskom analizom doznajemo postoji li povezanost između varijabli te ukoliko postoji, saznajemo na koji način jedna varijabla ovisi o drugoj varijabli ili više njih. Pomoću regresijske analize ključno je utvrditi može li se promatrana varijabla procjeniti pomoću opaženih vrijednosti drugih varijabli. Model linearne regresije pretpostavlja da je regresijska funkcija linearna u varijablama $x_1, \ldots, x_p$. Iako su opsežno razvijani u predkompjutersko doba statistike, i danas postoje snažni razlozi za proučavanje i upotrebu modela linearne regresije. Jednostavni su i često pružaju adekvatni i lako objašnjiv opis kako ulazne varijable utječu na izlaznu. Na poslijetku, mogu se primjeniti na transformirane ulazne podatke što značajno proširuje njihovo područje primjene.

1.1 Oblikovanje modela

Pretpostavimo da imamo niz od n sparenih mjerenja. Označavamo ih s $(x_i, y_i), i = 1, 2, \ldots, n$. Pritom su $x_i, i = 1, 2, \ldots, n$ vrijednosti nezavisne varijable, a $y_i, i = 1, 2, \ldots, n$ su odgovarajuće vrijednosti zavisne varijable. Nezavisna varijabla je neslučajna i ona se uobičajeno zadaje, a zavisna se varijabla opaža, to jest, mjeri. Cilj nam je utvrditi utjecaj nezavisne varijable na zavisnu. To je izraženo regresijskim modelom, odnosno funkcijom koja predstavlja zavisnost. Opći oblik regresijskog modela je:

$$y = f(x) + \varepsilon,$$

pri ćemu je ε Gaussova slučajna varijabla s očekivanjem nula i varijancom σ^2 koja označava odstupanja od zavisnosti koja nisu opisana modelom. Nazivamo ju *slučajna greška* i pišemo $\varepsilon \sim N(0, \sigma^2)$. U ovisnosti o obliku matematičke fukcije f kojom je model opisan, regresijski model može biti linearan ukoliko je f linearna funkcija i nelinearan ukoliko f

nije linearna funkcija. Mi ćemo se baviti linearnim modelima tako da od sada na dalje pretpostavljamo da je funkcija f linearna. Dakle, linearni regresijski model ima oblik:

$$y = \alpha x + \beta + \varepsilon,$$

pri ćemu su α i β nepoznati parametri modela koji se trebaju procijeniti, a x, y i ε su redom nezavisna varijabla, zavisna varijabla i slučajna greška, kako je gore navedeno. Opisani model ima jednu nezavisnu i jednu zavisnu varijablu te ga zbog toga nazivamo *model jednostavne* ili *jednostruke linearne regresije*.

Najčešće korišteni modeli su oni sa više od jedne nezavisne varijable budući da se rijetko koja pojava može opisati samo jednim utjecajem. Takav model nazivamo *model višestruke linearne regresije*. Njime ćemo se baviti u ostatku rada. Dakle, zavisna varijabla y ovisi o više nezavisnih varijabli $x_1, \ldots, x_p$ te koristimo linearni model regresije. Tada model višestruke linearne regresije kojim opisujemo podatke ima oblik:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon.$$

Ukoliko s $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ označimo vektor stupac nepoznatih parametara i s $x = (x_1, x_2, \ldots, x_p)^T$ vektor stupac nezavisnih varijabli, model možemo zapisati kao:

$$y = \beta^T x + \varepsilon. \tag{1.1}$$

Gotovo uvijek analizu neke pojave radimo pomoću više od jednog mjerenja ili opažanja. Kao i u slučaju jednostruke linearne regresije, označimo s n broj mjerenja. Tada je $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i), i = 1, 2, \ldots, n$ slučajni uzorak iz linearnog regresijskog modela kojeg opisuje n linearnih jednadžbi:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_p x_{np} + \varepsilon_n. \end{aligned}$$

Radi jednostavnosti koristimo matrični prikaz pa zbog toga redom označavamo:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

X je matrica dimenzije $n \times (p+1)$ u kojoj su u svakom retku vrijednosti nezavisnih varijabli za pojedino od n mjerenja sa jedinicom na prvoj poziciji. Pretpostavlja se da je $n \geqslant p+1$. Slično ćemo sa Y označiti vektor stupac duljine n koji sadrži opažene vrijednosti zavisne varijable te sa ε vektor stupac slučajnih grešaka koji je iste dimenzije n:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Vektorski zapis modela sada postaje:

$$Y = X\beta + \varepsilon \tag{1.2}$$

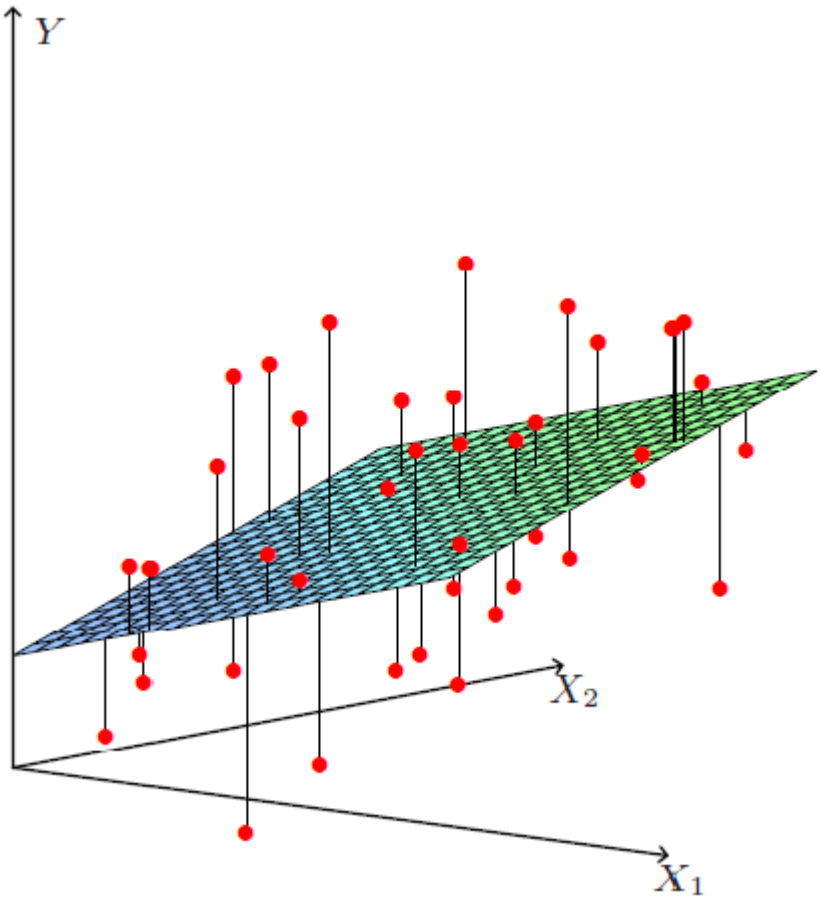
1.2 Procjena parametara modela

Najpoznatija metoda procjene parametara modela linearne regresije je *metoda najmanjih kvadrata*. Kao što joj ime kaže, u njoj izabiremo koeficijente $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ tako da minimiziramo sumu kvadrata reziduala

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{1.3}$$

Definicija 1. $\hat{\beta}$ je *najbolji procjenitelj* metodom najmanjih kvadrata uzorka (x_{ij}, y_i), $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$ ako $\hat{\beta}$ minimizira $RSS(\beta)$.

Sa statističkog stajališta, navedeni kriterij je prihvatljiv ako mjerenje (x_i, y_i) predstavlja nezavisne slučajne ishode iz njegove populacije. Čak ako x_{ij}-ovi, $j = 1, 2, \ldots, p$ nisu izvučeni slučajno, kriterij je i dalje opravdan ako su y_i-ovi uvjetno nezavisni uz dano x_{ij}, $j = 1, 2, \ldots, p$. Prilagodba najmanjih kvadrata intuitivno je zadovoljavajuća bez obzira na to kako nastaju podaci jer taj kriterij mjeri prosječan nedostatak uklapanja. Slika 1.1 prikazuje geometriju smještanja najmanjih kvadrata u $\mathbb{R}^{p+1}$ dimenzionalni prostor razapet parom (x, y).



Slika 1.1: Linearna prilagodba najmanjih kvadrata s $X \in \mathbb{R}^2$. Izvor [1, str. 45]

Budući da za slučajne greške vrijedi $\varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \ldots, n$, tada za vektor slučajnih greški vrijedi $\varepsilon \sim N(0, \sigma^2 I)$. Uvrštavanjem vektora $\hat{\beta}$ i $\hat{\varepsilon}$ koje dobivamo procjenom nepoznatih parametara i slučajnih greški u (1.2), dobivamo:

$$Y = X\hat{\beta} + \hat{\varepsilon},$$

to jest,

$$\hat{\varepsilon} = Y - X\hat{\beta},$$

gdje je $\hat{\varepsilon}$ procjena od ε koju zovemo *vektor reziduala* ili *vektor rezidualnih odstupanja*. U vektorskom zapisu, suma kvadrata reziduala ima oblik:

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta). \tag{1.4}$$

To je kvadratna funkcija koja ima $p + 1$ parametar. Minimiziramo je budući da koristimo metodu najmanjih kvadrata. Njenim deriviranjem po β-i dobivamo jednadžbe

$$\frac{\partial RSS}{\partial \beta} = -2X^T(Y - X\beta) \tag{1.5}$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2X^T X \tag{1.6}$$

Ako pretpostavimo da X ima puni stupčani rang, to jest $r(X) = p+1$, $X^T X$ je tada pozitivno definitna i prvu derivaciju izjednačujemo s nulom

$$X^T(Y - X\beta) = 0. \tag{1.7}$$

Na taj način dobivamo jedinstveno rješenje

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{1.8}$$

Sada imamo procjenjene parametre regresije te možemo izračunati predviđene vrijednosti zavisne varijable y za dani vektor nezavisnih varijabli x_0. Te vrijednosti dane su sa

$$\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}.$$

Pomoću toga računamo procjenu vektora zavisnih varijabli za svih n mjerenja. On je dan s

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y,$$

gdje je $\hat{y}_i = \hat{f}(x_i)$.
Matricu $H = X(X^T X)^{-1} X^T$ koja se pojavljuje u jednadžbi ponekad nazivamo "kapa" matrica jer stavlja kapu na Y. Ona je simetrična i idempotentna, odnosno vrijedi $H^2 = H$ i $H^T = H$. Iz $\hat{Y} = HY$ slijedi $\hat{Y} \sim N(Y, \sigma^2 H)$.

Geometrijska interpretacija procjene najmanjih kvadrata

Sada ćemo prikazati procjenu najmanjih kvadrata na drugačiji, geometrijski način, kao ortogonalnu projekciju vektora $\hat{Y}$ na potprostor razapet stupcima matrice X. U tu ćemo svrhu najprije iskazati nekoliko rezultata o projekciji.

Definicija 2. Neka su $x = (x_1, x_2, \ldots, x_n)$ i $y = (y_1, y_2, \ldots, y_n)$ dva vektora na vektorskom prostoru $\mathbb{R}^n$. Tada njihov *skalarni produkt* označavamo s $\langle x|y\rangle$ i definiramo kao

$$\langle x|y\rangle = \sum_{i=1}^{n} x_i y_i.$$

Napomena. Za $x \in \mathbb{R}^n$ vrijedi $\|x\| = \sqrt{\langle x|x\rangle} = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$, pri čemu je $\|x\|$ norma vektora x.

Definicija 3. Neka je $x \in \mathbb{R}^n$ i neka je M potprostor od $\mathbb{R}^n$. Tada kažemo da je x *ortogonalan* na M i pišemo $x \perp M$ ako vrijedi

$$\langle x|m\rangle = 0, \qquad \forall m \in M.$$

Neka je X konačnodimenzionalan vektorski prostor dimenzije n te neka je M potprostor dimenzije $k \leqslant n$. Tada na više ekvivalentnih načina možemo definirati projekciju $P : X \longrightarrow M$. P će biti linearan operator.

Neka su nam prvo zadani podaci $x_1, x_2, \ldots, x_n$ čiji je vektorski zapis $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$. Definiramo potprostor $M := \{\mu\mathbf{1} : \mu \in \mathbb{R}\} \leqslant \mathbb{R}^n$ gdje je $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^n$ te preslikavanje $P : \mathbb{R}^n \longrightarrow M$,

$$Px := \bar{x}\mathbf{1}.$$

Tada za preslikavanje P vrijede sljedeća svojstva:

1. P je linearni operator,
2. Px je jedinstveno rješenje problema

$$|x - Px|^2 = \min_{y \in M} |x - y|^2 = \min_{\mu \in \mathbb{R}} \sum_{i=1}^n (x_i - \mu)^2,$$

3. $P^2 = P$,
4. $x - Px \perp M$,
5. $x \in M \Leftrightarrow Px = x$.

Neka su nam sada zadani podaci $(x_i, y_i), i = 1, 2, \ldots, n$ čiji je vektorski zapis $x = (x_1, x_2, \ldots, x_n)^T, y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$. Definiramo potprostor $M := \{\alpha\mathbf{1} + \beta x : \alpha, \beta \in \mathbb{R}\} \leqslant \mathbb{R}^n$ te preslikavanje $P : \mathbb{R}^n \longrightarrow M$,

$$Py := \hat{\alpha}\mathbf{1} + \hat{\beta}x,$$

gdje su $\hat{\alpha}, \hat{\beta}$ koeficijenti dobiveni metodom najmanjih kvadrata. Tada za ovako definirano preslikavanje P vrijede sljedeća svojstva:

1. P je linearni operator,

2. Py je jedinstveno rješenje problema

$$|y - Py|^2 = \min_{z \in M} |x - z|^2 = \min_{\alpha,\beta \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2,$$

3. $P^2 = P$,

4. $y - Py \perp M$,

5. $y \in M \Leftrightarrow Py = y$.

