SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ida Faullend Heferer

MODELIRANJE TURISTIČKE POTROŠNJE KORIŠTENJEM METODA STATISTIČKOG UČENJA

Diplomski rad

Voditelj rada:
izv. prof. dr. sc. Nikola Sandrić

Zagreb, ožujak, 2023.

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. _____

Zahvaljujem mentoru izv. prof. dr. sc. Nikoli Sandriću na brojnim smjernicama, strpljenju i susretljivosti prilikom pisanja ovog diplomskog rada. Institutu za turizam, posebno doc. dr. sc. Damiru Krešiću i Zrinki Marušić, zahvaljujem na ustupljenim podacima.

Sadržaj

Uvod

Cilj ovoga rada je, u suradnji s Institutom za turizam, modelirati turističku potrošnju u Republici Hrvatskoj korištenjem metoda statističkog učenja. Ekonomski utjecaj turističkih tokova na gospodarstva često je zamjetan i predstavlja relevantan ključ za gospodarski rast. Poboljšanje učinaka turizma zahtijeva odgovarajuće podatke i alate kako bi se potaknula ponuda privatnog sektora i djelovanje kreatora politika. U tom kontekstu, ključno je utvrditi odrednice turističke potrošnje putem moćnih analitičkih modela, a mnoge od njih nudi upravo statističko učenje. U ovom radu posebnu pažnju posvetit ćemo metodama odabira prediktora u linearnim modelima u svrhu povećanja interpretabilnosti modela te metodama regularizacije u svrhu poboljšanja svojstva predviđanja modela.

Rad je strukturiran u četiri poglavlja. U prvom poglavlju dajemo uvod u područje statističkog učenja i pregled osnovnih pojmova koji će nam biti potrebni u ostatku rada. Iznosimo neke od glavnih zadaća statističkog učenja te standardnu podjelu na nadgledano učenje i nenadgledano učenje. Detaljno opisujemo postavke nadgledanog statističkog učenja, uvodeći terminologiju i problematiku te predstavljajući teorijski okvir metoda nadgledanog statističkog učenja. Naposljetku, kratko predstavljamo ključne koncepte koji se pojavljuju pri odabiru metode statističkog učenja.

U drugom dijelu rada predstavljamo klasičnu linearnu regresiju. Najprije se podsjećamo modela, a potom i metode najmanjih kvadrata kao najčešće korištene metode za njegovu prilagodbu. Predstavljamo kako se ocjenjuje preciznost procjene parametara linearnog modela kao i preciznost samog modela.

U trećem poglavlju uvodimo dvije važne klase metoda prilagobe linearnog modela kao alternativu metodi najmanjih kvadrata, odabir podskupa i regularizaciju. Prije predstavljanja samih metoda, uvodimo pojmove testne greške i očekivane testne greške, detaljno opisujemo pojmove pristranosti i varijance te njihov odnos poznat pod nazivom *bias-variance trade-off*. Nadalje, predstavljamo metodu unakrsne validacije kojom se procjenjuje očekivana testna greška.

U posljednjem poglavlju, opisane metode primjenjujemo na podatke Instituta za turizam o turističkoj potrošnji u Republici Hrvatskoj. Dajemo njihovu usporedbu i iznosimo najvažnije zaključke.

Poglavlje 1

Statističko učenje

1.1 Što je statističko učenje?

Pojam statističkog učenja odnosi se na niz metoda koje na osnovu skupa podataka daju okvir u svrhu predviđanja i statističkog zaključivanja. Osnovna pretpostavka je statistička priroda fenomena koji se proučava. Neke od glavnih zadaća statističkog učenja su

- klasifikacija - problem dodjeljivanja kategorije pojedinom objektu
- regresija - problem predviđanja vrijednosti nekog objekta ili stvaranja zaključaka o vezama između objekata
- rangiranje - problem rangiranja s obzirom na određeni kriterij
- klasteriranje - problem particioniranja prostora objekata na homogene skupove

Standardna podjela statističkog učenja je na nadgledano učenje i nenadgledano učenje. Kod nadgledanog učenja podaci se sastoje od objekata iz ulaznog skupa podataka i pripadne kategorije/vrijednosti/ranga iz izlaznog skupa podataka i na osnovu toga izvršava se *proces učenja*. Tu spadaju klasifikacija, regresija i rangiranje. Kod nenadgledanog učenja podaci se sastoje samo od objekata te je od interesa naći "zanimljive uzorke" među podacima. Ovdje spada klasteriranje.

Metode statističkog učenja koje ćemo koristiti u ovom radu za modeliranje turističke potrošnje spadaju u domenu nadgledanog učenja.

1.2 Nadgledano statističko učenje

Neka su U i I redom ulazni i izlazni skupovi u problemu učenja. Nadalje, neka su $\mathcal{U}$ i $\mathcal{I}$ pripadne σ–algebre te neka je $\mathbb{P}$ vjerojatnosna mjera na $(U \times I, \mathcal{U} \otimes \mathcal{I})$. U problemima

nadgledanog statističkog učenja traži se funkcija $f\colon U \to I$ koja predstavlja *najbolju* vezu između objekata iz U i pridruženih vrijednosti iz I. U većini slučajeva f nije poznata (ona ne mora ni postojati) te je cilj nadgledanog statističkog učenja konstruirati metode i algoritme koji daju njezinu *najbolju* procjenu.

Vrste varijabli i terminologija

Ulazna varijabla označava se sa X, a izlazna varijabla ima oznaku Y. U nastavku pretpostavljamo da je X slučajni vektor. Njegove komponente označavaju se sa $X_1, ..., X_p$ te se nazivaju *prediktori*, *kovarijate* ili *nezavisne varijable*. Y je slučajna varijabla, naziva se *odziv* ili *zavisna varijabla*. Opažene vrijednosti pišu se malim slovima, i–ta opažena vrijednost od X označava se sa x_i, dok je x_{ij} i–ta opažena vrijednost j–te komponente vektora prediktora X. i–ta opažena vrijednost od Y ima oznaku y_i. Matrica n opaženih p–dimenzionalnih vektora prediktora označava se sa $\mathbf{X}$, gdje je njezin i–ti redak ${x_i}^T$, budući da je uvriježena pretpostavka da su vektori stupci.

S obzirom na vrstu obilježja koje opisuju, varijable dijelimo na *kvalitativne* i *kvantitativne* varijable. Ukoliko je varijabla odziva Y kvalitativna varijabla, radi se o problemu *klasifikacije*, dok se slučaj kada je Y kvantitativna varijabla naziva problemom *regresije*.

Napomena 1.2.1. *Općenito, može se promatrati i više od jedne varijable odziva, no u ovom radu, gdje promatramo jednu odzivnu varijablu - turističku potrošnju, to nije potrebno.*

Skup $\mathcal{T} = \{(x_i, y_i)\colon i = 1, \ldots, n\}$ naziva se *skup za trening* gdje su (x_i, y_i), $i = 1, \ldots n$ realizacije slučajnog vektora (X, Y). [1] Metoda statističkog učenja primjenjuje se na skup za trening s ciljem procjene funkcije f. Pretpostavljamo da je veza između X i Y dana statističkim modelom $Y = f(X) + \varepsilon$, odnosno za elemente skupa za trening vrijedi $y_i = f(x_i) + \epsilon_i$. Ovdje je ε *slučajna greška* nezavisna od X i s očekivanjem 0, a ϵ_i, $i = 1, \ldots, n$ realizacije slučajnog uzorka za slučajnu grešku. S ovakvom formulacijom statističkog modela, f predstavlja cjelovitu informaciju koju X daje o Y. Slučajna greška ε obuhvaća sve druge moguće utjecaje na Y i potencijalnu grešku u mjerenju.

Napomena 1.2.2. *S obzirom na prirodu varijable turističke potrošnje, u nastavku rada Y je kvantitativna varijabla, za izlazni prostor I uzimamo $\mathbb{R}$ te daljnju teoriju izlažemo u tom duhu. Za $X_1, X_2, \ldots, X_p$ pretpostavljamo da su kvantitativne (realne) varijable ili kvalitativne varijable.*

[1] Potpuno precizno, elementi skupa za trening realizacije su slučajnog uzorka za slučajan vektor (X, Y), odnosno realizacije n slučajnih vektora koji su nezavisne jednakodistribuirane kopije slučajnog vektora (X, Y).

Parametarske metode

Metode koje ćemo obrađivati u ovom radu spadaju u parametarske metode. Parametarske metode podrazumijevaju pretpostavku o obliku funkcije f. Primjer jedne jednostavne pretpostavke o obliku funkcije f jest da je ona linearna u X, odnosno da vrijedi

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

S takvom pretpostavkom, problem procjene funkcije f znatno je pojednostavljen – umjesto procjenjivanja proizvoljne funkcije f, potrebno je procijeniti $(p+1)$ koeficijenata $\beta_0, \ldots, \beta_p$. Nakon odabira modela tj. oblika funkcije f, slijedi *prilagodba modela* podacima iz skupa za trening. U ovom primjeru, potrebno je naći vrijednosti parametara $\beta_0, \beta_1, \ldots, \beta_p$ takve da je

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

Jedna od najčešće korištenih metoda za prilagodbu linearnog modela je *metoda najmanjih kvadrata* koju ćemo u drugom poglavlju rada i detaljnije objasniti.

Dakle, metode poput upravo opisane zovu se parametarske metode jer se procjena funkcije f svodi na procjenu parametara pretpostavljenog oblika funkcije f. Prednost takvog pristupa je spomenuto pojednostavljenje problema, a nedostatak je što pretpostavka o obliku funkcije f može biti daleko od njezinog stvarnog oblika. Tom nedostatku može se doskočiti odabirom dovoljno *fleksibilnog* modela tj. modela koji obuhvaća razne moguće oblike funkcije f. Fleksibilni modeli, međutim, često zahtijevaju procjenu puno većeg broja parametara.

Teorijski okvir metoda nadgledanog statističkog učenja

Neka je X p-dimenzionalni slučajni vektor s vrijednostima u $\mathbb{R}^p$ i neka je Y slučajna varijabla s vrijednostima u $\mathbb{R}$ s konačnim matematičkim očekivanjem. Dosada smo u radu spominjali procjenu funkcije f koja daje vezu između ulaznih i izlaznih podataka te $f(X)$ smatramo predikcijom za Y uz dano X. Na početku odjeljka o nadgledanom učenju iskazali smo kako je cilj nadgledanog statističkog učenja konstruirati metode i algoritme koji daju *najbolju* procjenu funkcije f. U tu svrhu, potrebno je definirati *funkciju gubitka* kojom će se mjeriti koliko je procjena funkcije f dobra i na temelju toga moći i dati kriterij za odabir najbolje procjene za funkciju f.

Definicija 1.2.3. *Izmjerivo preslikavanje* $L : \mathbb{R}^2 \to [0, +\infty)$ *zove se* funkcija gubitka.

U slučaju kada je Y kvantitativna varijabla, tipično se za funkciju gubitka uzima $L\colon \mathbb{R}^2 \to [0, +\infty)$ definirana sa

$$L((y_1, y_2)) = (y_1 - y_2)^2.$$

Definiramo *kvadratnu grešku* od $f(X)$ pri procjeni za Y sa

$$L((Y, f(X)) = (Y - f(X))^2.$$

Nadalje, definiramo *srednjekvadratnu grešku* od $f(X)$ pri procjeni za Y sa

$$L(f) = \mathbb{E}[L(Y, f(X))] = \mathbb{E}[(Y - f(X))^2].$$

Definicija 1.2.4. *Neka je Y slučajna varijabla s konačnim matematičkim očekivanjem i X proizvoljan slučajni vektor. Funkcija ϕ definirana na Borelovoj σ-algebri $\mathcal{B}$ sa*

$$\phi(B) = \int\limits_{\{X \in B\}} Y \, d\mathbb{P}, B \in \mathcal{B}$$

je σ-aditivna i apsolutno neprekidna u odnosu na $\mathbb{P}_X$ pa iz Radon-Nikodymova teorema slijedi da postoji Borelova funkcija f(x) definirana za svaki $x \in \mathbb{R}^p$ i jednoznačno određena do na Borelov skup $\mathbb{P}_X$-mjere nula, relacijom

$$\int\limits_B r(x) \, d\mathbb{P}_X(x) = \int\limits_{\{X \in B\}} Y \, d\mathbb{P}$$

za svako $B \in \mathcal{B}$. Funkciju r(x) označavamo sa $\mathbb{E}[Y \mid X = x]$ i zovemo regresijska funkcija.

Napomena 1.2.5. *$\mathbb{P}_X$ je vjerojatnosna mjera na $\mathbb{R}^p$ definirana s $\mathbb{P}_X(B) := \mathbb{P}(X \in B)$, $B \subseteq \mathbb{R}^p$ Borelov skup (vidi str. 234 izvora [6]) i naziva se* distribucija *od X. Ostalih pojmova iz prethodne definicije poput σ-aditivnosti, konačne mjere, apsolutne neprekidnosti, Borelove funkcije, Borelove σ-algebre $\mathcal{B}$ te Radon-Nikodymova teorema moguće je podsjetiti se također u izvoru [6].*

Napomena 1.2.6. *Na str. 579 izvora [6] moguće je podsjetiti se definicije uvjetnog matematičkog očekivanja za danu σ – algebru $\mathcal{G}$, u oznaci $\mathbb{E}[Y \mid \mathcal{G}]$. Oznaka $\mathbb{E}[Y \mid X]$ predstavlja uvjetno očekivanje slučajne varijable Y s obzirom na σ – algebru $\sigma\{X\}$. Regresijsku funkciju interpretiramo kao* uvjetno očekivanje od Y za dano X = x. *Prema Teoremu 15.6 izvora [6], vrijedi $r(X) = \mathbb{E}[Y \mid X]$ gotovo sigurno.*

Navodimo, bez dokaza, nekoliko svojstava regresijske funkcije koja ćemo koristiti u radu. Neka su Y, Y_1, Y_2 realne slučajne varijable s konačnim matematičkim očekivanjem, tada vrijedi:

1. Ako je $Y = c$ (gotovo sigurno), gdje je c konstanta, tada je $\mathbb{E}[Y \mid X = x] = c$, $\mathbb{P}_X$ – gotovo sigurno.

2. $\mathbb{E}[a_1Y_1 + a_2Y_2 \mid X = x] = a_1\mathbb{E}[Y_1 \mid X = x] + a_2\mathbb{E}[Y_2 \mid X = x]$, $\mathbb{P}_X$ – gotovo sigurno, za a_1, a_2 proizvoljne konstante.

3. Ako je f Borelova funkcija takva da $f(X)Y$ ima konačno matematičko očekivanje, tada vrijedi $\mathbb{E}[f(X)Y \mid X = x] = f(x)\mathbb{E}[Y \mid X = x]$, $\mathbb{P}_X$ – gotovo sigurno.

4. Ako su Y i X nezavisne, tada vrijedi $\mathbb{E}[Y \mid X = x] = \mathbb{E}[Y]$, $\mathbb{P}_X$ – gotovo sigurno.

Koristeći vezu matematičkog očekivanja i uvjetnog matematičkog očekivanja, za srednjekvadratnu grešku od $f(X)$ pri procjeni za Y vrijedi

$$L(f) = \mathbb{E}[(Y - f(X))^2] = \int_{\mathbb{R}^p} \mathbb{E}[(Y - f(X))^2 \mid X = x] d\mathbb{P}_X(x).$$

U nastavku rada pretpostavljamo konačnost drugog momenta od $Y - f(X)$ te uvodimo oznaku $L_x(f) \coloneqq \mathbb{E}[(Y - f(X))^2 \mid X = x]$, greška od f u $X = x$.

Lema 1.2.7. *Za svaku $f\colon \mathbb{R}^p \to \mathbb{R}$ i proizvoljan $x \in \mathbb{R}^p$ vrijedi*

$$L_x(f) = L_x(r) + (r(x) - f(x))^2$$

gdje je r prethodno definirana regresijska funkcija.

Dokaz. Za sve f i za proizvoljan $x \in \mathbb{R}^p$ vrijedi

$$\begin{aligned}
L_x(f) &= \mathbb{E}[(Y - f(X))^2 \mid X = x] \\
&= \mathbb{E}[(Y - r(X) + r(X) - f(X))^2 \mid X = x] \\
&= \mathbb{E}[(Y - r(X))^2 + (r(X) - f(X))^2 + 2(Y - r(X))(r(X) - f(X)) \mid X = x] \\
&= \mathbb{E}[(Y - r(X))^2 \mid X = x] + (r(x) - f(x))^2 + (r(x) - f(x))(\mathbb{E}[Y \mid X = x] - r(x)) \\
&= \mathbb{E}[(Y - r(X))^2 \mid X = x] + (r(x) - f(x))^2 \\
&= L_x(r) + (r(x) - f(x))^2
\end{aligned}$$

gdje račun slijedi iz prethodno navedenih svojstava regresijske funkcije te činjenice da je $(r(x) - f(x))^2$ konstanta, a zadnja jednakost iz definicije od $L_x(f)$. □

Iz prethodne leme slijedi da je $L_x(f) \geq L_x(r)$, $\forall f$ i $\forall x \in \mathbb{R}^p$ tj. $r(x)$ je najbolja procjena za Y uz dano $X = x$, $\forall x \in \mathbb{R}^p$ u smislu srednjekvadratne greške $L((Y, f(X))) = (Y - f(X))^2$. Posebno, slijedi $L(f) \geq L(r)$, $\forall f$.

Napomena 1.2.8. *U ovoj napomeni sažimamo gore uvedene oznake, za proizvoljnu funkciju f:*

- $L(Y, f(X)) = (Y - f(X))^2$ - *kvadratna greška od* $f(X)$ *pri procjeni za* Y
- $L(f) = \mathbb{E}[(Y - f(X))^2]$ - *srednjekvadratna greška od* $f(X)$ *pri procjeni za* Y *ili očekivana greška predikcije*
- $L_x(f) = \mathbb{E}[(Y - f(X))^2 \mid X = x]$ - *greška od* f *u* $X = x$.

Lema 1.2.8 pokazuje kako je regresijska funkcija (teoretski koncept) najbolja procjena za Y uz dano $X = x$ i danu funkciju gubitka, a statističko učenje na temelju skupa za trening nastoji dati najbolje metode za *procjenu* regresijske funkcije. Tu procjenu procjenu u nastavku označavamo s f.
$L_x(r)$ zove se *ireducibilna greška* te se na nju ne može utjecati - čak i da za procjenu regresijske funkcije vrijedi $f = r$, i dalje postoji greška u procjeni za Y uvjetno na $X = x$ jer je prema pretpostavci $Y = f(X) + \varepsilon$ tj. X *ne sadrži* svu informaciju o Y. $(r(x) - f(x))^2$ zove se *reducibilna greška* te se nadgledano statističko učenje bavi tehnikama za odabir f s ciljem minimiziranja reducibilne greške.

