Metode strojnog učenja u određivanju tržišne vrijednosti igrača u nogometu

Čulin, Ivan

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ivan Čulin

METODE STROJNOG UČENJA U ODREĐIVANJU TRŽIŠNE VRIJEDNOSTI IGRAČA U NOGOMETU

Diplomski rad

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

Zagreb, veljača 2024.

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. _____

Obitelji.

Zahvaljujem mojoj mami na svim odricanjima i tati koji me vodio primjerom kroz život.
Hvala vam na neizmjernoj podršci kada je bilo najteže.
Hvala sestri Sandri, Jakovu i Marinu koji su mi bili izvor dodatne motivacije.
Hvala svim prijateljima koji su bili uz mene.
I na kraju, Kristin hvala ti što smo zajedno prolazili kroz ovo studentsko putovanje.

HŽV!

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Modeli linearne regresije i procjena parametara modela **3**
1.1 Oblikovanje modela 3
1.2 Procjena parametara modela 5
1.3 Gauss - Markovljev teorem 10
1.4 Višestruka regresija 12
1.5 Multivarijatna regresija 16

2 Ocjena preciznosti modela statističkog učenja **17**
2.1 Fleksibilnost modela 17
2.2 Pristranost, varijanca i kompleksnost modela 18

3 Odabir podskupa **21**
3.1 Odabir najboljeg podskupa 21
3.2 Stepwise unaprijed i unatrag 23
3.3 Stagewise regresija unaprijed 24

4 Metode sažimanja **27**
4.1 Ridge regresija 27
4.2 Lasso regresija 31
4.3 Usporedba metoda 33

5 Primjena opisanih metoda **37**
5.1 Prikupljanje i analiza podataka 37
5.2 Linearna regresija 39
5.3 Unakrsna validacija 41
5.4 Metode odabira podskupa 43

5.5 Metode sažimanja 47
5.6 Analiza rezultata 52

Bibliografija **53**

Uvod

Nogomet, kao globalni fenomen, ne samo da plijeni pažnju milijuna navijača diljem svijeta, već je i značajan ekonomski entitet. S rasponom transfera igrača koji dosežu značajne iznose, procjena tržišne vrijednosti nogometaša postala je ključna u donošenju odluka za klubove, menadžere igrača i ostale dionike u nogometnom ekosustavu.

U ovom radu istražuje se inovativan pristup procjeni tržišne vrijednosti nogometaša korištenjem metoda statističkog učenja. Statističko učenje pruža snažan alat za analizu velikih podataka i identifikaciju skrivenih uzoraka koji mogu utjecati na cijenu igrača. Kroz korištenje naprednih modela statističkog učenja, ovaj rad ima za cilj razviti modele koji će biti u mogućnosti preciznije procijeniti tržišnu vrijednost nogometaša na temelju različitih čimbenika kao što su performanse, statistički pokazatelji, dob, pozicija na terenu, i drugi relevantni faktori.

Ovaj rad sastoji se od pet poglavlja. Prvo poglavlje obuhvaća definiranje modela linearne regresije, njene vrste i opis glavne metode za procjenu njenih parametara. Drugo poglavlje fokusira se na osnovne pojmove statističkog učenja i glavne metode pri analizi i validaciji modela. Treće i četvrto poglavlje pružaju teorijski uvid u modele koji se shvaćaju kao unaprjeđenje modela linearne regresije. Naposljetku, peto poglavlje pruža opis problema i metoda koje su korištene, kao i numerički i grafički prikaz rezultata dobivenih u radu. Kao primarna literatura korištene su knjige (1) i (6). Kao suplement za bolje i dublje razumijevanje korišten je prilog (3).

Poglavlje 1

Modeli linearne regresije i procjena parametara modela

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

1.1 Oblikovanje modela

Neka je dan vektor prediktora (ulaznih varijabli ili varijabli poticaja) $X^T = (X_1, X_2, ..., X_p)$, kojim želimo predvidjeti vrijednost varijable odaziva (izlazne varijable) Y. Linearni regresijski model kao funkcija od X ima oblik

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon$$

gdje je ϵ Gaussov slučajni vektor s očekivanjem nula i varijancom σ^2 koja označava odstupanja od zavisnosti koja nisu opisana modelom. Nazivamo ju *slučajna greška* i pišemo $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. β_j, $j = 0, ..., p$ su nepoznati parametri modela odnosno koeficijenti. Varijable X_j mogu biti različite prirode:

1. kvantitativne varijable
2. transformacije kvantitativnih varijabli ($\log X, X^2, \sqrt{X}$)
3. polinomna reprezentacija ($X_2 = X_1^2, X_3 = X_1^3$)
4. dummy varijable, odnosno indikatori pripadnosti određenoj klasi. Ako statistička jedinica koju promatramo pripada nekoj klasi, vrijednost joj je 1, inače je 0. Najčešće se koriste kod podataka kao što su spol, rasa, politička opredjeljenost, itd. Da bi se modelirala kategorijalna varijabla koja može poprimiti k različitih vrijednosti, treba definirati $k-1$ dummy varijabli gdje je jedna kategorija, od ukupno k, referentna ili bazična kategorija.

Gotovo uvijek analizu neke pojave radimo pomoću više od jednog mjerenja. Tada je $(x_{i1}, x_{i2}, ..., x_{ip}, y_i), i = 1, 2, ..., n$ slučajni uzorak iz linearnog regresijskog modela kojeg opisuje n linearnih jednadžbi:

$$y_1 = \beta_0 + \beta_1 x_{11} + ... + \beta_p x_{1p} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + ... + \beta_p x_{2p} + \epsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + ... + \beta_p x_{np} + \epsilon_n.$$

Radi jednostavnosti koristimo matrični prikaz pa zbog toga redom označavamo:

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

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

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

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vektorski zapis modela postaje:

$$Y = X\beta + \epsilon \tag{1.1}$$

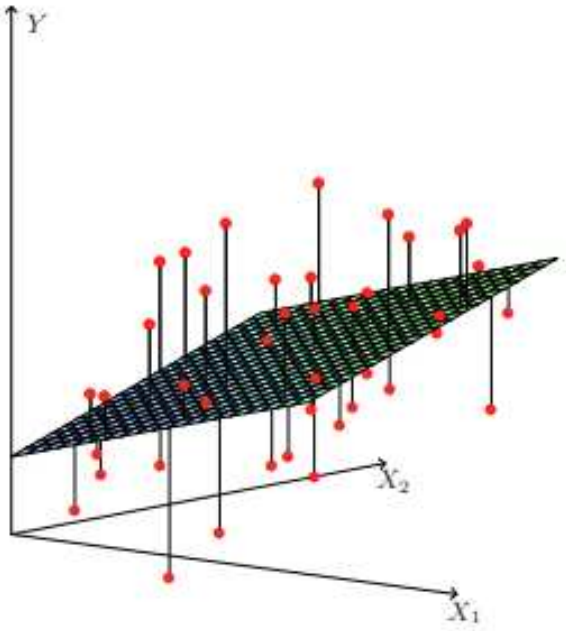
1.2 Procjena parametara modela

Cilj nam je pronaći nepoznate parametre β_j iz (1.1) koji će nam opisivati veličinu utjecaja pojedine prediktorske varijable X_j na Y. Najpoznatija metoda procjene parametara modela linearne regresije je *metoda najmanjih kvadrata*. Kao što joj ime kaže, koeficijente $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ odabiremo tako da minimiziramo sumu kvadrata reziduala:

$$RSS(\beta) = \sum_{j=1}^{n} (y_i - f(x_i))^2 = \sum_{j=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{1.2}$$

Definicija 1.2.1. *$\hat{\beta}$ je najbolji procjenitelj metodom najmanjih kvadrata uzorka (x_{ij}, y_i), $i = 1, 2, ..., n$, $j = 1, 2, ..., p$ ako $\hat{\beta}$ minimizira $RSS(\beta)$*

Sa statističkog stajališta, navedeneni kriterij je prihvatljiv ako mjerenje (x_i, y_i) predstavlja nezavisne slučajne ishode iz njegove populacije. Čak i ako x_i-ovi nisu izvučeni slučajno, kriterij je i dalje opravdan ako su y_i-ovi uvjetno nezavisni uz dani x_i. Slika 1.1 prikazuje geometriju smještanja najmanjih kvadrata u R^{p+1} dimenzionalni prostor:



Slika 1.1: Linearna prilagodba najmanjih kvadrata s $X \in R^2$. Izvor (1, str. 45)

U vektorskom zapisu, suma kvadrata reziduala ima oblik:

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta) \tag{1.3}$$

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

$$\frac{\partial RSS}{\partial \beta} = -2X^T (Y - X\beta) \tag{1.4}$$

$$\frac{\partial^2 RSS}{\partial \beta} = 2X^T X \tag{1.5}$$

Ako pretpostavimo (na trenutak) da X ima puni stupčani rang, odnosno da je $X^T X$ pozitivno definitna i prvu derivaciju izjednačimo s nulom

$$X^T (Y - X\beta) = 0 \tag{1.6}$$

Na taj način dobivamo jedinstveno rješenje

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{1.7}$$

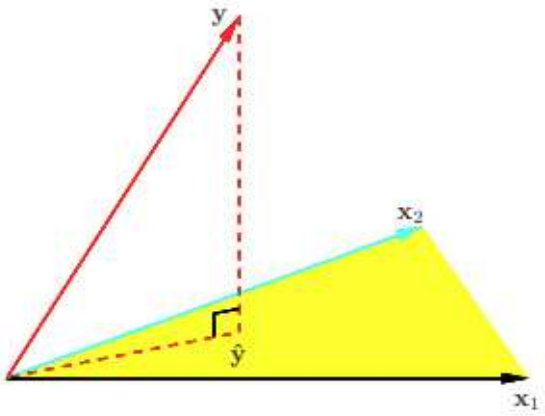
Sada imamo procjenjene parametre regresije te možemo izračunati predviđene vrijednosti zavisne varijable y za dani vektor nezavisnih varijabli x_0. Te vrijednosti dane su sa

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

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

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

gdje je $\hat{y} = \hat{f}(x_i)$.
Matricu $H = X(X^T X)^{-1} X^T$ koja se pojavljuje u jednadžbi ponekad nazivamo "kapa" matricom jer stavlja kapu na Y. Ona je simetrična i idempotentna, odnosno vrijedi $H^2 = H$ i $H^T = H$.
Reziduali $e = y - \hat{y}$ igraju važnu ulogu u istraživanju adekvatnosti modela i u otkrivanju odstupanja od temeljnih pretpostavki. Dakle, minimiziramo $RSS(\beta) = \|y - X\beta\|^2$ odabirom $\hat{\beta}$ tako da je vektor reziduala $y - \hat{y}$ okomit na potprostor razapet stupcima matrice X. $\hat{y}$ predstavlja ortogonalnu projekciju od y s obzirom na taj potprostor, što vidimo na slici 1.2.



Slika 1.2: N-dimenzionalna geometrija regresije najmanjih kvadrata s dva prediktora. Vektor $\hat{y}$ je ortogonalna projekcija izlaznog vektora y na hiperravninu razapetu ulaznim vektorima x_1 i x_2. $\hat{y}$ predstavlja vektor predikcije dobiven metodom najmanjih kvadrata. Izvor (1, str. 46)

Može se desiti da je neka nezavisna varijabla jako korelirana s nekom drugom (ili više njih) nezavisnom varijablom, odnosno da se neka x_i može skoro prikazati kao linearna kombinacija jednog ili više stupaca matrice X. Tada je matrica $X^T X$ loše uvjetovana ili skoro singularna što može dovesti do nestabilnosti metode najmanjih kvadrata u smislu da koeficijenti $\hat{\beta}$ nisu jedinstveno određeni. Dakle, cilj je izbaciti redudantne varijable koje

nam ne donose ništa novo u model.

Ovdje ćemo navesti neke činjenice o distribuciji podataka u modelu. Pretpostavili smo da su opažanja zavisne varijable y_i nekorelirana i imaju konstantnu varijancu σ^2 i da su x_i dani, to jest, nisu slučajni. Iz (1.7) slijedi da je varijacijsko - kovarijacijska matrica procjene parametara najmanjih kvadrata jednaka

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

Varijancu σ^2 obično procjenjujemo s

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

gdje je $Y - \hat{Y} = \hat{\epsilon}$, a nazivnik $N-p-1$ čini $\hat{\sigma}^2$ nepristranim procjeniteljem za σ^2. Kako bismo mogli donositi zaključke o parametrima i modelu, dodatno pretpostavimo da je uvjetno očekivanje od Y linearno u $X_1, X_2, ..., X_p$, te da su odstupanja od Y oko njegovog očekivanja aditivna i Gaussova. Stoga imamo

$$Y = E(Y|X_1, ..., X_p) + \epsilon$$

$$= \beta_0 + \sum_{j=1}^{p} X_j\beta_j + \epsilon,$$

gdje je ϵ slučajna varijabla s distribucijom $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Iz prethodnog slijedi

$$\hat{\beta} \sim \mathcal{N}(\beta, (X^T X)^{-1}\sigma^2), \tag{1.8}$$

te

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

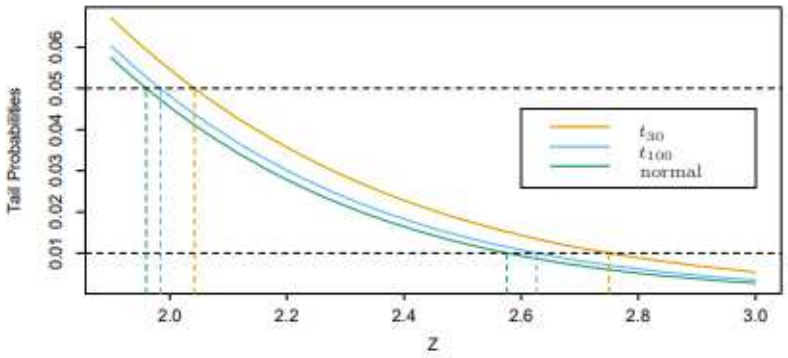
Dodatno, $\hat{\beta}$ i $\hat{\sigma}^2$ su statistički nezavisni.
Iz varijance reziduala dolazimo do kovarijance vektora zavisne varijable koja je jednaka

$$cov(Y) = E[(Y - X\beta)(Y - X\beta)^T] = E[\hat{\epsilon}\hat{\epsilon}^T] = \sigma^2 I. \tag{1.9}$$

Prikazane pretpostavke distribucije koristimo za testiranje hipoteza, te pouzdanih intervala za koeficijente $\beta_j, j = 0, ..., p$. Kako bismo testirali hipotezu da je pojedini koeficijent $\beta_j = 0$, formiramo standardizirani koeficijent ili Z - score:

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

gdje je v_j, j-ti dijagonalni element matrice $(X^T X)^{-1}$. Na temelju nulte hipoteze da je $\beta_j = 0$, z_j ima t_{n-p-1} distribuciju i zbog toga velika vrijednost od z_j vodi do odbacivanja navedene nulte hipoteze.
Ako $\hat{\sigma}$ zamijenimo s poznatom vrijednošću σ, z_j će imati standardnu normalnu distribuciju. Povećavanjem uzorka, razlika među repnim kvantilima standardne normalne i t-distribucije postaje zanemariva zbog čega uobičajeno koristimo kvantile normalne razdiobe. Navedenu situaciju možemo vidjeti na slici 1.3 gdje su istaknuti kvantili za testiranje hipoteza na razinama značajnosti $p = 0.01$ i $p = 0.05$.