Teorem 1.2.1 (Teorem o ortogonalnoj projekciji)**.** *Neka je M potprostor euklidskog prostora $\mathbb{R}^n$ i $x \in \mathbb{R}^n$ proizvoljan vektor. Tada vrijedi:*

1. $(\exists! \hat{x} \in M)|x - \hat{x}| = \inf_{y \in M} |x - y|$

2. $|x - \hat{x}| = \min_{y \in M} |x - y| \qquad \& \qquad \hat{x} \in M \Leftrightarrow x - \hat{x} \in M^{\perp} \qquad \& \qquad \hat{x} \in M.$

Tada je $\hat{x}$ ortogonalna projekcija od x na M.

Napomena. Neka je M potprostor prostora V. *Ortogonalni komplement* potprostora M je tada

$$M^{\perp} = \{x \in V : \langle x, v \rangle = 0, \forall v \in M\}.$$

Ortogonalni komplement $M^{\perp}$ je također potprostor od V.

Korolar 1.2.2. *Ako je M potprostor euklidskog prostora $\mathbb{R}^n$ i I je preslikavanje identiteta na $\mathbb{R}^n$, tada postoji jedinstveno preslikavanje $P_M : \mathbb{R}^n \longrightarrow M$ takvo da je $I - P_M : \mathbb{R}^n \longrightarrow M^{\perp}$. P_M je* projektor *od $\mathbb{R}^n$ na M.*

Projektor P_M na potprostor M od $\mathbb{R}^n$ ima sljedeća svojstva:

1. P_M je linearni operator,

2. $(\forall x \in \mathbb{R}^n)|x|^2 = |P_M x|^2 + |(I - P_M)x|^2$,

3. Svaki $x \in \mathbb{R}^n$ se na jedinstven način može prikazati kao zbroj elemenata iz M i elemenata iz $M^{\perp}$,

4. $(\forall x \in \mathbb{R}^n)x \in M \Leftrightarrow P_M x = x$,

5. $(\forall x \in \mathbb{R}^n)x \in M^{\perp} \Leftrightarrow P_M x = 0$,

6. $M_1 \leqslant M_2 \leqslant \mathbb{R}^n \Leftrightarrow P_{M_1} P_{M_2} = P_{M_1}$,

7. $P_M^2 = P_M$.

Za detalje vidi [4].

Teorem 1.2.3 (Teorem o projekciji)**.** *Neka je X konačnodimenzionalan vektorski prostor te neka je M potprostor od X. Neka je na X definiran skalarni produkt $\langle x|y\rangle$. Tada postoji jedinstvena funkcija $P : X \longrightarrow M$ koja zadovoljava sljedeće ekvivalentne uvjete:*

1. *P_X je vektor koji minimizira izraz $\|X - y\|^2, y \in M$*
2. *$P_X \in M$ koji zadovoljava $(X - P_X) \perp M$*

Tada je P linearni operator.

Označimo sada stupce matrice ulaznih podataka, odnosno nezavisnih varijabli, X redom s $X_0, X_1, \ldots, X_p$, pri čemu je $X_0 \equiv 1$. Vektori $X_0, X_1, \ldots, X_p$ razapinju potprostor od $\mathbb{R}^n$ koji je određen kao prostor stupaca matrice X. Minimiziramo $RSS(\beta) = \|Y - X\beta\|^2$ na način da odabiremo vektor $\hat{\beta}$ takav da vektor reziduala $Y - \hat{Y}$ bude ortogonalan na navedeni potprostor. Ta je ortogonalnost već izražena u (1.7) i zbog toga je rezultirajuća procjena $\hat{Y}$ upravo *ortogonalna projekcija Y* na prostor stupaca matrice X. Kapa matrica H računa ortogonalnu projekciju zbog čega je također znana kao matrica projekcije. Dakle, vrijedi $\hat{Y} = HY$. Zbog toga je vektor reziduala definiran kao $\hat{\varepsilon} = Y - \hat{Y} = MY$, pri čemu je $M = I - H$, za H i M ortogonalne projektore u $\mathbb{R}^n$ takve da vrijedi:

- $I = H + M$
- $r(H) = p + 1$
- $r(M) = n - p - 1$

Zbog navedenih svojstava vektor procjenjenih vrijednosti zavisne varijable $\hat{Y}$ je ortogonalan na vektor reziduala $\hat{\varepsilon}$, to jest, $\hat{Y} \perp \hat{\varepsilon}$. Iz toga slijedi $\hat{\varepsilon} \perp \mathbf{1}, X_1, \ldots, X_p$, gdje su $X_i, i = 1, \ldots, p$ vektori stupci nezavisnih varijabli, a $\mathbf{1} = (1, 1, \ldots, 1)^T$ je vektor stupac koji sadrži n jedinica.

Može se dogoditi da stupci matrice X nisu linearno nezavisni te tada X nije punog ranga. To bi se dogodilo, na primjer, kada bi dvije nezavisne varijable bile savršeno korelirane (npr. $X_1 = 5X_3$). Tada bi X^TX bila singularna matrica i koeficijenti dobiveni metodom najmanjih kvadrata $\hat{\beta}$ ne bi bili jedinstveno određeni. Međutim, procjenjene vrijednosti $\hat{Y} = X\hat{\beta}$ su i u tome slučaju projekcije od Y na prostor stupaca od X. Jedina je razlika da u ovom slučaju postoji više od jednog načina za izraziti tu projekciju u terminima vektora stupaca od X. Slučaj kada rang nije pun naješće se pojavljuje kada je jedna ili više nezavisnih varijabli suvišno uključena u model. Uobičajno postoji prirodan način

za rješavanje navedenog problema prikaza i to putem promjene izgleda modela ili izbacivanja suvišnih stupaca iz matrice X. Većina suvremenih softverskih programa za regresiju otkriva te viškove i automatski implementira strategiju za njihovo uklanjanje.

Distribucija podataka

Ovdje ćemo navesti neke činjenice o distribuciji podataka u modelu. Pretpostavili smo da su opažanja zavisne varijable y_i nekorelirana i imaju konstantnu varijancu σ^2 i da su x_i dani, to jest, nisu slučajni. Iz (1.8) slijedi da je varijanca procjene parametara najmanjih kvadrata jednaka

$$\mathrm{Var}(\hat{\beta}) = (X^T X)^{-1}\sigma^2.$$

Kao što ćemo vidjeti, vrijedi $\mathrm{E}[\hat{\beta}] = \beta$ pa je stoga

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2). \tag{1.9}$$

Iz toga zaključujemo da svaki pojedini procjenjeni parametar ima distribuciju

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j),$$

gdje je v_j j-ti dijagonalni element matrice $(X^T X)^{-1}$.

Reziduali su $\hat{\varepsilon} = Y - \hat{Y}$ i njihovu varijancu procjenjujemo kao

$$\hat{\sigma}^2 = \frac{RSS}{n - p - 1},$$

gdje je $RSS = \hat{\varepsilon}^T\hat{\varepsilon} = \sum_{i=1}^{n} \hat{\varepsilon}_i{}^2$ suma kvadrata reziduala. Dakle,

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

Također,

$$(n - p - 1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p-1}.$$

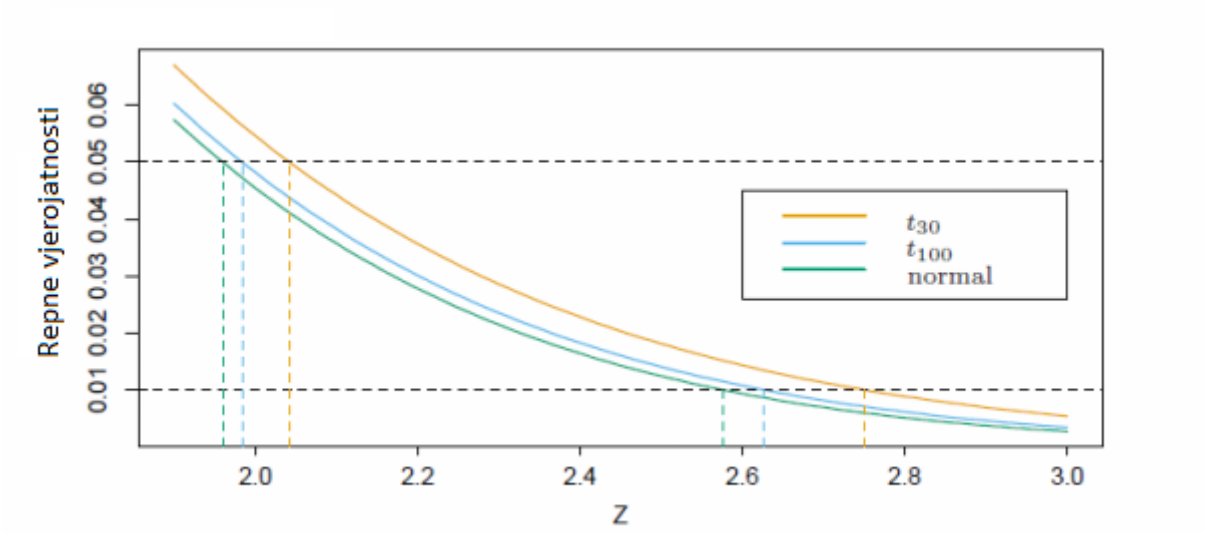
Dodatno, $\hat{\beta}$ i $\hat{\sigma}^2$ su statistički nezavisni.
Iz varijance reziduala dolazimo do kovarijance vektora zavisne varijable koja je jednaka

$$\mathrm{cov}(Y) = \mathrm{E}[(Y - X\beta)(Y - X\beta)^T] = \mathrm{E}[\hat{\varepsilon}\hat{\varepsilon}^T] = \sigma^2 I. \tag{1.10}$$

Ove pretpostavke distribucije naveli smo kako bismo oblikovali test hipoteza i intervala pouzdanosti za parametre β_j, $j = 0, 1, \ldots, p$. U svrhu testiranja hipoteze da je pojedini parametar β_j jednak nuli, tvorimo standardizirani parametar ili *Z-score*

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}},$$

gdje je v_j j-ti dijagonalni element matrice $(X^T X)^{-1}$, kako je već navedeno. Na temelju nulte hipoteze da je $\beta_j = 0$, z_j ima t_{n-p-1} distribuciju i zbog toga velika vrijednost od z_j vodi do odbacivanja navedene nulte hipoteze.
Ako $\hat{\sigma}$ zamijenimo s poznatom vrijednošću σ, z_j će imati standardnu normalnu distribuciju. Kako se veličina uzorka povećava, postaje zanemariva razlika među repnim kvantilima standardne normalne i t-distribucije, što je prikazano na slici 1.2, a zbog čega uobičajeno koristimo kvantile normalne razdiobe. Na slici su istaknuti kvantili za testiranje razina značajnosti $p = 0.01$ i $p = 0.05$, a primjećujemo da za t veći od 100 razlika između kvantila t i standardne normalne distribucije postaje neznatna.



Slika 1.2: Repni kvantili $\mathbb{P}(|Z| > z)$ za t_{30} i t_{100} distribuciju te za standardnu normalnu distribuciju. Izvor [1, str. 48]

Ukoliko trebamo istodobno testirati značajnost grupe koeficijenata, koristimo F statistiku

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1 - 1)},$$

gdje je RSS_1 suma kvadrata reziduala prilagodbe najmanjim kvadratima većeg modela sa p_1+1 parametara i RSS_0 je to isto samo za manji model s p_0+1 parametara, gdje je p_1-p_0 parametara postvaljeno na nulu. F statistika mjeri promjenu u sumi kvadrata reziduala za dodatni parametar u većem modelu i normalizirana je pomoću procjene od σ^2. *Z-score*-ovi z_j su ekvivalentni F statistici za izbacivanje pojedinog parametra β_j iz modela. S Gaussovim pretpostavkama i nultom hipotezom da je manji model ispravan, F statistika će imati $F_{p_1-p_0,n-p_1-1}$ distribuciju.

Procjenimo sada intervale pouzdanosti za parametre β_j. Iz (1.9) proizlazi da je približan $1-\alpha$ interval pouzdanosti za β_j jednak

$$\langle \hat{\beta}_j - t_{\frac{\alpha}{2}}\, \hat{se}(\hat{\beta}_j), \hat{\beta}_j + t_{\frac{\alpha}{2}}\, \hat{se}(\hat{\beta}_j) \rangle, \tag{1.11}$$

gdje je $\hat{se}(\hat{\beta}_j)$ drugi korijen j-tog dijagonalnog elementa matrice $\sigma^2(X^TX)^{-1}$.
Slično možemo dobiti približnu grupu pouzdanosti za cijeli vektor parametara β:

$$C_\beta = \{\beta | (\hat{\beta}-\beta)^T X^T X (\hat{\beta}-\beta) \leqslant \hat{\sigma}^2 \chi^{2(1-\alpha)}_{p+1}\}.$$

Ta grupa pouzdanosti za β generira odgovarajuću grupu pouzdanosti za stvarnu funkciju $f(X) = X^T\beta$, koja je jednaka $\{X^T\beta | \beta \in C_\beta\}$.
Za detalje vidi [1].