1.3 Ocjena preciznosti modela statističkog učenja

U ovom radu predstavit ćemo nekoliko metoda statističkog učenja kojima ćemo modelirati turističku potrošnju. Uvodeći pojam funkcije gubitka u prethodnom poglavlju, dali smo okvir za određivanje najbolje metode za dani skup za trening. Važno je napomenuti kako jedna metoda nije nužno uvijek najbolja - na određenom skupu podataka pojedina metoda može dati najbolje rezultate, ali neka druga metoda može dati bolje rezultate na sličnom, no drugačijem skupu podataka. Stoga je bitan zadatak statističkog učenja odrediti metodu koja će biti najbolja za proizvoljan skup podataka. Ovdje ćemo kratko predstaviti ključne koncepte koji se pojavljuju pri odabiru metode statističkog učenja za dani skup za trening. Kasnije, kod predstavljanja metoda i primjene na turističku potrošnju detaljnije ćemo objasniti kako se primjenjuju i u praksi.

Mjerenje kvalitete prilagodbe modela statističkog učenja

Kako bismo ocijenili kvalitetu metode statističkog učenja na danom skupu podataka, trebamo nekako moći mjeriti koliko dobro predikcije dobivene modelom odgovaraju opaženim podacima. Dakle, potrebno je kvantificirati koliko dobro vrijednost varijable odziva predviđena metodom statističkog učenja odgovara stvarnoj (opaženoj) vrijednosti varijable odziva. Najčešće korištena mjera u problemu regresije naziva se *srednjekvadratna greška*. Neka je f procjena za regresijsku funkciju dobivena nekom metodom statističkog učenja. Srednjekvadratna greška, u oznaci MSE (eng. *mean-squared error*), dana je sa

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

gdje je $f(x_i)$ predikcija koju f daje za $i-$tu opservaciju. Pretpostavimo da smo za prilagodbu metode statističkog učenja koristili skup za trening $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Tada možemo izračunati $f(x_1), f(x_2), \ldots, f(x_n)$. Ukoliko su dobivene vrijednosti približno jednake $y_1, y_2, \ldots, y_n$, vrijednost MSE će biti mala. Međutim, nas ne zanima vrijedi li $f(x_i) \approx y_i$, već vrijedi li $f(x_0) \approx y_0$, gdje je (x_0, y_0) prethodno neopažena (testna) opservacija koja nije korištena za prilagodbu metode statističkog učenja. Želimo odabrati onu metodu koja ima najnižu *testnu* MSE. Drugim riječima, kad bismo imali velik broj testnih opservacija, mogli bismo izračunati

$$\frac{1}{k}\sum_{j=1}^{k}(y_j - f(x_j))^2$$

što predstavlja prosječnu kvadratnu grešku predikcije na testnim opservacijama (x_j, y_j), $j = 1, 2, \ldots, k$. Htjeli bismo odabrati model za koji je gornja greška najmanja moguća. Problem je što ponekad nemamo toliko podataka pa samim time ni skup testnih podataka na raspolaganju. Tada možemo birati metodu statističkog učenja koja minimizira MSE skupa za trening. Fundamentalni problem u tom slučaju leži u tome što ne postoji garancija da će metoda s najmanjom MSE skupa za trening imati i najmanju testnu MSE.

MSE i fleksibilnost modela

Mnogi algoritmi statističkog učenja imaju parametre koji kontroliraju ono što zovemo fleksibilnost modela. Ne postoji precizna definicija fleksibilnosti modela, no jedan način za opisati fleksibilnost modela je koliko je ponašanje modela uvjetovano svojstvima samih podataka. Osnovno svojstvo statističkog učenja koje vrijedi neovisno o skupu podataka i metodi statističkog učenja koja se koristi jest da MSE skupa za trening monotono pada kako fleksibilnost modela raste, dok testna MSE poprima takozvani U-oblik. U slučaju kad metoda statističkog učenja daje malu MSE na skupu za trening, a veliku testnu MSE kažemo da se dogodio *overfitting* podataka. Do overfittinga dolazi jer konkretna metoda statističkog učenja previše traži uzorke u podacima za trening koji mogu biti samo rezultat slučajnosti, a ne svojstava nepoznate funkcije f. Testna MSE je tada velika zato što uzorci koje je metoda našla u skupu za trening jednostavno ne postoje u testnim podacima. Neovisno o overfittingu, gotovo uvijek očekujemo da će MSE skupa za trening biti manja od testne MSE jer većina metoda statističkog učenja direktno ili indirektno djeluje s ciljem minimiziranja MSE na skupu za trening. U praksi je relativno jednostavno izračunati MSE

skupa za trening, dok procjena testne MSE predstavlja znatno teži zadatak zbog nedostupnosti testnih podataka. Postoje različiti pristupi koji se u praksi koriste za odabir metode ili parametara pojedine metode s najmanjom testnom MSE. Jedna važna metoda koju ćemo predstaviti u trećem poglavlju ovog rada je unakrsna validacija.
U-oblik testne MSE posljedica je dvaju oprečnih svojstava metoda statističkog učenja - pristranosti i varijance. Njih također predstavljamo u trećem poglavlju rada.

Poglavlje 2

Linearna regresija

U prvom poglavlju rada pokazali smo da u slučaju kada prilagodbu regresijskog modela, gdje regresijskim modelom statističkog učenja jednostavno zovemo model s kvantitativnom varijablom odziva Y, provodimo na način da minimiziramo srednjekvadratnu grešku $L(f)$, minimum je upravo regresijska funkcija definirana sa $r(x) = \mathbb{E}[Y \mid X = x]$. Budući da je distribucija $\mathbb{P}$ s obzirom na koju je ovo uvjetno očekivanje definirano nepoznata, funkciju r želimo nekako procijeniti. U linearnom modelu ona se procjenjuje funkcijom linearnom u X, stoga i naziv linearna regresija (podrijetlo samog naziva *regresija* u ovom radu izostavljamo). Bitno je napomenuti da stvarna regresijska funkcija gotovo nikada nije linearna.

Unatoč tomu što je linearni model jednostavan, vrlo je koristan kako konceptualno, tako i praktično. Zbog svoje jednostavnosti, lako je interpretabilan i često daje dobre predikcije.

2.1 Linearni regresijski model

Linearni regresijski model pretpostavlja da je veza f između X i Y približno linearna, odnosno za model

$$Y = f(X) + \varepsilon$$

gdje je $X^T = (X_1, X_2, \ldots, X_p)$ vektor prediktora i želimo predvidjeti Y s vrijednostima u $\mathbb{R}$, imamo

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.$$

Ovdje X_j predstavlja j–ti prediktor, a β_j predstavlja nepoznati j–ti koeficijent ili parametar koji kvantificira vezu između j–tog prediktora X_j i varijable odziva Y. β_j interpretiramo

kao *prosječni* utjecaj promjene X_j za jediničnu mjeru na Y, držeći sve ostale prediktore fiksnima.

Prisjetimo se, matrica n opaženih p–dimenzionalnih vektora prediktora označava se sa $\mathbf{X}$, gdje je njezin i–ti redak ${x_i}^T$. Sada ćemo sa $\mathbf{X}$ označiti matricu dimenzija $n \times (p+1)$ gdje je njezin prvi stupac vektor jedinica, $\mathbf{1}$. Slučajan uzorak za odziv možemo modelirati nezavisnim slučajnim varijablama na sljedeći način:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

gdje je $i = 1, \ldots, n$ te su ε_i nezavisne jednakodistribuirane slučajne varijable s očekivanjem 0 i varijancom σ^2, dok kovarijate smatramo zadanima. U vektorskom obliku pišemo:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

gdje je $\mathbf{Y}$ slučajni vektor, $\mathbf{X}$ neslučajna matrica s linearno nezavisnim stupcima, a $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]$ vektor slučajnih pogrešaka s očekivanjem $\mathbf{0}$ i kovarijacijskom matricom $\sigma^2\mathbf{I}_n$. Sa $\mathbf{y}$ ćemo označavati vektor realizacija slučajnog vektora $\mathbf{Y}$.

Napomena 2.1.1. *Naglasimo ukratko razliku između modela za odziv Y i modela za slučajan uzorak odziva. Kod modela za odziv, Y je slučajna varijabla i model glasi* $Y = f(X) + \varepsilon$, *gdje je X slučajan vektor prediktora, a* ε *slučajna varijabla, tzv. slučajna greška. S druge strane, kod modela za slučajan uzorak od Y,* $\mathbf{Y}$ *je slučajan vektor, matrica* $\mathbf{X}$ *je neslučajna, a* ε_i, $i = 1, \ldots, n$ *slučajne su varijable koje predstavljaju slučajan uzorak od* ε *te su zbog toga po definiciji slučajnog uzorka nezavisne. Nadalje, pretpostavljamo da su* ε_i, $i = 1, \ldots, n$ *jednakodistribuirane s očekivanjem* 0 *i varijancom* σ^2.

2.2 Metoda najmanjih kvadrata

Nakon što smo specificirali linearni regresijski model, želimo ga prilagoditi podacima koji su nam dani u vidu skupa za trening $\mathcal{T} = \{(x_i, y_i) \colon i = 1, \ldots, n\}$. Skup za trening predstavlja n opaženih parova opservacija, za X i za Y. Budući da je linearni model parametarska metoda, kao što smo u prvom poglavlju nagovijestili, prilagodba modela svodi se na procjenu nepoznatih parametara $\beta_0, \beta_1, \ldots, \beta_p$. Postoji više metoda za procjenu parametara linearnog modela, a jedna od najčešće korištenih je *metoda najmanjih kvadrata* koju predstavljamo u ovom odjeljku.

Neka je $\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ predikcija za Y s obzirom na i–tu opservaciju vektora prediktora X. Razliku $e_i = y_i - \hat{y}_i$ nazivamo i–ti *rezidual*, dakle, on predstavlja razliku između i–te opservacije varijable odziva i i–te predviđene vrijednosti varijable odziva. Metoda najmanjih kvadrata zasniva se na odabiru parametara $\beta_0, \beta_1, \ldots, \beta_p$ koji minimiziraju sumu

kvadrata reziduala, odnosno koje minimiziraju

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i))^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

gdje je $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$. Ideja prilagodbe linearnog modela je naći *hiperravninu* koja je *najbliža* točkama iz skupa za trening, a *blizinu* u metodi najmanjih kvadrata mjerimo sumom kvadrata reziduala.
S notacijom uvedenom u prethodnom odjeljku, suma kvadrata reziduala može se zapisati kao

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

To je kvadratna funkcija u $p + 1$ parametara. Deriviranjem po β dobije se

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$
$$\frac{\partial^2 \mathrm{RSS}}{\partial \beta \partial \beta} = 2\mathbf{X}^T\mathbf{X}.$$

Izjednačavanjem prve derivacije s nulom, $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$, te uz pretpostavku da je $\mathbf{X}$ punog stupčanog ranga, stoga je $\mathbf{X}^T\mathbf{X}$ pozitivno definitna, dobivamo jedinstveno rješenje

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Dobiveno rješenje je u vidu konkretne realizacije $\mathbf{y}$. Zamijenimo li u gornjem računu $\mathbf{y}$ s $\mathbf{Y}$, rješenje možemo zapisati i kao slučajnu varijablu:

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

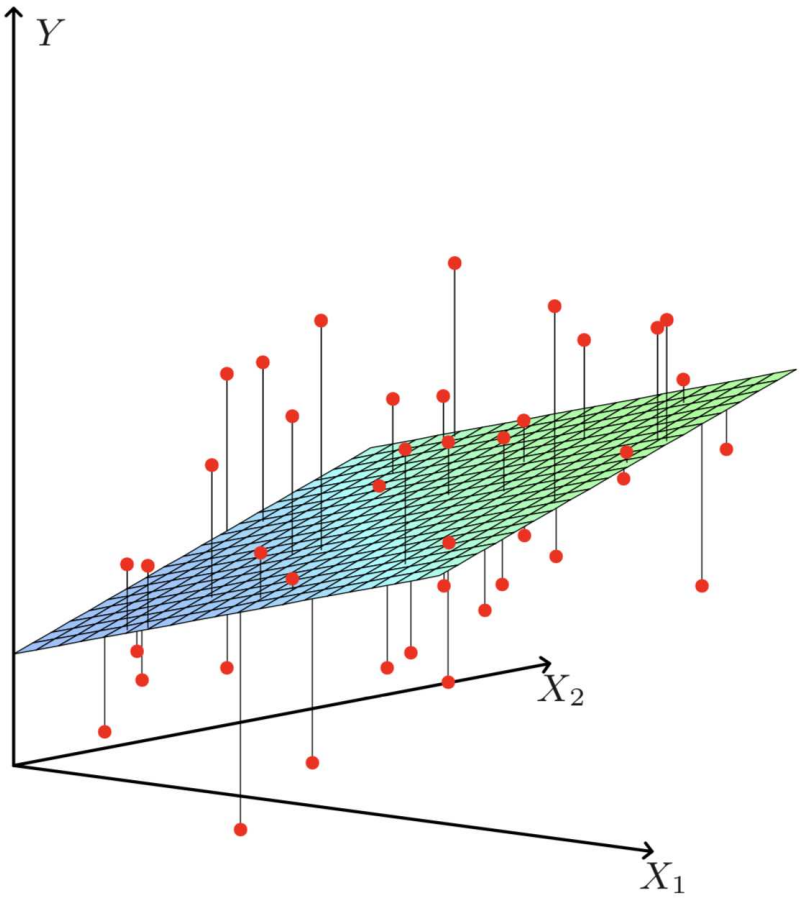
Dakle, slučajna varijabla $\hat{\beta}$ je procjenitelj za β.
Predviđene vrijednosti za opservacije prediktora iz skupa za trening dane su sa

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

iz čega slijedi da je $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$.

Matrica $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ je matrica ortogonalnog projektora te je minimum sume kvadrata reziduala upravo $\hat{\beta}$ za koji je dobivena predikcija $\hat{\mathbf{y}}$ ortogonalna projekcija vektora $\mathbf{y}$ na potprostor razapet stupcima matrice $\mathbf{X}$.

Napomena 2.2.1. *Ukoliko matrica* $\mathbf{X}$ *nije punog stupčanog ranga, parametri* $\hat{\beta}$ *dobiveni metodom najmanjih kvadrata nisu jedinstveno određeni.*



Slika 2.1: Metoda najmanjih kvadrata za X s vrijednostima $\mathbb{R}^2$. Na slici je graf funkcije f koja minimizira sumu kvadrata reziduala od Y. *Izvor:* [3]

2.3 Ocjena preciznosti procjene parametara

Prilagodbom linearnog modela metodom najmanjih kvadrata, dobili smo jednostavan način za predviđanje odziva Y na temelju prediktora $X_1, X_2, \ldots, X_p$. Budući da su $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ procjene za $\beta_1, \beta_2, \ldots, \beta_p$, hiperravnina dobivena metodom najmanjih kvadrata

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$

samo je procjena za

$$f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

te želimo izračunati *intervale pouzdanosti* za parametre $\beta_1, \beta_2, \ldots, \beta_p$ kako bismo ocijenili preciznost procjene, odnosno koliko je $\hat{Y}$ blizu *stvarnoj* $f(X)$. Nepreciznost u procjeni parametara je povezana s *reducibilnom greškom* spomenutom u prvom poglavlju.

Napomena 2.3.1. *Ne zaboravimo da je i sam linearan model, odnosno pretpostavka o linearnosti od f, samo aproksimacija stvarnosti, stoga je i to dodatan izvor potencijalne reducibilne greške koji nazivamo* pristranost modela*.*

Prije nego što opišemo svojstva dobivene procjene parametara, podsjećamo se pretpostavki o $\mathbf{X}$ i ε te iznosimo glavne rezultate kao posljedice tih pretpostavki. Pretpostavljamo sljedeće:

- ε_i, $i = 1, \ldots, n$ su nezavisne normalno distribuirane slučajne varijable s očekivanjem 0 i konstantnom varijancom σ^2 tj. $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n] \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$
- $\mathbf{X}$ je neslučajna matrica s linearno nezavisnim stupcima.
- model za slučajan uzorak za odziv dan je s $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$.

Napomena 2.3.2. *Osim zbog garancije jedinstvenosti procjene za koeficijente linearne regresije, linearnu nezavisnost stupaca matrice* $\mathbf{X}$ *pretpostavljamo i kako bismo mogli razlikovati zasebne utjecaje p prediktora.*

Dobiveni procjenitelj za β je nepristran, odnosno vrijedi $\mathbb{E}[\hat{\beta}] = \beta$. Zaista,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\
&= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \varepsilon)] \\
&= \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbb{E}[\varepsilon] \\
&= \beta.
\end{aligned}$$

Nadalje, kovarijacijska matrica od $\hat{\beta}$ jednaka je

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])^T] \\
&= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\varepsilon\varepsilon^T]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned}$$

Uobičajeno se za procjenitelj od σ^2 uzima

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

gdje je $n - p - 1$ u nazivniku kako bi $\hat{\sigma}^2$ bio nepristran procjenitelj za σ^2, $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$. Vrijedi $(n-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p-1}$ te $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. Također, pretpostavljamo da su $\hat{\beta}$ i $\hat{\sigma}^2$ nezavisni.

Napomena 2.3.3. *Pretpostavka o normalnoj distribuciji od ε je, jasno, teoretska. Čak i da ona nije ispunjena, možemo očekivati da će $\hat{\beta}$ imati približno normalnu distribuciju. Što je veći kardinalitet n skupa za trening, aproksimacija normalnom distribucijom je točnija, što je posljedica centralnog graničnog teorema.*

Iskazana svojstva sada koristimo kako bismo ocijenili preciznost procjene parametara. Izolirajući β_j iz $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, dobivamo $1 - 2\alpha$ pouzdan interval za β_j:

$$\left(\hat{\beta}_j - z^{1-\alpha}v_j^{\frac{1}{2}}\hat{\sigma}, \hat{\beta}_j + z^{1-\alpha}v_j^{\frac{1}{2}}\hat{\sigma}\right)$$

gdje je v_j j–ti dijagonalni element matrice $(\mathbf{X}^T\mathbf{X})^{-1}$, a $z^{1-\alpha}$ je $(1-\alpha)$– percentil normalne distribucije, na primjer $z^{1-0.025} = 1.96$. Kritično područje za vektor $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ jednako je:

$$C_\beta = \{\beta \mid (\hat{\beta}-\beta)^T\mathbf{X}^T\mathbf{X}(\hat{\beta}-\beta) \leq \hat{\sigma}^2\chi^2_{p+1}{}^{(1-\alpha)}\}$$

gdje je $\chi_l^{2(1-\alpha)}$ $(1-\alpha)$–percentil hi-kvadrat distribucije sa l stupnjeva slobode. Na primjer, $\chi_5^{2(1-0.05)} = 11.1$.