Slika 1.3: Repni kvantili $\mathbb{P}(\|Z\| > z)$ za t_{30} i t_{100} distribuciju. Primjećujemo da povećavanjem uzorka razlika između kvantila t i standardne normalne distribucije postaje neznatna. Izvor (1, str. 48)

Često moramo simultano testirati značajnost neke grupe koeficijenata. Na primjer, zanima nas može li se kategorička varijabla s k razina isključiti iz modela. Tada moramo testirati mogu li se koeficijenti dummy varijabli, njih $k-1$, korištenih za predstavljanje tih razina postaviti na nulu. U tom slučaju koristimo F statistiku:

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

gdje je RSS_1 suma kvadrata reziduala za procjenjene vrijednosti većeg modela sa p_1+1 parametara, a RSS_0 suma kvadrata reziduala manjeg modela sa p_0+1 parametara. F statistika mjeri promjenu sume kvadrata po dodanom parametru u većem modelu i normalizirana je pomoću procjene od σ^2. $Z-score$-ovi z_j su ekvivalentni F statistici za izbacivanje pojedinog parametra β_j iz modela. S Gaussovim pretpostavkama i nultom hipotezom koja tvrdi da je manji model dovoljan, F statistika ima $F_{p1-p0,N-p1-1}$ distribuciju.
Slično, možemo izolirati β_j, te pomoću (1.8) dobiti $1-2\alpha$ pouzdani interval za β_j:

$$(\hat{\beta}_j - z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma}, \hat{\beta}_j + z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma}). \tag{1.10}$$

Također, možemo dobiti približnu grupu pouzdanosti za cijeli vektor parametara β:

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

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

1.3 Gauss - Markovljev teorem

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

Definicija 1.3.1. *Procjenitelj S_n je nepristran procjenitelj za τ ako vrijedi*

$$E[S_n] = \tau.$$

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

- $E[\epsilon_i] = 0$, za sve $i = 1, 2, ..., n$,
- $Var[\epsilon_i] = \sigma^2$, za sve $i = 1, 2, ..., n$,
- $cov(\epsilon_i, \epsilon_j) = 0$, za sve $i \neq j$.

Tada su procjenitelji najmanjih kvadrata *nepristrani*:

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

Teorem 1.3.2. *Neka vrijede Gauss - Markovljevi uvjeti te neka je $\hat{\beta}$ procjenitelj od β dobiven metodom najmanjih kvadrata. Tada vrijedi*

$$cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{1.11}$$

Dokaz. Označimo $A = (X^T X)^{-1} X^T$. Tada je $\hat{\beta} = AY$. Sada pomoću (1.9) imamo

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

□

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

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

Definicija 1.3.3. *Neka je Y vektor opaženih vrijednosti zavisne varijable, to jest, varijable odaziva. Statistika $T = t(Y)$ je:*

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

$$T = c^T Y,$$

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

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

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

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

   - *linearan procjenitelj*
   - *nepristran procjenitelj*
   - *u klasi svih nepristranih linearnih procjenitelja ima najmanju varijancu.*

Budući da su predviđene vrijednosti zavisne varijable y oblika $f(x_0) = x_0^T\beta$, fokusirat ćemo se na procjenu bilo koje linearne kombinacije parametara $\theta = a^T\beta$. Prema (1.7), procjena najmanjih kvadrata od $a^T\beta$ je

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

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

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

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

Teorem 1.3.4. *(Gauss-Markov). Neka je $\hat{\beta}$ procjenitelj metodom najmanjih kvadrata za parametre linearnog regresijskog modela te neka je $L(\beta) = a^T\beta$. Ako vrijede Gauss - Markovljevi uvjeti, tada je statistika*

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

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

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

$$b^T X = a^T. \tag{1.12}$$

Sada iz (1.11) i (1.12) imamo:

$$Var(b^T Y) = b^T cov(Y)b = b^T(\sigma^2 I)b = \sigma^2 b^T b \tag{1.13}$$

i

$$Var(a^T\hat{\beta}) = a^T cov(\hat{\beta})a = \sigma^2 a^T(X^T X)^{-1}a = \sigma^2 b^T X(X^T X)^{-1}X^T b. \tag{1.14}$$

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

$$Var(b^T Y) - Var(a^T\hat{\beta}) = \sigma^2[b^T b - b^T X(X^T X)^{-1}X^T b] = \sigma^2 b^T[I - X(X^T X)^{-1}X^T]b \geq 0. \tag{1.15}$$

Time je dokazana tvrdnja teorema. □

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

$$MSE(\bar{\theta}) = E[\bar{\theta} - \theta]^2 = Var(\bar{\theta}) + [E[\bar{\theta}] - \theta]^2 \tag{1.16}$$

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

1.4 Višestruka regresija

Višestruki linearni regresijski model je linearni model sa $p > 1$ ulaznih varijabli. Procjenu koeficijenata metodom najmanjih kvadrata (1.7) je najlakše razumjeti na primjeru jednostruke linearne regresije ($p = 1$) bez slobodnog člana, odnosno:

$$y = x\beta + \epsilon.$$

Procjena metodom najmanjih kvadrata i reziduali tada izgledaju:

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

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

Koristeći vektorske zapise $y = (y_1, ..., y_n)^T, x = (x_1, ..., x_n)^T$ i definiciju skalarnog produkta gornje jednadžbe postaju

$$\hat{\beta} = \frac{\langle x, y\rangle}{\langle x, x\rangle} \tag{1.17}$$

$$r = y - x\hat{\beta}. \tag{1.18}$$

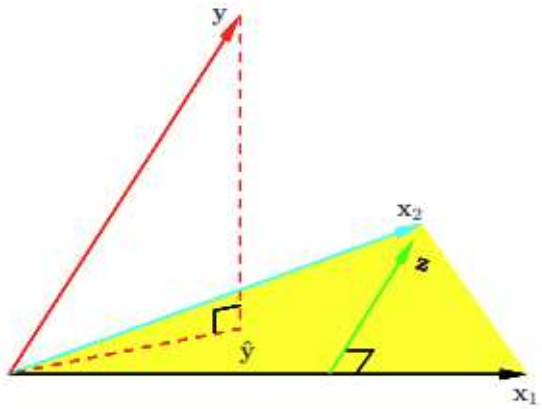
Vidjet ćemo da ova jednostavna jednovarijatna regresija daje temelj za višestruku linearnu regresiju. Pretpostavimom da su stupci matrice $X_1, ..., X_p$ ortogonalni, tj. vrijedi $\langle x_j, x_k\rangle = 0$ za svaki $j \neq k$. Tada su procjenitelji koeficijenata višestruke linaerne regresije $\hat{\beta}_j$ pomoću metode najmanjih kvadrata jednaki odgovarajućim procjeniteljima koeficijenata jednostruke linearne regresije, to jest

$$\hat{\beta}_j = \frac{\langle X_j, Y\rangle}{\langle X_j, X_j\rangle}. \tag{1.19}$$

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

$$\hat{\beta} = \frac{\langle X - \overline{X}\mathbf{1}, y\rangle}{\langle X - \overline{X}\mathbf{1}, X - \overline{X}\mathbf{1}\rangle}, \tag{1.20}$$

gdje je $\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n}$ i $\mathbf{1} = X_0$ je vektor s n jedinica. Procjena (1.20) rezultat je dvije primjene jednostruke regresije (1.19). Prvi je korak u tom postupku primjena jednostruke regresije od X na $\mathbf{1}$ da bi dobili rezidual $Z = X - \overline{X}\mathbf{1}$. Drugi je pak korak primjena jednostruke regresije od Y na rezidual Z kako bi dobili koeficijent $\hat{\beta}$. Na taj je način X, odnosno Y ortogonaliziran s obzirom na $\mathbf{1}$, odnosno Z. Dakle, drugi korak navedenog postupka je jednostruka regresija u kojoj se koriste ortogonalni prediktori $\mathbf{1}$ i Z.
Na slici 1.4 prikazan je navedeni proces za dvije općenite nezavisne varijable X_1 i X_2.



Slika 1.4: Regresija metodom najmanjih kvadrata ortogonalizacijom ulaza. Provedemo regresiju od X_2 na X_1, rezultirajući vektorom reziduala Z. Potom je opet primjenjena jednostruka regresija od Y na Z, te je dobiven koeficijent višestruke regresije od X_2. Pribrajajući projekcije Y na X_1 i Z dobivamo prilagodbu najmanjih kvadrata $\hat{Y}$. Izvor (1, str. 54)

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

1. Incijaliziraj $\mathbb{Z}_0 = X_0 = \mathbf{1}$
2. Za $j = 1, 2, ..., p$ primjeni jednostruku regresiju od X_j na $Z_0, Z_1, ..., Z_{j-1}$ kako bi dobio koeficijente $\hat{\gamma}_{lj} = \langle Z_j, X_j \rangle / \langle Z_l, Z_l \rangle, l = 0, 1, ..., j-1$ i vektor reziduala $Z_j = X_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} Z_k$
3. Primjeni jednostruku regresiju od Y na rezidual Z_p kako bi dobio procjenu $\hat{\beta}_p$.

Rezultat algoritma je procjenitelj:

$$\hat{\beta} = \frac{\langle Z_p, Y \rangle}{\langle Z_p, Z_p \rangle}. \tag{1.21}$$

Kako su ulazni podaci $Z_0, Z_1, ..., Z_{j-1}$ u drugom koraku ortogonalni, koeficijenti koji su ovdje dobiveni jednostrukom regresijom su zapravo koeficijenti višestruke regresije. Algoritam je poznat kao *Gram - Schmidt* postupak za višestruku regresiju.

U slučaju kada je X_p jako koreliran s nekime od ostalih X_k-ova, vektor reziduala Z_p će biti blizak nuli te iz (1.21) zaključujemo da će tada koeficijent $\hat{\beta}_p$ biti jako nestabilan. To će također vrijediti za sve varijable u koreliranom skupu. U toj bi situaciji svi Z − *score*-ovi mogli biti maleni što bi značilo da da bilo koja varijabla iz koreliranog skupa može biti

izbrisana, no ipak ih ne možemo sve izbrisati.

Iz jednadžbe (1.21) također dobivamo i alternativnu formulu za procjenitelj varijance:

$$Var(\hat{\beta}_p) = \frac{\sigma^2}{\langle Z_p, Z_p \rangle} = \frac{\sigma^2}{||Z_p||^2}. \tag{1.22}$$

Iz formule (1.22) zaključujemo da preciznost kojom možemo procijeniti $\hat{\beta}_p$ ovisi o duljini vektora reziduala Z_p.

prikažimo sada drugi korak algoritma u matričnoj formi:

$$X = Z\Gamma. \tag{1.23}$$

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

$$X = ZD^{-1}D\Gamma = QR, \tag{1.24}$$

odnosno QR dekompoziciju matrice X. Matrica Q je ortogonalna matrica dimenzije $n \times (p+1)$. QR dekompozicija prikazuje prikladnu ortogonalnu bazu za prostor stuapca matrice X. Iz ovog prikaza lako dolazimo do rješenja metode najmanjih kvadrata:

$$\hat{\beta} = R^{-1}Q^T Y. \tag{1.25}$$

Iz toga slijedi:

$$\hat{Y} = QQ^T Y. \tag{1.26}$$

Budući da je matrica R gornjetrokutasta, jednadžba (1.25) je lako rješiva.

1.5 Multivarijatna regresija

Pretpostavimo da imamo više varijabli odaziva $Y_1, ..., Y_k$ koje želimo procjeniti na temelju danih ulaznih varijabli $X_0, X_1, ..., X_p$. Pretpostavimo linearni model za svaki od njih:

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta jk + \epsilon_k = f_k(X) + \epsilon_k.$$

Ako pretpostavimo da imamo uzorak duljine N gornji izraz u matričnoj notaciji možemo zapisati kao

$$Y = XB + E,$$

gdje je Y $N \times K$ matrica odaziva s elementima y_{ik}, X je ulazna matrica dimenzije $N \times (p+1)$, B je $(p+1) \times K$ matrica parametara te E $N \times K$ matrica grešaka. Direktna generalizacija jednovarijatne sume kvadrata reziduala (1.2) dana je s:

$$RSS(\beta) = \sum_{k=1}^{K} \sum_{i=1}^{K} (y_{ik} - f_k(x_i))^2 = tr[(Y - XB)^T)(Y - XB)]. \tag{1.27}$$

Matrica procjenitelja ima istu formu kao i prije

$$\hat{B} = (X^T X)^B - 1 X^T Y. \tag{1.28}$$

Ako su greške $\epsilon = (\epsilon_1, ..., \epsilon_k)$ bilo bi prikladno modificirati (1.27). Preciznije, pretpostavimo da je $Cov(\epsilon) = \Sigma$. Dobijemo sljedeći izraz:

$$RSS(B; \Sigma) = \sum_{i=1}^{N} (y_i - f(x_i))^T \Sigma^{-1} (y_i - f(x_i))$$

što proizlazi direktno iz teorije o multivarijatnoj normalnoj distribuciji.

Poglavlje 2

Ocjena preciznosti modela statističkog učenja

2.1 Fleksibilnost modela

Kako bi smo ocijenili kvalitetu modela 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 odgovara stvarnoj vrijednosti varijable odziva. Najčešće korištena mjera u problemu regresije naziva se *srednjekvadratna greška*. Neka je f procjena za regresijsku funckiju dobivena nekim modelom. Srednjekvadratna greška, u oznaci MSE dana je sa:

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

gdje je $f(x_i)$ predikcija koju f daje za i-tu opservaciju. Ako je $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ skup za trening, tada znamo da f dobiven metodom najmanjih kvadrata minimizira MSE na danom treningu skupu. 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 modela regresije. Dakle naš cilj je odabrati onu metodu koja ima najnižu testnu MSE. Problem je što ponekad nemamo toliko podataka pa samim time ni skup testnih podataka na raspolaganju. Zbog toga je jedna od opcija koristiti model koji minimizira MSE na skupu za trening, ali ne postoji garancija da će ona imati i najmanju testnu MSE.

Mnogi algoritmi statističkog učenja (ridge regresija, lasso) imaju parametre koji kontroliraju ono što zovemo fleksibilnost modela. U slučaju kada metoda daje malu MSE na skupu za trening, a veliku testnu MSE kažemo da se dogodio overfitting podataka. To se događa jer metoda previše traži uzorke u podacima koji mogu biti samo uzrok slučajnosti, a ne 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. Zbog toga je osnovno svojstvo statističkog učenja da, neovisno o skupu podataka i metodi, MSE skupa za trening monotono pada, a testna MSE poprima takozvani $U-$ oblik. $U-$ oblik testne MSE posljedica je dvaju oprečnih svojstava metoda statističkog učenja, a to su pristranost i varijanca koje ćemo sada predstaviti.