Gram-Schmidt postupak za višestruku regresiju

U ovom ćemo odjeljku predstaviti algoritam za izračun koeficijenata višestruke regresije metodom najmanjih kvadrata. U tu svrhu prvo krećemo od jednostruke linearne regresije. Dakle, procjena najmanjim kvadratima i reziduali jednostrukog modela linearne regresije $Y = X\beta + \varepsilon$ su

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

i

$$r_i = y_i - x_i\hat{\beta}.$$

Koristeći vektorske zapise $Y = (y_1, \ldots, y_n)^T$, $X = (x_1, \ldots, x_n)^T$ i definiciju skalarnog produkta, gornje jednadžbe postaju

$$\hat{\beta} = \frac{\langle X, Y\rangle}{\langle X, X\rangle} \tag{1.12}$$

$$r = Y - X\hat{\beta}. \tag{1.13}$$

U nastavku ćemo gornje rezultate primjeniti na višestruku linearnu regresiju. Stoga pretpostavimo da su ulazni podaci $X_1, X_2, \ldots, X_p$, odnosno stupci matrice podataka X, ortogonalni. To znači da vrijedi

$$\langle X_j, X_k \rangle = 0 \qquad \text{za svaki} \qquad j \neq k.$$

Tada su procjenitelji koeficijenata višestruke linearne regresije $\hat{\beta}_j$ pomoću metode najmanjih kvadrata koristeći definiciju (1.12) jednaki odgovarajućim procjeniteljima koeficijenata jednostruke linearne regresije, to jest

$$\hat{\beta}_j = \frac{\langle X_j, Y \rangle}{\langle X_j, X_j \rangle}. \tag{1.14}$$

Dakle, u slučaju ortogonalnih ulaznih podataka, to jest nezvisnih varijabli, te nezavisne varijable nemaju utjecaja na međusobne procjenitelje parametara u modelu.
Ortogonalni ulazni podaci se gotovo nikada ne pojavljuju u opaženim podacima, već u projektiranim istraživanjima u kojima je ortogonalnost sprovedena. Zbog toga se ulazni podaci moraju ortogonalizirati kako bismo iskoristili gornje rezultate.
Pretpostavimo sada da imamo ishodište i jednu nezavisnu varijablu, to jest ulazni podatak, X. Tada koeficijenti od X dobiveni metodom najmanjih kvadrata imaju oblik

$$\hat{\beta} = \frac{\langle X - \bar{X}\mathbf{1}, y \rangle}{\langle X - \bar{X}\mathbf{1}, X - \bar{X}\mathbf{1} \rangle}, \tag{1.15}$$

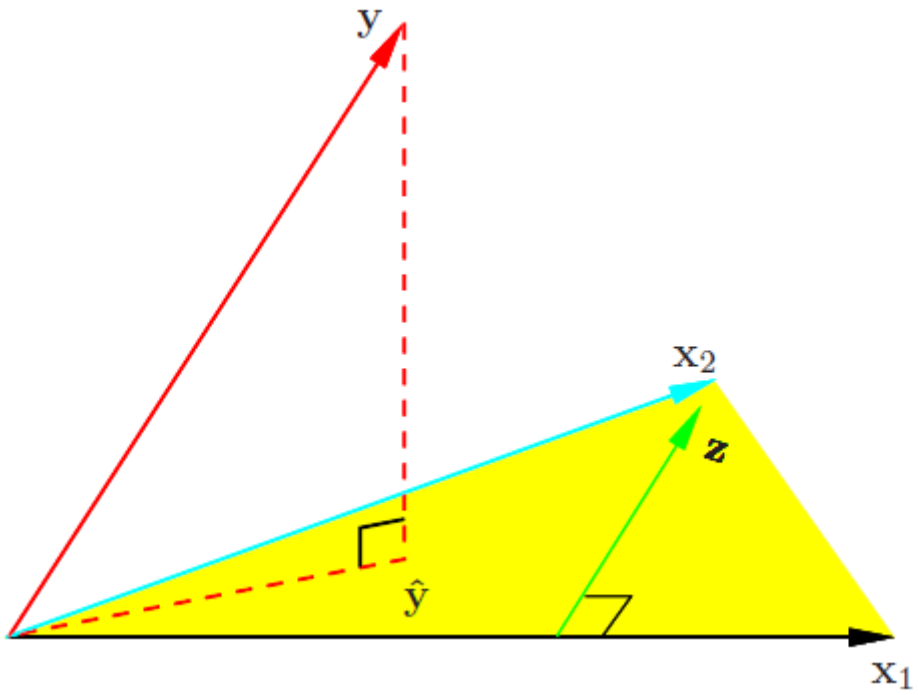
gdje je $\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$ i $\mathbf{1} = X_0$ je vektor s n jedinica.
Procjena (1.15) rezultat je dvije primjene jednostruke regresije (1.14). Prvi je korak u tom postupku primjena jednostruke regresije od X na $\mathbf{1}$ da bi dobili rezidual $Z = X - \bar{X}\mathbf{1}$. Drugi je pak korak primjena jednostruke regresije od Y na rezidual Z kako bi dobili koeficijent $\hat{\beta}$. Na taj je način X, odnosno Y ortogonaliziran s obzirom na $\mathbf{1}$, odnosno Z. Dakle, drugi korak navedenog postupka je jednostruka regresija u kojoj se koriste ortogonalni prediktori $\mathbf{1}$ i Z.
Na slici 1.3 je prikazan navedeni proces za dvije općenite nezavisne varijable X_1 i X_2. Primjećujemo da ortogonalizacija ne mijenja potprostor razapet s X_1 i X_2, već proizvodi ortogonalnu bazu za njihov prikaz.

Budući da želimo dobiti postupak za procjenu koeficijenata višestruke linearne regresije, generaliziramo gornji postupak do slučaja s p nezavisnih varijabli, a to je prikazano u sljedećem algoritmu *regresije uzastopnom ortogonalizacijom*:

1. Inicijaliziraj $Z_0 = X_0 = \mathbf{1}$.

2. Za $j = 1, 2, \ldots, p$ primjeni jednostruku regresiju od X_j na $Z_0, Z_1, \ldots, Z_{j-1}$ kako bi dobio koeficijente $\hat{\gamma}_{lj} = \langle Z_l, X_j \rangle / \langle Z_l, Z_l \rangle$, $l = 0, 1, \ldots, j-1$ i vektor reziduala $Z_j = X_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} Z_k$.

3. Primjeni jednostruku regresiju od Y na rezidual Z_p kako bi dobio procjenu $\hat{\beta}_p$.



Slika 1.3: Regresija metodom najmanjih kvadrata dobivena ortogonalizacijom nezavisnih varijabli. Primjenjena je jednostruka linearna regresija od X_2 na X_1, rezultirajući vektorom reziduala Z. Potom je opet primjenjena jednostruka linearna regresija od Y na Z te je dobiven koeficijent višestruke regresije od X_2. Pribrajajući projekcije Y-na na X_1 i Z dobivamo prilagodbu najmanjih kvadrata $\hat{Y}$. Izvor [1, str. 54]

Rezultat algoritma je

$$\hat{\beta}_p = \frac{\langle Z_p, Y \rangle}{\langle Z_p, Z_p \rangle}. \tag{1.16}$$

Kako su ulazni podaci $Z_0, Z_1, \ldots, Z_{j-1}$ u drugom koraku ortogonalni, koeficijenti koji su ovdje dobiveni jednostrukom regresijom su zapravo koeficijenti višestruke regresije. Algoritam je znan kao *Gram-Schmidt* postupak za višestruku regresiju. To je korisna numerička strategija za računanje procjena pomoću koje možemo dobiti cijelu prilagodbu višestruke regresije metodom najmanjih kvadrata.
Ako presložimo reziduale u drugom koraku primjetit ćemo da je svaki X_j linearna kombinacija Z_k-ova, $k \leqslant j$. Budući da su svi Z_j-ovi ortogonalni, oni tvore bazu za prostor stupaca matrice X zbog čega je projekcija najmanjih kvadrata na taj potprostor upravo $\hat{Y}$. S obzirom da sami Z_p obuhvaća X_p i to s koeficijentom 1, koeficijent (1.16) je zaista koeficijent dobiven višestrukom regresijom od Y na X_p. Također, preslagivanjem X_j, bilo koji od njih

može doći na zadnju poziciju što daje sličan rezultat. Stoga zaključujemo da koeficijent višestruke regresije $\hat{\beta}_j$ predstavlja dodatni doprinos od X_j na Y nakon što je X_j prilagođen za $X_0, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$.

U slučaju kada je X_p jako koreliran s nekime od ostalih X_k-ova, vektor reziduala Z_p će biti blizak nuli te iz (1.16) zaključujemo da će tada koeficijent $\hat{\beta}_p$ biti jako nestabilan. To će također vrijediti za sve varijable u koreliranom skupu. U toj bi situaciji svi *Z-score*-ovi mogli biti maleni što bi značilo da bilo koja varijabla iz koreliranog skupa može biti izbrisana, no ipak ih ne možemo sve izbrisati.

Iz jednadžbe (1.16) također dobivamo i alternativnu formulu za procjenitelje varijance:

$$\mathrm{Var}(\hat{\beta}_p) = \frac{\sigma^2}{\langle Z_p, Z_p \rangle} = \frac{\sigma^2}{||Z_p||^2}. \tag{1.17}$$

Iz formule (1.17) zaključujemo da preciznost kojom možemo procijeniti $\hat{\beta}_p$ ovisi o duljini vektora reziduala Z_p; to prikazuje koliko X_p nije objašnjen drugim X_k-ovima.

Prikažimo sada drugi korak algoritma u matričnoj formi:

$$X = Z\Gamma. \tag{1.18}$$

Ovdje je Z matrica čiji su stupci vektori Z_j u pravilnom redoslijedu, a Γ je gornjetrokutasta matrica s elementima $\hat{\gamma}_{kj}$. Uvrštavanjem dijagonalne matrice D kojoj je j-ti dijagonalni element $D_{jj} = ||Z_j||$, dobivamo

$$X = ZD^{-1}D\Gamma = QR, \tag{1.19}$$

odnosno QR dekompoziciju matrice X. Matrica Q je ortogonalna matrica dimenzije $n \times (p+1)$, to jest vrijedi $Q^TQ = I$, a matrica R je gornjetrokutasta matrica dimenzije $(p+1) \times (p+1)$. QR dekompozicija prikazuje prikladnu ortogonalnu bazu za prostor stupaca matrice X. Iz ovog prikaza lako dolazimo do rješenja metode najmanjih kvadrata:

$$\hat{\beta} = R^{-1}Q^TY. \tag{1.20}$$

Iz čega slijedi

$$\hat{Y} = QQ^TY. \tag{1.21}$$

Budući da je matrica R gornjetrokutasta, jednadžba (1.20) je lako rješiva.
Za detalje vidi [1].

Gauss-Markov teorem

Jedan od najpoznatijih rezultata u statistici dokazuje da procjena parametara β putem najmanjih kvadrata ima najmanju varijancu među svim linearnim nepristranim procjenama. Međutim, restrikcija na nepristrane procjene nije nužno mudra, što ćemo također precizirati i razjasniti. To će nas opažanje navesti da razmotrimo pristrane procjene poput ridge regresije koju ćemo kasnije definirati.
Prije iskaza Gauss-Markovljevog teorema iskazat ćemo i definirati neke pojmove koji su potrebni za njegovo razumijevanje kao i za njegov dokaz.

Definicija 4. Procjenitelj S_n je *nepristran procjenitelj* za τ ako vrijedi

$$\mathrm{E}[S_n] = \tau.$$

Pretpostavimo sada da za slučajne greške vrijede *Gauss-Markovljevi uvjeti*:

- $\mathrm{E}[\varepsilon_i] = 0$, za sve $i = 1, 2, \ldots, n$,
- $\mathrm{Var}[\varepsilon_i] = \sigma^2$, za sve $i = 1, 2, \ldots, n$,
- $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$, za sve $i \neq j$.

Tada su procjenitelji najmanjih kvadrata *nepristrani*:

$$E_\beta[\hat{\beta}] = \beta, \qquad \forall \beta \in \mathbb{R}^{p+1}.$$

Neka je $L : \mathbb{R}^{p+1} \longrightarrow \mathbb{R}$ linearni funkcional parametara:

$$L(\beta) = l^T \beta.$$

Teorem 1.2.4. *Neka vrijede Gauss-Markovljevi uvjeti te neka je $\hat{\beta}$ procjenitelj od β dobiven metodom najmanjih kvadrata. Tada vrijedi*

$$\mathrm{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}. \tag{1.22}$$

Dokaz. Označimo $A = (X^T X)^{-1} X^T$. Tada je $\hat{\beta} = AY$. Sada koristeći (1.10) imamo

$$\mathrm{cov}(\hat{\beta}) = A \,\mathrm{cov}(Y) A^T = \sigma^2 AIA^T = \sigma^2 AA^T = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

□

Definicija 5. Neka je Y vektor opaženih vrijednosti zavisne varijable, to jest, varijable odaziva. Statistika $T = t(Y)$ je:

1. *Linearni procjenitelj* za $L(\beta)$ ako je oblika

$$T = c^T Y,$$

za neki neslučajni vektor $c \in \mathbb{R}^n$.

2. *Nepristrani procjenitelj* za $L(\beta)$ ako je

$$E_\beta[T] = L(\beta), \qquad \forall \beta \in \mathbb{R}^{k+1}.$$

3. *Najbolji linearni nepristrani procjenitelj* za $L(\beta)$ ako je za $L(\beta)$ on:

   - linearan procjenitelj
   - nepristran procjenitelj
   - u klasi svih nepristranih linearnih procjenitelja za $L(\beta)$ ima najmanju varijancu.

Budući da su predviđene vrijednosti zavisne varijable y oblika $f(x_0) = x_0^T\beta$, fokusirat ćemo se na procjenu bilo koje linearne kombinacije parametara $\theta = a^T\beta$. Prema (1.8), procjena najmanjih kvadrata od $a^T\beta$ je

$$\hat{\theta} = a^T\hat{\beta} = a^T(X^TX)^{-1}X^TY$$

Kako je X matrica koja sadrži vektor jedinica i vektore stupce nezavisnih varijabli, odnosno mjerenja, koja su dana, to jest, fiksna, i cijela matrica X je fiksna. Zbog toga je $\tilde{\theta}$ zapravo linearna funkcija c_0^TY u varijabli vektora opaženih vrijednosti varijable odaziva Y.
Pretpostavimo da je linearan model ispravan. Tada je procjena $a^T\hat{\beta}$ nepristrana budući da vrijedi

$$\mathrm{E}[a^T\hat{\beta}] = \mathrm{E}[a^T(X^TX)^{-1}X^TY] = a^T(X^TX)^{-1}X^TX\beta = a^T\beta.$$

Gauss-Markovljev teorem tvrdi da ukoliko imamo bilo koji drugi linearni procjenitelj $\tilde{\theta} = c^TY$ koji je nepristran za $a^T\beta$, tada mora vrijediti $\mathrm{Var}(a^T\hat{\beta}) \leqslant \mathrm{Var}(c^TY)$:

Teorem 1.2.5 (Gauss-Markov)**.** *Neka je $\hat{\beta}$ procjenitelj metodom najmanjih kvadrata za parametre linearnog regresijskog modela te neka je $L(\beta) = a^T\beta$. Ako vrijede Gauss-Markovljevi uvjeti, tada je statistika*

$$T = a^T\hat{\beta}$$

najbolji linearni nepristrani procjenitelj za $L(\beta)$.