Sjetimo se, čak i kad bismo znali točan oblik funkcije r, zbog člana slučajne greške ε u modelu $Y = r(X) + \varepsilon$, ne bismo u točnosti mogli predvidjeti vrijednost varijable odziva. Takvu grešku smo u prvom poglavlju nazvali *ireducibilnom greškom*. Da bismo odredili koliko se Y razlikuje od $\hat{Y}$ koristimo *intervale predikcije* koji nam kažu u kojem rasponu možemo očekivati vrijednost odziva za dosad neopaženu opservaciju. Interval predikcije je uvijek širi od intervala pouzdanosti jer obuhvaća i grešku u procjeni od r (reducibilnu grešku) kao i slučajnu grešku ε (ireducibilnu grešku). Interval predikcije daje raspon za Y, dok interval pouzdanosti daje raspon za $\mathbb{E}[Y \mid X = x]$.

2.4 Ocjena preciznosti linearnog modela

U ovom odjeljku želimo ocijeniti preciznost linearnog modela, odnosno koliko je *dobra* prilagodba linearnog modela.

Prije toga, zanima nas postoji li uopće veza između prediktora i varijable odziva. Kako bismo odgovorili na to pitanje, testiramo nultu hipotezu

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0$$

u odnosu na alternativnu hipotezu

$$H_a : \text{postoji } \beta_j \neq 0$$

Testna statistika je F-statistika

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

gdje je $\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ i $\text{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. Ukoliko su pretpostavke linearnog modela točne, može se pokazati da vrijedi $\mathbf{E}[\text{RSS}/(n-p-1)] = \sigma^2$, a ako ne odbacujemo H_0 tada je $\mathbf{E}[(\text{TSS} - \text{RSS})/p] = \sigma^2$. Stoga, kad ne postoji veza između varijable odziva i prediktora očekujemo vrijednost F-statistike blizu 1. S druge strane, ukoliko odbacujemo H_0, tada je $\mathbf{E}[(\text{TSS} - \text{RSS})/p] > \sigma^2$ pa očekujemo vrijednost F-statistike veću od 1. Naravno, zaključke testa provodimo na temelju računa kritičnog područja ili p–vrijednosti za određenu razinu značajnosti. Ponekad želimo testirati jesu li vrijednosti parametara nekog podskupa parametara jednake 0. Preciznije, testiramo nultu hipotezu da je reducirani model dovoljan

$$H_0 : \beta_{(p-q+1)} = \beta_{(p-q+2)} = \ldots = \beta_p = 0$$

u odnosu na alternativnu hipotezu da je potreban puni model. Reducirani model koristi sve varijable *osim* tih q. Neka je RSS_0 suma kvadrata reziduala reduciranog modela. Testna statistika je F-statistika dana sa

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}$$

Nakon što smo odbacili nultu hipotezu u korist alternativne, možemo ocijeniti koliko je dobra prilagodba linearnog modela (*eng. goodness-of-fit*). Dvije najčešće korištene mjere za ocjenu preciznosti linearnog modela su *standardna greška reziduala*, u oznaci RSE (*eng. residual standard error*), i *koeficijent determinacije*, u oznaci R^2.

Standardna greška reziduala

Standardna greška reziduala je procjena za standardnu devijaciju od ε. Definira se kao

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Ukoliko su predikcije dobivene prilagodbom linearnog modela, $\hat{y}_i$, blizu stvarnih opaženih vrijednosti, y_i, odnosno $\hat{y}_i \approx y_i$, za $i = 1, \ldots, n$, tada će standardna greška reziduala biti mala i možemo zaključiti da je prilagodba linearnog modela dosta dobra. S druge strane, ako je $\hat{y}_i$ jako daleko od y_i, za jednu ili više opservacija, tada RSE može biti jako velik indicirajući da prilagodba linearnog modela za dani skup za trening nije dobra.

Koeficijent determinacije

Koeficijent determinacije predstavlja alternativnu mjeru prilagodbe modela. Označavamo ga s R^2 i računamo po formuli

$$\mathrm{R}^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}.$$

TSS mjeri ukupno odstupanje od aritmetičke sredine opaženih odziva, dok RSS mjeri koji dio odziva nije objašnjen modelom. Tada $\frac{\mathrm{RSS}}{\mathrm{TSS}}$ daje postotak *varijabilnosti* u Y koji nije objašnjen modelom, a R^2 onda upravo mjeri postotak varijabilnosti u Y koji jest objašnjen modelom. Zbog toga vrijednost R^2 blizu 1 znači da je velik dio varijabilnosti u odzivu objašnjen linearnim regresijskim modelom. Vrijednost blizu nule najčešće ukazuje na to da linearan model nije dobar za dane podatke. Što smatramo *dobrom* vrijednosti statistike R^2 ovisi o primjeni i konkretnom problemu.

Poglavlje 3

Odabir i regularizacija linearnog modela

U prethodnom poglavlju prilagodbu linearnog modela proveli smo koristeći metodu najmanjih kvadrata kao jednu od najčešće korištenih metoda za tu svrhu. U ovom poglavlju predstavljamo dvije važne klase metoda prilagodbe linearnog modela kao alternativu metodi najmanjih kvadrata:

- odabir podskupa - ovaj pristup uključuje odabir p prediktora za koje se smatra da su najviše povezani s odzivom te se prilagodba linearnog modela provodi na temelju reduciranog skupa parametara
- regularizacija - kod metoda regularizacije uvodi se *parametar regularizacije* λ kojim se nastoji smanjiti velike koeficijente β u linearnom modelu s ciljem pojednostavljivanja modela i rješavanja određenih problema statističkog učenja.

Gornji pristupi predstavljaju unaprjeđenja linearnog modela s obzirom na interpretabilnost i preciznost predikcija. Prije nego što ih detaljno opišemo, objasnit ćemo odnos pristranosti i varijance te ćemo, budući da želimo odabrati *najbolji* model, objasniti metodu *unakrsne validacije* - metodu na temelju koje se često u praksi procjenjuje *očekivana testna greška* te bira najbolji model.

3.1 Pristranost, varijanca i kompleksnost modela

U prvom poglavlju uveli smo pojam funkcije gubitka i kao najčešći izbor u slučaju kvantitativnog odziva naveli smo funkciju $L\colon \mathbb{R}^2 \to [0, +\infty)$ definiranu sa $L((y_1, y_2)) = (y_1 - y_2)^2$. Tada smo kvadratnu grešku od $f(X)$ definirali kao $L((Y, f(X))) = (Y - f(X))^2$. Sada pomoću tih pojmova želimo precizno definirati dva važna pojma - *testnu grešku* i *očekivanu testnu grešku*.

Napomena 3.1.1. *Testna greška u literaturi se naziva još i greška predikcije, greška generalizacije, greška generalizacije na nezavisnom testnom skupu, stvarna testna greška i slično.*

Definicija 3.1.2. *Testna greška definirana je sa*

$$Err_{\mathcal{T}} = \mathbb{E}\left[L(Y, f_{\mathcal{T}}(X))\right]$$

gdje je $\mathcal{T}$ fiksan skup za trening, a (X, Y) nezavisan i jednakodistribuiran kao slučajan uzorak za skup za trening.

Naglasimo, u gornjoj definiciji (X, Y) je slučajan te predstavlja dosad neopažene podatke, a skup za trening $\mathcal{T}$ je fiksan te se pojam testne greške odnosi na grešku predikcije s obzirom na taj konkretan skup koji je korišten za dobivanje f. Testnu grešku interpretiramo kao prosječnu grešku predikcije modela na novim podacima koji dolaze iz iste distribucije (i nezavisni su) kao i slučajan uzorak za skup za trening.

Pri definiciji skupa za trening u prvom poglavlju rada spomenuli smo kako su elementi skupa za trening ustvari realizacije slučajnog uzorka za slučajan vektor (X, Y). Na taj slučajan uzorak gledamo kao na uređenu $n-$torku slučajnih pokusa. Neka je $(\Omega_j, \mathcal{F}_j, \mathbb{P}_j)$ vjerojatnosni prostor koji je model $j-$toga slučajnog pokusa ($j = 1, 2 \ldots, n$). Prirodni prostor elementarnih događaja tog slučajnog pokusa je skup $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$. Nadalje, s $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \cdots \times \mathcal{F}_n$ označavamo $\sigma-$algebru koju zovemo produkt $\sigma-$algebri $\mathcal{F}_1, \ldots, \mathcal{F}_n$. Može se pokazati da postoji jedinstvena vjerojatnosna mjera $\mathbb{P}$ na $(\Omega, \mathcal{F})$ takva da vrijedi $\mathbb{P}(A_1 \times A_2 \cdots \times A_n) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2) \ldots \mathbb{P}_n(A_n)$, $A_j \in \mathcal{F}_j$, $j = 1, \ldots, n$. Vjerojatnosni prostor $(\prod_{j=1}^n \Omega_j, \prod_{j=1}^n \mathcal{F}_j, \prod_{j=1}^n \mathbb{P}_j)$ zovemo produkt vjerojatnosnih prostora $(\Omega_j, \mathcal{F}_j, \mathbb{P}_j)$, $j = 1, \ldots, n$.

Na slučajan uzorak za skup za trening možemo gledati kao na slučajan element na produktu vjerojatnosnih prostora. Njegove različite realizacije u pravilu će rezultirati drugačijom funkcijom $f_{\mathcal{T}}$, dakle ona je u ovisnosti o slučajnom uzorku za skup za trening slučajna i time također slučajan element na istom vjerojatnosnom prostoru. [1] Sada možemo uvesti definiciju očekivane testne greške.

Definicija 3.1.3. *Očekivana testna greška definirana je sa*

$$Err = \mathbb{E}^n\left[Err_{\mathcal{T}}\right]$$

gdje je $Err_{\mathcal{T}}$ prethodno definirana testna greška, $\mathbb{E}^n$ očekivanje na produktnom vjerojatnosnom prostoru $(\prod_{j=1}^n \Omega_j, \prod_{j=1}^n \mathcal{F}_j, \prod_{j=1}^n \mathbb{P}_j)$, a slučajnost dolazi od slučajnog odabira skupa za trening $\mathcal{T}$.

[1] Drugim riječima, za različite (slučajne) skupove za trening dobije se različita funkcija f - na primjer, u linearnom modelu dobiju se različiti parametri $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. Radi se o jednom modelu (linearnom modelu), ali o različitim dobivenim linearnim funkcijama.

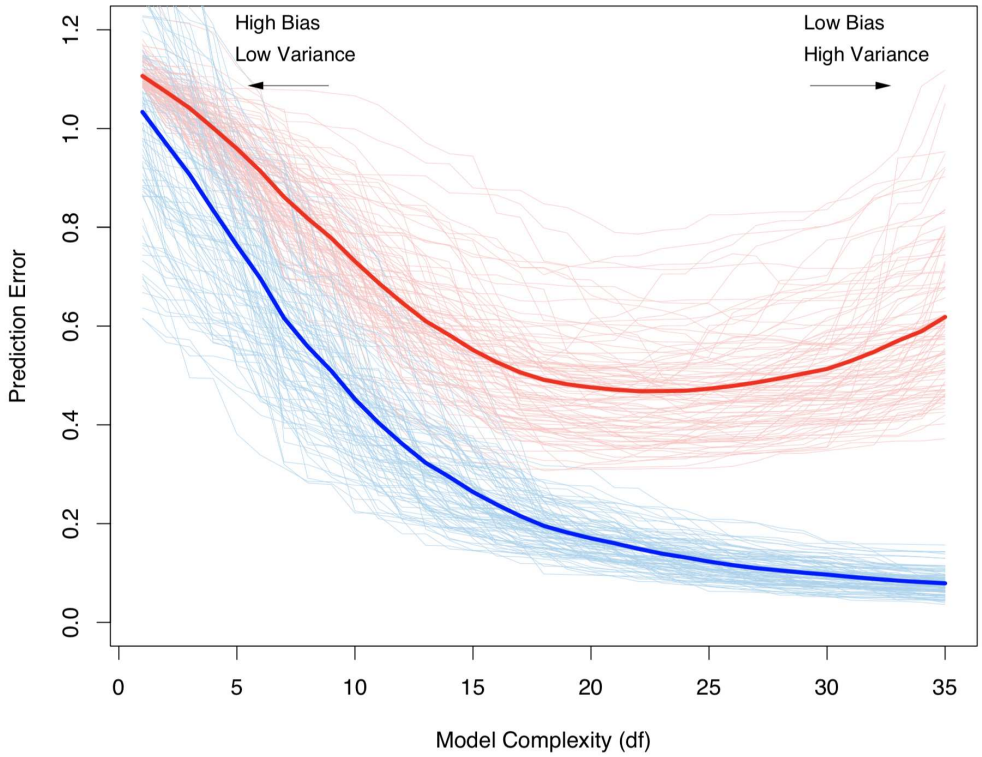
U statističkom učenju cilj je procijeniti testnu grešku tj. grešku funkcije f koju smo dobili za jedan konkretan skup za trening. Ispostavlja se da to općenito nije lagano stoga većina metoda statističkog učenja ustvari procjenjuje *očekivanu testnu grešku*.

Napomena 3.1.4. *Napomenimo da općenito oba pojma, testna greška, $Err_{\mathcal{T}}$, i očekivana testna greška $\mathbb{E}^n[Err_{\mathcal{T}}]$, mogu biti od interesa, ovisno o pogledu. Testna greška nam govori kako se model dobiven pomoću konkretnog skupa za trening ponaša na novim podacima, prosječno. Očekivana testna greška pak gleda očekivanje te greške s obzirom na razne (sada slučajne) skupove za trening. Statističko učenje više se bavi prilagodbom modela i njegovom ocjenom, stoga je od većeg interesa testna greška. Teoretičare, s druge strane, može više zanimati očekivana testna greška jer ih ne zanima jedan dobiveni model, već općenito ponašanje modela.*

Definicija 3.1.5. *Greška predikcije na skupu za trening $\mathcal{T} = \{(x_i, y_i) : i = 1, \ldots, n\}$ naziva se* greška skupa za trening *i definira se kao*

$$\overline{err} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$$

U slučaju kad za funkciju gubitka koristimo funkciju $L : \mathbb{R}^2 \to [0, +\infty)$ definiranu sa $L((y_1, y_2)) = (y_1 - y_2)^2$, greška skupa za trening često se označava sa MSE (eng. *mean squared error*). Svjetlo crvene krivulje na sljedećoj slici prikazuju testnu grešku $\mathrm{Err}_{\mathcal{T}}$ metode statističkog učenja (konkretno *LASSO*) za 100 simuliranih skupova podataka $\mathcal{T}$ duljine 50. Debela crvena krivulja predstavlja njihov prosjek tj. Err. Svjetlo plave krivulje prikazuju greške predikcije na skupovima za trening, $\overline{\mathrm{err}}$. Debela plava krivulja predstavlja očekivanu grešku predikcije na skupu za trening, dakle $\mathbb{E}[\overline{\mathrm{err}}]$.



Slika 3.1: Ponašanje testne greške, $\mathrm{Err}_{\mathcal{T}}$, i greške predikcije na skupu za trening, $\overline{\mathrm{err}}$, u ovisnosti o kompleksnosti modela. *Izvor:* [2]

Možemo zaključiti kako greška za trening neće uvijek biti dobra procjena testne greške budući da greška za trening monotono pada kako kompleksnost modela raste, dok kod testne greške uočavamo U-oblik pripadne krivulje. Što je veća kompleksnost metode statističkog učenja (modela), manja je njezina pristranost, ali veća varijanca. Idealno, ne bismo htjeli da dobivena funkcija f daje jako različite rezultate za različite skupove za trening. Kada je to slučaj, kažemo da model ima veliku varijancu. Općenito, fleksibilniji/kompleksniji modeli imaju veću varijancu. Pristranost statističkog modela odnosi se na grešku koja proizlazi iz pretpostavke o vezi između odziva i kovarijata. Linearni model primjer je nefleksibilnog modela za koji očekujemo visoku pristranost. Fleksibilniji/kompleksniji modeli imaju manju pristranost. Dakle, varijanca i pristranost dva su oprečna svojstva. Odnos pristranosti i varijance u statističkom učenju naziva se *bias-variance trade-off*.

U Lemi 1.2.8. pokazali smo kako je regresijska funkcija, r, najbolja procjena za Y uz dano $X = x$ i danu funkciju gubitka. Sjetimo se da je regresijska funkcija teoretski koncept te statističko učenje na temelju skupa za trening nastoji dati najbolje metode za njezinu

procjenu koju kroz rad označavamo s f.

Lema 3.1.6. *Vrijedi*

$$\begin{aligned} Err &= \mathbb{E}^n[\mathbb{E}[(Y - f_{\mathcal{T}}(X))^2]] \\ &= \mathbb{E}[(\mathbb{E}^n[f_{\mathcal{T}}(X)] - r(X))^2] + \mathbb{E}[\mathbb{E}^n[(f_{\mathcal{T}}(X) - \mathbb{E}^n[f_{\mathcal{T}}(X)])^2]] + Var(\varepsilon) \end{aligned}$$

gdje je (X, Y) *nezavisan od* $\mathcal{T}$.