2.2 Pristranost, varijanca i kompleksnost modela

Kako bismo lakše razumjeli ključne pojmove u ovom poglavlju uvodimo pojam funkcije gubitka.

Definicija 2.2.1. *Izmjerivo preslikavanje $L : \mathbb{R}^2 \longrightarrow [0, +\infty)$ zove se funkcija gubitka. U slučaju kada je Y kvantitativna varijabla, tipično se za funkciju gubitka uzima $L((y_1, y_2)) = (y_1 - y_2)^2$.*

Tada kvadratnu grešku od $f(X)$ definiramo kao $L((Y, f(X))) = (Y - f(X))^2$. Sada pomoću tih pojmova definiramo bitne pojmove - *testna greška i očekivana testna greška*.

Definicija 2.2.2. *Testna greška definirana je sa*

$$Err_\tau = \mathbb{E}[L(Y, f_\tau(X))|\tau]$$

gdje je τ fiksan skup za trening, a (X, Y) nezavisan i jednakodistribuiran kao slučajan uzorak za skup za trening.

Definicija 2.2.3. *Očekivana testna greška definirana je sa*

$$Err = \mathbb{E}^n[Err_\tau]$$

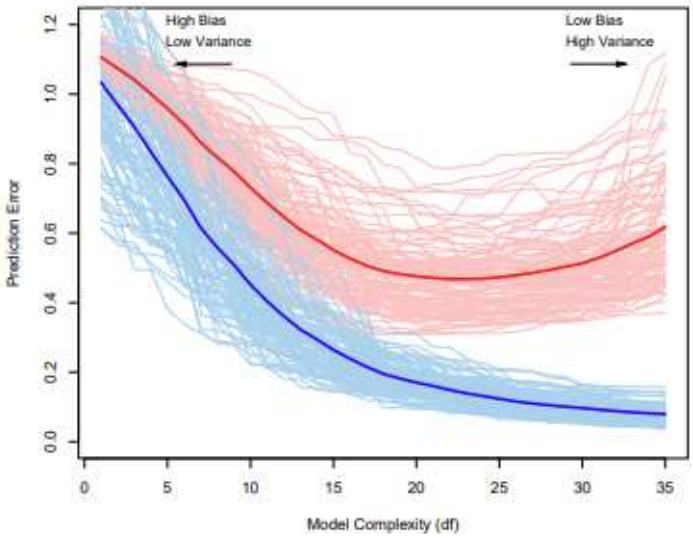
gdje je Err_τ prethodno definiriana testna greška, $\mathbb{E}^n$ očekivanje na produktnom vjerojatnosnom prostoru.

Testna greška nam govori kako se model dobiven pomoću konkretnog skupa za trening ponaša na novim podacima prosječno. S druge strane, očekivana testna greška gleda očekivanje te greške s obzirom na razne skupove za trening.

Definicija 2.2.4. *Greška predikcije na skupu za trening $\tau = \{(x_i, y_i) : i = 1, ..., n\}$ naziva se greška skupa za trening i definira se kao*

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

Na slici 2.1 crvene krivulje prikazuju testnu grešku Err_τ za 100 simuliranih skupova podataka τ duljine 50. Debela crvena krivulja predstavlja njihov prosjek tj. Err. Svijetlo plave krivulje prikazuju greške predikcije na skupovima za trening, $\overline{err}$. Debela plava krivulja predstavlja očekivanu grešku predikcije na skupu za trening, dakle $\mathbb{E}[\overline{err}]$.



Slika 2.1: Ponašanje testne greške, Err_τ i greške predikcije na skupu za trening $\overline{err}$, u ovisnosti o kompleksnosti modela. Izvor (1, str. 220)

Iz gornje slike 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 modela, manja je njezina pristranost, ali veća varijanca. Idealno bi bilo da dobivena funkcija ne daje jako različite rezultate za različite skupove za trening. Ako se to dogodi kažemo da model ima veliku varijancu. Pristranost se pak odnosi na grešku koja proizlazi iz pretpostavke o vezi između varijable odziva i prediktora. Linearni model primjer je nefleksibilnog modela za koji očekujemo visoku pristranost, dok će fleksibilniji modeli imati manju pristranost. Dakle vidimo da su pristranost i varijanca dva suprotna pojma. Odnos pristranosti i varijance se u statističkom učenju naziva bias-variance tradeoff. U nastavku iskazujemo teorem koji se često naziva *dekompozicija očekivane testne greške*.

Teorem 2.2.5. *Vrijedi*

$$\begin{aligned} Err &= \mathbb{E}^n[\mathbb{E}[(Y - f_\tau(X))^2]] \\ &= \mathbb{E}[\mathbb{E}^n[f_\tau(X)] - r(X))^2] + \mathbb{E}[\mathbb{E}^n[(f_\tau(X) - \mathbb{E}^n[f_\tau(X)])^2]] + Var(\epsilon) \end{aligned} \tag{2.1}$$

gdje je $r(x)$ regresijska funkcija, odnosno $r(x) = \mathbb{E}[Y|X = x]$, a (X, Y) je nezavisan od τ.

U gronjoj dekompoziciji u točki $x \in \mathbb{R}^p$ prvi član predstavlja kvadrat pristranosti, drugi član varijancu od $f(x)$, a treći član ireducibilnu grešku. Iz toga zaključujemo, ako želimo minimizirati očekivanu testnu grešku trebamo pronaći model sa malom pristranosti i malom varijancom.

Poglavlje 3

Odabir podskupa

Kao što smo već napomenuli, procjena najmanjih kvadrata često nije najbolja procjena zbog čega koristimo i pristrane modele. Dva su glavna razloga zbog čega nismo zadovoljni tom procjenom:

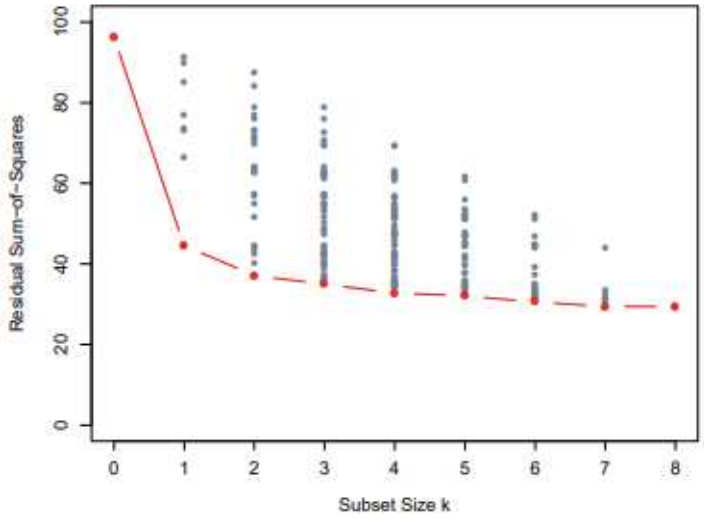
- Prvi razlog je preciznost, odnosno točnost procjene. Procjenitelji najmanjih kvadrata često imaju malu pristranost, ali veliku varijancu. Točnost procjene se ponekad može poboljšati smanjivanjem ili postavljanjem pojedinih koeficijenata na nulu. Radeći to, žrtvujemo malo pristranosti kako bi smanjili varijancu procjenjenih vrijednosti i poboljšali cjelokupnu točnost predviđanja.

- Drugi razlog je tumačenje. Često imamo velik broj procjenitelja, ali bismo htjeli odrediti manji podskup koji će prikazati najjače djelovanje. Odnosno želimo reducirati model tako da u njemu ostanu samo one nezavisne varijable čiji je učinak na zavisnu varijablu najveći te čijom će se linearnom kombinacijom lako opisati utjecaj novog mjerenja na ishod. U tom smislu, voljni smo žrtvovati neke sitne detalje kako bismo dobili širu sliku.

3.1 Odabir najboljeg podskupa

Kod konstrukcije regresijskog modela, uklanjanje irelevantnih varijabli učinit će model lakšim za interpretaciju i manje sklonim prekomjernom prilagođavaju podacima, samim time više generaliziranim.
Prva od metoda odabira podskupa varijabli koju ćemo opisati je Metoda odabira najboljeg podskupa (eng Best - Subset Selection method). Njen cilj je pronaći podskup nezavisnih varijabli X_i koje najbolje predviđaju varijablu odaziva Y uzimajući u obzir sve moguće kombinacije prediktora. Dakle, ako imamo na raspolaganju p prediktora, ova metoda kreira

modele s k varijabli gdje je redom $k \in \{1, 2, ..., p\}$ i za svaki k pronalazi model s najmanjom sumom kvadrata reziduala. Na slici 3.1 prikazujemo sve modele podskupova za poznati primjer modela raka prostate kod kojeg je $p = 8$. Donje crvene točke predstavljaju modele koji su prikladni za odabir putem metode odabira najboljeg podskupa. Bitno je napomenuti da najbolji podskup duljine 2 ne mora sadržavati varijablu koja je bila u najboljem podskupu duljine 1 i tako dalje. Također, donja crvena linija nužno opada, stoga ju ne možemo koristiti za odabir najbolje veličine podskupa. Odgovor na pitanje kako odabrati k uključuje kompromis između pristranosti i varijance (*bias - variance tradeoff*), zajedno sa subjektivnom željom za štednjom.



Slika 3.1: Svi mogući modeli podskupa za primjer raka prostate. Za svaku veličinu podskupa prikazujemo zbroj kvadrata reziduala za svaki model te veličine. Izvor (1, str. 58)

Prednosti ove metode su generaliziranje regresijskog modela uklanjanjem nepotrebnih prediktora dajući jednostavan i lako razumljiv model. Metoda pruža ponovljiv i objektivan način smanjenja broja prediktora u usporedbi s ručnim odabirom varijabli kojima se može manipulirati kako bi se služilo vlastitim hipotezama i interesima.
Mana ovog algoritma je da broj modela koji se moraju uzeti u obzir raste eksponencijalno s brojem prediktora u modelu.
Ostali pristupi o kojima raspravljamo u ovom poglavlju funkcioniraju na sličan način.

3.2 Stepwise unaprijed i unatrag

S obzirom da otkrivanje najboljeg modela među svim mogućim podskupovima prediktora može biti jako sporo (posebno za p puno veći od 40), pokušavamo pronaći brži način traženja najefikasnijeg modela.

Stepwise selekcija se pojavljuje u 2 oblika: unaprijed i unazad. Obje imaju za cilj maksimiziranje korelacije između Y i $\hat{Y}$ koristeći onoliko malo prediktora koliko je za to potrebno. Da bi se to postigla potrebna je neka vrsta pravila odlučivanja o tome mijenja li se R do željenog stupnja dodadavanjem ili uklanjanjem prediktora, budući da znamo da će dodavanje ili uklanjanje prediktora, osim u vrlo neobičnim okolnostima, promijeniti R u određenoj mjeri. R ovdje označava linearni korelacijski koeficijent koji govori o korelaciji i smjeru linearne povezanosti između dvije varijable i poprima vrijednosti u segmentu$[-1, 1]$.

Stepwise unaprijed započinje tako da se u model prvo ubaci samo slobodan član, a zatim od zadanog skupa od k varijabli pronalazi i ubacuje u model varijablu (nazovimo je P_1) s najvećom apsolutnom korelacijom s Y. Zatim od preostalih $k-1$ varijabli metoda pronalazi varijablu (P_2) kada se doda u model koji sadrži samo P_1. Metoda nastavlja istim postupkom sve dok ne dobijemo model sa svih k prediktora. Dakle, na kraju ćemo imati izgenerirano k modela, pa možemo odabrati model koji balansira između velike vrijednosti R i malog broja prediktora. Varijable koje su dodane kasnije u model najčešće ne dorpinose toliko povećanju od R, pa se varijable obično ne smatraju vrijednim zadržavanja u predikcijskom modelu.

Opisani postupak obično se ne provodi konstruiranjem svih k modela, već, umjesto toga korištenjem testa statističke značajnosti za odlučivanje treba li modelu dodati varijablu ili potpuno zaustaviti postupak odabira. U prvom koraku metode, variajbla P_1 se bira samo u slučaju da je korelirana s Y pomoću statistički značajnog kriterija, kao što je naprimjer p - vrijednost manja od 0.05. U slučaju da ne postoji niti jedna varijabla značajno korelirana s Y među njih k, postupak staje s modelom bez prediktora. Pretpostavljajući da je jedan pronađen, drugi korak bira prediktora, iz preostalih $k-1$ varijabli, koji najviše povećava R(do statistički značajne razine) kada se doda modelu koji sadrži samo P_1. Ako takve varijable ne postoje, postupak se zaustavlja na modelu koji ima P_1 kao jedini prediktor. Ako se pronađe takav, proces se nastavlja sve dok više ne postoje varijable koje prethodno nisu dodane u model, takve da povećavaju R do statistički značajne razine ili dok se ne iscrpe svi prediktori.

Ovaj se postupak može dodatno poboljšati dopuštanjem uklanjanja varijabli koje su prethodno već dodane u model. U kasnijim koracima metode stepenaste slekcije unaprijed, moguće je da varijabla koja je značajno povećala R u nekom od prethodnih koraka postane beznačajno povezana s Y nakon dodavanja drugih varijabli u model nakon nje. U tom slučaju uklanjanje te varijable ne bi znatno smanjilo R, pa ta varijabla postaje kandidat

za uklanjanje. Ovakvo usavršavanje *stepwise regresije unaprijed* je zapravo kombinacija prethodno opisane *stepwise regresije unaprijed* i *stepwise regresije unatrag* koju ćemo opisati u nastavku.

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

Odabir *stepwise unatrag* započinje punim modelom nako čega ih uklanja jednog po jednog koristeći neki od kriterija za uklanjanje. Prvi način je pomoću linearnog korelacijskog koeficijenta R. U stepwise selekciji unaprijed, varijabla se dodaje u trenutnom koraku ako najviše povećava R, dok se u stepwise selekciji unatrag uklanja ona koja najmanje snizi R u odnosu na ostale prediktore u modelu. Drugi način za uklanjanje prediktora iz modela je koristeći $z - score$. U tom slučaju je kandidat za izbacivanje varijabla s najmanjim $z - scoreom$. Za razliku od stepwise unaprijed koja se može uvijek koristiti, stepwise unatrag se može koristiti samo u slučaju kada je $n > p$.