Dokaz. Neka je $b^T Y$ proizvoljan nepristran linearan procjenitelj od $a^T\beta$. Budući da je $b^T Y$ nepristran procjenitelj od $a^T\beta$ vrijedi $a^T\beta = \mathrm{E}(b^T Y) = b^T X\beta$, za sve β iz čega pak zaključujemo

$$b^T X = a^T. \tag{1.23}$$

Sada iz (1.22) i (1.23) imamo:

$$\mathrm{Var}(b^T Y) = b^T \,\mathrm{cov}(Y)b = b^T(\sigma^2 I)b = \sigma^2 b^T b \tag{1.24}$$

i

$$\mathrm{Var}(a^T\hat{\beta}) = a^T \,\mathrm{cov}(\hat{\beta})a = \sigma^2 a^T(X^T X)^{-1}a = \sigma^2 b^T X(X^T X)^{-1}X^T b. \tag{1.25}$$

Zbog toga i činjenice da je $M = I - X(X^T X)^{-1}X^T$ pozitivna semidefinita matrica vrijedi

$$\mathrm{Var}(b^T Y) - \mathrm{Var}(a^T\hat{\beta}) = \sigma^2[b^T b - b^T X(X^T X)^{-1}X^T b] = \sigma^2 b^T[I - X(X^T X)^{-1}X^T]b \geqslant 0. \tag{1.26}$$

Time je dokazana tvrdnja teorema. □

Promotrimo sada srednju kvadratnu pogrešku procjenitelja $\tilde{\theta}$ koji procjenjuje θ:

$$MSE(\tilde{\theta}) = \mathrm{E}[\tilde{\theta} - \theta]^2 = \mathrm{Var}(\tilde{\theta}) + [\mathrm{E}[\tilde{\theta}] - \theta]^2 \tag{1.27}$$

Prvi izraz srednje kvadratne pogreške je varijanca procjenitelja, dok je drugi izraz njegova kvadratna pristranost. Po Gauss-Markovljevom teoremu procjenitelj najmanjih kvadrata ima najmanju varijancu među svim linearnim nepristranim procjeniteljima što povlači da on ima najmanju srednju kvadratnu pogrešku među njima. Međutim, može postojati pristran procjenitelj sa manjom srednjom kvadratnom pogreškom. Takav procjenitelj zamijenjuje malo pristranosti za veće smanjenje u varijanci. Metode koje smanjuju ili postavljaju na nulu nekog od koeficijenata najmanjih kvadrata mogu rezultirati pristranom procjenom. S obzirom da želimo procjenitelja sa što je manjom mogućom srednjom kvadratnom pogreškom, u nastavku rada razmotrit ćemo neke od pristranih procjena, poput varijabilnog odabira podskupa i ridge regresije, budući da se oni uobičajeno koriste.

Poglavlje 2

Odabir podskupa

Kao što smo već napomenuli procjena najmanjih kvadrata često nije najbolja procjena zbog čega koristimo i pristrane modele. Točnije, odabir pravog modela svodi se na odabir ispravne ravnoteže između pristranosti i varijance.

Dva su razloga zbog čega često nismo zadovoljni procjenom najmanjih kvadrata. Prvi od njih je preciznost, odnosno točnost procjene. Procjenitelji najmanjih kvadrata imaju malo odstupanje, ali često veliku varijancu. Smanjivanjem ili postavljanjem nekih koeficijenata na nulu može se poboljšati preciznost procjene. Radeći to žrtvujemo malo pristranosti kako bismo smanjili varijancu procjenjenih vrijednosti i kako bi na taj način poboljšali cjelokupnu preciznost predviđanja.

Drugi je razlog tumačenje. Procjene najmanjih kvadrata uobičajno imaju velik broj procjenitelja, no mi bismo htjeli odrediti manji podskup koji prikazuje najjače djelovanje. Odnosno želimo reducirati model tako da u njemu ostanu samo one nezavisne varijable čiji je učinak na zavisnu varijablu najveći te čijom će se linearnom kombinacijom lako opisati utjecaj novog mjerenja na ishod. U tom smislu, voljni smo žrtvovati malu pristranost kako bismo dobili širu sliku.

Prilikom odabira podskupa zadržavamo samo podskup varijabli i eliminiramo ostale varijable iz modela. Regresiju najmanjim kvadratima tada koristimo za procjenu koeficijenata zadržanih nezavisnih varijabli. Postoji nekoliko različitih strategija za odabir podskupa koje u nastavku opisujemo.

2.1 Odabir najboljeg podskupa

Prva od strategija odabira podskupa varijabli koju ćemo opisati je regresija koja koristi najbolji podskup. Takva regresija za svaki $k \in \{0, 1, \ldots, p\}$ pronalazi podskup varijabli duljine k koji daje najmanju sumu kvadrata reziduala. Djelotvoran algoritam koji radi navedeno je procedura *skokova i granica* (Furnival i Wilson, 1974.). Pomoću tog algoritma moguće je

izabrati najbolji podskup veličine do 40. Prilikom takvog odabira, najbolji podskup jedne veličine ne mora nužno sadržavati varijablu koja je u najboljem podskupu manje veličine. Pitanje odabira prave veličine podskupa k svodi se na razmjenu između pristranosti i varijance, kao i na uvažavanje subjektivne želje za što manjim brojem varijabli. Nekoliko je kriterija koje možemo koristiti prilikom tog odabira, no najčešće odabiremo najmanji model koji minimizira procjenitelja očekivane greške procjene.
Za detalje vidi [1].

2.2 Stepwise unaprijed i unatrag

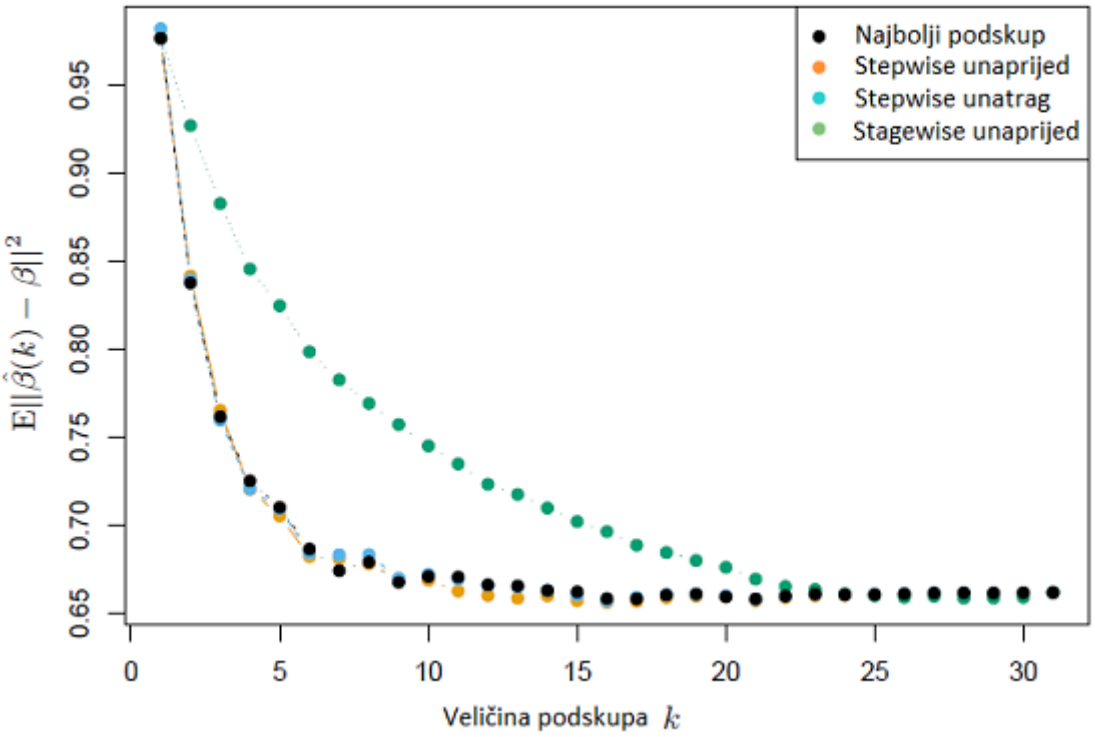
Traženje najboljeg modela po svim mogućim podskupovima varijabli postaje nepraktično za veliki broj nezavisnih varijabli. Stoga umjesto toga tražimo dobar put kroz podskupove.

Odabir *stepwise unaprijed* započinje s praznim modelom u kojeg se uzastopno dodaje varijabla koja najviše poboljšava prilagodbu. To se radi na način da se u model prvo doda varijabla koja ima najmanju p-vrijednost kada je ona jedini prediktor. Nakon toga se u svakom koraku dodaje varijabla koja ima najmanju p-vrijednost u odnosu na ostale prediktore koje smo već dodali. Budući da je mnogo kandidata za prediktora, može djelovati da ovakva procedura zahtijeva puno računanja. Međutim, tome nije tako jer pametni algoritmi za aktualizaciju mogu koristiti QR dekompoziciju trenutne prilagodbe kako bi brzo utvrdili slijedećeg kandidata.

Kako je stepwise unaprijed *pohlepan algoritam*, on proizvodi pohlepan niz modela zbog čega bi se mogao smatrati manje optimalnim od odabira najboljeg podskupa. Međutim, on može biti bolji iz računskih razloga. Naime, za veliki broj nezavisnih varijabli teško je izračunati niz najboljih podskupova, no uvijek možemo izračunati niz stepwise unaprijed odabirom. Također, za odabir najboljeg podskupa svake duljine se plaća cijena u varijanci, dok je stepwise unaprijed ograničenija pretraga zbog čega će imati manju varijancu, ali možda veću pristranost.

Odabir *stepwise unatrag* započinje punim modelom nakon čega uzastopno izbacuje varijablu koja ima najmanji utjecaj na prilagodbu. U ovom je slučaju kandidat za izbacivanje varijabla s najmanjim z-scoreom. Za razliku od stepwise unaprijed koji se može uvijek koristiti, odabir stepwise unatrag moguće je koristiti samo u slučaju kada je $n > p$.

Na slici 2.1 su prikazani rezultati proučavanja male simulacije u svrhu usporedbe regresije najboljim podskupom sa manje zahtjevnim alternativama, stepwise unaprijed i unatrag odabirom. Vidimo da je njihovo djelovanje vrlo slično, što se često događa. Slika 2.1 sadrži i djelovanje stagewise unaprijed odabira kojem je potrebno više vremena kako bi postigao



Slika 2.1: Usporedba četiriju metoda odabira podskupa na simuliranom problemu linearne regresije. Prikazana je srednja kvadratna greška procjenjenog koeficijenta $\hat{\beta}(k)$ za svaku veličinu podskupa. Izvor [1, str. 59]

minimalnu grešku. Taj ćemo odabir opisati u slijedećem odjeljku.

Postoje softverski paketi koji provode hibridne strategije stepwise odabira. Takve strategije u svakom koraku razmatraju pomake unaprijed i unatrag, odnosno dodavanje nove varijable u model i izbacivanje postojeće iz modela. Nakon toga odabiru bolju od te dvije opcije. Jedan primjer takve procedure je *step()* funkcija u R paketu koja koristi Akaike informacijski kriterij (AIC) za validaciju odabira. Taj kriterij je relativna mjera, a temelji se na ocjeni kompleksnosti modela. Kada u modelu ima k parametara i L je maksimalna vrijednost funkcije vjerojatnosti, AIC vrijednost se računa na način:

$$AIC = 2k - 2ln(L).$$

Za računanje AIC vrijednosti u R paketu, koristimo funkciju *AIC()*. Prilikom validacije modela, bolji je onaj čija je AIC vrijednost niža.

2.3 Stagewise unaprijed regresija

Stagewise unaprijed regresija je ograničenija od stepwise unaprijed regresije. Ona počinje kao stepwise unaprijed regresija s početkom jednakim $\bar{Y}$ i centriranim prediktorima čiji su svi koeficijenti inicijalizirani na 0. U svakom koraku algoritam prepoznaje varijablu koja ima najveću korelaciju sa trenutnim rezidualom. Potom računa koeficijent jednostavne linearne regresije od reziduala na izabranoj varijali te ga nakon toga dodaje trenutnom koeficijentu za tu varijablu. Proces se nastavlja sve dok nijedna od varijabli nije korelirana sa rezidualom.

U ovome procesu, za razliku od stepwise unaprijed regresije, nijedna od ostalih varijabli nije prilagođena u trenutku kada se varijabla dodaje u model. Posljedica toga je da stagewise unaprijed regresija može zahtijevati puno više od p koraka da bi postigla prilagodbu najmanjim kvadratima te je u povijesti bila odbačena jer je bila neučinkovita.

Kao što smo već spomenuli, stagewise unaprijed regresija je sadržana na slici 2.1. U tom joj je primjeru potrebno više od 1000 koraka da sve korelacije budu manje od 10^{-4}. U istom su primjeru označene greške za zadnji korak u kojem je k koeficijenata različito od nule, budući da je duljina podskupa upravo k. Iako se greška podudara s najboljom prilagodbom, u ovome je slučaju potrebno više vremena za njeno dostizanje.

Poglavlje 3

Metode sažimanja

Odabirom podskupa prediktora te odbacivanjem ostalih, metode koje koriste odabir podskupa stvaraju interpretabilan model čija je greška predviđanja vjerojatno manja od greške predviđanja potpunog modela. Međutim, u tom postupku varijable su ili zadržane ili odbačene što čini proces diskretnim, zbog čega je konačan rezultat često izložen velikoj varijanci, a zbog čega pak navedeni postupak ne smanjuje grešku predviđanja potpunog modela.
Za razliku od odabira podskupa, metode sažimanja su neprekidnije i nisu toliko sklone velikoj varijabilnosti. Stoga ćemo u nastavku opisati dvije takve metode.