Dokaz. Za proizvoljan (fiksan, dosad neviđen) $x \in \mathbb{R}^p$ vrijedi

$$\mathbb{E}^n[\mathbb{E}[(Y - f_{\mathcal{T}}(X))^2 | X = x]] = \mathbb{E}^n[\mathbb{E}[(r(X) + \varepsilon - f_{\mathcal{T}}(X))^2 | X = x]] \tag{3.1}$$

$$= \mathbb{E}^n[\mathbb{E}[(r(x) + \varepsilon - f_{\mathcal{T}}(x))^2]] \tag{3.2}$$

$$= \mathbb{E}^n[\mathbb{E}[(r(x) - f_{\mathcal{T}}(x))^2]] + \mathbb{E}^n[\mathbb{E}[\varepsilon^2]] + 2\mathbb{E}^n[\mathbb{E}[(r(x) - f_{\mathcal{T}}(x))\varepsilon]] \tag{3.3}$$

$$= \mathbb{E}^n[(r(x) - f_{\mathcal{T}}(x))^2] + \underbrace{\mathbb{E}[\varepsilon^2]}_{=\mathrm{Var}(\varepsilon)} + 2\mathbb{E}^n[(r(x) - f_{\mathcal{T}}(x)) \underbrace{\mathbb{E}[\varepsilon]}_{=0}] \tag{3.4}$$

$$= \mathbb{E}^n[(r(x) - f_{\mathcal{T}}(x))^2] + \mathrm{Var}(\varepsilon) \tag{3.5}$$

gdje smo u drugoj jednakosti koristili nezavisnost slučajne greške ε od X, a u daljnjem računu linearnost matematičkog očekivanja te činjenicu da je $r(x) - f_{\mathcal{T}}(x)$ konstanta za očekivanje $\mathbb{E}$.
Pogledajmo sada kako možemo dalje raščlaniti $\mathbb{E}^n[(r(x) - f_{\mathcal{T}}(x))^2]$:

$$\mathbb{E}^n[(r(x) - f_{\mathcal{T}}(x))^2] = \mathbb{E}^n[((r(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)]) - (f_{\mathcal{T}}(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)]))^2] \tag{3.6}$$

$$= \mathbb{E}^n[(\mathbb{E}^n[f_{\mathcal{T}}(x)] - r(x))^2] + \mathbb{E}^n[(f_{\mathcal{T}}(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])^2] \tag{3.7}$$

$$- 2\mathbb{E}^n[(r(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])(f(_{\mathcal{T}}x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])] \tag{3.8}$$

$$= \underbrace{(\mathbb{E}^n[f_{\mathcal{T}}(x)] - r(x))^2}_{\text{pristranost od } f_{\mathcal{T}}(x)} + \underbrace{\mathbb{E}^n[(f_{\mathcal{T}}(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])^2]}_{\text{varijanca od } f_{\mathcal{T}}(x)} \tag{3.9}$$

$$- 2(r(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])\mathbb{E}^n[(f_{\mathcal{T}}(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])] \tag{3.10}$$

$$= (\mathbb{E}^n[f_{\mathcal{T}}(x)] - r(x))^2 + \mathbb{E}^n[(f_{\mathcal{T}}(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])^2] \tag{3.11}$$

$$- 2(r(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])(\mathbb{E}^n[f_{\mathcal{T}}(x)] - \mathbb{E}^n[f_{\mathcal{T}}(x)]) \tag{3.12}$$

$$= (\mathbb{E}^n[f_{\mathcal{T}}(x)] - r(x))^2 + \mathbb{E}^n[(f_{\mathcal{T}}(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])^2] \tag{3.13}$$

gdje smo na više mjesta koristili da je $r(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)]$ konstanta te linearnost matematičkog očekivanja. Uvrštavanjem (3.13) u (3.5) dobivamo:

$$\mathbb{E}^n[\mathbb{E}[(Y - f_{\mathcal{T}}(X))^2 | X = x]] = (\mathbb{E}^n[f_{\mathcal{T}}(x)] - r(x))^2 + \mathbb{E}^n[(f_{\mathcal{T}}(x) - \mathbb{E}^n[f_{\mathcal{T}}(x)])^2] + \mathrm{Var}(\varepsilon)$$

te uzimanjem očekivanja $\mathbb{E}$ s obje strane slijedi:

$$\mathbb{E}^n[\mathbb{E}[(Y - f_{\mathcal{T}}(X))^2]] = \mathbb{E}[(\mathbb{E}^n[f_{\mathcal{T}}(X)] - r(X))^2] + \mathbb{E}[\mathbb{E}^n[(f_{\mathcal{T}}(X) - \mathbb{E}^n[f_{\mathcal{T}}(X)])^2]] + \text{Var}(\varepsilon)$$

što smo i htjeli dokazati. □

Jednakost iz iskaza leme često se naziva *dekompozicija očekivane testne greške*. Prvi član u gornjoj dekompoziciji očekivane testne greške u točki $x \in \mathbb{R}^p$ predstavlja ireducibilnu grešku, drugi član predstavlja kvadrat pristranosti, a treći član varijancu od $f(x)$. Iz gornje leme zaključujemo da kako bismo minimizirali očekivanu testnu grešku trebamo *pronaći* onaj model koji će istovremeno imati malu varijancu i malu pristranost.

Kompleksnost modela najčešće je određena parametrom koji označavamo α. Tipično želimo naći onu vrijednost parametra α koja daje model s minimalnom očekivanom testnom greškom.

Dvije glavne zadaće statističkog učenja su odabir modela i potom ocjena preciznosti dobivenog modela. U tu svrhu, uz snažnu pretpostavku da imamo dovoljno velik skup podataka, skup podataka možemo podijeliti na tri manja skupa: skup za trening koji koristimo za prilagodbu različitih modela, skup za validaciju koji koristimo kako bismo odredili koji model je najbolji i potom testni skup kojim procjenjujemo testnu grešku odabranog modela. Kako najčešće dostupan skup podataka nije dovoljno velik da bismo ga efikasno podijelili na tri dijela, postoje metode koje na drugi način izlaze tomu na kraj. Jedna takva metoda je unakrsna validacija koju opisujemo u sljedećem odjeljku.

3.2 Unakrsna validacija

Unakrsna validacija jedna je od najjednostavnijih i često korištenih metoda za procjenu očekivane testne greške. Kad bismo imali dovoljno velik skup podataka, tada bismo taj skup mogli podijeliti na dva dijela - skup za trening pomoću kojeg bismo prilagodili model i skup za validaciju koji bismo potom iskoristili za ocjenu dobivenog modela. Najčešće nemamo dovoljno velik skup podataka te unakrsna validacija nastoji tom problemu doskočiti uzorkovanjem skupa za trening.

K-struka unakrsna validacija

Ideja K-struke unakrsne validacije je podijeliti skup za trening na K dijelova približno jednake veličine i potom $K - 1$ dobivenih grupa iskoristiti za prilagobu modela, a preostalu jednu grupu za validaciju modela. Preciznije, neka je $\kappa : \{1, \ldots, n\} \to \{1, ..., K\}$ indeksna funkcija koja za dani element skupa podataka daje jednu od K grupa kojoj pripada. Procedura je sljedeća:

- podijeliti skup za trening na K dijelova
- za $k \in 1, \ldots, K$ prilagoditi model koristeći sve podatke osim $k-$te grupe podataka, za svaki k oznaka za pripadni dobiveni model neka je $f^{-k}(x)$
- procjena testne greške dana je sa

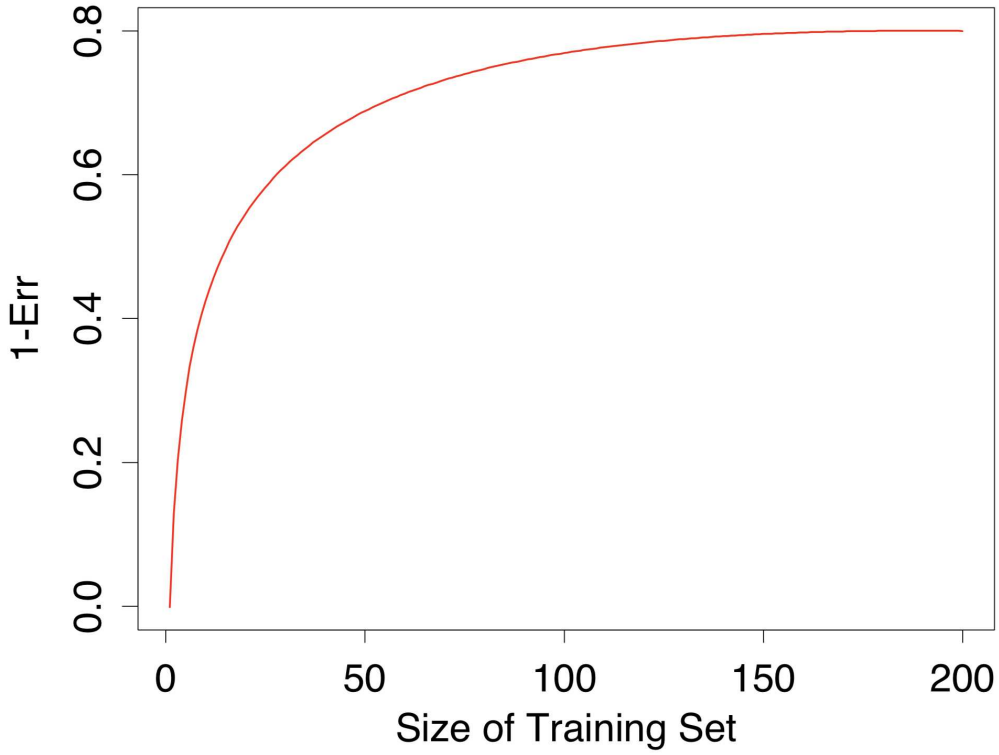
$$\mathrm{CV}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f^{-\kappa(i)}(x_i)).$$

Neka je $f(x, \alpha)$ niz modela različite fleksibilnosti koja je određena parametrom podešavanja α i neka je $f^{-k}(x, \alpha)$ model dobiven uz dani α i bez $k-$te grupe podataka. Definiramo

$$\mathrm{CV}(f, \alpha) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f^{-\kappa(i)}(x_i, \alpha))$$

Odabiremo onaj model $f(x, \hat{\alpha})$ za koji $\hat{\alpha}$ minimizira gore definiranu $\mathrm{CV}(f, \alpha)$ te potom prilagođavamo model $f(x, \hat{\alpha})$ na cijelom skupu za trening.
Uočimo kako smo upravo opisali na koji način koristimo unakrsnu validaciju za odabir modela one fleksibilnosti koja minimizira dobivenu procjenu testne greške. U tom slučaju ne zanima nas stvarna vrijednost procjene testne greške, već točka minimuma $\hat{\alpha}$ za dobivene procjene u ovisnosti o α. U praksi stvarnu testnu grešku ne znamo, no kod simuliranih podataka možemo ju izračunati i usporediti s procjenom dobivenom unakrsnom validacijom. Tada se pokazuje da unatoč tomu što procjena testne greške dobivena unakrsnom validacijom ponekad podcjenjuje testnu grešku, unakrsnom validacijom u većini slučajeva možemo dosta precizno odrediti optimalnu fleksibilnost modela. Također, imajmo na umu da unakrsna validacija uspješno procjenjuje samo *očekivanu* testnu grešku, s obzirom da se temelji na promjeni skupa za trening u svakom koraku algoritma.

Na kraju, postavlja se pitanje odabira K. Slučaj K = n u literaturi se često naziva *LOOCV - Leave One out Cross Validation* i on daje približno nepristranu procjenu testne greške s obzirom da je skup koji koristimo za prilagodbu približno jednak skupu za trening. Međutim, varijanca u tom slučaju može biti velika budući da svaki od K = n modela prilagođavamo na gotovo jednakim podacima. Dakle, pri odabiru K potrebno je voditi računa o odnosu pristranosti i varijance. Za K = 5 unakrsna validacija ima manju varijancu, ali ovisno o veličini skupa za trening velika pristranost može postati problem. Unakrsna validacija za procjenu koristi skupove približne veličine $\frac{\mathrm{K}-1}{\mathrm{K}}$n. Za n = 200 imamo $\frac{\mathrm{K}-1}{\mathrm{K}}$n = 160 i malu pristranost, dok za n = 50 imamo $\frac{\mathrm{K}-1}{\mathrm{K}}$n = 40 i veliku pristranost. Sljedeća slika prikazuje hipotetsku ovisnost $1 - \mathrm{Err}$ o veličini skupa za trening za K = 5. Ukoliko *krivulja učenja* ima velik nagib s obzirom na veličinu skupa za trening, tada će peterostruka unakrsna validacija precijeniti vrijednost stvarne testne greške. Kako god, K = 5 ili K = 10 smatraju se u praksi dobrim kompromisom s obzirom na odnos pristranosti i varijance.



Slika 3.2: Hipotetska ovisnost 1 − Err o veličini skupa za trening za K = 5. *Izvor:* [2]

3.3 Odabir podskupa

Uz pretpostavku da je stvarna veza između varijable odziva i prediktora linearna, metoda najmanjih kvadrata ima malu pristranost. Ukoliko je i broj opservacija n puno veći od broja varijabli p pokazuje se da ima i malu varijancu te će stoga imati i dobre performanse na prethodno neviđenim podacima. S druge strane, u slučaju kad n nije puno veći od p, varijanca modela dobivenog prilagodbom metode najmanjih kvadrata može biti velika rezultirajući lošim predikcijama za nove podatke. Smanjivanjem broj parametara p, može se znatno smanjiti varijanca bez prevelikog utjecaja na pristranost što dovodi do preciznijih predikcija za podatke koji nisu korišteni za prilagodbu modela. Također, uzimanjem manjeg podskupa prediktora za koje se pokazuje da imaju najveći utjecaj na odziv poboljšavamo interpretabilnost modela.

Odabir najboljeg podskupa

Kod metode najboljeg podskupa za svaki $k \in \{1, \ldots, p\}$ prilagođavamo model za sve moguće kombinacije k prediktora i biramo onaj koji minimizira RSS. Potom za sve dobivene modele biramo onaj koji je sveukupno najbolji s obzirom na neki kriterij . Najčešće biramo onaj koji minimizira procjenu očekivane testne greške. Algoritam je sljedeći:

- Neka je $\mathcal{M}_0$ oznaka za model koji ne sadrži prediktore. On za svaku opservaciju kao predikciju daje aritmetičku sredinu skupa za trening.
- Za k = $1, 2, \ldots, p$:
  - prilagodi $\binom{p}{k}$ modela koji koriste točno k prediktora
  - odaberi onaj od dobivenih $\binom{p}{k}$ modela koji minimizira RSS i označi ga s $\mathcal{M}_k$.
- Odaberi najbolji od dobivenih p modela $\mathcal{M}_0, \ldots, \mathcal{M}_p$ unakrsnom validacijom.

Uočimo da, na primjer, najbolji podskup duljine 2 ne mora uključivati varijablu prediktora koja je u najboljem podskupu duljine 1. Nadalje, u drugom koraku gornjeg algoritma bira se najbolji model za skup za trening za svaku moguću veličinu podskupa te se problem biranja 2^p modela svodi na problem biranja $p+1$ modela. Među tih $p+1$ modela bira se onaj koji minimizira očekivanu testnu grešku, a jedna od metoda kojom to možemo provesti je unakrsna validacija. Uočimo i da RSS monotono pada kako raste broj prediktora pa nju niti ne bi imalo smisla koristiti kao kriterij jer bismo u tom slučaju uvijek birali sve prediktore. Razlog naravno leži u tomu što na temelju RSS biramo model s najmanjom greškom na skupu za trening dok nam je od interesa model s najmanjom testnom greškom. Problem metode odabira najboljeg podskupa je što za p veći od 40 postaje računarski neizvediv čak i za jako brza moderna računala. Stoga, u narednom odjeljku predstavljamo računarski povoljnije metode.

Postupni odabir prediktora

Kao što smo spomenuli, metoda odabira najboljeg podskupa ne može se primijeniti kad je p jako velik. Također, kad je p velik zbog iznimno velikog skupa modela koji promatramo (2^p) puno je veća vjerojatnost overfitting-a i velike varijance dobivenih procjena za koeficijente. Čak i kad je računarski izvedivo provesti metodu odabira najboljeg podskupa, ponekad zbog velike pripadne varijance to neće biti najbolji pristup. Metode postupnog odabira prediktora nastoje naći najbolji model u znatno manjem skupu modela u odnosu na metodu odabira najboljeg podskupa. U nastavku objašnjavamo metodu postupnog odabira prediktora unaprijed i postupnog odabira prediktora unatrag koje se razlikuju u tome krećemo li od modela bez prediktora ili od modela koji uključuje sve varijable prediktora.

Postupni odabir prediktora unaprijed

Ideja metode odabira postupnog prediktora unaprijed je početi s modelom bez prediktora te postupno dodavati u model jedan po jedan prediktor. U svakom koraku dodaje se ona varijabla prediktora koja najbolje poboljšava prilagodbu modela. Pritom najbolje mjerimo sa RSS. Algoritam glasi:

- Neka je $\mathcal{M}_0$ model bez prediktora.
- Za svaki k = $0, \ldots, p-1$
  - Napravi prilagodbu svih $p-k$ modela koji dodaju u model $\mathcal{M}_k$ jedan dodatni prediktor.
  - Odaberi najbolji (RSS) od tih $p-k$ modela i označi ga s $\mathcal{M}_{k+1}$
- Odaberi najbolji model među $\mathcal{M}_0, \ldots, \mathcal{M}_p$ koristeći unakrsnu validaciju.

Dok metoda odabira najboljeg podskupa za $p = 20$ zahtijeva prilagodbu 1048576 modela, metoda postupnog odabira prediktora unaprijed zahtijeva prilagodbu 211 modela. Metoda postupnog odabira prediktora unaprijed je *pohlepan algoritam* jer u svakom koraku (dakle, lokalno) bira najbolji model s ciljem pronalaska globalno najboljeg modela. U tom pogledu može se činiti suboptimalan u odnosu na odabir najboljeg podskupa. Na primjer, pretpostavimo da za skup podataka sa $p = 3$ prediktora, najbolji model s jednim prediktorom sadrži X_1, dok najbolji model s dva prediktora sadrži X_2 i X_3. Tada metoda postupnog odabira prediktora unaprijed neće davati najbolji model budući da model $\mathcal{M}_1$ sadrži X_1 pa i model $\mathcal{M}_2$ mora sadržavati X_1. Ipak, zbog računarske prednosti i manje varijance, ona se preferira.

Postupni odabir prediktora unatrag

Za razliku od metode postupnog odabira prediktora unaprijed, metoda postupnog odabira prediktora unatrag kreće od punog modela te potom, korak po korak, odbacuje onu varijablu prediktora koja najmanje utječe na odziv. Kako bismo mogli provesti prilagodbu punog modela, ovdje je nužno da duljina skupa za trening n bude veća od broja varijabli p.

- Neka je $\mathcal{M}_p$ model koji sadrži svih p prediktora.
- Za svaki k = $p, p-1, \ldots, 1$
  - Napravi prilagodbu svih k modela koji sadrže sve prediktore osim jednog iz modela $\mathcal{M}_k$, dakle ukupno $k-1$ prediktora.
  - Odaberi najbolji (RSS) od tih k modela i označi ga s $\mathcal{M}_{k-1}$
- Odaberi najbolji model među $\mathcal{M}_0, \ldots, \mathcal{M}_p$ koristeći unakrsnu validaciju.

3.4 Metode regularizacije: Ridge i LASSO regresija

Odbacivanjem dijela prediktora metode odabira podskupa daju model koji je interpretabilan te ima potencijalno manju grešku predikcije nego puni model. Ipak, budući da se radi o diskretnom procesu gdje se varijable prediktora zadržavaju ili odbacuju, često imaju veliku varijancu te ne smanjuju grešku predikcije punog modela. Ideja metoda regularizacija je koristiti svih p prediktora te penalizirati velike koeficijente s ciljem pronalaska jednostavnijeg modela, a koji i dalje daje sve potrebne informacije o odzivu. Pritom se pokazuje da se i varijanca smanjuje.

Sjetimo se, funkcija gubitka za linearnu regresiju dana je sa $\mathcal{L} = \frac{1}{n}\|y - \mathbb{X}\beta\|^2$. Kod metoda regularizacije uvodi se dodatan član te ona izgleda ovako:

$$\mathcal{L} = \frac{1}{n}\|y - \mathbb{X}\beta\|^2 + \lambda R(\beta)$$

Dodatni član $\lambda R(\beta)$ naziva se *član penalizacije*, a $\lambda \in [0, +\infty)$ je *parametar podešavanja*. Skup svih procjenitelja $\{\hat{\beta}(\lambda) : \lambda \in [0, +\infty)\}$ naziva se *regularizacijski put* procjenitelja.