Osim korištenja testova stastističke značajnosti odluka o ubacivanja ili izbacivanju parametara iz modela se može temeljiti i na AIC kriteriju. Kada u modelu ima k parametara i L je maksimalna vrijednost funckije vjerojatnosti, AIC vriejdnost se računa na način:

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

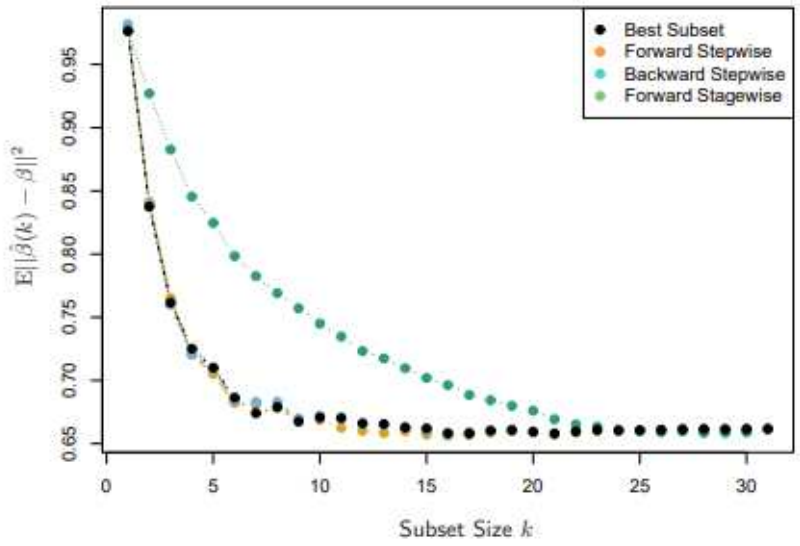
U svakom koraku dodavanja ili izbacivanja varijable, radnja će biti izvršena nad onom koja minimizira AIC vrijednost.

3.3 Stagewise regresija unaprijed

Stagewise regresija unaprijed je ograničenija od stepwise regresije unaprijed. Ona počinje kao stepwise regresija unprijed sa slobodnim članom jednakim $\overline{Y}$ i centriranim prediktorom s koeficijentima 0. U svakom koraku ova metoda identificira varijablu koja je najviše u korelaciji s trenutnim rezidualom. Zatim izračunava koeficijent jednostavne linearne regresije reziduala na odabranoj varijabli i dodaje je trenutnom koeficijentu odabrane varijable. Postupak se nastavlja sve dok nijedna varijabla nema značajne korelacije s rezidualima. U ovom procesu, za razliku od stepwise regresije unaprijed, nijedna od ostalih varijabli nije prilagođena u trenutku kada se varijabla dodaje u model. Posljedica toga je da stagewise regresija unaprijed može zahtijevati puno više od p koraka da bi pos-

tigla prilagodbu najmanjim kvadratima, pa je u povijesti zbog toga bila odbačena zbog neučinkovitosti.

Na slici 3.2 su prikazani rezultati proučavanja male simulacije u svrhu usporedbe regresije najboljim podskupom sa manje zahtijevnim alternativama, stepwise unaprijed i unatrag, te stagewise unaprijed odabirom. Vidimo da je njihovo djelovanje vrlo slično što se često događa. Također uočavamo da je kod stagewise regresije unaprijed potrebno više vremena kako bi postigla minimalnu grešku.



Slika 3.2: Usporedba četiriju metoda odabira podskupa na simuliranom problemu linearne regresije. Prikazana je srednja kvadratna greška procjenjenog koeficijenta $\hat{\beta}(k)$ za svaku veličinu podskupa. Izvor (1, str. 59)

Poglavlje 4

Metode sažimanja

Zadržavanjem određenog skupa prediktora i izbacivanjem ostalog, metode odabira podskupa generiraju model koji je interpretabilan i ima vjerojatno manju grešku predviđanja od punog modela. Međutim, jer je to diskretan proces, varijable su ili zadržane ili odbačene procesom što rezultira visokom varijancom pa ne smanjuje grešku punog modela. Za razliku od metoda odabira podskupa, metode sažimanja su neprekidnije, te stoga nisu sklone tako visokoj varijabilnosti.

4.1 Ridge regresija

Ridge regresija sažima regresijske koeficijente stavljanjem penalizacije na njihovu veličinu. Koeficijenti regresije minimiziraju penaliziranu sumu kvadrata reziduala,

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}\beta_j^2\}. \tag{4.1}$$

Ovdje je $\lambda \geq 0$ parametar složenosti koji kontrolira količinu sažimanja. Što je λ veći, koeficijenti će se više saženi. Koeficijenti ridge regresije se sažimaju prema nuli i jedan prema drugome. Ridge problem se može zapisati i kao:

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, za \sum_{j=1}^{p}\beta_j^2 \leq t. \tag{4.2}$$

Ovakav zapis izričito ograničava veličinu parametara. Postoji $1-1$ korespondencija između parametra λ u (4.1) i t u (4.2).
U slučaju kada ima mnogo koreliranih varijabli u linearnom regresijskom modelu, njihovi

koeficijenti mogu postati slabo određeni i izloženi visokoj varijanci. Ako postoji jako pozitivan koeficijent za jednu varijablu on se može "poništiti" slično velikim negativnim koeficijentom na njegovom koreliranom paru. Nametanjem ograničenja veličine koeficijenta, kao u (4.2), postižemo ublažavanje problema.

Rješenja dobivena ridge regresijom ovise o skaliranju ulaznih podataka, tako da se oni normalno standardiziraju prije rješavanja (4.1). Dodatno, uočimo da je slobodni član β_0 izostavljen iz izraza penaliziranja. To je zbog toga što bi procedura ovisila o početnoj točki izabranoj za Y kada bi slobodni član bio penaliziran. Odnosno, dodavanje konstante c svakom ishodu y_i ne bi pojednostavilo rezultat u pomaku predviđanja za isti iznos c. Može se pokazati da rješenje (4.1) može biti podjeljeno u dva dijela. Koriste se centrirani ulazni podaci što znači da se svaki x_{ij} zamijeni s $x_{ij} - \overline{x}_j$. Tada procjenjujemo β_0 sa $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Ostali koeficijenti se procjenjuju ridge regresijom bez slobodnog člana, koristeći centrirane x_{ij}. Ubuduće pretpostavljamo da je centraliziranje napravljeno, pa ulazna matrica X ima p (umjesto $p + 1$) stupaca.

Zapišimo (4.1) u matričnoj formi:

$$RSS(\lambda) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta, \tag{4.3}$$

iz čega se lako vidi da su rješenja ridge regresijom dana s:

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY, \tag{4.4}$$

gdje je I $p\times p$ jedinična matrica. Izborom kvadratičnog penaliziranja $\beta^T\beta$ dobiveno rješenje ridge regresijom ponovno bi bila linearna funkcija od Y. Prije invertiranja, rješenje dodaje pozitivnu konstantu dijagonali matrice X^TX i to čini problem nesingularnim, čak i ako X^TX nije punog ranga. U slučaju ortogonalnih ulaznih podataka procjene $\hat{\beta}^{ridge}$ su zapravo skalirane verzije problema najmanjih kvadrata, odnosno

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

Ridge regresija također se može izvesti kao srednja vrijednost od aposteriorne distribucije procjenitelja koeficijenata, s prikladno odabranom apriornom distribucijom. Aposteriorna distribucija način je da sažmemo ono što znamo o neizvjesnim veličinama u Bayesovoj analizi. To je kombinacija apriorne distribucije i funkcije vjerodostojnosti (eng. likelihood function), koja nam govori koje su informacije sadržane u opaženim podacima. Drugim riječima, aposteriorna distribucija sažima ono što znamo nakon što su podaci opaženi. Dakle pretpostavimo $Y_i \sim \mathcal{N}(\beta_0 + x_i^T\beta, \sigma^T)$, a koeficijenti β_j su apriorno distribuirani kao $\mathcal{N}(0, \tau^2)$, nezavisno jedan od drugoga. Zatim (negativna) log-aposteriorna gustoća od β, pretpostavljajući da su τ^2 i σ^2 poznati, jednaka je izrazu (4.1), za $\lambda = \frac{\sigma^2}{\tau^2}$. Pa je tako $\hat{\beta}^{ridge}$ mod aposteriorne distribucije. Budući da je distribucija Gaussova, to je ujedno i aposteriorna srednja vrijednost.

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

$$X = UDV^T, \tag{4.5}$$

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

$$X\hat{\beta}^{ls} = X(X^TX)^{-1}X^TY = UU^TY. \tag{4.6}$$

U^TY su koordinate vektora Y s obzirom na ortonormiranu bazu U. Sada su rješenja ridge regresije dana s

$$\begin{aligned} X\hat{\beta}^{ridge} &= X(X^TX + \lambda I)^{-1}X^TY \\ &= UD(D^2 + \lambda I)^{-1}DU^TY \\ &= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^TY. \end{aligned} \tag{4.7}$$

gdje su u_j stupci matrice U. Također vrijedi $\frac{d_j^2}{d_j^2+\lambda}$ budući da je $\lambda \geq 0$. Dakle ridge regresija, kao i linearna regresija računa koordinate od Y s obzirom na ortonormiranu bazu U, te zatim sažima dobivene koordinate faktorom $\frac{d_j^2}{d_j^2+\lambda}$. To znači da su na koordinate vektora baze s manjim d_j^2 primijenjene veće količine sažimanja.
Međutim, što mala vrijednost od d_j^2 zapravo znači? Dekompozicija singularne vrijednosti centrirane matrice X samo je drugi način izražavanja glavnih komponenti (eng. principal components) varijabli iz X Matrica kovarijance dana je sa:

$$S = X^TX/n$$

pa iz (4.5) imamo

$$X^TX = VD^2V^T, \tag{4.8}$$

što predstavlja svojstvenu dekompoziciju matrice X^TX (i matrice S, do na faktor n). Svojstvene vektore v_j (stupce matrice V) još nazivamo smjerovi glavnih komponenti od X. Prva glavna komponenta smjera je v_1 i ona ima svojstvo $z1 = Xv_1$, te ima najveću uzoračku

varijancu od svih normiranih linearnih kombinacija stupaca iz X. Ta uzoračka varijanca jednaka je:

$$Var(z_1) = Var(Xv_1) = \frac{d_1^2}{n}, \tag{4.9}$$

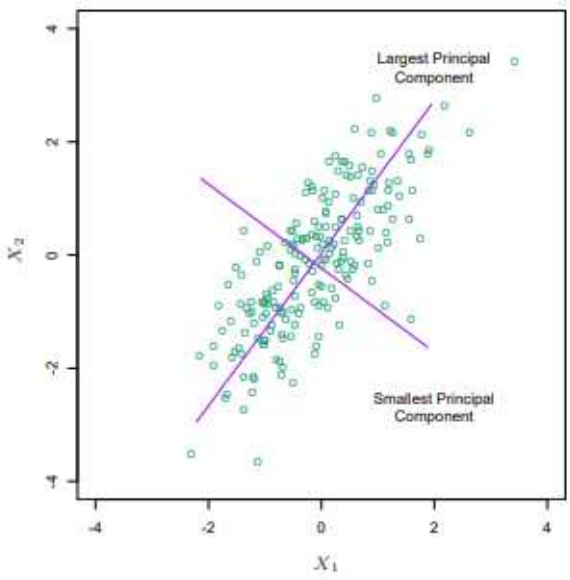
također možemo pisati i $z_1 = Xv_1 = u_1 d_1$. Dobivena varijabla z_1 zove se prva glavna komponenta od X zbog čega je onda u_1 normalizirana prva glavna komponenta. Sljedeće glavne komponente z_j imaju maksimalnu varijancu d_j^2/n te su ortogonalne s ranijima. Zadnja glavna komponenta ima najmanju varijancu. Zbog tako male singularne vrijednosti d_j odgovaraju onim smjerovima u prostoru stupaca matrice X koji imaju malu varijancu te ridge regresija najviše sažima upravo te smjerove.

Definirajmo sada funkciju efektivnih stupnjeva slobode prilagodbe ridge regresijom:

$$\begin{aligned} df(\lambda) &= tr[X(X^T X + \lambda I)^{-} - 1 X^T] \\ &= tr(H_\lambda) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}. \end{aligned} \tag{4.10}$$

Ovo je monotona padajuća funkcija u λ. Obično, u linearnoj regresijskoj prilagodbi s p varijabli stupanj slobode je p, odnosno broj slobodnih koeficijenata. Ideja je da iako svih p koeficijenata neće biti nula, oni su ograničeni, tj. kontrolirani s λ. Uočimo da je $df(\lambda) = p$ kada $\lambda = 0$ i $df(\lambda) \to 0$ kada $\lambda \to \infty$. Imamo još jedan stupanj slobode za slobodan član, no on je apriorno bio ukonjen.

Slika 4.1 prikazuje glavne komponente nekih podataka u dvije dimenzije. Kada razmotrimo prilagodbu linearne plohe duž te domene struktura podataka omogućava nam da preciznije odredimo njen nagib u dugom smjeru nego u kratkom. Ridge regresija štiti od potencijalne velike varijance gradijenta koji su procjenjeni u kratkim smjerovima. Tome je tako jer se podrazumijeva da će odaziv u smjerovima koji imaju veliku varijancu ulaznih varijabli biti sklon najvećoj varijabilnosti.



Slika 4.1: Glavne komponente nekih ulaznih podataka. Najveća glavna komponenta je smjer koji maksimizira varijancu prikazanih podataka, a najmanja glavna komponenta minimizira varijancu. Ridge regresija sažima koeficijente nisko - varijabilnih komponenti više nego onih visoko - varijabilnih. Izvor (1, str. 67)

4.2 Lasso regresija

Lasso je također metoda metoda sažimanja, baš kao i ridge. Između njih postoje neke male, ali značajne razlike. Lasso procjenitelj definiran je kao

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, za \sum_{j=1}^{p} |\beta_j| \leq t. \tag{4.11}$$

Baš kao i u ridge regresiji, može se reparametrizirati konstantu β_0 standardizacijom prediktora. Rješenje za $\hat{\beta}_0$ je $\overline{Y}$, i nakon toga prilagođavamo model bez slobodnog člana. Lasso problem možemo zapisati i u *Lagrangeovoj formi*:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin}\{\frac{1}{2}\sum_{i=1}^{N} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}. \tag{4.12}$$

Kao što vidimo postoje velike sličnosti s postavljanjem ograničenja na veličinu koeficijenata u ridge regresiji. L_2 ridge kazna $\sum_{j=1}^{p} \beta_j^2$ je zamiijenjena s L_1 lasso kaznom $\sum_{j=1}^{p} |\beta_j|$.