3.1 Ridge regresija

Ridge regresija metoda je sažimanja koja sažima koeficijente regresije stavljanjem kazne na njihovu veličinu. Zbog toga koeficijenti ridge regresije minimiziraju penaliziranu sumu kvadrata reziduala,

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{3.1}$$

Ovdje je $\lambda \geqslant 0$ parametar složenosti koji upravlja iznosom sažimanja tako da vrijedi: što je veća vrijednost λ, veći je iznos sažimanja. Koeficijenti ridge regresije sažimaju se prema nuli i jedan prema drugome.

Svaka od dosad spomenutih metoda, kao i one koje ćemo tek obraditi, ima parametar složenosti koji je odabran da minimizira procjenu greške predviđanja baziranu na desetorostrukoj unakrsnoj validaciji. Unakrsna validacija djeluje tako da na slučajan način dijeli ulazne podatke na deset jednakih dijelova. Potom se devet desetina podataka prilagođava odabranom metodom, a na preostaloj desetini računa se greška predviđanja. Postupak se

uzastopno ponavlja za svaku desetinu podataka i deset dobivenih procjena greške predviđanja se usrednji. Iz toga dobivamo krivulju procjena greške predviđanja kao funkciju od parametra složenosti.

Drugačiji, no ekvivalentan zapis gore navedenog ridge problema je

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \qquad \text{tako da} \qquad \sum_{j=1}^{p} \beta_j^2 \leqslant t. \tag{3.2}$$

Ovakav zapis jasnije predočava ograničenje veličine parametara. Postoji 1-1 korespodencija između parametra λ u (3.1) i parametra t u (3.2).
U slučaju kada postoji puno koreliranih varijabli u linearnom regresijskom modelu, njihovi koeficijenti mogu postati slabo određeni i biti izloženi velikoj varijanci. U takvim situacijama se znatno velik pozitivan koeficijent na jednoj varijabli može poništiti sličnim velikim negativnim koeficijentom na njegovom koreliranom paru. Nametanjem ograničenja na veličinu koeficijenata, kao u (3.2), taj se problem ublažava.

Rješenja dobivena ridge regresijom nisu ekvivarijantna s obzirom na skaliranje ulaznih podataka zbog čega uobičajeno prije rješavanja (3.1) standardiziramo ulazne podatke. Također, slobodni član β_0 je izostavljen iz izraza penaliziranja. Tome je tako zato što bi procedura ovisila o početnoj točki izabranoj za Y kada bi slobodni član bio penaliziran. Odnosno, dodavanje konstante c svakom ishodu y_i ne bi pojednostavilo rezultat u pomaku predviđanja za isti iznos c.

Nakon reparametrizacije koristeći centrirane ulazne podatke, rješenje (3.1) se može razdvojiti na dva dijela. Koriste se centrirani ulazni podaci što znači da svaki x_{ij} biva zamijenjen s $x_{ij} - \bar{x}_j$. Tada procjenjujemo β_0 sa $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Nakon toga se preostali koeficijenti procjenjuju ridge regresijom bez slobodnog člana, koristeći centrirane x_{ij}. Ubuduće pretpostavljamo da je centriranje provedeno tako da matrica ulaznih podataka X ima p stupaca, radije nego $p + 1$.

Zapišimo sada (3.1) u matričnom obliku:

$$RSS(\lambda) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta. \tag{3.3}$$

Iz gornjeg zapisa problema slijedi da su rješenja dobivena ridge regresijom:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y, \tag{3.4}$$

pri čemu je I matrica identiteta dimenzije $p \times p$.

Izborom kvadratične kazne $\beta^T\beta$, rješenje dobiveno ridge regresijom opet je linearna funkcija od Y. Rješenje dodaje pozitivnu konstantu na dijagonalu od $X^T X$ prije invertiranja.

Zbog toga je problem regularan, čak i u slučaju kada $X^T X$ nije punog ranga.
Kada su ulazni podaci ortonormalni, procjene koeficijenata dobivene ridge regresijom su skalirane verzije procjene koeficijenata metodom najmanjih kvadrata, odnosno vrijedi:

$$\hat{\beta}^{ridge} = \frac{\hat{\beta}}{1+\lambda}.$$

Dodatan pogled na funkcioniranje ridge regresije daje nam *dekompozicija singularnih vrijednosti* (SVD) centrirane matrice ulaznih podataka X. Matrica X je dimenzije $n \times p$ i njena dekompozicija singularnih vrijednosti ima oblik:

$$X = UDV^T, \tag{3.5}$$

gdje su U i V ortogonalne matrice dimenzija $n \times p$, odnosno $p \times p$. Stupci matrice U razapinju prostor stupaca od X, dok stupci matrice V razapinju prostor redaka matrice X. D je dijagonalna matrica dimenzije $p \times p$, s elementima na dijagonali $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_p \geqslant 0$. Te dijagonalne elemente zovemo *singularne vrijednosti* od X. Za matricu X vrijedi da je singularna ukoliko vrijedi $d_j = 0$ za jednu ili više singularnih vrijednosti.
Vektor prilagodbe najmanjih kvadrata možemo zapisati koristeći dekompoziciju singularnih vrijednosti kao

$$X\hat{\beta}^{nk} = X(X^T X)^{-1} X^T Y = UU^T Y. \tag{3.6}$$

Pri tome su $U^T Y$ koordinate vektora Y s obzirom na ortonormalnu bazu U.
Rješenja ridge regresije sada su oblika

$$\begin{aligned} X\hat{\beta}^{ridge} &= X(X^T X + \lambda I)^{-1} X^T Y \\ &= UD(D^2 + \lambda I)^{-1} DU^T Y \\ &= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T Y. \end{aligned} \tag{3.7}$$

Ovdje su u_j stupci matrice U. λ je parametar složenosti koji, kao što smo već napomenuli, upravlja iznosom sažimanja. Za njega u slučaju ridge regresije vrijedi $\lambda \geqslant 0$ iz čega proizlazi $d_j^2/(d_j^2 + \lambda) \leqslant 1$. Dakle, kao i linearna regresija, ridge regresija računa koordinate od Y s obzirom na ortonormalnu bazu U. Nakon toga ridge regresija sažima izračunate koordinate za faktor $d_j^2/(d_j^2 + \lambda)$. Stoga je veći opseg sažimanja primjenjen na one koordinate vektora ortonormalne baze uz koje je manja vrijednost d_j^2.

Sljedeće što nas zanima je što znači mala vrijednost od d_j^2? Dekompozicija singularnih vrijednosti centrirane matrice X samo je drugi način izražavanja *glavnih komponenti*

varijabli u matrici X. Primjer matrice kovarijance dan je s $S = X^T X/n$ s čime iz (3.5) imamo

$$X^T X = VD^2V^T. \tag{3.8}$$

Posljednji izraz je *svojstvena dekompozicija* matrice $X^T X$ (također i matrice S, do na faktor n). Stupci matrice V su svojstveni vektori te se također nazivaju i *glavne komponente* smjerova od X. Prva glavna komponenta smjera je v_1 i ona ima svojstvo $z_1 = Xv_1$ te z_1 ima najveću uzoračku varijancu među svim normaliziranim linearnim kombinacijama stupaca matrice X. Ta je uzoračka varijanca jednaka

$$\text{Var}(z_1) = \text{Var}(Xv_1) = \frac{d_1^2}{n}. \tag{3.9}$$

Također vrijedi i

$$z_1 = Xv_1 = u_1 d_1.$$

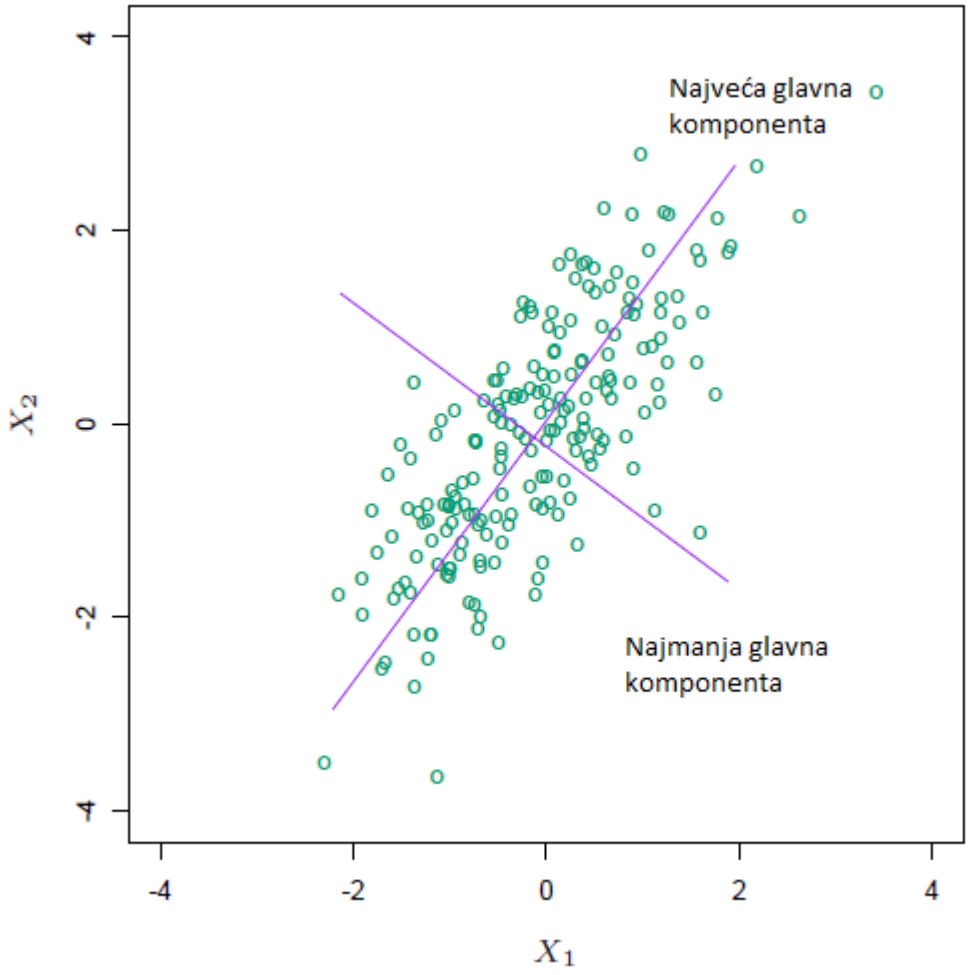
Izvedena varijabla z_1 zove se *prva glavna komponenta* od X zbog čega je u_1 normalizirana prva glavna komponenta. Sljedeće glavne komponente z_j imaju maksimalnu varijancu d_j^2/n te su ortogonalne s ranijima. Zadnja glavna komponenta ima najmanju varijancu. Zbog toga male singularne vrijednosti d_j odgovaraju onim smjerovima u prostoru stupaca matrice X koji imaju malu varijancu te ridge regresija najviše sažima upravo te smjerove.

Slika 3.1 prikazuje glavne komponente točaka nekih podataka u dvije dimenzije. Kada razmotrimo prilagodbu linearne plohe duž te domene, struktura podataka omogućava nam da preciznije odredimo njen nagib u dugom smjeru, nego u kratkom. Ridge regresija štiti od potencijalne velike varijance gradijenta koji su procjenjeni u kratkim smjerovima. Tome je tako jer se podrazumijeva da će odaziv u smjerovima koji imaju veliku varijancu inputa biti sklon najvećem variranju.

Osvrnimo se sada na *efektivne stupnjeve slobode*. Za prilagodbu ridge regresijom definiramo ih na slijedeći način:

$$\begin{aligned} df(\lambda) &= \text{tr}[X(X^T X + \lambda I)^{-1} X^T] \\ &= \text{tr}(H_\lambda) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}. \end{aligned}$$

Dakle, *efektivni stupanj slobode* prilagodbe ridge regresijom je montono padajuća funkcija od λ. U prilagodbi linearnom regresijom s p varijabli stupanj slobode prilagodbe jednak je upravo p, to jest, broju slobodnih parametara. Iako će u prilagodbi ridge regresijom svaki od p koeficijenata vjerojatno biti različit od nule, svaki od njih je kontroliran s λ te je na taj način izložen ograničenju. Zbog toga će u ridge regresiji stupanj slobode biti manji, što



Slika 3.1: Glavne komponente točaka nekih ulaznih podataka. Izvor [1, str. 67]

je veći parametar složenosti λ. Na kraju primjetimo još da je $df(\lambda) = p$ kada je $\lambda = 0$, dok $df(\lambda) \to 0$ u slučaju kada $\lambda \to \infty$.

3.2 Lasso regresija

Baš kao i ridge regresija, lasso regresija je metoda sažimanja. Između njih postoje neke suptilne, no značajne razlike koje ćemo navesti. Procjena lasso regresijom definirana je s

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \qquad \text{tako da} \qquad \sum_{j=1}^{p} |\beta_j| \leqslant t. \tag{3.10}$$

Kao i u slučaju ridge regresije, možemo reparametrizirati konstantu β_0 standardiziranjem prediktora. Tada je rješenje lasso regresije za $\hat{\beta}_0$ jednako $\bar{Y}$, a nakon toga prilagođavamo model bez slobodnog člana.