Najčešći izbori za $R(\beta)$ su:

- $R(\beta) = \|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2$
- $R(\beta) = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

Prvi izbor naziva se *ridge* regresija, a drugi izbor *lasso* regresija.

Ridge regresija

Koeficijenti dobiveni ridge regresijom minimiziraju tzv. penaliziranu RSS i dani su sa:

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

Ekvivalentni način zapisa ridge regresije je:

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

$$\sum_{j=1}^{p} \beta_j^2 \leq t$$

Kad je $\lambda = 0$ ridge regresija ekvivalentna je linearnoj regresiji uz metodu najmanjih kvadrata. Što je λ veći to je veći utjecaj člana penalizacije te su procjene koeficijenata dobivene ridge regresijom sve bliže nuli.

Formulacija problema ridge regresije najčešće je takva da ne uključuje penaliziranje koeficijenta β_0. Prije provođenja ridge regresije stupci matrice $\mathbf{X}$ se centriraju tj. svaki x_{ij} zamijeni se sa $x_{ij} - \bar{x}_j$. Tada je procjena za β_0 jednaka $\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$.

Problem ridge regresije u matričnom obliku glasi:

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$

te je rješenje dano u obliku

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

gdje je $\mathbf{I}$ jedinična $p \times p$ matrica.
Tradicionalni opisi ridge regresije počinju upravo s gornjom definicijom procjenitelja. Proučimo dva primjera.

Primjer 3.4.1. *Pogledajmo matricu*

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 1 & 0 \end{pmatrix}$$

Kako je prvi stupac matrice suma druga dva stupca, rang matrice $\mathbf{X}$ *jednak je 2. Općenitije, ako* $n \times p$ *matrica* $\mathbf{X}$ *nema puni (stupčani) rang, tada ni* $p \times p$ *matrica* $\mathbf{X}^T\mathbf{X}$ *nema puni rang. Naime, ako je rang matrice* $\mathbf{X}$ *manji od p tada postoji netrivijalan vektor* $\mathbf{v} \in \mathbf{R}^p$ *takav da je* $\mathbf{Xv} = \mathbf{0}$*. Množenjem s* $\mathbf{X}^T$ *slijedi da je* $\mathbf{X}^T\mathbf{Xv} = \mathbf{0}$*. Kako je* $\mathbf{v} \neq \mathbf{0}$*, zaključujemo da je i* $\mathbf{X}^T\mathbf{X}$ *singularna.*

Sjetimo se definicije procjenitelja metodom najmanjih kvadrata, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Vidimo da je dani procjenitelj dobro definiran ako i samo ako inverz $(\mathbf{X}^T\mathbf{X})^{-1}$ postoji. Hoerl i Kennard 1970. kao ad-hoc rješenje za singularnost od $\mathbf{X}^T\mathbf{X}$ predlažu da se $\mathbf{X}^T\mathbf{X}$ zamijeni s $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp}$. To rješava problem singularnosti budući da se pozitivna matrica $\lambda\mathbf{I}_{pp}$ dodaje pozitivnoj semidefinitnoj matrici $\mathbf{X}^T\mathbf{X}$, što sve skupa čini pozitivno definitnu matricu, a takve matrice su invertibilne.

Primjer 3.4.2. *Pogledajmo sada matricu* $(\mathbf{X}^T\mathbf{X})^{-1}$ *i dodajmo joj* $\lambda = 1$

$$\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp} = \begin{pmatrix} 5 & 2 & 2 \\ 2 & 7 & -4 \\ 2 & -4 & 7 \end{pmatrix}$$

Lako se vidi da su njezine svojstvene vrijednosti 11, 7 i 1. Stoga, $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp}$ *nema svojstvenih vrijednosti jednakih nuli pa postoji njezin inverz.*

U slučaju kad matrica $\mathbf{X}^T\mathbf{X}$ ima puni rang te vrijedi $p < n$, postoji linearna veza između ridge procjenitelja i procjenitelja metodom najmanjih kvadrata. Definirajmo linearni operator $\mathbf{W}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{X}$. Tada se ridge procjenitelj $\hat{\beta}(\lambda)$ može prikazati kao $\mathbf{W}_\lambda\hat{\beta}$. Procjena za odziv sada je dana analogno kao u slučaju metode najmanjih kvadrata:

$$\hat{\mathbf{Y}}(\lambda) = \mathbf{X}\hat{\beta}(\lambda) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}(\lambda)\mathbf{Y}.$$

Kod prilagodbe metodom najmanjih kvadrata procjena za odziv $\mathbf{Y}$ ortogonalna je projekcija na potprostor razapet stupcima od $\mathbf{X}$. To znači da je procjena za odziv dobivena metodom najmanjih kvadrata točka u prostoru kovarijata najbliža opservaciji. Drugim riječima, u prostoru kovarijata ne nalazi se točka koja *bolje* (u danom smislu) objašnjava odziv. To je prikazano na sljedećoj slici zajedno sa crvenom isprekidanom linijom koja prikazuje prilagodbu ridge regresijom. Prilagodba ridge regresijom parametrizirana je sa $\{\lambda : \lambda \in [0, +\infty)\}$ gdje je svaka točka na tom pravcu presjek regularizacijskog puta $\hat{\beta}(\lambda)$ i okomitog pravca $x = \lambda$. Prilagodba ridge regresijom $\hat{Y}(\lambda)$ ide od procjene $\hat{Y} = \hat{Y}(0)$ ka nul-modelu u kojem kovarijate ne objašnjavaju odziv. Jasno je da za svaki $\lambda > 0$ ridge procjena $\hat{Y}(\lambda)$ neće biti ortogonalna projekcija od Y. U tom smislu, ridge prilagodba ne objašnjava najbolje odziv, ali rješava problem singularnosti.

Već smo spomenuli kako procjene parametara dobivene ridge regresijom konvergiraju u nulu kako λ teži beskonačnosti neovisno o skupu podataka. Promotrimo očekivanje ridge procjenitelja.

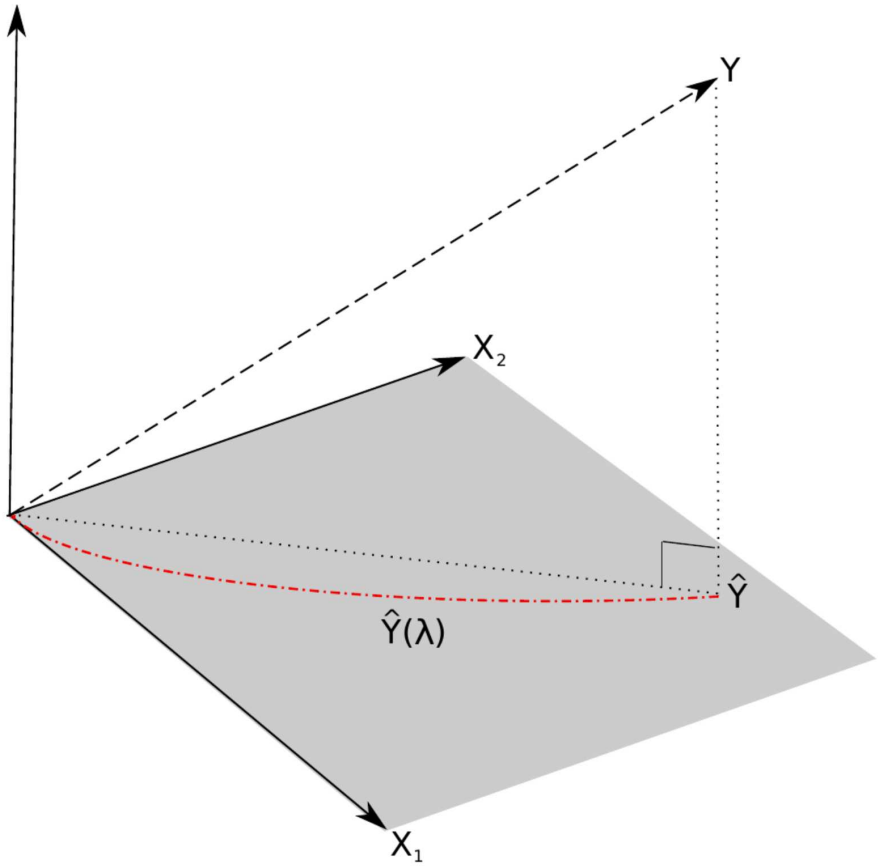
$$\begin{aligned}\mathbb{E}\left[\hat{\beta}(\lambda)\right] &= \mathbb{E}\left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y}\right] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbb{E}\left[\mathbf{Y}\right] \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\beta.\end{aligned}$$

Očito, $\mathbf{E}[\hat{\beta}(\lambda)] \neq \beta$ za svaki $\lambda > 0$. Dakle, ridge procjenitelj je pristran. Nadalje, uz pretpostavku linearne veze između procjenitelja metodom najmanjih kvadrata i ridge procjenitelja, za varijancu ridge procjenitelja dobivamo sljedeći izraz:

$$\begin{aligned}\mathrm{Var}\left[\hat{\beta}(\lambda)\right] &= \mathrm{Var}(\mathbf{W}_\lambda\hat{\beta}) = \mathbf{W}_\lambda\mathrm{Var}(\hat{\beta})\mathbf{W}_\lambda^T = \sigma^2\mathbf{W}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{W}_\lambda^T \\ &= \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{X}\left[(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\right]^T\end{aligned}$$

gdje smo koristili činjenice da $\mathrm{Var}(\mathbf{AY}) = \mathbf{A}\mathrm{Var}(\mathbf{Y})\mathbf{A}^T$ za neslučajnu matricu $\mathbf{A}$, $\mathbf{W}_\lambda$ neslučajnu te $\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.
Varijanca ridge procjenitelja teži u nulu kako λ teži beskonačno:

$$\lim_{\lambda\to\infty}\mathrm{Var}\left[\hat{\beta}(\lambda)\right] = \lim_{\lambda\to\infty}\sigma^2\mathbf{W}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{W}_\lambda^T = \mathbf{0}.$$



Slika 3.3: Procjena za Y dobivena metodom najmanjih kvadrata, $\hat{Y}$, i procjena za Y dobivena ridge regresijom, $\hat{Y}(\lambda)$, u hiperravnini razapetoj kovarijatama. *Izvor:* [8]

Uočimo da množenje varijabli prediktora konstantom c može značajno utjecati na procjene koeficijenata dobivenih ridge regresijom, što nije slučaj kod metode najmanjih kvadrata. Naime, vrijednost $X_j\hat{\beta}_{j,\lambda}$ ne ovisi samo o koeficijentu λ već i o skaliranju j-tog (čak i drugih) prediktora. Stoga je običaj prije provođenja ridge regresije standardizirati varijable prediktora:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

Kako je nazivnik gornjeg izraza procjena standardne devijacije j-tog prediktora, standardna devijacija standardiziranih prediktora jednaka je 1 te dobiveni koeficijenti ne ovise o skali prediktora.

Općenito, kada je stvarna veza između odziva i kovarijata približno linearna, procjene dobivene metodom najmanjih kvadrata imat će malu pristranost, ali moguću veliku varijancu. Velika varijanca znači da mala promjena u podacima rezultira u velikoj promjeni u procjenama za koeficijente. Pokazali smo da ridge regresija smanjuje varijancu, uz naravno porast pristranosti za koji se u praksi pokazuje da nije prevelik.

LASSO regresija

Uočimo kako model dobiven ridge regresijom uvijek uključuje svih p prediktora. S porastom vrijednosti λ član penalizacije $\lambda \sum_{j=1}^{p} \beta_j^2$ dovest će do smanjivanja koeficijenata $\beta_1, \ldots, \beta_p$, ali, osim u slučaju da je $\lambda = \infty$, niti jedan od procijenjenih koeficijenata neće biti jednak nuli. Kada je broj prediktora p velik uključivanje svih prediktora u model može utjecati na njegovu interpretabilnost. LASSO regresija metoda je statističkog učenja koja u tom pogledu predstavlja alternativu ridge regresiji. Koeficijenti dobiveni LASSO regresijom dani su sa:

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

Promjena l_2 norme (ridge regresija) u l_1 normu (LASSO regresija) kod člana penalizacije može se činiti tek detalj. Međutim, upravo u tome leži razlog zašto vektor procijenjenih parametara β dobiven LASSO regresijom može sadržavati nule ili puno nula. Dakle, LASSO penalizacija daje $\hat{\beta}_j(\lambda_l) = 0$ za neke j i velike λ_l, dok ridge regresija daje procjenu $j-$tog parametra $\hat{\beta}_j(\lambda_r) \neq 0$. Ovdje smo sa λ_l označili parametar podešavanja pripadan LASSO regresiji, a sa λ_r parametar podešavanja pripadan ridge regresiji. Ukoliko je neki (ili više njih) $\hat{\beta}_j(\lambda_l) = 0$ tada LASSO zapravo provodi selekciju varijabli. Jednako kao i ridge regresija, LASSO regresija ima ekvivalentni zapis:

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

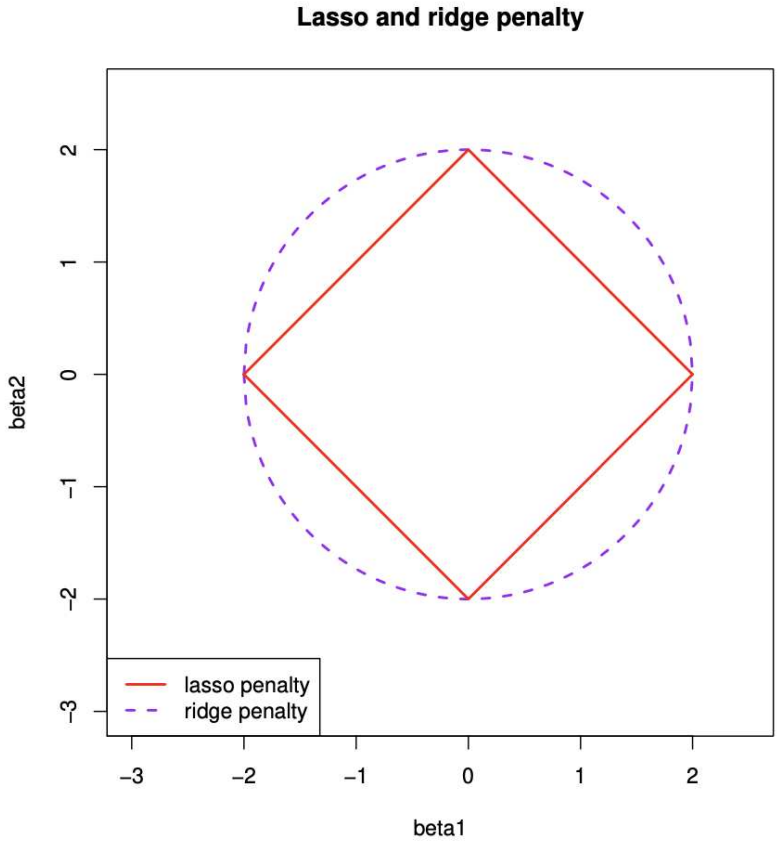
$$\sum_{j=1}^{p} |\beta_j| \leq t.$$

Uočimo kako se zapravo radi o problemu najmanjih kvadrata, ali s uvjetom na parametre $\beta_1, \ldots, \beta_p$ - dok kod metode najmanjih kvadrata svaki β_j teoretski može poprimiti bilo koju vrijednost u skupu realnih brojeva između $-\infty$ i ∞, ovdje su vrijednosti ograničene na neki skup vrijednosti. Ključna razlika između ridge i LASSO regresije je upravo u toj domeni vrijednosti koje β_j mogu poprimiti. Uvjeti na parametre rezultiraju dvjema različitim kuglama, l_1 i l_2:

$$\{\beta \in \mathbb{R}^p : |\beta_1| + |\beta_2| + \cdots + |\beta_p| \leq t\}$$

$$\{\beta \in \mathbb{R}^p : \beta_1^2 + \beta_2^2 + \cdots + \beta_p^2 \leq s\}.$$

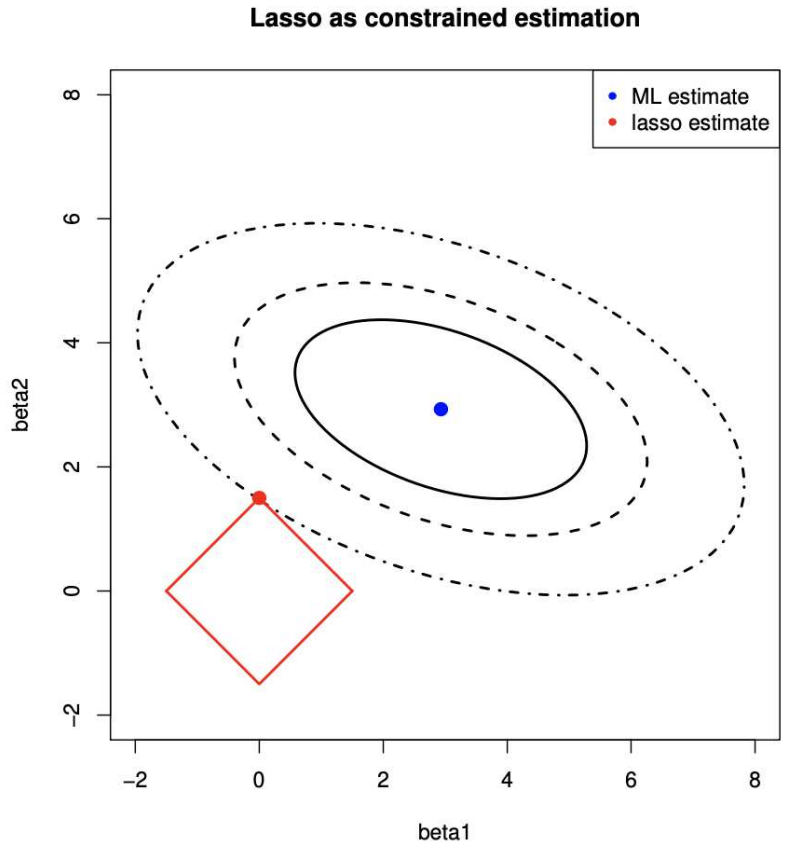
Sljedeća slika prikazuje ograničenja na parametre za $p = 2$ i $s = t = 2$. U Euklidskom prostoru uvjet za ridge regresiju formira krug, a uvjet za LASSO regresiju dijamant.