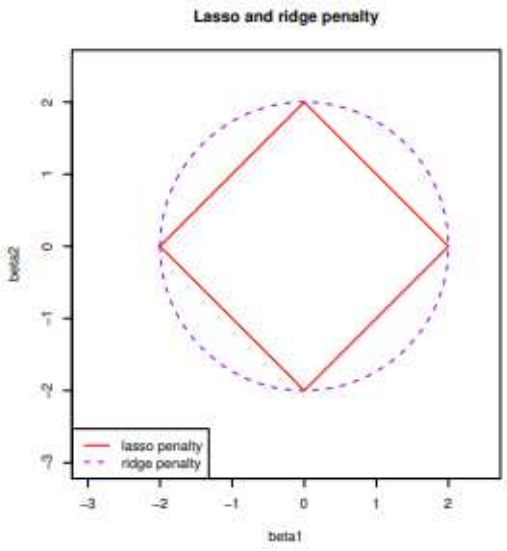
Potonje ograničenje dovodi do nelinearnosti njegovog rješenja u y_i zbog čega ne postoji zatvorena forma rješenja kao što je bio slučaj kod ridge regresije. Računanje rješenja Lasso regresije je kvadratičan problem programiranja, međutim dostupni su učinkoviti algoritmi za rješavanje cjelokupnog puta rješenja. Zbog prirode ograničenja, uzimanjem dovoljno malog t uzrokovat ćemo da neki koeficijenti budu točno jednaki nuli. Zbog toga lasso regresija radi neku vrstu neprekidnog odabira podskupa. Ako odaberemo t koji je veći od $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$ (gdje je $\hat{\beta}_j = \hat{\beta}_j^{ls}$, koeficijenti procjenjeni metodom najmanjih kvadrata), tada su i lasso procjenitelji $\hat{\beta}_j$. Međutim, s druge strane, ako je $t = t_0/2$, tada su koeficijenti sažeti u prosjeku 50%. Baš kao i veličina podskupa u metodi odabira podskupa, ili kazna kod ridge regresija, t mora biti izabran tako da minimizira procjenu očekivane greške predviđanja.

Da zaključimo, možemo uočiti kako su ridge i lasso regresija zapravo problemi najmanjih kvadrata, ali s uvjetima na parametre β_j. Ključna razlika imeđu ridge i lasso regresije je u domeni vrijednosti koje β_j može poprimiti. Uvjeti na parametre će rezultirati dvjema različitim kuglama l_1 i l_2:

$$\{\beta \in \mathbb{R}^p : |\beta_1| + ... + |\beta_p| <= t\}$$

$$\{\beta \in \mathbb{R}^p : \beta_1^2 + ... + \beta_p^2 <= s\}$$

Slika 4.2 prikazuje ograničenja za parametre $p = s = t = 2$.



Slika 4.2: Primjer efekta procjenitelja iz tablice 4.1 s obzirom na nerestringirane procjenitelje. Izvor (3, str. 110)

4.3 Usporedba metoda

U ovom poglavlju uspoređujemo 3 pristupa obrađena u prethodnim poglavljima: odabir podskupa, ridge regresija i lasso.

U slučaju ortonormirane ulazne matrice X sve 3 procedure imaju eksplicitna rješenja. Svaka metoda primjenjuje jednostavnu transformaciju na procjenitelj najmanjih kvadrata $\hat{\beta}_j$ kao što je detaljnije prikazano u tablici 4.1.

Ridge regresija radi proporcionalno sažimanje. Lasso translatira svaki koeficijent za konstantni faktor λ, smanjujući ih prema nuli, dok metoda odabira najboljeg podskupa odbacuje sve varijable s manjim koeficijentima od M-tog najvećeg.

Kod neortogonalnog slučaja odnos između metoda dočaran je slikom 4.4. Slika prikazuje metodu lasso (lijevo) i ridge regresiju(desno) kada imamo samo dva prediktora. Eliptične konture, centrirane u punom procjenitelju dobivenim metodom najmanjih kvadrata prikazuju područja jednakih suma kvadrata reziduala, $RSS(\beta_1, \beta_2)$ Ograničeno područje ridge regresije je disk $\beta_1^2 + \beta_2^2 \leq t$, dok je to za metodu lasso dijamant $|\beta_1| + |\beta_2| \leq t$. Obje metode daju rješenje prikazanog optimizacijskog problema tamo gdje eliptične konture pogađaju ograničena područja. Za razliku od diska, dijamant ima uglove. Ako se rješenje pojavi u uglu, onda postoji parametar β_j koji je jedank nuli. Kada je $p > 2$ dijamant postaje romboid koji ima mnogo stranica, uglova i ravnih rubova. Zbog toga je puno više mogućnosti da procjenjeni parametri budu jednaki nuli.

Također, možemo generalizirati ridge regresiju i lasso, i gledati ih kao Bayesovske procjenitelje. Pogledajmo kriterij:

$$\tilde{\beta} = \underset{\beta}{argmin}\{\sum_{i=1}^{N} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q\} \tag{4.13}$$

za $q \geq 0$. Konture za konstantne vrijednosti od $\sum_j |\beta_j|^q$ su prikazane na slici 4.5.

Na $|\beta_j|^q$ možemo gledati i kao log apriornu gustoću od β_j. Vrijednost $q = 0$ odgovara odabiru podskupa varijabli, s obzirom da kazna broji samo ne- nul koeficijente. $q = 1$ odgovara lassu, a $q = 2$ ridge regresiji. Uočimo da za $q \leq 1$ apriorna distribucija nema uniformni smjer, već koncentrira više mase u smjeru koordinatnih osi. Apriorna distribucija koja odgovara slučaju kada je $q = 1$ (lasso) je nezavisna dvostruko eksponencijalna distribucija za svaki ulaz, s gustoćom:

$$(1/2\tau)exp(-|\beta|/\tau)$$

za $\tau = 1/\lambda$. $q = 1$ je najmanji takav q za koji je ograničeno područje konveksno. Nekonveksno ograničena područja čine optimizacijski problem puno težim.

Promatrajući kriterij (4.13), mogli bismo pokušati iskoristiti i ostale vrijednosti od q osim 0, 1 i 2. Iako bi se moglo razmisliti o procjeni q-a iz podataka, iskustvo nam govori

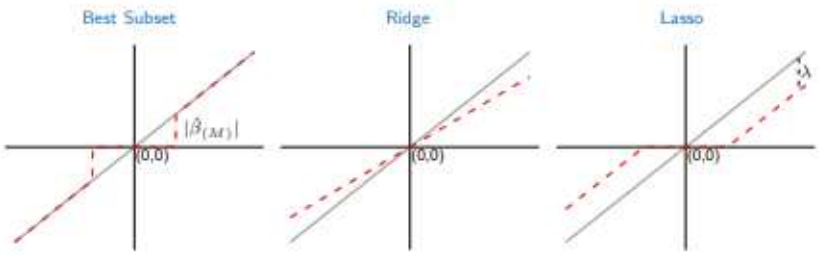
da se ne isplati truditi za višak nastale varijance. Vrijednosti od $q \in (1, 2)$ nam sugeriraju kompromis između lasso i ridge regresije. Za $q > 1$, $|\beta_j|^q$ je diferencijabilna u 0, pa nema sposobnost postavljanja koeficijenata na nulu, što je slučaj kod lasso metode ($q = 1$). Djelomično zbog toga razloga, uvedeno je penaliziranje elastičnom mrežom:

$$\lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1 - \alpha)|\beta_j|),$$

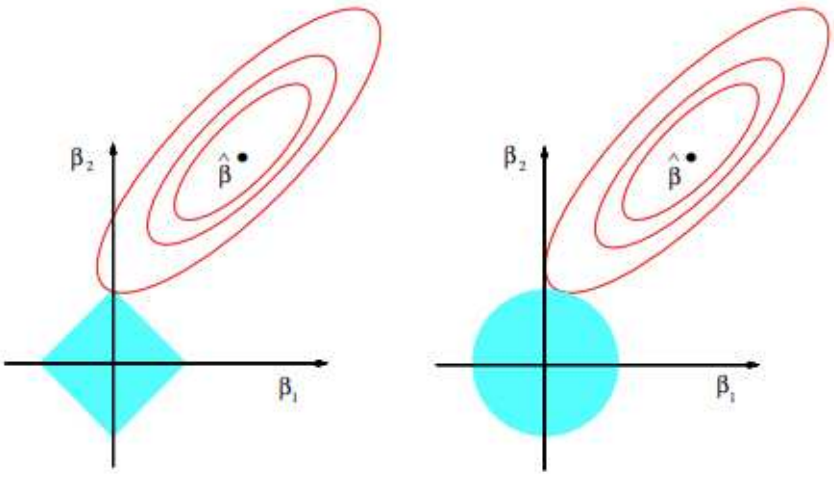
drugačiji kompromis između ridge regresije i lassa. Slika 4.6 uspoređuje L_q kaznu za $q = 1.2$ i kaznu elastičnom mrežom za $\alpha = 0.2$. Elastična mreža bira varijable kao lasso, te sažima koeficijente koreliranih prediktora kao ridge regresija.

Tablica 4.1: Procjenitelji za koeficijente β_j u slučaju ortonormiranih stupaca matrice X. M i λ su konstante odabrane odgovarajućim tehnikama, sign označava predznak argumenta, x_+ označava pozitivno dio od x, a I označava karakterističnu funkciju navedenog skupa.

Metoda	Formula
Najbolji podskup (veličine M)	$\hat{\beta}_j \cdot I(\|\hat{\beta}_j\| \geq \|\hat{\beta}_{(M)}\|$
Ridge	$\hat{\beta}_j/(1 + \lambda)$
Lasso	$sign(\hat{\beta}_j)(\|\hat{\beta}_j - \lambda\|)_+$

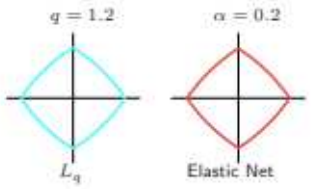


Slika 4.3: Primjer efekta procjenitelja iz tablice 4.1 s obzirom na nerestringirane procjenitelje. Izvor (1, str. 71)



Slika 4.4: Procjene za ridge regresiju (desno) i lasso(lijevo). Plava područja su ograničena područja pripadnih metode, a crvene elipse konture funkcije greške procjenitelja metodom najmanjih kvadrata. Izvor (1, str. 71)

Slika 4.5: Konture za konstantne vrijednosti od $\sum_j |\beta_j|^q$ za pripadne vrijednosti q. Izvor (1, str. 72)



Slika 4.6: Konture za konstantne vrijednosti od $\sum_j |\beta_j|^q$ za $q = 1.2$ (lijevo) i penaliziranje elastičnom mrežom $(1 - \alpha)|\beta_j|$ za $\alpha = 0.2$ (desno). Izvor (1, str. 73)

Poglavlje 5

Primjena opisanih metoda

5.1 Prikupljanje i analiza podataka

Nakon teorijske obrade modela statističkog učenja vrijeme je da se posvetimo i njihovoj primjeni u procjeni tržišne vrijednosti nogometaša. Prije nego što krenemo s primjenom i analizom rezultata obrađenih modela moramo se upoznati s podacima. Podaci su prikupljeni sa stranica Kaggle (4) i Statsbomb (5). Inicijalni podaci koje smo prikupili su bili spremljeni u 5 tablica. Svaka od tih tablica je sadržavala neke podatke koji će nam biti od značaja, a konačnu tablicu korištenu za modeliranje smo dobili agregiranjem, grupiranjem, spajanjem i čiščenjem gore spomenutih tablica. Cijeli kod kojim je rađen taj postupak pisan je u Pythonovom paketu Pandas. Konačni skup podaka koji smo dobili sastoji se od varijable odziva koja predstavlja tržišnu vrijednost nogometaša i od 20 prediktora koje prikazujemo u tablici 5.1. Sada ćemo detaljnije opisati značenje nekih varijabli. *Club points* prediktor nam govori koliko je uspjeha imao klub za koji igrač nastupa u ligi prvaka ili europskoj ligi, dok *league coefficient* mjeri koliko je snažna liga u kojoj se igračev klub nalazi. Važno je da imamo oba prediktora jer samo jedan od njih ne može jedinstveno odrediti snagu kluba (club points > 0 samo za klubove koji sudjeluju u europskim natjecanjima, dok u recimo engleskoj ligi ima puno jakih klubova koji ne sudjeluju u europskim natjecanjima). Igrači su po *Age group* parametru podjeljeni u 4 kategorije : ispod 20 godina, između 20 i 24 godina, između 25 i 29 godina, te više od 30 godina. Razlog zašto smo ih podijelili u grupe je taj što očekujemo da godine igrača neće linearno utjecati na tržišnu vrijednost igrača, već da će igrači maksimalnu vrijednost doseći oko dvadeset pete godine. Također vidimo da imamo podatke o igraču za svaku sezonu, ali također imamo i podatke o golovima, nastupima i drugim parametrima i za proteklu sezonu. Razlog tome je što tržišna vrijednost igrača neće ovisiti samo o njegovim performansama te sezone nego i o proteklim sezonama. Također vidimo da imamo nekoliko kategorijskih varijabli koje ćemo tretirati kao dummy varijable. Pojam dummy varijable smo definirali u prvome poglavlju.

Naš konačni dataset se sastoji od ukupno 25787 podataka od čega ćemo 80% koristiti kao skup za trening, a 20% kao testni skup. Dio tih podataka možemo vidjeti na slici 5.1.