Lasso problem zapisujemo i u *Lagrangeovoj formi* koja je ekvivalentna gornjoj formi:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.11}$$

Sličnost s ridge regresijom je u postavljanju ograničenja na veličinu koeficijenata, dok je razlika u prirodi tog ograničenja za svaku od metoda. Preciznije, L_2 kazna ridge regresije koja je izražena s $\sum_{j=1}^{p} \beta_j^2$ zamijenjena je L_1 lasso kaznom izraženom s $\sum_{j=1}^{p} |\beta_j|$. Potonje ograničenje dovodi do nelinearnosti njegovog rješenja u y_i zbog čega ne postoji zatvorena forma rješenja kao što je bio slučaj kod ridge regresije. Rješenje lasso problema je kvadratičan problem programiranja. Međutim, dostupni su učinkoviti algoritmi za rješavanje cjelokupnog puta rješenja. Zbog prirode ograničenja, uzimanjem dovoljno malog t uzrokovat ćemo da neki koeficijenti budu točno jednaki nuli. Zbog toga lasso regresija radi neku vrstu neprekidnog odabira podskupa.
Ako odaberemo t koji je veći od $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$, pri čemu su $\hat{\beta}_j = \hat{\beta}_j^{nk}$ koeficijenti procjenjeni metodom najmanjih kvadrata, tada će procjene koeficijenata lasso regresijom biti upravo $\hat{\beta}_j$. Međutim, u slučaju kada je t jednak recimo $t = t_0/2$, koeficijenti najmanjih kvadrata će se lasso regresijom sažeti za oko 50% u prosjeku. Parametar t mora biti odabran tako da minimizira procjenu očekivane greške predviđanja, baš kao što se odabire i veličina podskupa kod odabira podskupa varijabli te parametar složenosti, odnosno sankcije, kod ridge regresije.

3.3 Usporedba odabira podskupa, ridge i lasso regresije

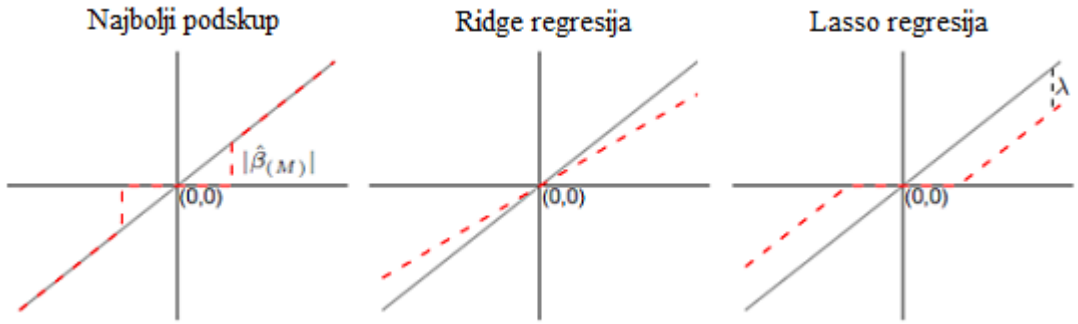
Opisali smo tri pristupa za ograničavanje modela linearne regresije koje ćemo sada usporediti: odabir podskupa varijabli, ridge regresiju i lasso regresiju.

Kada je matrica ulaznih podataka X ortonormalna, svaka od metoda primjenjuje jednostavne trasformacije na koeficijente procjenjene metodom najmanjih kvadrata, $\hat{\beta}_j$. Zbog toga sva tri pristupa u tom slučaju imaju eksplicitna rješenja koja su dana u tablici 3.1.

Procjenitelj	Formula
Najbolji podskup (veličine M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geqslant \lvert\hat{\beta_{(M)}}\rvert)$
Ridge regresija	$\hat{\beta}_j/(1+\lambda)$
Lasso regresija	$sign(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

Tablica 3.1: Procjenitelji $\hat{\beta}_j$ u slučaju ortonormalne matrice X. Izvor [1, str. 71]

Učinak navedenih metoda za ograničavanje prikazan je na slici 3.2.

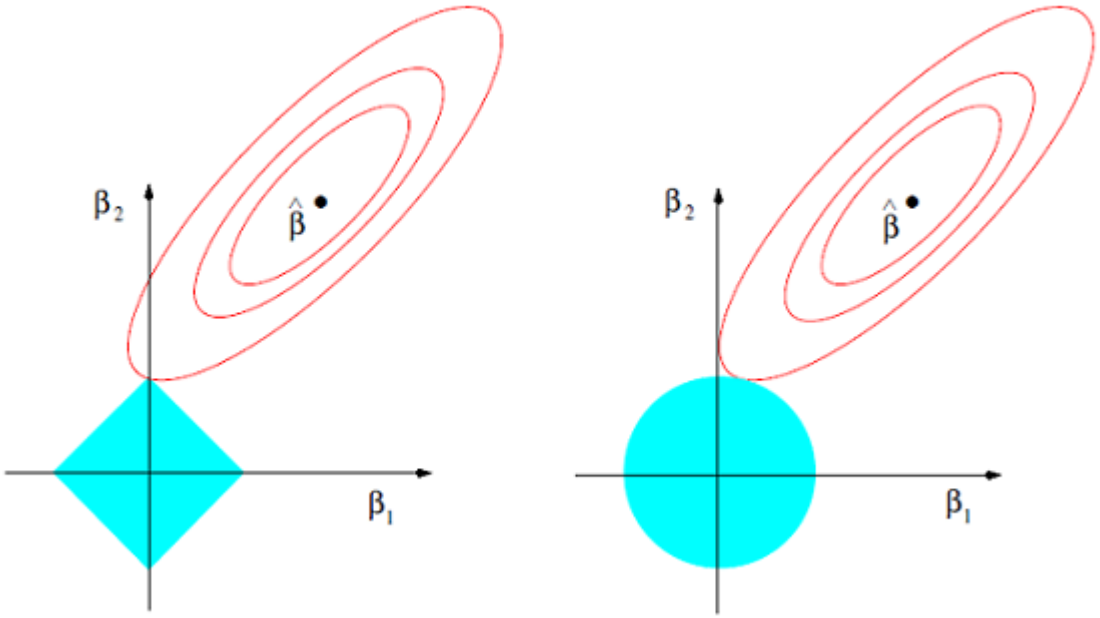


Slika 3.2: Procjenitelji $\hat{\beta}_j$ u slučaju ortonormalne matrice X prikazani su crvenim isprekidanim linijama. Siva linija $y = x$ prikazuje neograničenu procjenu za usporedbu. Izvor [1, str. 71]

Primjećujemo da ridge regresija proporcionalno sažima koeficijente. Lasso regresija translatira svaki od koeficijenata za konstantni faktor λ te odbacuje koeficijente blizu nule. Na poslijetku, metoda odabira najboljeg podskupa odbacuje sve varijable čiji su pripadni koeficijenti manji od M-tog najvećeg.

Slučaj kada matrica X nije ortonormalna prikazan je na slici 3.3 za lasso (lijevo) i ridge regresiju (desno) sa samo dva parametra.

Eliptične konture koje prikazuju sumu kvadrata reziduala centrirane su kod potune procjene najmanjim kvadratima. Za ridge regresiju, područje ograničenja je krug $\beta_1^2 + \beta_2^2 \leqslant t$, dok je to za lasso regresiju dijamant $|\beta_1| + |\beta_2| \leqslant t$. Obje metode pronalaze prvu točku u kojoj eliptične konture pogađaju područje ograničenja. Za razliku od kruga, dijamant ima kuteve te ukoliko se rješenje pojavljuje na kutu, tada to rješenje ima jedan parametar β_j



Slika 3.3: Plave površine su područja ograničenja $|\beta_1| + |\beta_2| \leqslant t$, odnosno $\beta_1^2 + \beta_2^2 \leqslant t^2$, dok su crvene elipse konture funkcije greške najmanjih kvadrata. Izvor [1, str. 71]

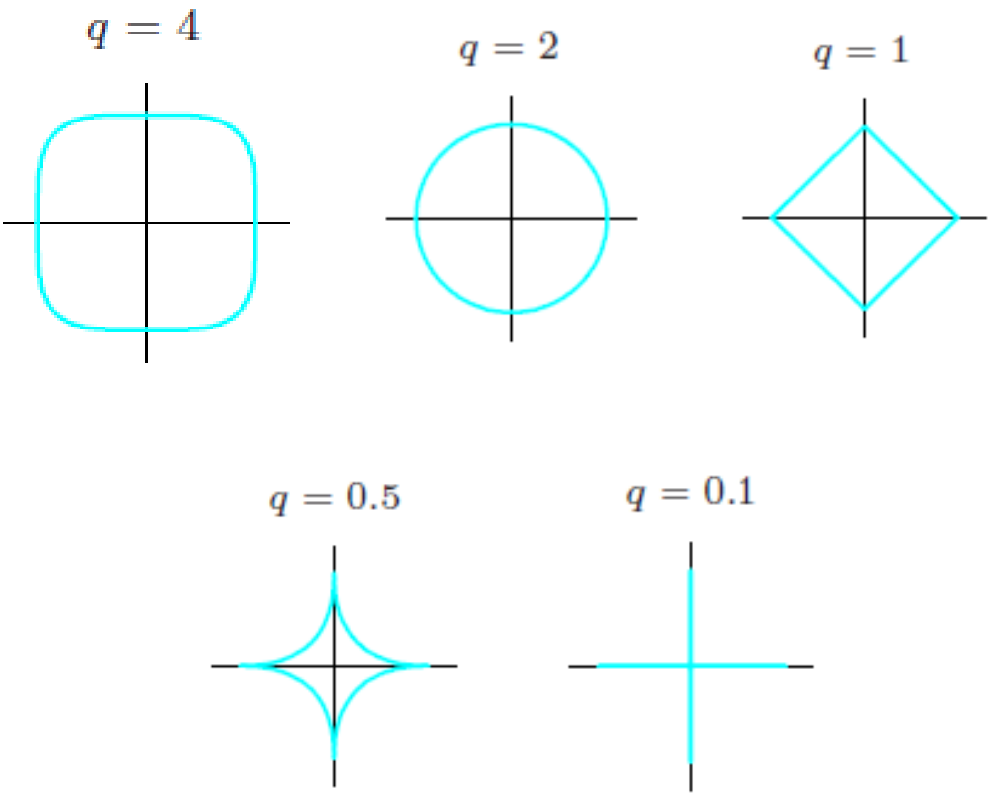
jednak nuli. U slučaju kada je $p > 2$ dijamant postaje romboid te ima puno kuteva zbog čega postoji još mnogo mogućnosti da procjenjeni parametri budu jednaki nuli.

Ridge i lasso regresiju možemo generalizirati te ih promatrati kao Bayesove procjene. Promotrimo kriterij

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \tag{3.12}$$

za $q \geqslant 0$. Konture $\sum_{j=1}^{p} |\beta_j|^q$ su prikazane na slici 3.4 za različite vrijednosti q, ponovno za slučaj s dva parametra.

Ako razmišljamo o $|\beta_j^q|$ kao o apriori gustoći od β_j, tada su to također konture apriori distribucije parametara. Vrijednost $q = 0$ odgovara odabiru podskupa varijabli, budući da kazna u toj metodi zbraja broj parametara različitih od nule. Vrijednost $q = 1$ odgovara lasso regresiji, dok $q = 2$ odgovara ridge regresiji. U slučaju $q \leqslant 1$ apriori gustoća nije uniformna u pravcu, ali koncentrira veću masu u smjerovima koordinata.
Dakle, s ovog su gledišta odabir najboljeg podskupa te ridge i lasso regresija Bayesove procjene s različitim apriori distribucijama. Za detalje vidi [1].



Slika 3.4: Konture konstantne vrijednosti od $\sum_{j=1}^{p} |\beta_j|^q$ za dane vrijednosti q. Izvor [1, str. 72]

Poglavlje 4

Primjena linearne regresije

Nakon teorijske obrade višestruke linearne regresije, u ovom ćemo se poglavlju posvetiti njenoj primjeni. Linearna regresija ima bitnu ulogu u raznim procjenama, a zbog njene prilagodljivosti i lakog računanja na transformiranim podacima spektar njene primjene je širok. U ovom slučaju posvetit ćemo se primjeni linearne regresije u svrhu predviđanja razine prodaje određenog proizvoda u ovisnosti o demografskim i drugim varijablama. Naime, ljudi koji žive u različitim regijama vrlo vjerojatno potražuju različite proizvode. Kao najjednostavniji primjer toga možemo spomenuti potražnju za ogrijevnim sredstvima. Ona je u kontinentalnim predjelima vrlo vjerojatno veća od potražnje u primorskim predjelima. Jasan i logičan uzrok toga je blaža klima primorskih područja u usporedbi s klimom kontinentalnih predjela. Baš kao i područje stanovanja, na potražnju za nekim proizvodom utječe dugi niz čimbenika, kao što su spol osobe, njena starost, bračni status, broj djece, mjesto stanovanja u vidu gradskog ili seoskog područja te kuće ili stana kao i još mnogi drugi čimbenici. Budući da sve te katakteristike doprinose različitim potražnjama, a suvremeni trgovački lanci uglavnom posluju na području cijele države, od velike je važnosti mogućnost predviđanja potražnje, to jest prodaje, u ovisnosti o demografskim podacima. Na taj način moguće je predvidjeti koji proizvodi bi mogli zanimati kupce čime se može pospješiti rad trgovine, a što za posljedicu ima smanjenje gubitaka poslovanja te povećanje dobiti.

Kako bi sve navedeno mogli provesti u djelo, potrebno je konstantno sakupljati podatke o dosadašnjoj prodaji. Podaci se nakon toga analiziraju, prilagođavaju i ugrađuju u model, a što je manja razlika predviđene i stvarne prodaje, model je bolji i uspjeh veći.

4.1 Modeliranje

Za razvoj modela su sa stranice Kaggle skinuti podaci o internetskoj prodaji proizvoda. Vidi [7]. Podaci sadrže mnoštvo varijabli, od kojih odabiremo samo nama potrebne, kao

što su podaci o prodaji, regiji stanovanja, starosti, spolu, broju djece i tome slični. Također, za odabrane varijable uklanjamo opažanja kod kojih nije zabilježena količina prodaje. Sveukupno imamo deset varijabli, od kojih je jedna zavisna, a ostalih devet je nezavisno. Varijable su kategorijske pa ih pretvaramo u *dummy*, to jest, identifikatorske varijable. Za svaku od varijabli postoji 519 opažanja.

Podatke dijelimo na dva skupa. Prvi od njih je skup za učenje koji sadrži 400 opažanja, a koji koristimo za stvaranje modela. Drugi od skupova je skup za validaciju koji sadrži 119 opažanja, a na kojem testiramo model koji smo razvili na temelju opažanja iz skupa za učenje. Nakon navedenih priprema, prelazimo na izradu modela. Željeli bismo razviti model kojim bi mogli predvidjeti jednu varijablu na temelju utjecaja preostalih varijabli. Metoda kojom ćemo sve to provesti je višestruka linearna regresija. Varijabla koju želimo predvidjeti, odnosno zavisna varijabla, je godišnja prodaja proizvoda. Nezavisnih je varijabli, kao što je gore navedeno, devet te se svaka od njih odnosi na demografske čimbenike.

