Metode procjene i odabira linearnog regresijskog modela

Mucak, Josip

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Josip Mucak

METODE PROCJENE I ODABIRA LINEARNOG REGRESIJSKOG MODELA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Huzak

Zagreb, rujan, 2021.

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. _____

Ovaj diplomski rad posvećujem svojim roditeljima. Zahvaljujem im na bezuvjetnoj podršci, ljubavi, vjeri, razumijevanju i odricanjima od prvoga dana. Hvala mojoj obitelji, prijateljima i kolegama koji su studentske dane učinili ljepšima i ležernijima. Posebno hvala mome mentoru prof. dr. sc. Miljenku Huzaku na strpljenu, pristupačnosti i stručnim savjetima prilikom izrade ovoga rada.

Sadržaj

Uvod

Često nas u različitim statističkim istraživanjima, odnosno predviđanjima zanimaju konkretne, kvantitivne (najčešće realne) izlazne vrijednosti procjene podataka. Regresijski modeli, odnosno problem regresije, daje nam upravo takva moguća rješenja. Regresijska analiza, kao dio takozvanog nadziranog statističkog učenja, jedna je od najkorištenijih statističkih tehnika za analiziranje višefaktorskih podataka. Njezina popularnost i efikasnost proizlaze iz konceptualno logičnog procesa korištenja jednadžbe kao objekta pomoću kojeg opisujemo ovisnost željene varijable (tzv. varijable odaziva) te uz nju vezanog skupa nezavisnih prediktorskih varijabli (varijabli poticaja). Matematička, odnosno statistička teorija koja se nalazi u pozadini regresijske analize je, iako relativno nova, elegantna i jako dobro razvijena što pridonosi korisnosti i efikasnosti regresijskih metoda u praktičnim primjenama.

Ovaj rad fokusirati će se na prvotnu vrstu regresijske analize, zvanu linearna regresija. Linearni regresijski modeli pretpostavljaju da je povezanost između zavisne varijable koju procjenjujemo i prediktorskih, nezavisnih varijabli linearna, odnosno da se može prikazati pomoću linearne funkcije. Model dobiven linearnom regresijom je, kao takav, vrlo jednostavan te na vrlo intuitivnoj i interpretabilnoj razini opisuje kako ulazne varijable utječu na izlaznu. Zanimljivo je da, u procesu predviđanja, ovakvi modeli ponekad mogu nadjačati neke kompleksnije, atraktivnije nelinearne modele. Također, možemo im i dodatno povećati okvir primjene transformiranjem ulaznih podataka, npr. logaritmiranjem. U radu će se opisivati te uspoređivati neke od osnovnih metoda dobivanja linearnih regresijskih modela. U prvom poglavlju prikazat će se postupak dobivanja koeficijenata modela metodom najmanjih kvadrata. Drugo poglavlje uvest će nas u metode odabira "najboljeg" podskupa prediktivnih varijabli. U trećem poglavlju opisat će se metode sažimanja kao što su regresija grebenom, metoda laso i regresija najmanjim kutom. Metode će se uspoređivati na primjeru iz knjige [3].

Poglavlje 1

Modeli linearne regresije i metoda najmanjih kvadrata

1.1 Metoda najmanjih kvadrata

Neka ja 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

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

gdje su $\beta_j, j = 0, ..., p$ nepoznati parametri modela, odnosno koeficijenti. Varijable X_j mogu biti različite prirode:

- kvantitativne varijable;
- transformacije kvantitivnih, npr. $\log X$, X^2, $\sqrt{X}$;
- interakcije između varijabli, npr. $X_3 = X_1 \cdot X_2$;
- polinomna reprezentacija, npr. $X_2 = X_1^2, X_3 = X_1^3$;
- dummy varijable, odnosno indikatori pripadnosti određenoj klasi (ili događaju). Ako statistička jedinica koja se opaža pripada toj nekoj klasi (ili je prisutan odabrani događaj), vrijednost joj je 1, a inače je 0. Najčešće se koriste kod podataka kao što su spol, rasa, politička opredijeljenost, itd. Broj dummy varijabli potrebnih za modeliranje određene kategorijalne varijable ovisi o broju vrijednosti koje ta varijabla može poprimiti. Da bi se modelirala kategorijalna varijabla koja može poprimiti k

različitih vrijednosti, treba definirati $k-1$ dummy varijabli gdje je jedna kategorija, od ukupno njih k, referentna ili bazična kategorija.