Naziv varijable	Tip varijable	Opis varijable
position	kategorijska	golman, branič, vezni ili napadač
foot	kategorijska	je li igrač lijevak ili dešnjak
height in cm	kvantitativna	visina igrača u centimetrima
year	kvantitativna	sezona u kojoj je evaluacija napravljena
starting 11	kvantitativna	broj utakmica koje je igrač započeo u prvih 11
goals	kvantitativna	broj golova igrača
yellow cards	kvantitativna	broj žutih kartona igrača
red cards	kvantitativna	broj crvenih kartona igrača
assists	kvantitativna	broj asistencija igrača
matches played	kvantitativna	broj odigranih utakmica
international match	kvantitativna	broj međunarodnih utakmica
club points	kvantitativna	jakost kluba za kojeg igrač nastupa
matches last season	kvantitativna	broj nastupa igrača prošle sezone
starts last season	kvantitativna	broj startova igrača prošle sezone
yellow cards last season	kvantitativna	broj žutih kartona prošle sezone
red cards last season	kvantitativna	broj crvenih kartona prošle sezone
goals last season	kvantitativna	broj golova prošle sezone
assists last season	kvantitativna	broj asistencija prošle sezone
league coefficient	kvantitativna	jakost lige u kojoj igrač nastupa
age group	kategorijska	dobna kategorija u koju igrač spada

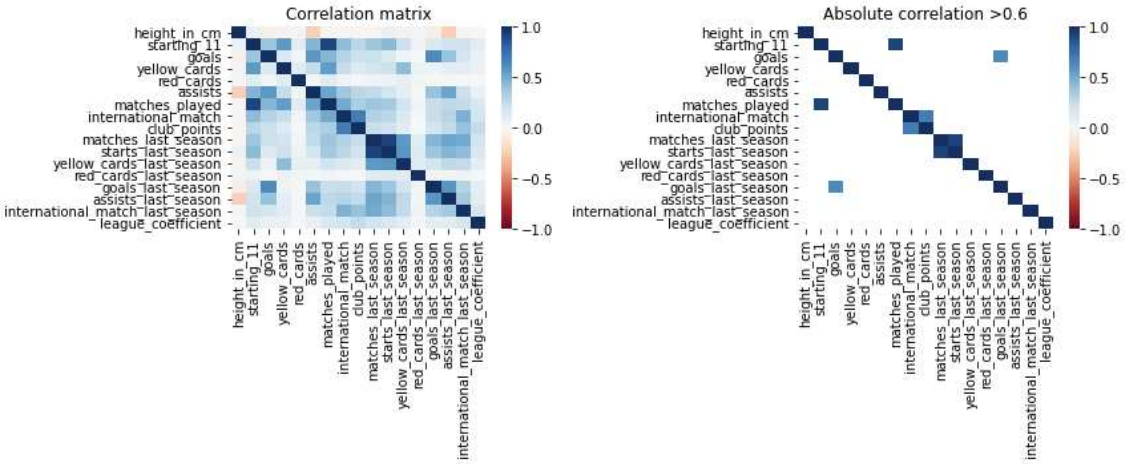
Tablica 5.1: tablica prediktora

starting_11	goals	yellow_cards	red_cards	assists	...	club_points
57	34	8	0	20	...	8
53	23	6	0	17	...	8
53	11	7	0	9	...	24
58	2	8	0	6	...	8
58	2	19	0	2	...	5
56	9	8	0	7	...	1
58	0	3	0	1	...	24
54	2	4	0	11	...	8
51	4	3	0	0	...	8
47	2	8	0	6	...	12

Slika 5.1: Podskup redaka i stupaca originalnog dataseta za razvoj modela

5.2 Linearna regresija

Prva metoda statističkog učenja koju ćemo ispitati na našim podacima je klasična linearna regresija. Potencijalni problem s kojim bismo se mogli susresti prilikom modeliranja linearne regresije, s obzirom na naše podatke je multikolinearnost među podacima. Zbog toga postoji mogućnost da naš model bude osjetljv na male promjene u podacima, što može rezultirati velikom testnom greškom. Stoga na sljedeće 3 slike prikazujemo matrice korelacije kvantitativnih nezavisnih varijabli:



Slika 5.2: Korleacijska matrica i korelacijska matrica gdje su zamaskirane sve apsolutne korelacije manje od 0.6

Iako bi za točniju provjeru multikolinearnosti trebali izračunati VIF metrike, iz gore navedenih slika se može vidjeti da postoji svega par varijabli koje su značajnije korelirane. No, bez obzira na ove spoznaje, provest ćemo sve ranije opisane modele, te ćemo iz tih rezultata moći više zaključiti kako multikoreliranost utječe na grešku u modelu. Sve modele smo dobili koristeći paket R, uz glavne smjernice iz (7). Prilagođavanjem našeg linearnog modela podacima skupa za trening dobivamo sljedeće rezultate:

```
Coefficients:
                                 Estimate   Std. Error t value         Pr(>|t|)
(Intercept)                  -1195.600802    59.841734 -19.979          < 2e-16 ***
positionDefender                 0.874536     0.206108   4.243 0.000022142929 02 ***
positionGoalkeeper               1.446916     0.317648   4.555 0.000005266288 26 ***
positionMidfield                 1.471383     0.178458   8.245          < 2e-16 ***
footright                       -0.144473     0.134693  -1.073           0.2835
height_in_cm                     0.018004     0.009949   1.810           0.0704 .
year                             0.588675     0.029645  19.857          < 2e-16 ***
starting_11                      0.283025     0.015214  18.604          < 2e-16 ***
goals                            0.501844     0.021891  22.925          < 2e-16 ***
yellow_cards                    -0.040525     0.026338  -1.539           0.1239
red_cards                       -0.062078     0.202962  -0.306           0.7597
assists                          0.378742     0.031363  12.076          < 2e-16 ***
matches_played                  -0.221697     0.014474 -15.317          < 2e-16 ***
international_match              0.221064     0.030796   7.178 0.000000000000 73 ***
club_points                      1.135241     0.029901  37.966          < 2e-16 ***
matches_last_season             -0.104310     0.014844  -7.027 0.000000000002 17 ***
starts_last_season               0.149980     0.015621   9.601          < 2e-16 ***
yellow_cards_last_season        -0.042719     0.026255  -1.627           0.1037
red_cards_last_season            0.008398     0.206429   0.041           0.9675
goals_last_season                0.391193     0.021753  17.984          < 2e-16 ***
assists_last_season              0.273457     0.031374   8.716          < 2e-16 ***
international_match_last_season  0.667823     0.025163  26.540          < 2e-16 ***
league_coefficient               0.537391     0.010633  50.540          < 2e-16 ***
age_group25-29                  -2.086688     0.142776 -14.615          < 2e-16 ***
age_groupover 30                -6.066501     0.161293 -37.612          < 2e-16 ***
age_groupunder 20                2.992884     0.648104   4.618 0.000003899687 05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Stupac "Estimate" daje procjene koeficijenata β dobivenih metodom najmanjih kvadrata. Stupac "Std. Error" procjenjuje standardnu devijaciju procjenjenih koeficijenata, dok je treći stupac vrijednost t - statistike, koja je zapravo jednaka količniku prva dva stupca. t - statistika u ovom slučaju je zapravo rezultat t testa koji testira koliko su značajni koeficijenti uz određeni prediktor, odnosno jesu li oni jednaki nuli. Što je apsolutna vrijednost t - statistike veća, p - vrijednost tog testa će biti manja i samim tim odbacujemo nultu hipotezu da je koeficijent jednak nula na većoj razini značajnosti. Vidimo da odbacujemo hipotezu da je koeficijent uz prediktor jednak 0 na razini značajnosti od 0.001 za 20 od 26 prediktora. Prediktori za čije koeficijente ne odbacujemo hipotezu da su jednaki 0 na spomenutoj razini značajnosti su: footright(dummy varijabla koju smo dobili iz varijable foot), height in cm, yellow cards, red cards, yellow cards last season i red cards last season.

Još jedna mjera koja nam govori koliko dobar je naš model je *prilagođeni* R^2. To je mjera koja nam govori koliko je varijabilnosti u zavisnoj varijabli objašnjeno prediktorima u modelu sa više prediktora. U našem modelu je ta vrijednost jednaka 0.5618. Međutim, ovo su sve podaci koji nam govore koliko je naš model dobro prilagođen skupu podataka za trening, a cilj metoda statističkog učenja je minimiziranje greške testnog skupa. Mjeru koju koristimo za validaciju greške testnog skupa u svim modelima je *RMSE(root mean square error)*. Za ovaj model testna greška jednaka je 7.773 (u milijunima eura). S obzirom da je raspon vrijednosti varijable odziva jako visok (vrijednosti variraju od 0.025 do 200), te da je greška na skupu za trening neznačajno manja (7.41), možemo biti zadovoljni kako je model reagirao na testnim podacima, odnosno nije došlo do "overfittinga". S obzirom da sve daljnje metode koje ćemo koristiti podešavaju određene parametre modela, morat ćemo koristiti postupak unakrsne validacije kojeg opisujemo u sljedećem poglavlju.

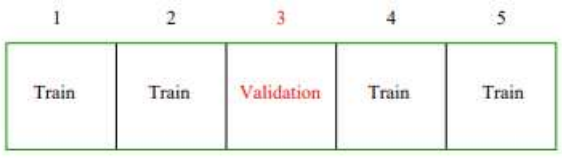
5.3 Unakrsna validacija

Unakrsna validacija jedna je od najčešće korištenih metoda u procjeni očekivane testne greške. U idealnom slučaju, kada imamo dovoljno velik skup podataka, možemo ga podijeliti na dva dijela - skup za trening pomoću kojeg prilagođavamo model i skup za validaciju koji koristimo za ocjenu dobivenog modela. Međutim, u praksi, često nemamo dovoljno velik skup podataka. Unakrsna validacija nastoji doskočiti tom problemu uzorkovanjem skupa za trening.

K-struka unakrsna validacija

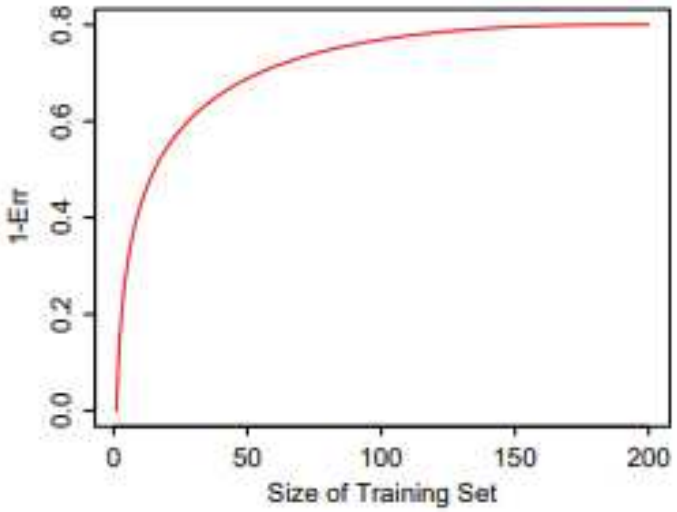
Ideja K-struke unakrsne validacije je podijeliti skup za trening na K dijelova približno jednakih veličina i potom $K-1$ dobivenih grupa iskoristiti za prilagodbu modela, a preostalu jednu grupu za validaciju modela, za koju izračunu MSE. Spomenuti proces se ponovi k puta, svaki put je drugi dio tretiran kao grupa za validaciju. U konačnici taj proces rezultira s K procjena testnih grešaka, $MSE_1, MSE_2, ..., MSE_K$. Tada je procjena testne greške dana sa:

$$CV = \frac{1}{K}\sum_{i=1}^{K} MSE_i.$$



Slika 5.3: Primjer za $K = 5$

Unakrsna validacija se uglavnom koristi kada imamo niz modela različite fleksibilnosti koja je određena parametrom podešavanja α. Tada se za svaki od tih modela provodi unakrsna validacija, te se odabire onaj model koji minimizira gore definiranu procjenjenu testnu grešku. Nakon toga se model prilagođava na cijelom skupu za trening. U tom slučaju ne zanima nas stvarna vrijednost procjene testne greške, već točka minimuma u kojoj se ta greška postiže. U praksi stvarnu testnu grešku ne znamo, no kod simuliranih podataka možemo ju izračunati i usporediti s procjenom dobivenom unakrsnom validacijom. Tada se da pokazuje da unatoč tome što procjena testne greške dobivena unakrsnom validacijom ponekad podcjenjuje testnu grešku, u većini slučajeva možemo dosta dobro procijeniti fleksibilnost modela. Međutim, ostaje nam pitanje koji K moramo odabrati. Slučaj kada je $K = n$ se u literaturi naziva *LOCV - Leave One out Cross Validation* i on daje prilično nepristranu procjenu testne greške s obzirom da je skup koji koristimo za prilagodbu približno jednak skupu za trening. Međutim, varijanca može postati velika budući da svaki od n modela prilagođavamo na gotovo jednakim podacima. Za $K = 5$ unakrsna validacija ima manju varijancu, ali ovisno o veličini skupa za trening pristranost može postati problem. Kako god, u praksi se $K = 5$ i $K = 10$ smatraju dobrim kompromisom s obzirom na odnos pristranosti i varijance. Sljedeća slika nam prikazuje hipotetsku ovisnost $1 - Err$ o veličini skupa za trening za $K = 5$.

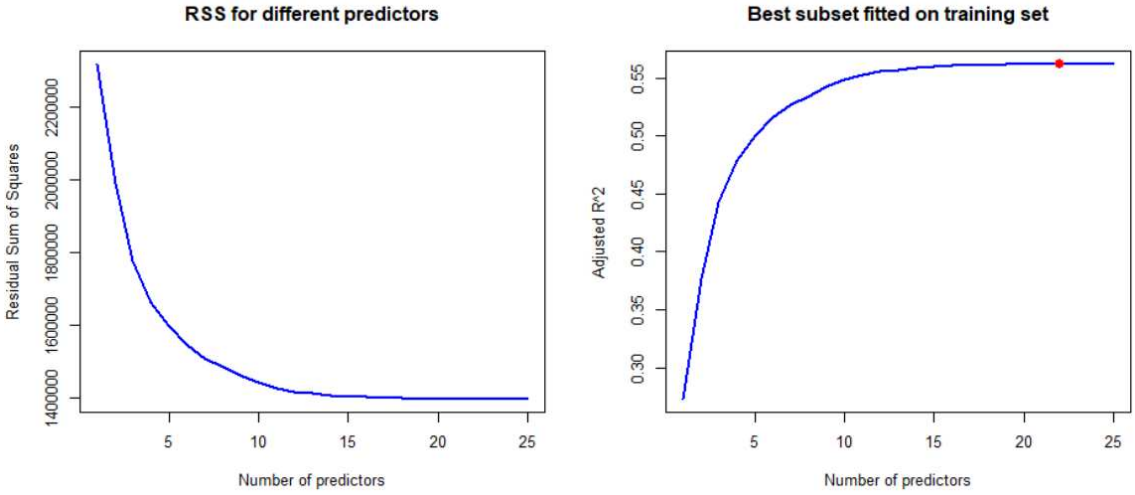


Slika 5.4: Hipotetska ovisnost očekivane testne greške o veličini skupa za trening za $K = 5$. Izvor (1, str. 243)

5.4 Metode odabira podskupa

Metoda odabira najboljeg podskupa

Prisjetimo se, kod metode najboljeg podskupa za svaki $m \in \{1, ..., p\}$, gdje je p broj prediktora, prilagođavamo model za sve moguće kombinacije m prediktora i biramo onaj koji minimizira RSS. S obzirom da za ovaj model koristimo k-struku unakrsnu validaciju, za svaki k ćemo izabrati p modela i izračunati testnu grešku na validacijskom. Nakon toga, za svaki m imamo k testnih grešaka, te računanjem aritmetičke sredine dobijamo procjenu očekivane testne greške za svaki m. Na kraju, biramo onaj model koji ima najmanju očekivanu procjenu testne greške. U našem slučaju je $p = 25$ jer imamo 25 varijabli prediktora (uključujući i *dummy* varijable). Prije nego krenemo s postupkom unakrsne validacije provjerimo kako se svaki od 25 najboljih modela prilagodio podacima skupa za treniranje. Za to ćemo koristiti metrike RSS (Residual sum of squares) i prethodno objašnjeni prilagođeni R^2.



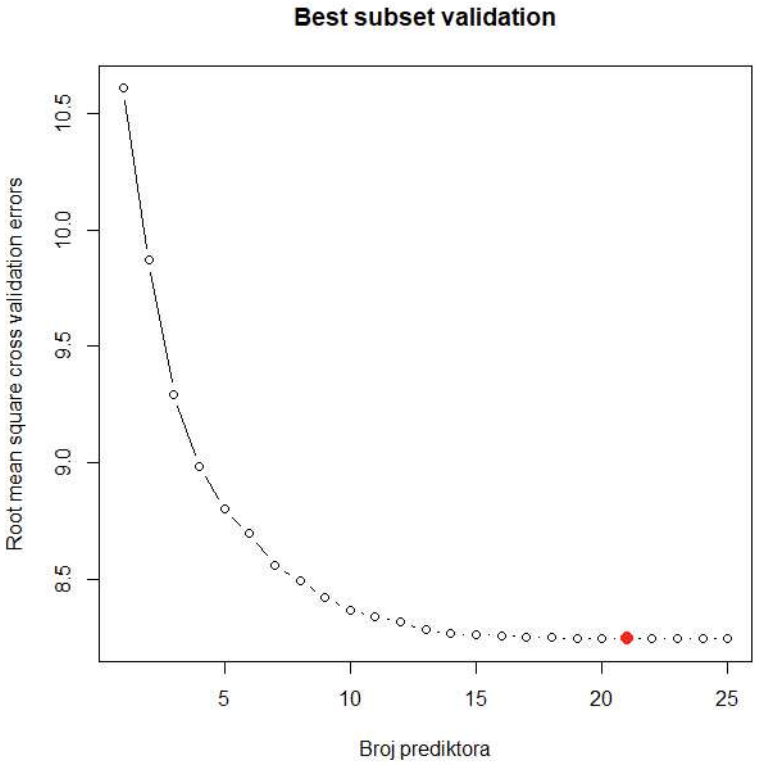
Slika 5.5: RSS i prilagođeni R^2 za najbolje podskupove različite veličine

Na temelju ovih slika, vidimo kako je prilagođeni R^2 najbolji za model sa 22 prediktora i on iznosi 0.5619. Međutim, to ne mora ništa značiti jer je nama od glavnog interesa testna greška, tako da krećemo s postupkom k-struke unakrsne validacije. Nakon provedenog postupke rezultati dobiveni unakrsnom validacijom su spremljen u matricu dimenzija 25×10 čiji je izgled u našem slučaju prikazan na slici 5.6.