Za svaki model koji razvijemo izračunat ćemo pogrešku predviđanja kako bismo vidjeli da li i koliko dobro odabrani model predviđa zavisnu varijablu. U tu svrhu koristimo WMAPE (*Weighted Mean Absolute Percentage Error*) mjeru. Formula za izračun WMAPE mjere je

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

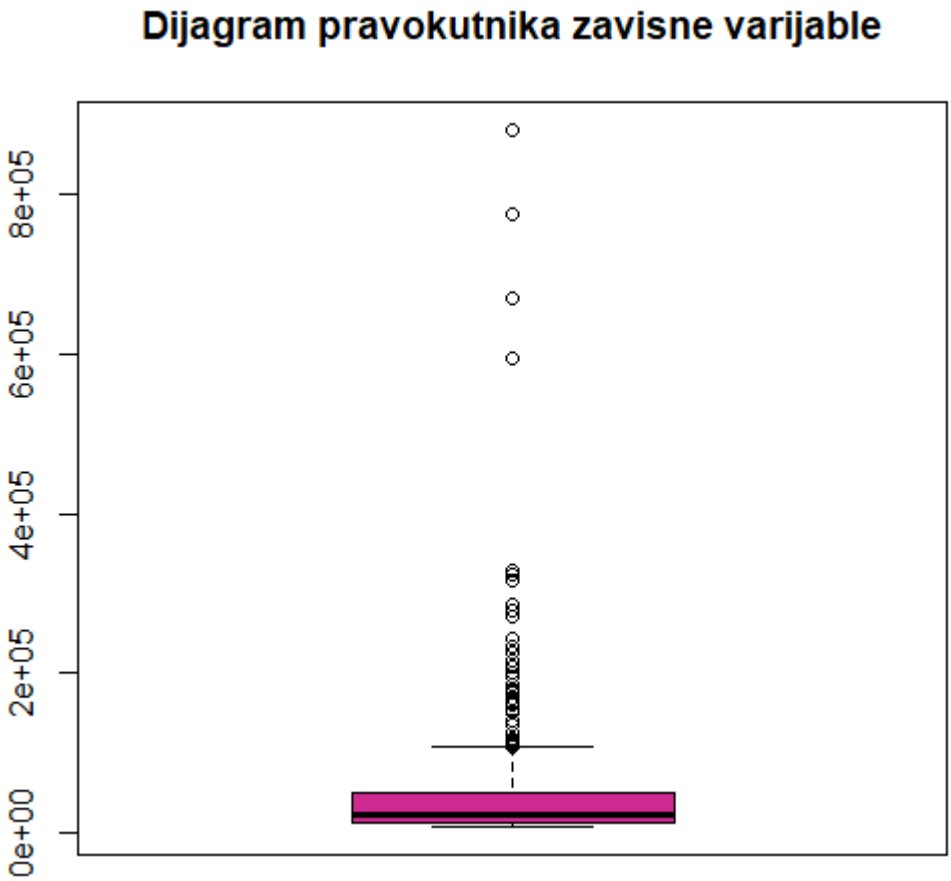
pri čemu je y_i opažena vrijednost zavisne varijable u i-tom opažanju, $\hat{y}_i$ je njena predviđena vrijednost, a n je broj opažanja.

Prije nego započnemo s izradom modela, posvetit ćemo se opisnoj statistici zavisne varijable. Prvo što računamo su frekvencije i relativne frekvencije varijable ”prodaja” te crtamo njen pripadni histogram. Potom računamo aritmetičku sredinu, varijancu, standardnu devijaciju, medijan, mod i raspon uzorka. Na poslijetku računamo karakterističnu petorku i interkvartil uzorka te crtamo njegov dijagram pravokutnika koji je prikazan na slici 4.1.

Sada kada smo gotovi sa opisnom statistikom varijable ”prodaja”, okrećemo se ispitivanju normalne distribuiranosti iste varijable. U tu svrhu crtamo njen normalni vjerojatnosni graf koji je prikazan na slici 4.2. Iz grafa 4.2 zaključujemo da varijabla ”prodaja” u određenoj mjeri prati normalnu distribuciju što je i bilo potrebno.

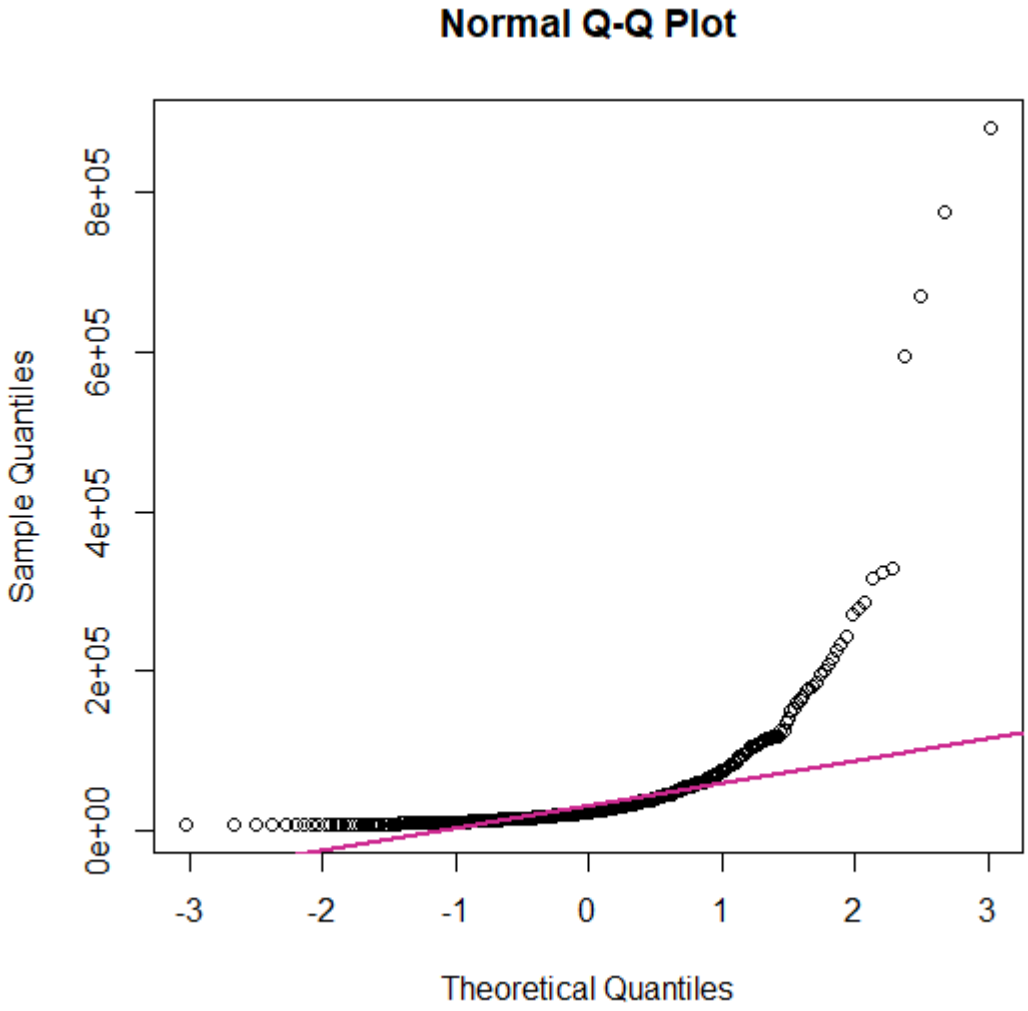
Sada napokon prelazimo na izradu modela. Definirat ćemo različite modele u smislu ubacivanja i izbacivanja značajnih, odnosno neznačajnih varijabli. Druga razlika među modelima nastat će različitim kodiranjem dummy varijabli, što ćemo u nastavku precizirati.

Počinjemo od punog linearnog modela u kojem je jedna varijabla zavisna, i to je ”prodaja”, a ostalo su prediktori, to jest, nezavisne varijable. Dodatno, u tom početnom modelu kao graničnu vrijednost za varijablu ”starost” uzimamo trideset godina, dok za varijablu ”broj djece” kao graničnu vrijednost uzimamo nula djece. To znači da od varijable ”starost” koja je izvorno numerička, jer označava broj godina kupca, nastaje dummy varijabla



Slika 4.1: Dijagram pravokutnika zavisne varijable ”prodaja”.

s dva stupnja. Ta dvostupanjska dummy varijabla poprima dvije vrijednosti, nula i jedan i to: nula za sva opažanja u kojima kupac ima manje ili jednako od trideset godina, a jedan za sva opažanja u kojima kupac ima strogo više od trideset godina. Analogno, numerička varijabla ”broj djece” koja označava broj djece kupca, postaje dvostupanjska dummy varijabla koja poprima vrijednost nula za sva opažanja u kojima kupac ima nula, odnosno nema djece, dok vrijednost jedan poprima za sva opažanja u kojima kupac ima strogo više od nula djece, to jest, opažanja u kojima kupac ima djece. Sada kada smo grupirali podatke i odredili koje varijable koristimo, a to su u ovom slučaju sve varijable budući da koristimo puni model, provodimo višestruku linearnu regresiju. Model definiramo pomoću R-ove funkcije *lm()*. Potom na varijablu oblika *lm* primjenjujemo R-ovu funkciju *summary()* koja daje rezultate višestruke linearne regresije. Od tih rezultata najbitnije su nam vrijednosti koeficijenta determinacije R^2 i prilagođenog R^2 koji označavaju postotak varijance zavisne



Slika 4.2: Normalni vjerojatnosni graf zavisne varijable

varijable koji se može objasniti odabranim modelom. Za detalje vidi [6]. U ovom slučaju njihove su vrijednosti 3.82% za R^2 te 1.6% za priladođeni R^2.

Kako funkcija *summary()* daje uvid u značajnost pojedine nezavisne varijable u modelu, u sljedećem koraku iz modela izbacujemo varijablu koja nije značajna te gradimo novi model, a to sve u svrhu pronalaska modela s što većom vrijednošću R^2, odnosno radi poboljšanja modela. Navedeno je upravo stepwise unatrag odabir.

Drugi se model razlikuje od prvog u tome što je iz njega izbačena neznačajna varijabla ”zaposlen” te je za graničnu vrijednost varijable ”starost” i ”broj djece” postavljeno četrdeset godina, odnosno dvoje djece čime je dobiveno novo grupiranje navedenih varija-

bli. Vrijednost R^2 za ovaj model je 5.01%, dok je prilagođeni R^2 jednak 3.06%.

U trećem smo modelu opet izbacili neznačajnu varijablu, u ovom slučaju varijablu "grad_selo" te smo za granične vrijednosti varijabli "starost" i "broj djece" stavili pedeset godina, odnosno četvero djece. Vrijednosti R^2 i prilagođenog R^2 u ovom su slučaju redom jednake 4.25%, odnosno 2.54%.

Četvrti je model manji od trećeg za varijablu "obrazovanje", dok je za graničnu vrijednost varijabli "starost" i "broj djece" postavljeno 45 godina, odnosno troje djece. Vrijednosti R^2 i prilagođenog R^2 jednake su 4.28%, odnosno 2.82%.

Na poslijetku, u petom modelu vraćamo varijablu "obrazovanje", a za granične vrijednosti varijabli "starost" i "broj djece" stavljamo 35 godina, odnosno opet dvoje djece. Vrijednosti R^2 i prilagođenog R^2 jednake su 5.09%, odnosno 3.4%.

Svi modeli imaju male vrijednosti R^2 i prilagođenog R^2, no te su vrijednosti najveće kod zadnjeg modela zbog čega je taj model najbolji.

4.2 Provjere ispravnosti modela

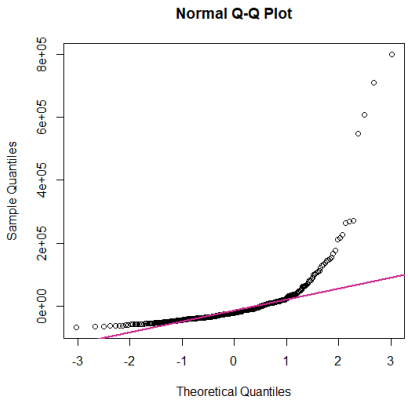
Da bismo mogli koristiti višestruku linearnu regersiju, moraju biti zadovoljene pretspostavke linearnog odnosa između varijabli poticaja i odaziva, normalne distribucije grešaka te njihovoj nezavisnosti i homogenosti.

Normalnu distribuciju grešaka provjeravamo normalnim vjerojatnosnim grafovima pripadnih reziduala za svaki od pet definiranih modela. Grafovi su prikazani na slici 4.3. Iz njih zaključujemo da su greške linearnih modela zaista normalno distribuirane budući da normalni vjerojatnosni grafovi reziduala prate normalnu distribuciju, no razlike u varijancama pogrešaka ima.

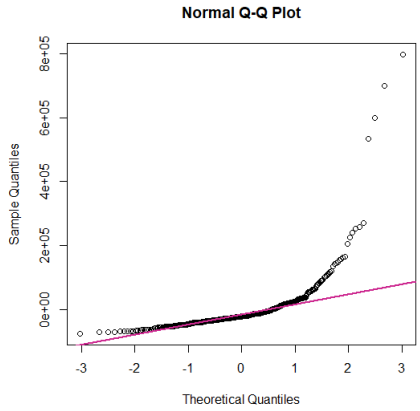
Linearnost podataka ispitujemo *residual-fit* plotom, odnosno grafičkim prikazom reziduala i procijenjenih vrijednosti zavisne varijable. Residual-fit plotovi za svaki od pet navedenih modela prikazani su na slici 4.4. Budući da su podaci simetrično raspršeni oko apscisa, iz residual-fit plotova zaključujemo da linearnost postoji.