Slika 3.4: LASSO uvjet ($|\beta_1| + |\beta_2| \leq 2$) i ridge uvjet ($\beta_1^2 + \beta_2^2 \leq 2$). *Izvor:* [8]

Selekcijsko svojstvo LASSO regresije posljedica je toga da vrhovi dijamanta sijeku koordinatne osi. Budući da kod ridge regresije nemamo oštre vrhove jer je domena krug tamo se to neće dogoditi. Procjenitelj dobiven LASSO regresijom je onaj β unutar dijamanta za koji se postiže najmanja RSS te je to ujedno točka unutar dijamanta najbliža procjenitelju dobivenim metodom najmanjih kvadrata bez uvjeta (obična linearna regresija). Ukoliko

je ta točka baš neki od vrhova dijamanta, tada će jedna od koordinata vektora β biti nula. U višim dimenzijama ($p > 2$) može ih biti i više jednako nula. Na sljedećoj slici je to i prikazano, ponovo za $p = 2$:



Slika 3.5: LASSO procjenitelj kao procjenitelj dobiven metodom najmanjih kvadrata uz uvjet. *Izvor:* [8]

Veće vrijednosti penalizacijskog parametra λ kod LASSO regresije dovode do većeg broja nul-elemenata u vektoru procjenitelja β (lokalno to ne mora biti monotono, no ta rasprava je van dosega ovog rada). Dakle, za dovoljno velike vrijednosti parametra λ, procijenjeni regresijski model sadrži samo podskup prediktora. U visokim dimenzijama broj parametara koje LASSO odabire u odnosu na ukupan broj parametara je uobičajeno mali te je dobiven model tzv. *sparse* model.

Što se tiče usporedbe ridge i LASSO regresije, ovisno od slučaja do slučaja jedna će metoda davati bolje rezultate od druge. Općenito, LASSO je pogodnija u slučajevima kad relativno mali broj prediktora ima velike pripadne koeficijente β_j, dok su pripadni koefi-

cijenti za ostale prediktore blizu nule ili jednaki nuli. Ridge regresija će pak davati bolje rezultate kad je varijabla odziva funkcija brojnih prediktora s pripadnim koeficijentima približno jednake veličine. Naravno, *a priori* nam broj prediktora povezan s odzivom nije poznat te kako bismo za određeni skup podataka procijenili koja od ove dvije metode daje bolje rezultate možemo koristiti unakrsnu validaciju. Kao i ridge regreesija, LASSO regresija smanjuje varijancu u odnosu na metodu najmanjih kvadrata u zamjenu za ne tako velik porast pristranosti te stoga može rezultirati boljim predikcijama. Za razliku od ridge regresije, LASSO provodi odabir varijabli te stoga dobiven model može biti lakši za interpretirati.

Sjetimo se, već smo napomenuli da ćemo parametar λ birati unakrsnom validacijom. U sljedećem poglavlju samu proceduru pri primjeni i detaljnije opisujemo.

Poglavlje 4

Modeliranje turističke potrošnje

4.1 Uvodno o podacima Instituta za turizam

U ovom poglavlju primijenit ćemo metode statističkog učenja koje smo dosad predstavili na podatke Instituta za turizam o turističkoj potrošnji. Podaci su prikupljeni u razdoblju između petog mjeseca 2019. i trećeg mjeseca 2020. godine i rezultat su upitnika od strane Instituta za turizam. Varijabla odziva Y je turistička potrošnja, a varijable prediktora koje ćemo koristiti za predviđanje turističke potrošnje prikazujemo u sljedećoj tablici:

Prediktor	Naziv prediktora	Tip varijable	Moguće vrijednosti
X_1	vrsta objekta	kvalitativna	1, 2, 3 ili 4
X_2	kategorija objekta	kvalitativna	1, 2, 3, 4 ili 5
X_3	prijevoz	kvalitativna	1, 2, ... ili 9
X_4	rezervacija smještaja	kvalitativna	1, 2, 3, 4, ili 6
X_5	glavni motiv odmorišnog puta	kvalitativna	1, 2, ..., 13
X_6	izvor informacija 1	kvalitativna	0 ili 1
X_7	izvor informacija 2	kvalitativna	0 ili 1
X_8	izvor informacija 3	kvalitativna	0 ili 1
X_9	izvor informacija 4	kvalitativna	0 ili 1
X_{10}	izvor informacija 5	kvalitativna	0 ili 1
X_{11}	izvor informacija 6	kvalitativna	0 ili 1
X_{12}	izvor informacija 7	kvalitativna	0 ili 1
X_{13}	izvor informacija 8	kvalitativna	0 ili 1
X_{14}	dob	kvantitativna	$\mathbb{N}$
X_{15}	edukacija	kvalitativna	1, 2, 3
X_{16}	mjesečna primanja kućanstva	kvalitativna	1, 2, ... ili 10

Opišimo detaljnije značenje naših varijabli prediktora. Vrste objekta su redom hotel, kamp, soba/apartman/kuća, OPG. Varijabla kategorija objekta ima pet razina i one predstavljaju moguće zvjezdice od 1 do 5. Prijevoz redom čine automobil, automobil s kamp-kućicom, kamper, autobus, motocikl, bicikl, zrakoplov, brod/trajekt, jahta/jedrilica. Rezervacija smještaja označava redom rezervaciju posredstvom turističke/putničke agencije osobnim kontaktom, posredstvom turističke/putničke agencije online booking-om, izravno sa smještajnim objektom osobnim kontaktom, izravno sa smještajnim objektom online booking-om i bez rezervacije unaprijed. Glavni motiv odmorišnog puta sastoji se od 13 mogućih kategorija, more, priroda, selo, gradovi, touring/sightseeing, kultura i umjetnost, zabava i festivali, manifestacije i događanja, planinarenje/hodanje, cikloturizam/mountain biking, ostali sportovi i rekreacija, gastronomija, wellness/toplice. Za dosadašnje varijable turist je mogao zaokružiti samo jednu kategoriju, s obzirom da je za izvor informacija u upitniku Instituta za turizam bilo moguće zaokružiti više opcija, svaki od izvora informacija modeliramo jednom binarnom kvalitativnom varijablom i ukupno njih 8 - brošure/oglasi/plakati, članci u novinama ili časopisima, radio/televizija/film/video, preporuke rodbine ili prijatelja, turistički sajmovi/izložbe, preporuke turističke agencije ili kluba/katalog, internet i prijašnji boravak. Dob je kvantitativna varijabla i za nju je odlučeno ne stvarati kategorije. Edukacija ima tri kategorije srednja škola ili niže, viša škola te fakultet i viši stupnjevi. Na kraju varijabla mjesečnih primanja kućanstva sastoji se od 10 kategorija, prva je do 500 eura, a zadnja (deseta) 5001 euro i više.

U daljnoj analizi koristimo podatke turista kojima je glavni razlog putovanja odmor, oni čine 76% ukupnih podataka.

Kao i obično, prije nego što metode primijenimo na podatke, potrebno je *pročistiti* skup podataka. Jedan od glavnih problema predstavljaju vrijednosti koje nedostaju, tzv. NA vrijednosti. Postoji više pristupa te u same detalje odabira načina nećemo ulaziti, no za takve vrijednosti birali smo aritmetičku sredinu, medijan ili mod uzorka konkretnog obilježja. Dakle, naši podaci na kojima ćemo primijeniti u radu predstavljene metode statističkog učenja sačinjeni su od opaženih vrijednosti za 16 varijabli prediktora za turiste kojima je glavni razlog putovanja odmor te opaženih vrijednosti odzivne varijable - turističke potrošnje.[1] Veličina tog skupa podataka je 10324. 70% koristit ćemo za skup za trening, a 30% kao testni skup.

U prvom poglavlju naveli smo kako pri modeliranju turističke potrošnje varijable prediktora mogu biti kvantitativne ili kvalitativne, dok je varijabla odziva, budući da se radi o turističkoj potrošnji, kvantitativna. Uočimo kako je 15 od 16 varijabli upravo kvalitativnih. U narednom odjeljku kratko objašnjavamo tretiranje kvalitativnih varijabli u modeliranju.

[1] Preciznije, turističke potrošnje na razini ispitanika izražene u eurima, isključujući troškove prijevoza u dolasku i odlasku.

Kvalitativne varijable

Najjednostavniji primjer kvalitativne varijable je binarna kvalitativna varijabla ili kvalitativna varijabla s dvije razine. Binarna varijabla X s dvije razine, na primjer a i b, može se prikazati kao:

$$D = \begin{cases} 0, & \text{if } X = a \\ 1, & \text{if } X = b. \end{cases}$$

D je takozvana *dummy* varijabla: ona je jednaka nula, odnosno jedan za dvije moguće razine kategorijalne varijable. U našem slučaju takve varijable su varijable koje predstavljaju neki izvor informacija $X_6, \ldots, X_{13}$. Vrijednost 0 označava da turist nije koristio dani izvor informacija, a vrijednost 1 označava da ga je koristio. Općenito, ukoliko razina a odgovara 0, tada se ona interpretira kao *referentna razina* s kojom se uspoređuje razina b. To je glavna poanta uvođenja *dummy* varijabli: jedna razina kategorijalne varijable je referentna razina, a ostale se uspoređuju s njom. U linearnom regreesijskom modelu, koeficijent koji stoji uz *dummy* varijablu predstavlja prosječni utjecaj promjene razine kategorijalne varijable s a na b na Y, držeći sve ostale prediktore fiksnima. Pogledajmo i slučaj kategorijalnih varijabli s više od dvije razine, na primjer kvalitativna varijabla s X s tri razine a, b i c. Uzmemo li a kao referentnu razinu, tada se varijabla X može prikazati pomoću dvije *dummy* varijable na sljedeći način:

$$D_1 = \begin{cases} 0, & \text{if } X = b \\ 1, & \text{if } X \neq b. \end{cases}$$

i

$$D_2 = \begin{cases} 0, & \text{if } X = c \\ 1, & \text{if } X \neq c. \end{cases}$$

Interpretacija je slična kao i u slučaju kategorijalne varijable s dvije razine. Koeficijent koji stoji uz D_1 predstavlja prosječni utjecaj promjene razine kategorijalne varijable s a na b na Y, držeći sve ostale prediktore fiksnima, a koeficijent koji stoji uz D_2 prosječni utjecaj promjene s a na c. Općenito, ako kvalitativna varijabla ima j razina, broj potrebnih *dummy* varijabli je $j - 1$.

4.2 Primjena linearne regresije

U ovom odjeljku primijenit ćemo klasični linearni model na naše podatke koristeći R.

Najprije naš skup podataka dijelimo na skup za trening i testni skup tako da 70% skupa podataka čini skup za trening, a 30% podataka testni skup:

```
set.seed(1)
sample <- sample (c(TRUE, FALSE), nrow(podaci),
           replace = TRUE, prob = c(0.7, 0.3))
x_train <- podaci[sample, -17]
x_test <- podaci[!sample, -17]
y_train <- podaci[sample,17]
y_test <- podaci[!sample,17]
podaci_train <- podaci[sample, ]
```

Potom linearni model prilagođavamo na skupu za trening, koristeći funkciju `lm()` iz R-a:

```
lm1 <- lm(formula = Ukupno ~ ., data = podaci_train)
```

Pozivanjem naredbe `summary(lm1)` dobivamo sljedeće:

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             92.612926  24.583464   3.767 0.000166 ***
vrsta_objekta2         -43.340651   3.285296 -13.192  < 2e-16 ***
vrsta_objekta3         -29.895035   2.025969 -14.756  < 2e-16 ***
vrsta_objekta4         -20.198919  10.873174  -1.858 0.063255 .
vrsta_objekta5         -43.844937   4.972181  -8.818  < 2e-16 ***
kategorija_objekta2    -17.132558  21.318046  -0.804 0.421618
kategorija_objekta3    -19.131673  21.027392  -0.910 0.362935
kategorija_objekta4     -4.357042  21.000991  -0.207 0.835650
kategorija_objekta5     23.082493  21.289266   1.084 0.278299
prijevoz2              -11.634280   5.597474  -2.078 0.037700 *
prijevoz3               -4.827060   4.312809  -1.119 0.263076
prijevoz4                5.025036   3.708402   1.355 0.175447
prijevoz5               12.731569   7.294458   1.745 0.080963 .
prijevoz6               -8.785031  24.445581  -0.359 0.719328
prijevoz7               16.111227   2.380252   6.769 1.40e-11 ***
prijevoz8                3.722802  10.321538   0.361 0.718347
prijevoz9              116.390289  18.095479   6.432 1.34e-10 ***
rezervacija_smjestaja2   3.847900   2.526803   1.523 0.127845
rezervacija_smjestaja3   3.909311   3.066052   1.275 0.202340
rezervacija_smjestaja4  -3.003300   2.954848  -1.016 0.309475
rezervacija_smjestaja6   4.300173   4.960845   0.867 0.386068
gl_mot_odm_put2          2.069396   2.438469   0.849 0.396107
gl_mot_odm_put3         -5.682457   4.788433  -1.187 0.235383
gl_mot_odm_put4         -1.409193   3.265919  -0.431 0.666129
gl_mot_odm_put5          0.317493   3.075582   0.103 0.917783
gl_mot_odm_put6        -11.541578   5.145810  -2.243 0.024933 *
```

```
gl_mot_odm_put7         16.904525   6.256042   2.702 0.006906 **
gl_mot_odm_put8         24.020835   5.885224   4.082 4.52e-05 ***
gl_mot_odm_put9        -19.530360  11.932828  -1.637 0.101739
gl_mot_odm_put10        28.379319  12.876170   2.204 0.027555 *
gl_mot_odm_put11        24.085489   8.395666   2.869 0.004132 **
gl_mot_odm_put12         4.022585   7.521394   0.535 0.592792
gl_mot_odm_put13       -11.152901   4.414911  -2.526 0.011552 *
izvor_informacija11      1.629361   3.583600   0.455 0.649359
izvor_informacija21     -4.281181   4.908888  -0.872 0.383168
izvor_informacija31      9.159288   3.942628   2.323 0.020199 *
izvor_informacija41     -0.033075   2.009754  -0.016 0.986870
izvor_informacija51     11.647043   4.023436   2.895 0.003806 **
izvor_informacija61     -4.002816   2.637613  -1.518 0.129162
izvor_informacija71     -0.320479   1.927237  -0.166 0.867934
izvor_informacija81     -0.806429   2.431634  -0.332 0.740170
DOB                     -0.007911   0.067538  -0.117 0.906759
EDUKACIJA2               6.365410   2.414263   2.637 0.008393 **
EDUKACIJA3               5.354811   2.496467   2.145 0.031990 *
PRIMANJA2               14.577253  14.408548   1.012 0.311712
PRIMANJA3               13.511819  13.981616   0.966 0.333877
PRIMANJA4               18.863545  13.796557   1.367 0.171585
PRIMANJA5               20.577275  13.758019   1.496 0.134787
PRIMANJA6               23.007539  13.523088   1.701 0.088920 .
PRIMANJA7               19.661648  13.808088   1.424 0.154513
PRIMANJA8               26.848857  13.843930   1.939 0.052493 .
PRIMANJA9               28.010730  13.930713   2.011 0.044393 *
PRIMANJA10              55.302852  13.833588   3.998 6.46e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Prvi stupac "Estimate" daje procjene koeficijenata β u linearnom modelu. Kad ne bismo nijednu varijablu uključili u model, tada bi prosječna turistička potrošnja iznosila 92.612926 eura ("Intercept"). Drugi stupac "Std. Error" daje procjenu standardne devijacije koeficijenata. Treći stupac daje vrijednost t - statistike, on je jednak koeficijentu uz prediktor podijeljenom standardnom greškom (dakle, $\frac{\text{"Estimate"}}{\text{"Std. Error"}}$). t - statistika potom se koristi za računanje p - vrijednosti, što je četvrti stupac u gornjem rezultatu. Na temelju p - vrijednosti zaključujemo je li neka varijabla statistički značajna na određenoj razini značajnosti ili nije.

U skladu s prethodno opisanim tretmanom kvalitativnih varijabli, uočavamo kako funkcija `lm()` za kvalitativnu varijablu s j razina stvori $j-1$ binarnu varijablu. Na primjer, varijabla vrsta objekta je kvalitativna varijabla s 5 razina. U gornjem ispisu vidimo kako je

funkcija `lm()` stvorila 4 binarne varijable: vrsta_objekta2, vrsta_objekta3, vrsta_objekta4 i vrsta_objekta5. Vrijednost binarne varijable vrsta_objekta2 je 0 ukoliko turist nije odabrao vrstu objekta u kojoj boravi označenu s 2 (kamp), a 1 ukoliko ju je odabrao. Uočimo kako je referentna razina upravo vrsta objekta 1 (hotel) i za nju `lm()` ne stvara binarnu varijablu - ako turist boravi u hotelu, preostalih četiri binarnih varijabli poprima vrijednost 0. Jednadžba linearne regresije za turističku potrošnju sada glasi:

$$\begin{aligned} Y = 92.612926 - 43.340651\text{vrsta\_objekta2} + \cdots - 43.844937\text{vrsta\_objekta5} \\ -17.132558\text{kategorija\_objekta2} + \cdots + 23.082493\text{kategorija\_objekta5} \\ + \cdots + 14.577253\text{PRIMANJA2} + \cdots + 55.302852\text{PRIMANJA10}. \end{aligned}$$

Uvrstimo li u gornju jednadžbu odgovore određenog turista, dobit ćemo modelom predviđenu prosječnu turističku potrošnju u eurima za turista s tim odgovorima.

Koeficijenti uz pojedinu binarnu varijablu koja predstavlja kategoriju kvalitativne varijable predstavljaju razliku između predviđene vrijednosti za tu kategoriju i predviđene vrijednosti za referentnu kategoriju te kvalitativne varijable. t- statistika i pripadne p- vrijednosti bazirane su na nultoj hipotezi da su koeficijenti jednaki nula. Za $p-$ vrijednost veću od 0.05 (standardna razina značajnosti 5%) odbacujemo nultu hipotezu, a za $p-$ vrijednost manju od 0.05 ne odbacujemo nultu hipotezu. Pogledajmo u gornjem ispisu neku veliku p-vrijednost (dakle, blizu 1), npr. za gl_mot_odm_put5 pripadna p-vrijednost jednaka je 0.917783 - to znači da nema neke razlike u prosječnoj potrošnji turista koji je za glavni motiv odmorišnog puta odabrao 1 (more) i turista koji je za glavni motiv odmorišnog puta odabrao 5 (touring/sightseeing) (držeći sve ostale prediktore fiksnima!). To je i u skladu s normom koeficijenta uz gl_mot_odm_put5, norma je malena, koeficijent iznosi 0.317493. Analognim zaključivanjem na temelju $p-$ vrijednosti zaključujemo kako su prediktori koji utječu na turističku potrošnju:

- vrsta objekta
- 2., 5., 7. i 9. kategorija prijevoza s obzirom na referentnu kategoriju
- 6., 7., 8., 10., 11. i 13. kategorija glavnog motiva odmorišnog putovanja s obzirom na referentnu kategoriju
- izvor informacija 3
- izvor informacija 5
- edukacija
- 6., 8., 9. i 10. kategorija primanja s obzirom na referentnu kategoriju

Osim gornje tablice naredba `summary(lm1)` daje i donje podatke:

```
Residual standard error: 71.81 on 7136 degrees of freedom
Multiple R-squared:  0.1451,Adjusted R-squared:  0.1389
F-statistic: 23.29 on 52 and 7136 DF,  p-value: < 2.2e-16
```

Standardna greška reziduala ("Residual standard error") iznosi 71.81, što znači da se predviđene vrijednosti prosječno razlikuju za 71.81 eura od stvarnih vrijednosti za turističku potrošnju. Budući da je prosječna turistička potrošnja 102.52 eura, to nam daje do znanja da rezultati koje smo dobili nisu precizni. Općenito, želimo što manju standardnu grešku reziduala. "Multiple R-squared" daje postotak varijabilnosti u odzivu koji je objašnjem modelom. U našem slučaju on iznosi 14.51% što je loš rezultat. S obzirom da se u nijansama razlikuju u izračunu, slično interpretiramo i "Adjusted R-squared". Na kraju, na temelju p - vrijednosti za F - statistiku koja je jako blizu nule, zaključujemo da na svim standardnim razinama značajnosti postoji veza između turističke potrošnje i varijabli prediktora.