```
       1        2        3        4        5        6        7        8        9       10       11       12       13       14       15
11.715145 10.905621 9.908644 9.723619 9.472014 9.662936 9.367151 9.335522 9.274914 9.206247 9.174812 9.139322 9.091071 9.084182 9.057845
10.984918 10.193108 9.608945 9.373585 9.181045 9.048509 8.853722 8.775451 8.671330 8.627690 8.603316 8.602909 8.576458 8.537443 8.549639
10.495060  9.923897 9.371028 9.106578 8.930412 8.784399 8.644330 8.546299 8.500719 8.441950 8.399086 8.369080 8.337401 8.328192 8.304581
10.389490  9.872185 9.349742 9.037986 8.818015 8.685996 8.542314 8.476420 8.400760 8.317359 8.284125 8.257368 8.251535 8.231228 8.234117
10.814560  9.962461 9.399843 9.044794 8.928001 8.777961 8.684349 8.629407 8.557523 8.494745 8.481892 8.440004 8.387439 8.405538 8.402403
10.190219  9.565292 9.059219 8.776855 8.594625 8.480300 8.386019 8.377386 8.293728 8.179619 8.180192 8.169021 8.153613 8.099830 8.096086
10.701835  9.676424 9.139904 8.818769 8.644082 8.463966 8.385971 8.290583 8.220688 8.181949 8.155927 8.113060 8.099289 8.074868 8.068442
10.738863 10.087873 9.614136 9.250239 9.089666 8.906239 8.854824 8.791886 8.739408 8.710133 8.674130 8.660424 8.594309 8.601721 8.591354
 9.924792  9.120467 8.645687 8.204990 8.028104 7.919357 7.779608 7.692006 7.629357 7.598621 7.578639 7.550856 7.504362 7.525275 7.503912
10.108064  9.388502 8.861733 8.500073 8.348683 8.243191 8.142702 8.037745 7.958702 7.918129 7.887245 7.867052 7.850166 7.818092 7.815939
       16       17       18       19       20       21       22       23       24       25
9.063559 9.063728 9.063324 9.047902 9.049606 9.047914 9.047350 9.048163 9.048765 9.048844
8.546181 8.541431 8.548214 8.528160 8.535040 8.534030 8.537410 8.539244 8.539570 8.539633
8.300928 8.290858 8.295394 8.297922 8.298368 8.298425 8.301049 8.301103 8.301155 8.301197
8.232566 8.220015 8.230344 8.217795 8.218161 8.222179 8.222258 8.222366 8.222373 8.222372
8.402058 8.394642 8.387468 8.372686 8.370545 8.368053 8.368643 8.368869 8.368975 8.368987
8.092256 8.083954 8.080632 8.082565 8.078386 8.075509 8.075069 8.075347 8.075544 8.075565
8.056795 8.043266 8.034315 8.027753 8.028675 8.021020 8.023185 8.021838 8.021944 8.021987
8.591959 8.585262 8.585793 8.590389 8.594015 8.593030 8.594157 8.594303 8.594360 8.594367
7.511081 7.505326 7.508990 7.509121 7.504355 7.502715 7.501258 7.501405 7.501438 7.501441
7.810162 7.800150 7.798209 7.793111 7.792697 7.792076 7.791164 7.791645 7.791950 7.792185
```

Slika 5.6: Greške na skupu validaciju u svakom koraku k-struke unakrsne validacije

Na mjestu (i, j) gornje matrice nalazi RMSE modela na skupu za vallidaciju s i prediktora čiji je skup za validaciju j-ta grupa. Na kraju se procjena testne greške za svaki broj prediktora dobiva računanjem prosjeka po stupcima, te konačni rezultat prikazujemo na sljedećoj slici:



Slika 5.7: Prosjek grešaka na skupovima za validaciju

S obzirom da smo dobili najmanju procjenjenu testnu grešku za modele s 21 prediktorom to će biti naš konačni model. Sada ga prilagođavamo na cijelom skupu za treniranje i računamo grešku na testnom skupu. Uočimo kako model na kojem je procjenjena testna greška najmanja (21 prediktor) ima različit broj prediktora od modela s najmanjim prilagođenim R^2 (22 prediktora)! Nakon prilagodbe modela dobijamo sljedeće koeficijente za određene prediktore:

(Intercept)	positionDefender	positionGoalkeeper	positionMidfield
-1202.32194551	0.88033942	1.40596522	1.33808970
height_in_cm	year	starting_11	goals
0.02049729	0.59180620	0.28005515	0.49955507
assists	matches_played	international_match	club_points
0.39644413	-0.22755094	0.19401617	1.20780235
matches_last_season	starts_last_season	yellow_cards_last_season	goals_last_season
-0.11502089	0.16275168	-0.06712034	0.41173464
assists_last_season	international_match_last_season	league_coefficient	age_group25-29
0.28648021	0.66176618	0.53550497	-2.19274760
age_groupover 30	age_groupunder 20		
-6.12432040	3.44166706		

Uočimo kako u analizi nema prediktora *foot right*, *yellow cards*, *red cards* i *red cards last season*. Testna greška koju dobijemo je 7.778 što je neznačajno više od modela linearne regresije (7.773), ali ipak vidimo da odabirom najboljeg podskupa nismo smanjili grešku našeg modela.

Stepwise unaprijed i unatrag

Procedura najboljeg podskupa je jako spora i računalno zahtjevna metoda. S obzirom da ona razvija model za svaki podskup prediktora ukupno smo morali razviti $\sum_{i=1}^{25} \binom{25}{i} = 33554431$ modela. S druge strane, stepwise unaprijed i unatrag su manje računalno zahtjevne metode. Stepwise unaprijed kreće bez prediktora, te se postupno u model dodavaju jedan po jedan prediktor. U svakom koraku se u model dodaje onaj prediktor koji najbolje poboljšava prilagodbu modela. Primijetimo, ukoliko se varijabla X nalazi u modelu s jednim prediktorom, ona se mora nalaziti i u svim ostalim modelima, dok kod metode odabira najboljeg podskupa nemamo takva ograničenja. Slično, stepwise unatrag kreće s punim modelom i i korak po korak odbacuje onaj prediktor koji najmanje utječe na odziv. Stoga je za svaku od ove dvije metode potrebno razviti ukupno $\sum_{i=1}^{25} i = 325$ modela. S obzirom da za naš skup podataka, obje metode daju potpune iste rezultate kao i metoda odabira najboljeg podskupa, nećemo ponovno stavljati rezultate, već ćemo na iduće dvije slike prikazati kojim su redoslijedom prediktori ubacivani, odnosno izbacivani iz modela.

	club_points	goals	league_coefficient	international_match_last_season	age_groupover 30	starts_last_season	assists	year	goals_last_season	matches_last_season
1	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
8	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
9	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
12	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
13	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
14	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
15	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
16	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
17	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
18	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
19	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
20	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
21	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
22	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
23	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
24	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
25	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Slika 5.8: Prvih 10 prediktora koji su ubačeni u stepwise unaprijed model

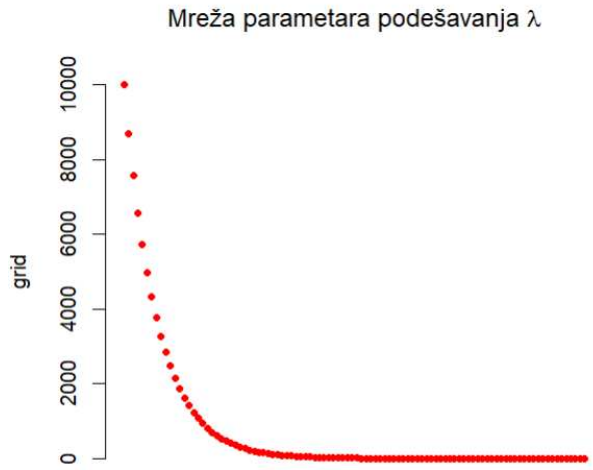
	red_cards	footright	red_cards_last_season	yellow_cards	height_in_cm	yellow_cards_last_season	positionDefender	positionGoalkeeper	age_groupunder 20	positionMidfield
25	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
24	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
23	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
22	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
21	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
20	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
19	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
18	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
17	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
16	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
15	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
14	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
12	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Slika 5.9: Prvih 10 prediktora koji su izbačeni iz stepwise unatrag modela

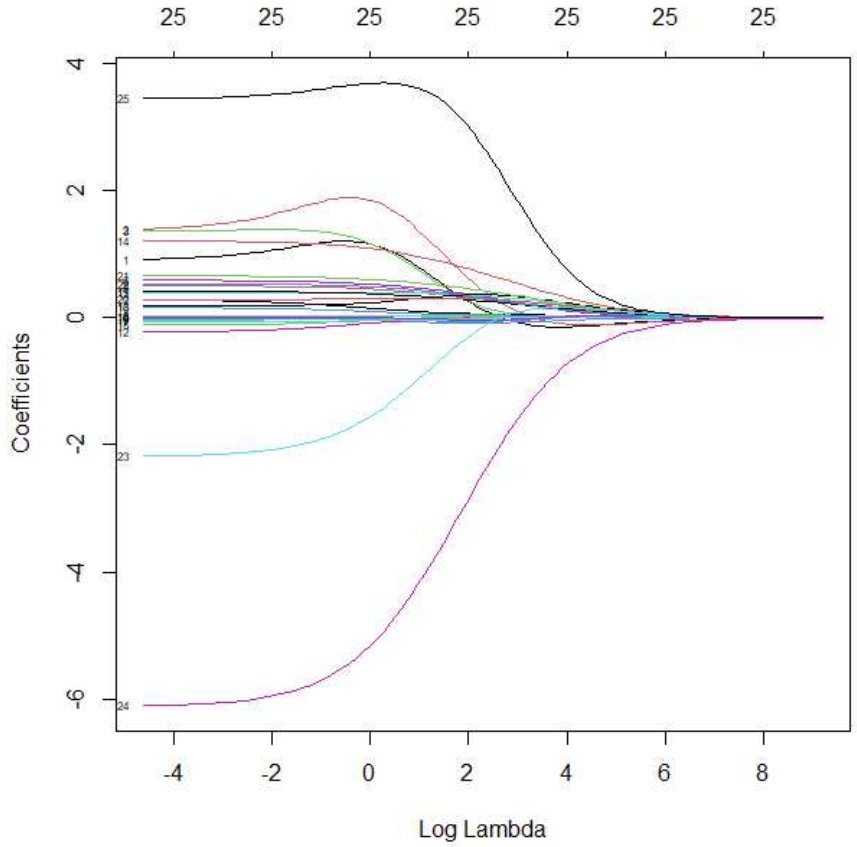
5.5 Metode sažimanja

Ridge regresija

Sljedeća metoda koju provodimo je Ridge regresija. Ona, za razliku od prethodno provedenih metoda nije metoda selekcije (ne izbacuje nužno prediktore iz modela), već metoda sažimanja, odnosno sažima koeficijente kako bi se smanjila varijanca modela. Ridge regresija je pogodna za rješavanje problema multikolinearnosti, za koju smo u analizi varijabli uočili da postoji među nekim varijablama. Za validaciju ove metode ponovno koristimo k-struku unakrsnu validaciju, te ćemo nastojati minimizirati testnu grešku podešavanjem parametra fleksibilnosti λ. Na idućoj slici možemo vidjeti mrežu parametara λ koju ćemo koristiti u našem modelu.

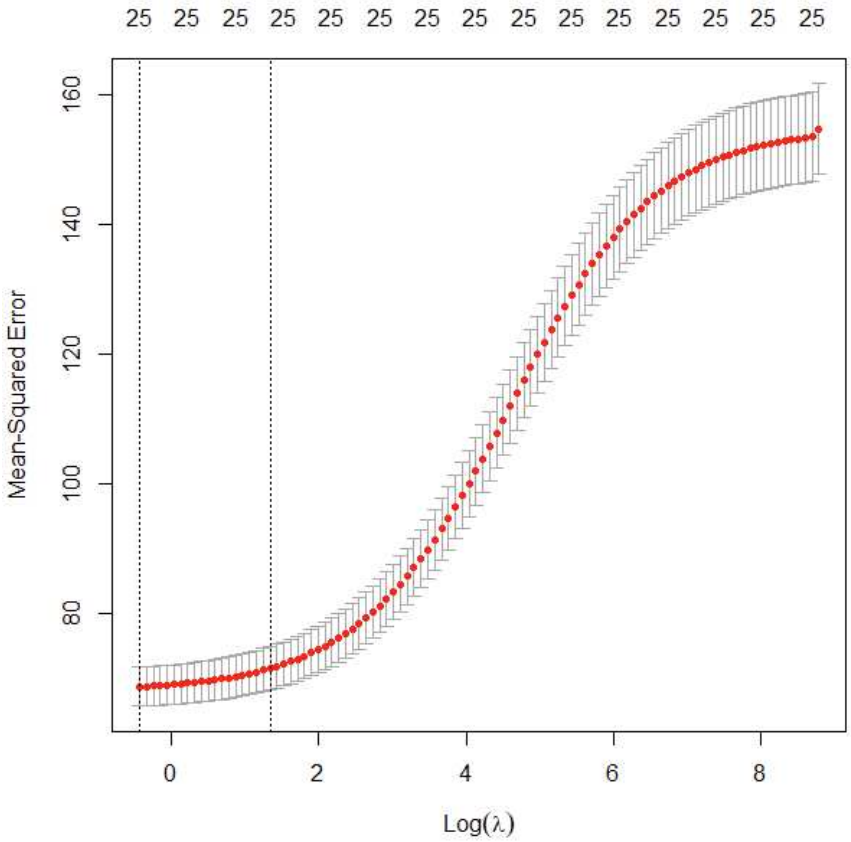


Sada provodimo Ridge regresiju. Za početak vizulizirajmo sažimanje koeficijenata u ovisnosti o parametru λ:



Slika 5.10: Veličina koeficijenata s obzirom na λ