Nakon što smo se uvjerili da možemo koristiti višestruku linearnu regresiju za dane podatke, preostaje nam izračunati pogreške modela. To znači da na skupu za validaciju moramo primjeniti procjenjene koeficijente β kako bismo dobili predviđene vrijednosti zavisne varijable za taj skup. U tu svrhu koristimo R-ovu funkciju *predict()* koja za parametre uzima odabrani model i skup podataka za validaciju. Za detalje vidi [6]. Na kraju iz predviđenih i opaženih vrijednosti zavisne varijable računamo pogrešku modela pomoću WMAPE mjere.

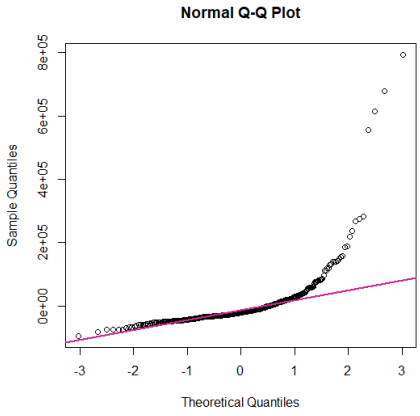
Pogreške za svaki od pet modela redom su jednake 88.67% za prvi, 89.2% za drugi, 89.47% za treći, 89.09% za četvrti te 88.97% za peti model. Iako su sve pogreške velike, najmanje su one od prvog i zadnjeg modela zbog čega, kao i zbog činjenice da je R^2 zadnjeg modela



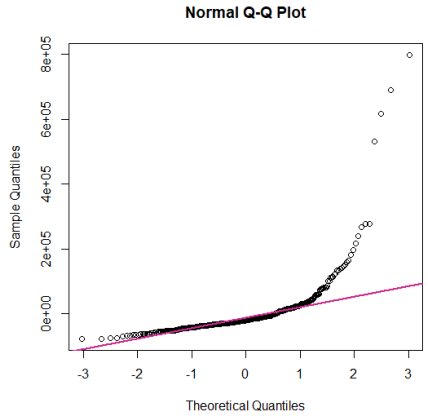
(a) Prvi model



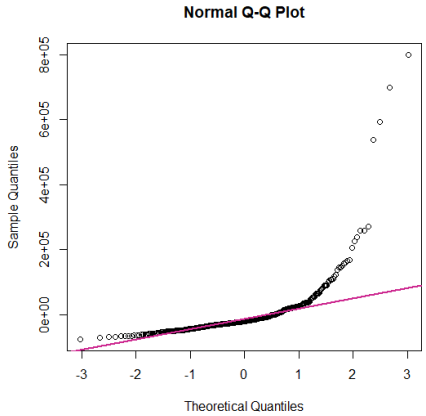
(b) Drugi model



(c) Treći model

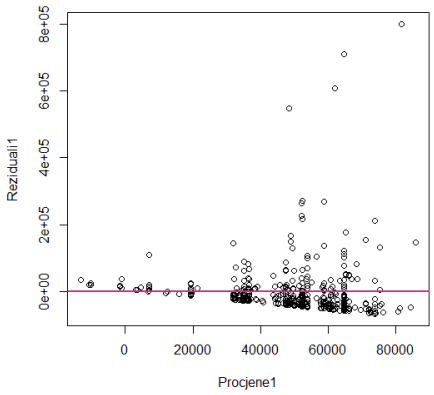


(d) Četvrti model

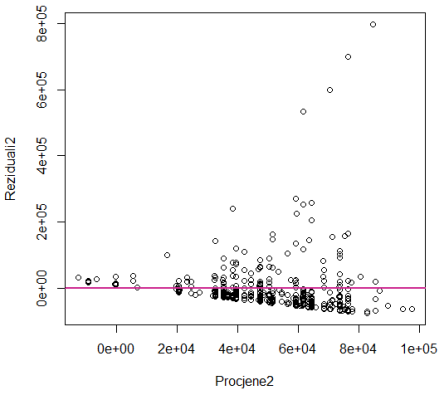


(e) Peti model

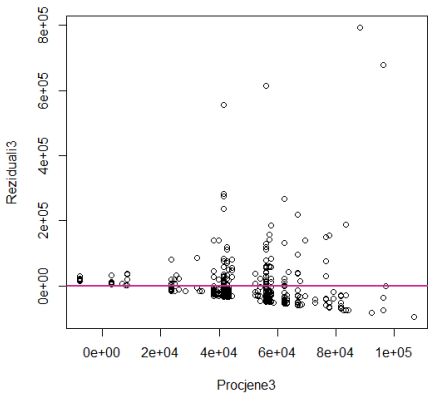
Slika 4.3: Normalni vjerojatnosni grafovi reziduala za svaki od pet linearnih modela



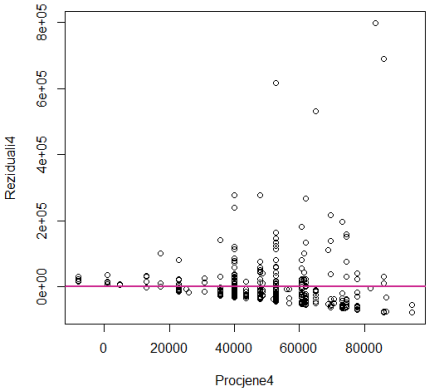
(a) Prvi model



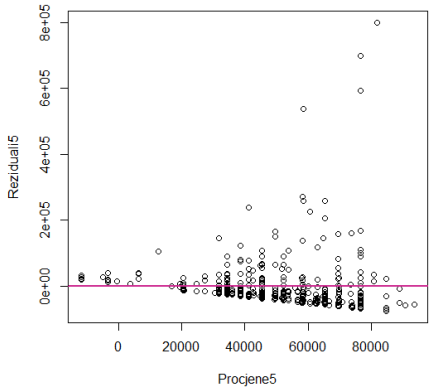
(b) Drugi model



(c) Treći model



(d) Četvrti model



(e) Peti model

Slika 4.4: Residual-fit plotovi za svaki od pet linearnih modela

najveći, zaključujemo da je upravo zadnji model najbolji.

Nakon što smo izračunali koeficijente determinacije i pogreške, prikazat ćemo koeficijente najboljeg od definiranih modela:

```
Call:
lm(formula = prodaja ~ starost35 + spol + regija + u_braku +
    djeca2 + stan_kuca + obrazovanje)

Coefficients:
(Intercept)    starost35         spol       regija      u_braku       djeca2
      35988         6865       -11215        -8249       -24013        24190
  stan_kuca  obrazovanje
      45953        -4210
```

Kao što smo vidjeli, koeficijent determinacije najboljeg linearnog modela je malen, a njegova pogreška je velika zbog čega na poslijetku zaključujemo da postoje modeli kojima se obrađeni podaci mogu bolje modelirati.

Poglavlje 5

Dodatak

U dodatku prilažemo kod pomoću kojeg smo proveli modeliranje. Kod je napravljen u programskom jeziku R.

```
rm(list=ls())

podaci=read.table("skup_za_ucenje.txt")
podaci

prodaja=podaci[,1]
starost30=podaci[,2]
starost35=podaci[,3]
starost40=podaci[,4]
starost45=podaci[,5]
starost50=podaci[,6]
spol=podaci[,7]
regija=podaci[,8]
u_braku=podaci[,9]
djeca0=podaci[,10]
djeca1=podaci[,11]
djeca2=podaci[,12]
djeca3=podaci[,13]
djeca4=podaci[,14]
zaposlen=podaci[,15]
grad_selo=podaci[,16]
stan_kuca=podaci[,17]
obrazovanje=podaci[,18]
```

```
n=length(prodaja)
n

wmape=function(y,ykapa){
gornji=0
donji=0
for(i in 1:length(y)){
gornji=gornji+abs((ykapa[i]-y[i])/y[i])*y[i]
donji=donji+y[i]
}
return(gornji/donji)
}

tablica_frekvencija=data.frame(table(prodaja))
tablica_frekvencija

tablica_frekvencija=data.frame(tablica_frekvencija,
tablica_frekvencija[2]/sum(tablica_frekvencija[2]))
tablica_frekvencija

names(tablica_frekvencija)[3]="relative_frekvencije"
names(tablica_frekvencija)[2]="frekvencije"
tablica_frekvencija

hist(prodaja, probability=TRUE)

mean(prodaja)

var(prodaja)

sd(prodaja)

statmod=function(x){
z=table(as.vector(x))
names(z)[z==max(z)]
}
statmod(prodaja)

raspon_uzorka=max(prodaja)-min(prodaja)
```

```
raspon_uzorka

median(prodaja)

quantile(prodaja,0.25,type=6)

quantile(prodaja,0.75,type=6)

quantile(prodaja,type=6)

IQR(prodaja)

boxplot(prodaja, col="maroon3",main="Dijagram
_pravokutnika_zavisne_varijable")

qqnorm(prodaja)
qqline(prodaja, col="maroon3", lwd=2)

model1=lm(prodaja~starost30+spol+regija+u_braku+
djeca0+zaposlen+grad_selo+stan_kuca+obrazovanje)
model1
sum1=summary(model1)
sum1

model2=lm(prodaja~starost40+spol+regija+u_braku+
djeca2+grad_selo+stan_kuca+obrazovanje)
model2
sum2=summary(model2)
sum2

model3=lm(prodaja~starost50+spol+regija+u_braku+
djeca4+stan_kuca+obrazovanje)
model3
sum3=summary(model3)
sum3

model4=lm(prodaja~starost45+spol+regija+u_braku+
djeca3+stan_kuca)
model4
```

```
sum4=summary(model4)
sum4

model5=lm(prodaja~starost35+spol+regija+u_braku+
djeca2+stan_kuca+obrazovanje)
model5
sum5=summary(model5)
sum5

qqnorm(model1$res)
qqline(model1$res, col="maroon3", lwd=2)

qqnorm(model2$res)
qqline(model2$res, col="maroon3", lwd=2)

qqnorm(model3$res)
qqline(model3$res, col="maroon3", lwd=2)

qqnorm(model4$res)
qqline(model4$res, col="maroon3", lwd=2)

qqnorm(model5$res)
qqline(model5$res, col="maroon3", lwd=2)

Procjene1=model1$fit
Reziduali1=model1$res
plot(Procjene1,Reziduali1)
abline(h=0, col="maroon3", lwd=2)

Procjene2=model2$fit
Reziduali2=model2$res
plot(Procjene2,Reziduali2)
abline(h=0, col="maroon3", lwd=2)

Procjene3=model3$fit
Reziduali3=model3$res
plot(Procjene3,Reziduali3)
abline(h=0, col="maroon3", lwd=2)
```

```
Procjene4=model4$fit
Reziduali4=model4$res
plot(Procjene4,Reziduali4)
abline(h=0, col="maroon3", lwd=2)

Procjene5=model5$fit
Reziduali5=model5$res
plot(Procjene5,Reziduali5)
abline(h=0, col="maroon3", lwd=2)

validacija1=read.table("test1.txt")
validacija1
procjena1=predict.lm(object=model1,newdata=validacija1)
test_pogreska1=wmape(validacija1$prodaja,procjena1)
test_pogreska1

validacija2=read.table("test2.txt")
validacija2
procjena2=predict.lm(object=model2,newdata=validacija2)
test_pogreska2=wmape(validacija2$prodaja,procjena2)
test_pogreska2

validacija3=read.table("test3.txt")
validacija3
procjena3=predict.lm(object=model3,newdata=validacija3)
test_pogreska3=wmape(validacija3$prodaja,procjena3)
test_pogreska3

validacija4=read.table("test4.txt")
validacija4
procjena4=predict.lm(object=model4,newdata=validacija4)
test_pogreska4=wmape(validacija4$prodaja,procjena4)
test_pogreska4

validacija5=read.table("test5.txt")
validacija5
procjena5=predict.lm(object=model5,newdata=validacija5)
test_pogreska5=wmape(validacija5$prodaja,procjena5)
test_pogreska5
```

Bibliografija

[1] Jerome Friedman, Trevor Hastie i Robert Tibshirani, *The Elements of Statistical Learning*, Springer Series in Statistics, New York, 2009.

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

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

[4] Miljenko Huzak, *Statistika*, predavanja, PMF-MO

[5] Tomislav Šmuc, *Strojno učenje*, predavanja, PMF-MO, 2013.

[6] `https://www.rdocumentation.org/`

[7] `https://www.kaggle.com/c/online-sales/data`

Sažetak

Prvi cilj ovoga rada je teorijski predstaviti model višestruke linearne regresije kao i metode odabira podskupa i metode sažimanja koje se koriste za njegovo poboljšanje. Kako bismo to napravili, rad smo započeli s teorijskom obradom modela jednostruke linearne regresije.

Drugi je cilj pokušati statistički opisati vezu između godišnje prodaje određenog proizvoda i demografskih čimbenika. U tu smo svrhu koristili metodu višestruke linerane regresije koju smo prethodno teorijski obradili. Podaci su prvo prilagođeni, a nakon toga je provedena opisna statistika, modeliranje podataka i analiza modela. Na kraju rada iskazan je zaključak o optimalnosti korištenja modela linearne regresije za dane podatke.

Summary

The first goal of this paper is to theoretically present the multiple linear regression model as well as the subset selection methods and the shrinkage methods that are used for its improvement. In order to do so, we have started the paper with theoretical analysis of a univariate linear regression model.

The second goal is to try to describe the relationship between annual sales of a particular product and demographic factors. For this purpose, we used the method of multiple liner regression which we have previously theoretically presented. Data was first adjusted, and then descriptive statistics, data modeling, and model analysis were performed. At the end of the paper, a conclusion was reached on the optimum utilization of linear regression model for given data.

Životopis

Martina Škrnjug rođena je 29. ožujka 1993. godine u Zagrebu. Završava osnovnu školu Sesvetski Kraljevec te potom i Srednju školu Sesvete, smjer opća gimnazija, u Sesvetama. Godine 2011. upisuje Preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. Spomenuti studij završava 2015. godine te tako stječe naziv univ. bacc. math. Iste godine upisuje diplomski sveučilišni studij, smjer Financijska i poslovna matematika.