Odredimo i testnu grešku dobivenog modela na testnom skupu:

```
lm1_pred <- predict(lm1, x_test)
mean(t((lm1_pred - y_test)^2))
```

Dobivamo veliku testnu grešku jednaku 3259.078.

Na temelju svih navedenih rezultata zaključujemo kako dobiveni linearni model nije optimalan stoga prethodno opisane interpretacije utjecaja pojedinih varijabli na prosječnu turističku potrošnju treba uzeti sa rezervom.

4.3 Primjena metode odabira najboljeg podskupa i metoda postupnog odabira prediktora

Predikcije za turističku potrošnju dobivene linearnim modelom nisu se pokazale previše preciznima. U ovom odjeljku želimo primjenom metode odabira najboljeg podskupa i metoda postupnog odabira prediktora unaprijed i unazad odabrati dobre prediktore (u danom smislu) za predikciju turističke potrošnje. U R-u za sve tri metode - metoda odabira najboljeg podskupa, metoda postupnog odabira prediktora unaprijed i metoda postupnog odabira prediktora unazad - koristimo funkciju `regsubsets` iz paketa `leaps`.

Primjena metode odabira najboljeg podskupa

Sjetimo se, kod metode najboljeg podskupa za svaki $k \in \{1, \ldots, p\}$ prilagođavamo model za sve moguće kombinacije k prediktora i biramo onaj koji minimizira RSS. Potom od dobivenih p modela biramo onaj koji je sveukupno najbolji, gdje najbolje mjerimo najmanjom procjenom očekivane testne greške.

Funkcija `regsubsets` iz paketa `leaps` provodi metodu odabira najboljeg podskupa na način da za dani broj prediktora $k \in \{1, \ldots, p\}$ vraća najbolji model s k prediktora, gdje je najbolje mjereno s RSS. U našem slučaju dobit ćemo 16 najboljih modela (s najmanjom RSS) za dani k, dakle najbolji model s 1 varijablom prediktora, najbolji model s 2 varijable prediktora, ..., najbolji model s 16 varijabli prediktora. Koristeći unakrsnu validaciju odabrat ćemo onaj od 16 modela koji ima najmanju procjenu očekivane testne greške. Spremit ćemo rezultat u `reg.summary` objekt.

```
regfit.full <-
regsubsets(Ukupno ~ ., podaci_train, nvmax = 16, method = "exhaustive")

reg.summary <- summary(regfit.full)
```

Komponente objekta `reg.summary` možemo vidjeti pozivanjem naredbe `names()`, dobivamo:

```
names(reg.summary)
```

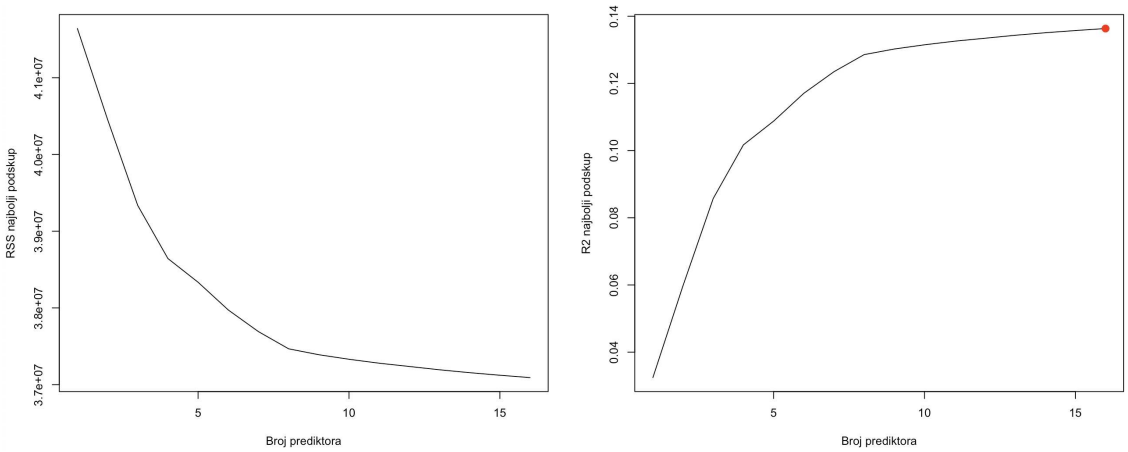
```
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

Nama su od interesa RSS i R^2 te crtamo njihov graf u ovisnosti o broju prediktora koje model koristi:

```
par(mfrow = c(2,2))
plot(reg.summary$rss, xlab = "Broj prediktora",
ylab = "RSS najbolji podskup", type = "l")

plot(reg.summary$adjr2, xlab = "Broj prediktora",
ylab = "R2 najbolji podskup", type = "l")

#točka gdje R2 dostiže maksimum
adjr2.max <- which.max( reg.summary$adjr2 )
points(adjr2.max, reg.summary$adjr2[adjr2.max],
col = "red", pch = 20, cex = 2)
```



Slika 4.1: Opažene vrijednosti statistika RSS i R^2 u ovisnosti o broju prediktora u modelu dobivenom metodom odabira najboljeg podskupa.

Na temelju ovih slika, imajući na umu kako želimo model sa što manjom RSS i što većim R^2, naslućujemo kako bi model sa svih 16 prediktora trebao biti najbolji u tom smislu. Zbog velikog broja prediktora, ovdje nećemo provoditi unakrsnu validaciju (prespora je) kako bismo potvrdili da model s 16 prediktora ima i najmanju procjenu očekivane testne greške.

Pozivanjem naredbe `reg.summary$which[16,]` možemo vidjeti koje kategorije pojedinog prediktora model sa 16 prediktora bira i zaključiti da se rezultati u velikoj mjeri podudaraju s onima dobivenim primjenom linearne regresije u prethodnom odjeljku.

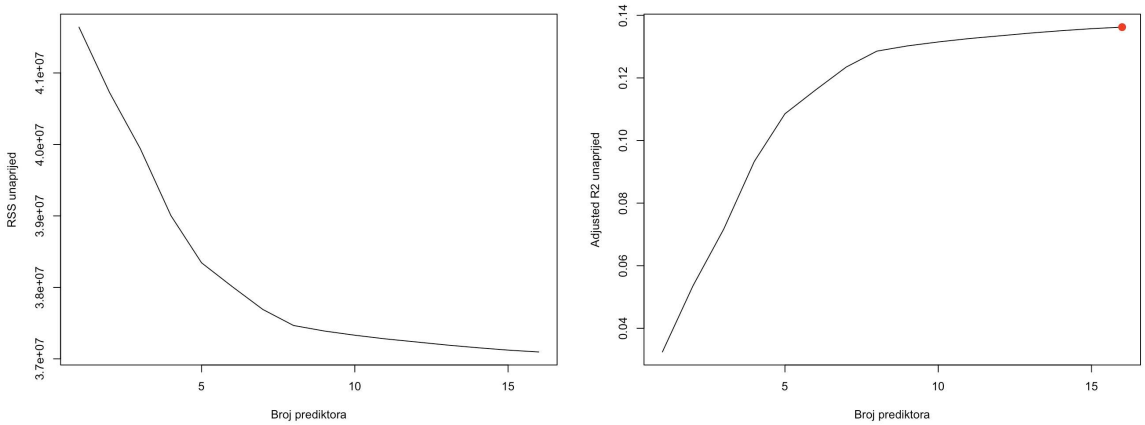
Na kraju, uočimo kako su opažene vrijednosti statistike RSS jako velike, a opažene vrijednosti statistike R^2 bliže nuli. Takvi rezultati upućuju na nelinearnost veze između prediktora i odziva pa zapravo niti jedan od 16 dobivenih modela neće davati precizne procjene za naše podatke.

Primjena metoda postupnog odabira prediktora unaprijed i unazad

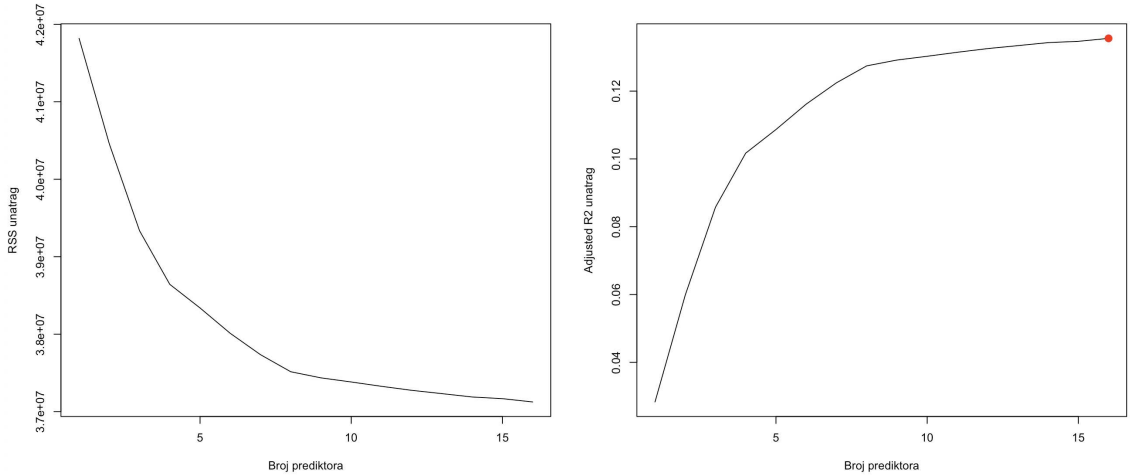
Procedura odabira najboljeg podskupa u R-u bila je jako spora, jedan od razloga zasigurno je velik broj prediktora. Uočimo da iako je $p = 16$, zbog *dummy* varijabli potrebnih za reprezentaciju naših 15 kvalitativnih prediktora, ukupan broj varijabli prešao je 40. Metode postupnog odabira prediktora nastoje naći najbolji model u znatno manjem skupu modela. Sada u R-u pozivanjem iste funkcije `regsubsets()` kao i za metodu odabira najboljeg podskupa provodimo metodu postupnog odabira prediktora unaprijed i postupnog odabira prediktora unazad.

```
regsubsets(Ukupno ~ ., podaci_train, nvmax = 16, method = "forward")
regsubsets(Ukupno ~ ., podaci_train, nvmax = 16, method = "backward")
```

Rezultati su gotovo isti kao i u slučaju metode odabira najboljeg podskupa. Razlika je u brzini procedure - ove metode su, očekivano, znatno brže. Radi potpunosti, za obje metode prilažemo grafove opaženih vrijednosti statistika RSS i R^2 u ovisnosti o broju prediktora.



Slika 4.2: Opažene vrijednosti statistika RSS i R^2 u ovisnosti o broju prediktora u modelu dobivenom metodom postupnog odabira prediktora unaprijed.



Slika 4.3: Opažene vrijednosti statistika RSS i R^2 u ovisnosti o broju prediktora u modelu dobivenom metodom postupnog odabira prediktora natrag.

Kao i u slučaju metode odabira najboljeg podskupa, zaključujemo kako dobiveni rezultati upućuju na odsustvo linearne veze. U nastavku provodimo ridge i LASSO regresiju.

4.4 Primjena ridge i LASSO regresije

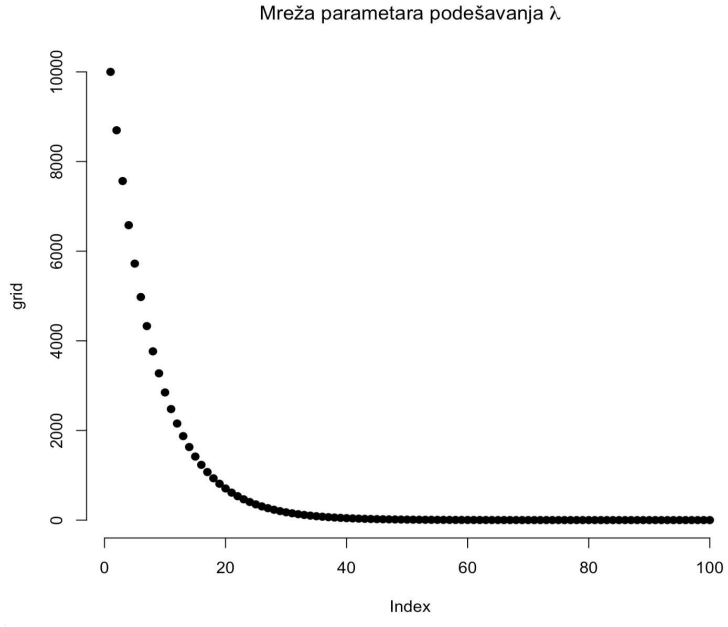
Paket `glmnet` u R-u pruža funkcionalnost za prilagodbu ridge i LASSO regresije.

Kako bismo proveli ridge regresiju, najprije stvaramo matricu iz našeg skupa podataka pomoću funkcije `model.matrix()` iz koje uklanjamo odsječak na ipsilon osi iz rezultirajuće matrice budući da je on automatski uključen (`intercept = T`) pri pozivanju funkcije `glmnet()` iz istoimenog paketa.

```
x_train <- model.matrix(Ukupno ~ . -1, podaci_train)
y_train <- podaci_train$Ukupno
```

Odredimo mrežu parametara podešavanja i prikažimo njezin graf:

```
grid <- 10^seq(4, -2, length = 100)
plot(grid, bty = "n", pch = 19,
     main = expression(paste("Mreža parametara podešavanja ", lambda)))
```



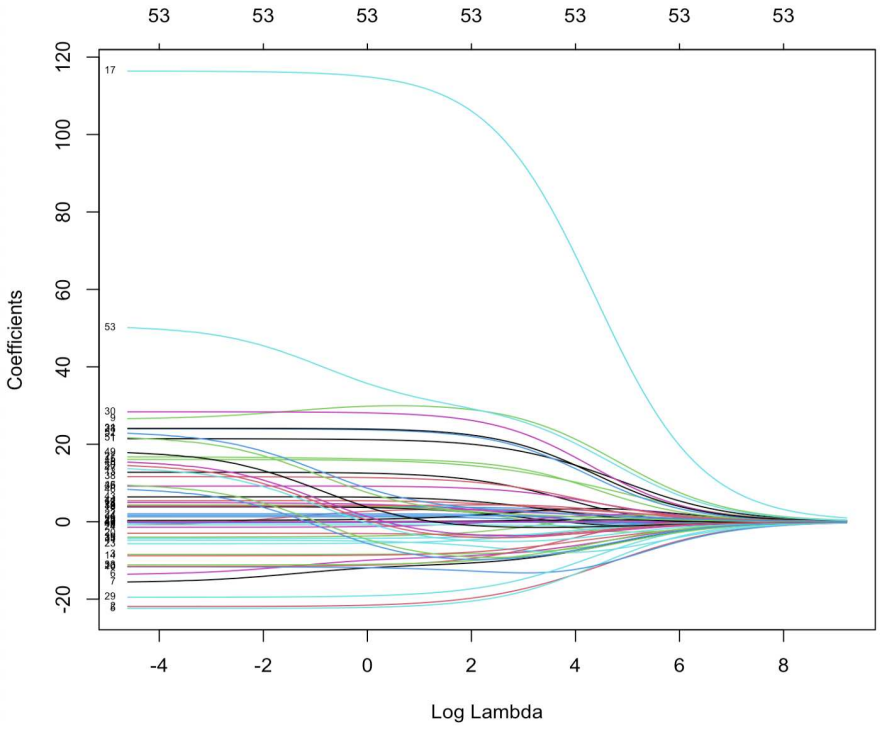
Slika 4.4: Mreža parametara podešavanja λ.

Sada pozivamo funkciju `glmnet()` i provodimo ridge regresiju. Ridge regresija zadana je odabirom `alpha = 0` pri pozivu funkcije. Napomenimo kako se kovarijate standardiziraju po defaultu (inače bi "kazna" regularizacijom bila nepoštena).