Na y-osi je prikazana veličina koeficijenata svih prediktora, dok je na x-osi prikazana logaritamska vrijednost parametra λ zajedno s brojem prediktora koji ostaju u modelu. Uočimo da za svaki λ imamo maksimalan broj prediktora (25) iz razloga jer ridge regresija ne provodi odabir prediktora kao što smo već objasnili. Na slici 5.11 je prikazana procjena očekivane testne greške provedena unakrsnom validacijom. Crvena linija prikazuje očekivanu testnu grešku (MSE, RMSE se dobije korjenovanjem MSE) za različite parametre λ. Prva vertikalna linija prikazuju redom λ_{min}, odnosno onaj λ koji minimizira očekivanu testnu grešku. Druga vertikalna linija prikazuje λ_{1se}, odnosno onaj λ čiji je MSE udaljen za jednu standardnu grešku od MSE od λ_{min}. Za optimalni λ uzimamo λ_{min} koji je u našem slučaju jednak 0.65, dok se u praksi za optimalni λ često uzima i λ_{1se} kako bi se smanjila varijanca modela.



Slika 5.11: Procjena očekivane testne greške za ridge regresiju k-strukom unakrsnom validacijom

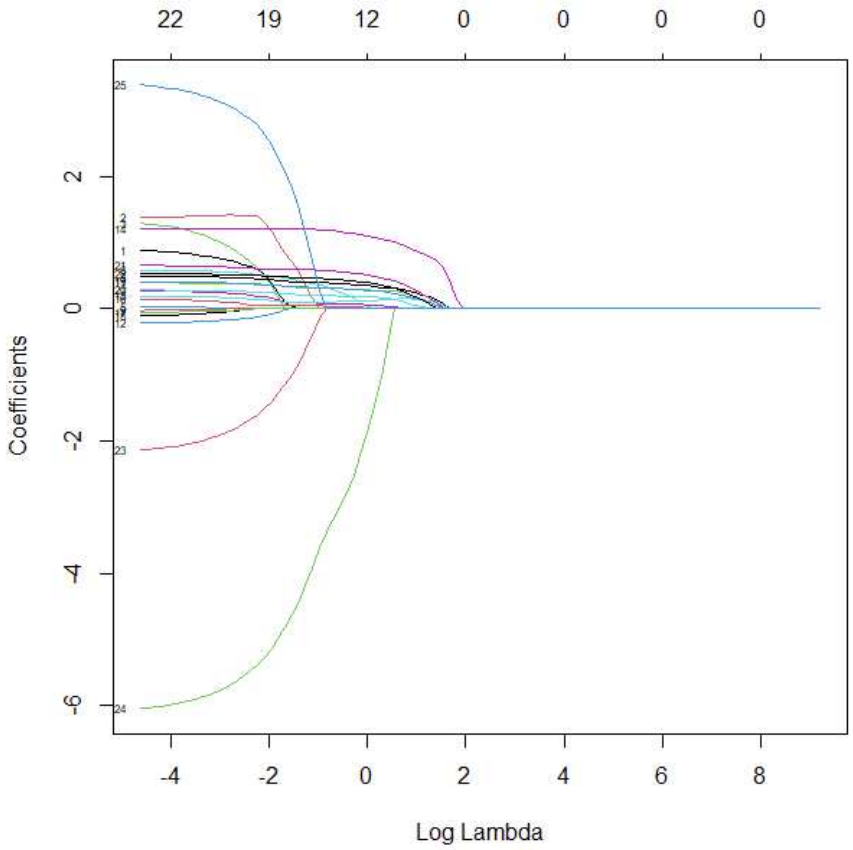
Sljedeći korak je prilagodba optimalnog modela na cijelom skupu podataka za trening. Nakon prilagodbe dobijamo sljedeće koeficijente za određene prediktore:

(Intercept)	positionDefender	positionGoalkeeper
-1106.426450857	1.201023457	1.889737412
positionMidfield	footright	height_in_cm
1.284458576	-0.015287098	0.028518427
year	starting_11	goals
0.543123557	0.174103873	0.470306817
yellow_cards	red_cards	assists
-0.020000845	-0.017126895	0.398774445
matches_played	international_match	club_points
-0.113380867	0.218283336	1.128595544
matches_last_season	starts_last_season	yellow_cards_last_season
-0.057890117	0.109799774	-0.048858587
red_cards_last_season	goals_last_season	assists_last_season
0.007184065	0.389193942	0.294002110
international_match_last_season	league_coefficient	age_group25-29
0.616788818	0.500754534	-1.738778788
age_groupover 30	age_groupunder 20	
-5.442954892	3.652916132	

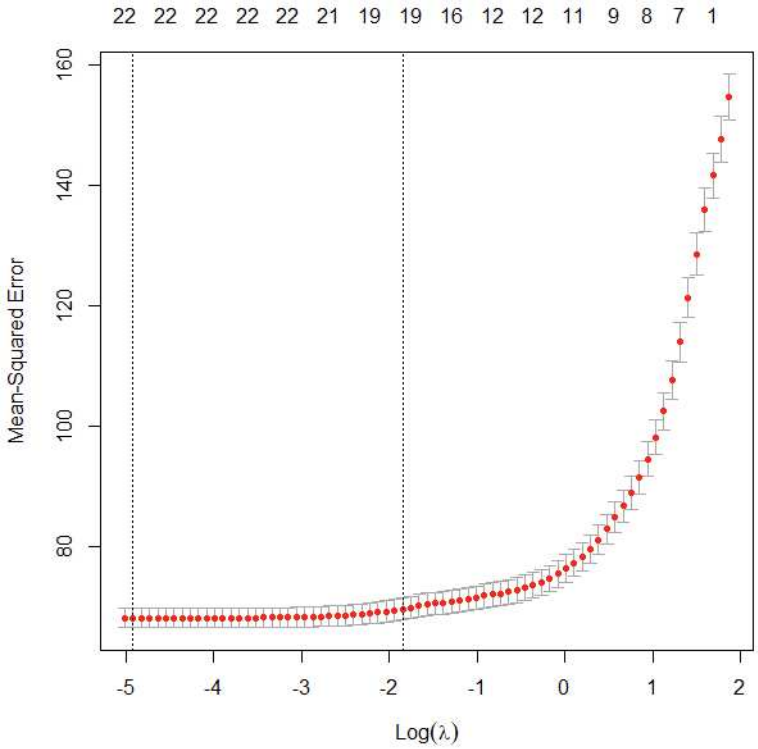
Uočimo kako su koeficijenti dobiveni ridge regresijom u pravilu nešto manji nego običnom linearnom regresijom, ali zbog veičine parametra λ ta razlika nije velika. Testna greška koju dobijemo je 7.745 što je nešto bolje od ostalih dosad provedenih modela. Za kraj nam ostaje još posljednji model kojeg testiramo, a to je Lasso.

Lasso

Isto kao i Ridge, Lasso je metoda sažimanja, uz razliku što uz sažimanje Lasso vrši i odabir prediktora. Razlog tome je razlika u načinu regularizacije koeficijenata u ove dvije metode. Postupak provođenja lasso modela je isti kao i kod ridge regresije, pa u nastavku prilažemo slike veličine koeficijenata i procjene očekivane testne greške s obzirom na parametar regularizacije λ.



Slika 5.12: Veličina koeficijenata s obzirom na λ



Slika 5.13: Procjena očekivane testne greške za lasso regresiju k-strukom unakrsnom validacijom

Na gornjim slikama uočavamo kako se povećanjem parametra λ smanjuje broj prediktora u modelu. Za lasso regresiju dobivamo da je optimalni parametar $\lambda = 0.71$. sa sljedećim koeficijentima uz prediktore:

(Intercept)	positionDefender	positionGoalkeeper
-1188.03261963	0.88000327	1.38867475
positionMidfield	footright	height_in_cm
1.29502139	0.00000000	0.02029564
year	starting_11	goals
0.58469241	0.27509109	0.49458232
yellow_cards	red_cards	assists
-0.02532181	0.00000000	0.39289321
matches_played	international_match	club_points
-0.21644758	0.18956255	1.20766687
matches_last_season	starts_last_season	yellow_cards_last_season
-0.10169133	0.14821743	-0.05146693
red_cards_last_season	goals_last_season	assists_last_season
0.00000000	0.40809078	0.28283294
international_match_last_season	league_coefficient	age_group25-29
0.65506880	0.53196344	-2.13275629
age_groupover 30	age_groupunder 20	
-6.04742328	3.38147638	

Uočavamo da su u procesu sažimanja koeficijenata, koeficijenti uz *foot right*, *red cards*, *red cards last season* poprimili vrijednost 0. Testna greška jednaka je 7.767.

5.6 Analiza rezultata

U ovom radu istražuje se utjecaj različitih faktora na tržišnu vrijednost nogometaša. Za validaciju rezultata korišteno je nekoliko metoda statističkog učenja, a to su: linearna regresija, metoda odabira najboljeg podskupa, stepwise regresija unaprijed, stepwise regresija unatrag, ridge regresija i lasso.

Nakon pripreme podataka i analize svih 20 prediktora zaključili smo da postoji multikolinearnost između nekih od njih. Zbog toga smo slutili da bi modeli sažimanja možda mogli davati bolje rezultate od linearne regresije, ali svejedno smo proveli sve opisane metode, te se njihovi rezultati nalaze u 5.2

Metoda	Prilagođeni R^2	Testna greška (RMSE)
Linearna regresija	0.5619	7.773
Metode odabira podskupa	0.5618	7.778
Ridge regresija	0.57	7.745
Lasso regresija	0.569	7.767

Tablica 5.2: Rezultati različitih modela

Uočavamo, da najbolje rezultate daje metoda ridge regresije, iako je s obzirom na raspon vrijednosti varijable odziva razlika među greškama zanemariva. Također, RMSE na skupu za trening nije puno manji od istog na testnom skupu, pa također možemo zaključiti da je došlo do "underfittinga" podataka, odnosno da naš model ima malu varijancu, a veliku pristranost. Jedna od potencijalnih mogućnosti za smanjenje pristranosti je dodavanje dodatnih prediktora u model. Također, uočimo da smo nevezano za poziciju igrača, promatrali iste metrike (broj golova, broj asistencija...), što nije najprecizniji pristup s obzirom da broj postignutih golova sigurno više utječe na vrijednost napadača, dok bi na vrijednost obrambenog igrača pozitivno utjecale neke druge metrike kao što je naprimjer broj utakmica bez primljenog gola.

Na kraju, zaključujemo kako postoji linearna veza između naših prediktora i tržišne vrijednosti nogometaša, ali za još preciznije rezultate preporučamo korištenje robusnijih i fleksibilnijih modela. Jedna od ideja za unaprjeđenje modela je korištenje različitih modela linearne regresije ovisno o poziciji igrača. Opcija je također i kombiniranje modela linearne regresije s modelima statističkog učenja koji omogučavaju grananje prediktora kao što je naprimjer model slučajnih šuma.

Bibliografija

[1] J.Friedmann, R. Tibshirani, T. Hastie, *The Elements od Statistical Learning*, Springer, New York, 2009.

[2] M. Huzak, *Vjerojatnost i matematička statistika*, predavanja, 2006., dostupno na http://aktuari.math.pmf.unizg.hr/docs/vms.pdf.

[3] W. N. van Wieringen, *Lecture notes on ridge regression*, 2021, https://arxiv.org/pdf/1509.09169.pdf

[4] https://www.kaggle.com/datasets/mexwell/football-data-from-transfermarkt?resource=download

[5] https://statsbomb.com/what-we-do/soccer-data/

[6] G .James, R. Tibshirani, T. Hastie, D. Witten *An introduction to statistical learning*, Springer, New York, 2013.

[7] https://philippbroniecki.com/ML2017.io/day5.html

[8] I. Faullend Heferer *Modeliranje turističke potrošnje korištenjem metoda statističkog učenja*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2023.

[9] J. Mucak *Metode procjene i odabira linearnog regresijskog modela*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2021.

[10] A. Ng. *Machine Learning*, Stanford University, 2018.

Sažetak

Ovaj diplomski rad bavi se predviđanjem tržišne vrijednosti nogometaša koristeći metode statističkog učenja. Na početku rada podsjetili smo se linearnog regresijskog modela i najpoznatije metode njegove prilagodbe - metode najmanjih kvadrata. U nastavku definiramo osnovne pojmove statističkog učenja, te uvodimo metode odabira prediktora u linearnim modelima i metode regularizacije (ridge i lasso regresija).

Svaku od metoda smo primijenili na podatke o nogometašima. Rezultati ukazuju na postojanje određene linearne veze između prediktora i varijable odziva uz očite znakove potrebe za povećanjem kompleksnosti modela. Iako sve metode dovode do približno sličnih rezultata, kao najpouzdanija metoda pokazala se ridge regresija s obzirom da je njena pripadna testna greška najmanja.

Summary

The main goal of this thesis is to predict the market value of football players using methods of statistical learning. At the beginning of the paper, we introduce the linear regression model and the least squares method. Furthermore, we defined the basic concepts of statistical learning and introduced methods for predictor selection in linear models and regularization methods (ridge and lasso regression).

Each of these methods was applied to football player data. The results indicate the existence of a certain linear relationship between predictors and the response variable with obvious signs of the need to increase the model's complexity. Although all methods lead to approximately similar results, ridge regression proved to be the most reliable method since its corresponding test error is the lowest.

Životopis

Rođen sam 17.9.1999. u Splitu gdje sam pohađao osnovnu i srednju školu. Tijekom osnovnoškolskog i srednjoškolskog obrazovanja aktivno sam se bavio nogometom i sudjelovao na raznim natjecanjima iz matematike. Nakon završetka srednje škole 2018. godine upisao sam preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Završetkom preddiplomskog studija stekao sam titulu sveučilišnog prvostupnika matematike 2021. godine kada sam upisao diplomski sveučilišni studij Matematička statistika na istom fakultetu.

Tijekom studija bio sam aktivni član fakultetske futsal sekcije i demonstrator iz kolegija Statistika. Od siječnja 2024. godine zaposlen sam kao podatkovni znanstvenik u turističkoj kompaniji *Maistra d.d.*