```
ridge.mod <- glmnet(x_train, y_train, alpha = 0, lambda = grid)
```

Vizualizirajmo smanjenje koeficijenata u ovisnosti o λ:

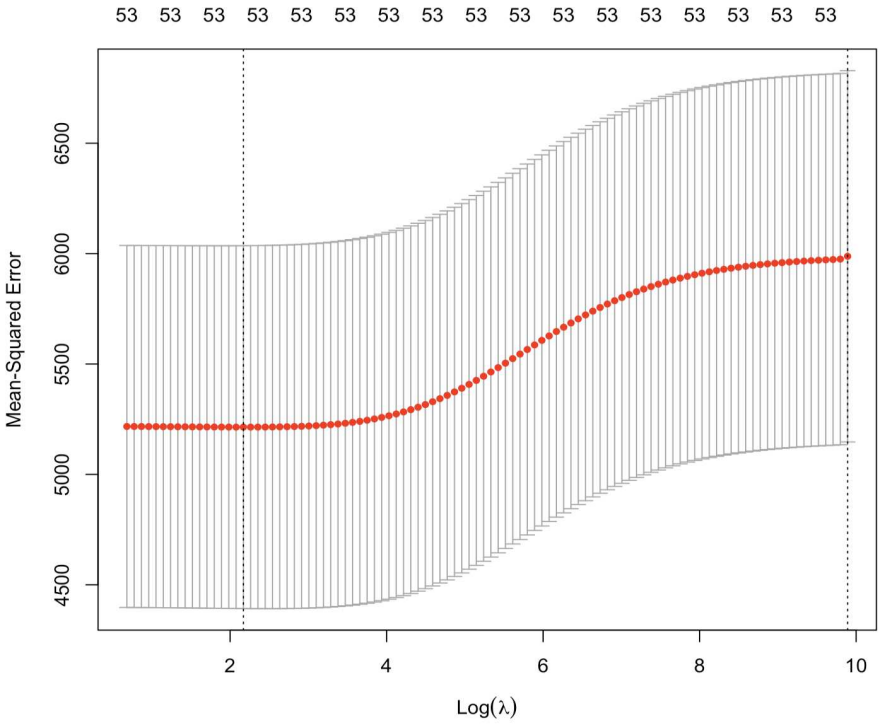
```
plot(ridge.mod, xvar = "lambda", label = TRUE)
```



Slika 4.5: Smanjenje koeficijenata s obzirom na veći λ.

Unakrsnom validacijom odredimo procjenu očekivane testne greške (tzv. CV greška):

```
set.seed(1)
cv.out <- cv.glmnet(x_train, y_train, alpha = 0)
plot(cv.out)
```



Slika 4.6: Procjena očekivane testne greške za ridge regresiju unakrsnom validacijom. Na x- osi su stupnjevi slobode, a dvije vertikalne crte označavaju λ koji minimizira CV grešku, odnosno λ dobiven uz pravilo jedne standardne pogreške.

Ponovno uočavamo jako veliku procjenu za očekivanu testnu grešku. Najbolji model (u ovisnosti o λ) dobiva se za onaj λ koji minimizira CV grešku:

```
bestlam <- cv.out$lambda.min
```

Dobivamo da je λ koji minimizira CV grešku jednak 8.753454.
Odabrani model (dakle, za $\lambda = 8.753454$) dan je s:

```
out <- glmnet(x_train, y_train, alpha = 0)
predict(out, type = "coefficients", s = bestlam)[,1]
```

```
 (Intercept)                89.0177457
vrsta_objekta1              19.7608640
vrsta_objekta2             -19.4489114
vrsta_objekta3              -8.2004077
vrsta_objekta4               0.9268746
```

vrsta_objekta5	-20.2295404
kategorija_objekta2	-8.6198062
kategorija_objekta3	-10.499569
kategorija_objekta4	3.2098592
kategorija_objekta5	28.6709813
prijevoz2	-12.8373908
prijevoz3	-6.4295875
prijevoz4	2.9785824
prijevoz5	10.5565092
prijevoz6	-7.6981063
prijevoz7	14.2637779
prijevoz8	3.6776649
prijevoz9	104.3874326
rezervacija_smjestaja2	2.9502366
rezervacija_smjestaja3	2.2013569
rezervacija_smjestaja4	-3.4539328
rezervacija_smjestaja6	2.0575389
gl_mot_odm_put2	1.8309570
gl_mot_odm_put3	-4.5065547
gl_mot_odm_put4	-0.2727905
gl_mot_odm_put5	1.7369001
gl_mot_odm_put6	-9.2044016
gl_mot_odm_put7	15.0425263
gl_mot_odm_put8	21.6760110
gl_mot_odm_put9	-17.0128491
gl_mot_odm_put10	25.7662065
gl_mot_odm_put11	22.0753195
gl_mot_odm_put12	4.3362308
gl_mot_odm_put13	-9.9062390
izvor_informacija11	1.5163675
izvor_informacija21	-3.8437179
izvor_informacija31	8.4986630
izvor_informacija41	0.1906326
izvor_informacija51	10.0654409
izvor_informacija61	-2.5508299
izvor_informacija71	-0.1622059
izvor_informacija81	-0.8342446
DOB	0.0076289
EDUKACIJA2	5.1082834
EDUKACIJA3	4.6310418
PRIMANJA2	-9.1538381
PRIMANJA3	-9.7218892

```
PRIMANJA4                  -4.9773615
PRIMANJA5                  -3.4359471
PRIMANJA6                  -1.0619076
PRIMANJA7                  -4.0614299
PRIMANJA8                   2.4938627
PRIMANJA9                   3.4439368
PRIMANJA10                 28.7395789
```

Usporedimo li s linearnim modelom, vidimo da je odsječak na ispilon osi ostao približno jednak: kod linearnog modela on je bio jednak 92.612926, a ovdje je jednak 89.0177457. On, podsjetimo se, predstavlja procjenu za prosječnu turističku potrošnju kad ne bismo koristili niti jedan prediktor. Procjene za koeficijente β nešto su manje što je očekivana posljedica uvođenja parametra penalizacije $\lambda = 8.753454$. Odlučimo li se koristiti ridge regresiju kao metodu procjene prosječne turističke potrošnje, interpretacija koeficijenata uz varijable ostaje ista kao i u slučaju linearnog modela. Tako, na primjer, veliki (i pozitivan) koeficijent (jednak 104.3874326) uz binarnu varijablu prijevoz9 (jahta/jedrilica), upućuje na puno veću prosječnu turističku potrošnju kod turista koji je koristio jahtu/jedrilicu kao prijevozno sredstvo u odnosu na turista koji je putovao automobilom (prijevoz1, referentna razina kategorijalne varijable prijevoz). S druge strane, negativan i relativno velik koeficijent uz binarnu varijablu gl_mot_odm_put9 (hodanje/planinarenje) daje naslutiti kako turist kojem je glavni motiv odmorišnog putovanja hodanje/planinarenje prosječno manje potroši od turista kojem je glavni motiv odmorišnog putovanja more (gl_mot_odm_put1). Uočimo i kako je koeficijent uz varijablu dob ovdje još manji nego u slučaju linearne regresije te i ridge regresija upućuje na irelevantnost utjecaja dobi pri procjeni prosječne turističke potrošnje.

Izračunajmo testnu grešku odabranog modela:

```
podaci_test <- podaci[!sample,]
x_test <- model.matrix(Ukupno ~ . -1, podaci_test)
y_test <- podaci_test$Ukupno
ridge.pred <- predict(ridge.mod, s = bestlam, newx = x[test, ])
mean((ridge.pred - y_test)^2)
```

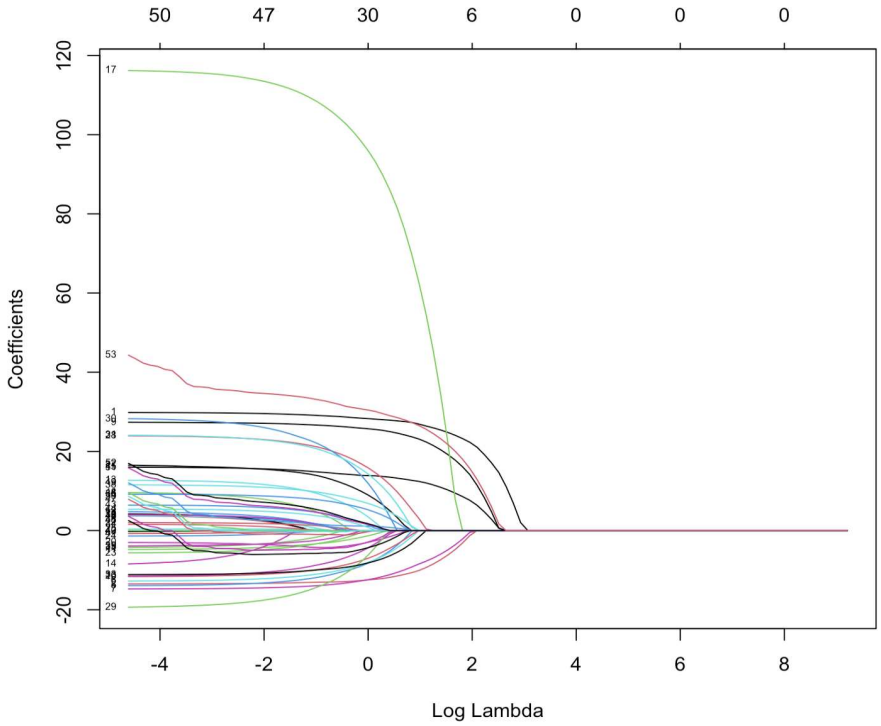
Dobivamo testnu grešku od 3232.948, što je manje nego u slučaju linearnog modela gdje smo dobili testnu grešku jednaku 3259.078. U tom smislu, ispostavlja se da ridge regresija daje preciznije predikcije na dosad neviđenim podacima.

Provedimo sada i LASSO regresiju na analogan način. LASSO regresija zadana je odabirom `alpha = 1` pri pozivu funkcije `glmnet()`.

```
lasso.mod <- glmnet(x_train, y_train, alpha = 1, lambda = grid)
```

Vizualizirajmo smanjenje koeficijenata u ovisnosti o λ i u slučaju LASSO regresije (za razliku od ridge regresije, LASSO regresija radi odabir kovarijata):

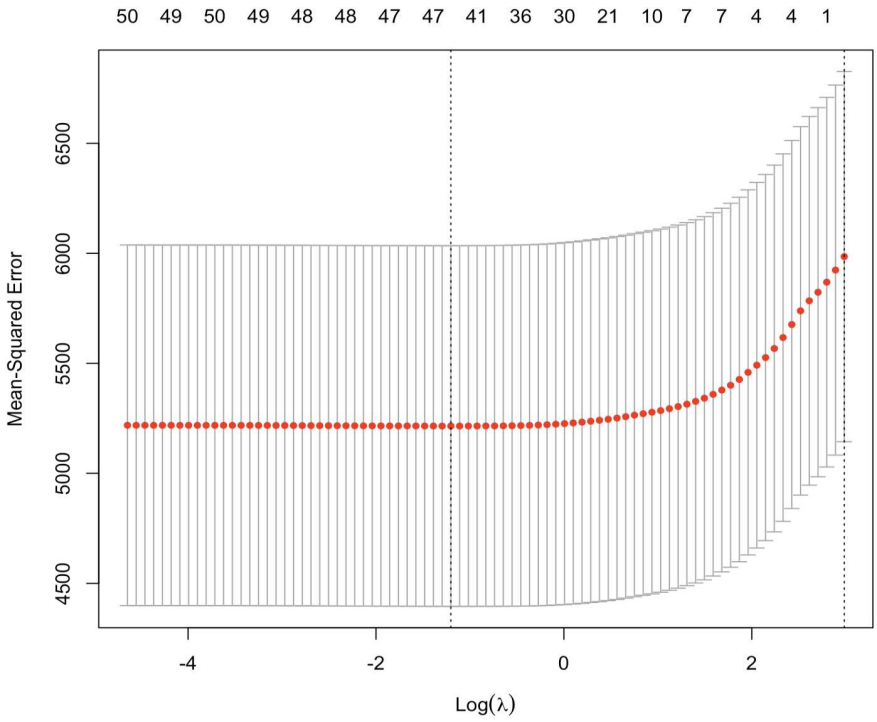
```
plot(lasso.mod, xvar = "lambda", label = TRUE)
```



Slika 4.7: Smanjenje koeficijenata s obzirom na veći λ.

Kao i kod ridge regresije , unakrsnom validacijom odredimo procjenu očekivane testne greške:

```
set.seed(1)
cv.out <- cv.glmnet(x_train, y_train, alpha = 1)
plot(cv.out)
```



Slika 4.8: Procjena očekivane testne greške za LASSO regresiju unakrsnom validacijom.

Ponovno uočavamo jako veliku procjenu za očekivanu testnu grešku.
Dobivamo da je λ koji minimizira CV grešku jednak 0.3002845.
Model za $\lambda = 0.3193503$ dan je s:

```
out <- glmnet(x_train, y_train, alpha = 1)
predict(out, type = "coefficients", s = bestlam)[,1]
```

```
 (Intercept)                81.5437322
vrsta_objekta1              29.2161369
vrsta_objekta2             -13.1186080
vrsta_objekta3               0.0000000
vrsta_objekta4               5.7447805
vrsta_objekta5             -12.2699495
kategorija_objekta2        -11.3579971
kategorija_objekta3        -14.0939974
kategorija_objekta4          0.0000000
kategorija_objekta5         26.7811863
prijevoz2                  -10.2210730
```

prijevoz3	-3.5769046
prijevoz4	2.2083812
prijevoz5	10.1179366
prijevoz6	0.0000000
prijevoz7	15.0601940
prijevoz8	0.3597679
prijevoz9	109.9223468
rezervacija_smjestaja2	2.01883781
rezervacija_smjestaja3	1.11678854
rezervacija_smjestaja4	-4.11911252
rezervacija_smjestaja6	0.44102551
gl_mot_odm_put2	1.42795249
gl_mot_odm_put3	-3.78315097
gl_mot_odm_put4	-0.29784271
gl_mot_odm_put5	0.32492905
gl_mot_odm_put6	-9.25160134
gl_mot_odm_put7	14.41907337
gl_mot_odm_put8	21.71134601
gl_mot_odm_put9	-15.31138733
gl_mot_odm_put10	23.18137176
gl_mot_odm_put11	21.35672185
gl_mot_odm_put12	2.10470393
gl_mot_odm_put13	-10.18924529
izvor_informacija11	0.26286715
izvor_informacija21	-2.13610266
izvor_informacija31	8.11028051
izvor_informacija41	0.00000000
izvor_informacija51	10.11182976
izvor_informacija61	-2.74617219
izvor_informacija71	0.00000000
izvor_informacija81	-0.04960783
DOB	0.0000000
EDUKACIJA2	4.9679431
EDUKACIJA3	4.0914952
PRIMANJA2	-4.7558347
PRIMANJA3	-5.7830077
PRIMANJA4	-0.9119992
PRIMANJA5	0.0000000
PRIMANJA6	1.8428134
PRIMANJA7	-0.1856021
PRIMANJA8	5.1365448
PRIMANJA9	5.9109640

```
PRIMANJA10                  33.6100944
```

Uočimo kako su kod LASSO regresije procjene za neke koeficijente jednake upravo 0. To je posljedica biranja kovarijata karakteristična za LASSO regresiju. Usporedimo li s rezultatima metode odabira najboljeg podskupa, uočavamo neka podudaranja s odabirom varijabli, npr. potpuno jednaki rezultati su za odbacivanje binarnih varijabli prijevoz6, informacija41, informacija71 i PRIMANJA6. Ponovno uočavamo na primjer velik i pozitivan koeficijent uz binarnu varijablu prijevoz9 uz analognu interpretaciju kao i kod linearne i ridge regresije. Koeficijent uz varijablu dob jednak je nula što je u skladu s rezultatima svih dosad primijenjenih metoda.

Za testnu grešku odabranog modela ($\lambda = 0.3002845$) dobivamo da je jednaka 3242.049, što je nešto veća testna greška u usporedbi s ridge regresijom (3232.948) te manja testna greška od one dobivene za linearnu regresiju (3259.078).

Rezultati svih metoda koje smo primijenili upućuju na nelinearnost veze između varijabli odziva i varijable prediktora te daljnje proučavanje tog problema izlazi van dosega ovog rada. Sve metode dovode do približnih zaključaka vezano uz utjecaj pojedinih prediktora na odziv. S obzirom da je testna greška najmanja kod metoda ridge i LASSO regresije, za njih bismo se najprije odlučili pri donošenju zaključaka. Ipak, ponovno napominjemo da zbog naslućene nelinearnosti sve rezultate treba uzeti s rezervom.

Bibliografija

[1] G. Brida, J. i R. Scuderi, *Determinants of tourist expenditure: a review of microeconometric models*, (2012), https://core.ac.uk/download/pdf/211603242.pdf.

[2] T. Hastie, R. Tibshirani i J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2009.

[3] G. James, D. Witten, T. Hastie i R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2021.

[4] Z. Marušić, S. Čorak, N. Ivandić, I. Beroš i M. Ambrušec, *Stavovi i potrošnja turista u Hrvatskoj – TOMAS 2019.*, (2020), https://www.iztzg.hr/files/file/RADOVI/KNJIGE/TOMAS-Hrvatska-2019.pdf.

[5] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, Cambridge, Massachussets, 2012.

[6] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[7] B. Schölkopf i A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Cambridge, Massachussets, 2018.

[8] Wessel N. van Wieringen, *Lecture notes on ridge regression*, (2021), https://arxiv.org/pdf/1509.09169.pdf.

[9] Y. Wanga i M. C. G. Davidson, *A review of micro-analyses of tourist expenditure. Current Issues in Tourism Vol. 13, No. 6, 507 –524*, (2010), https://www.unica.it/static/resources/cms/documents/Wangdavidson_2010_Areviewofmicroanalysesoftouristexpenditure.pdf.

Sažetak

U ovom radu modelirala se turistička potrošnja u Republici Hrvatskoj koristeći metode statističkog učenja. Na početku rada dali smo uvod u statističko učenje, uveli terminologiju, centralne pojmove najbolje veze i funkcije gubitka te cijeli okvir za daljnje predstavljanje metoda koje smo koristili. Podsjetili smo se linearnog regresijskog modela i najčešće metode njegove prilagodbe - metode najmanjih kvadrata. Definirali smo pojmove testne greške i očekivane testne greške te predstavili metodu unakrsne validacije kao metodu za procjenu očekivane testne greške. Koristeći lemu o dekompoziciji očekivane testne greške, objasnili smo odnos pristranosti i varijance u statističkom učenju. Potom smo uveli metode odabira prediktora u linearnim modelima te metode regularizacije (ridge i LASSO regresija). Na kraju smo predstavljene metode primijenili na podatke Instituta za turizam o turističkoj potrošnji u Republici Hrvatskoj. Rezultati svih metoda koje smo primijenili upućuju na nelinearnost veze između varijabli odziva i varijable prediktora. Sve metode dovode do približno jednakih zaključaka vezano uz utjecaj pojedinih prediktora na odziv. S obzirom da je testna greška najmanja kod metoda ridge i LASSO regresije, za njih bismo se najprije odlučili pri donošenju zaključaka.

Summary

In this thesis, we model tourist expenditure in the Republic of Croatia using statistical learning methods. The thesis begins with an introduction to the statistical learning terminology, its concepts and its main tools. We then further revise the linear regression model and the least squares method. Next, we define the test error and the expected test error of a prediction and we present the cross-validation method as a tool for estimating the expected test error. We also provide a decomposition of the expected test error which explains the relationship between bias and variance in the statistical learning theory. We next discuss subset selection methods in linear models and regularization methods (ridge and LASSO). Finally, we apply the presented methods to the data provided by the Institute for Tourism on tourist expenditure in the Republic of Croatia. The results of all the methods we apply suggest a non-linear relationship between the response variable and the predictor variables. However, they also lead to a similar conclusion on the relationship between the response variable and the regressors, with regularization methods giving lower test error.

Životopis

Rođena sam u Zagrebu 15. svibnja 1997. Osnovnu školu pohađala sam u OŠ Izidora Kršnjavoga, u Zagrebu. Maturirala sam 2015. godine završivši zagrebačku Drugu gimnaziju. a 2016. godine upisala sam Preddiplomski studij Matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Završetkom preddiplomskog studija, 2019. godine upisala sam Diplomski studij Financijska i poslovna matematika na istom fakultetu.