Neovisno o podrijetlu X_j, model je linearan u parametrima[3]. Cilj nam je pronaći nepoznate parametre β_j koji će nam opisivati veličinu utjecaja pojedine prediktorske varijable X_j na Y. Najpoznatija metoda procjene koeficijenata $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ je metoda najmanjih kvadrata. Neka je zadan skup opažanja $(x_1, y_1)...(x_N, y_N)$, gdje je svaki $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ vektor mjerenja za i-ti slučaj budući da u primjenama obično imamo više opažanja svake varijable. Koeficijente β odabiremo minimizirajući rezidual sume kvadrata:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2.$$

Označimo s X $N\times(p+1)$ matricu ulaza (dizajna) koja ima sljedeći oblik: $\begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_{N1} & \dots & x_{Np} \end{bmatrix}$,
te s y N-dimenzionalni vektor $(y_1, y_2, ..., y_N)^T$. Tada problem možemo zapisati kao:

$$RSS(\beta) = (y - X\beta)^T (y - X\beta).$$

Diferenciranjem po β dobivamo:

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

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

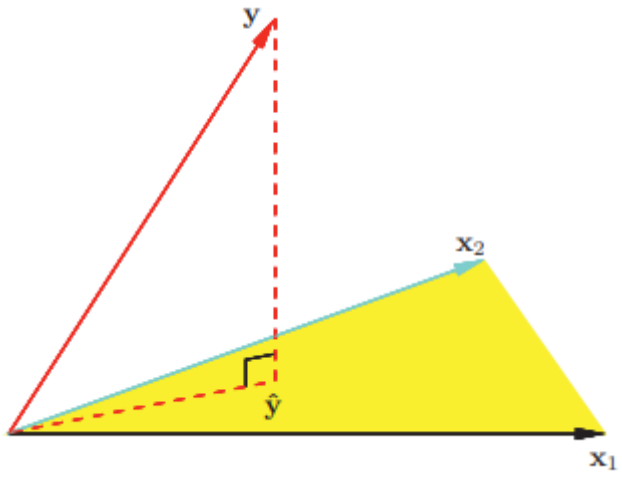
Prepostavimo da je X punog stupčanog ranga tako da je matrica $X^T X$ pozitivno definitna.U tom slučaju pretpostavljamo da je $N > p$. Izjednačavanjem obje jednadžbe s nula dobivamo jedinstveno rješenje sustava:

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

Procijenjene vrijednosti tada su jednake:

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

Reziduali $e = y - \hat{y}$ igraju važnu ulogu u istraživanju adekvatnosti modela i u otkrivanju odstupanja od temeljnih pretpostavki [1]. 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 1.1 u prostoru $\mathbb{R}^N$.



Slika 1.1: 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.

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.

Do sada smo napravili minimalne pretpostavke o stvarnoj distribuciji podataka[3]. Da bismo utvrdili svojstva uzorkovanja $\hat{\beta}$, pretpostavimo da su Y_i nekorelirane i da imaju konstantnu varijancu σ^2, te da su x_i fiksni (ne slučajni).

Varijacijsko – kovarijacijska matrica procjene koeficijenata dobivenih metodom najmanjih kvadrata je dana s:

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

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 nazivnik $N-p-1$ čini $\hat{\sigma}^2$ nepristranim procjeniteljem za σ^2.

Kako bismo mogli donositi zaključke o parametrima i modelu, dodatno pretpostavljamo da je uvjetno očekivanje od Y linearno u $X_1, ...X_p$, te da su odstupanja od Y oko njegovog očekivanja aditivna te Gaussova. Stoga imamo

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

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

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

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

te

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

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

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

gdje je v_j j-ti dijagonalni element matrice $(X^T X)^{-1}$[3]. Pod pretpostavkom nulte hipoteze koja tvrdi da je neki $\beta_j = 0$, z_j je distribuiran kao t-distribucija sa $N-p-1$ stupnjeva slobode pa će stoga velike apsolutne vrijednosti od z_j ishoditi odbacivanjem nulte hipoteze. Ako je $\hat{\sigma}$ zamijenjena poznatom vrijednošću σ, tada Z_j ima standardnu normalnu razdiobu. Razlika između repnog kvantila t-raspodjele i standardnog normalnog postaje zanemariva kako se veličina uzorka povećava, tako da obično koristimo normalne kvantile[3].

Često moramo simultano testirati značajnost neke grupe koeficijenata. Na primjer, zanima nas može li kategorička varijabla s k razina biti isključena 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:

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

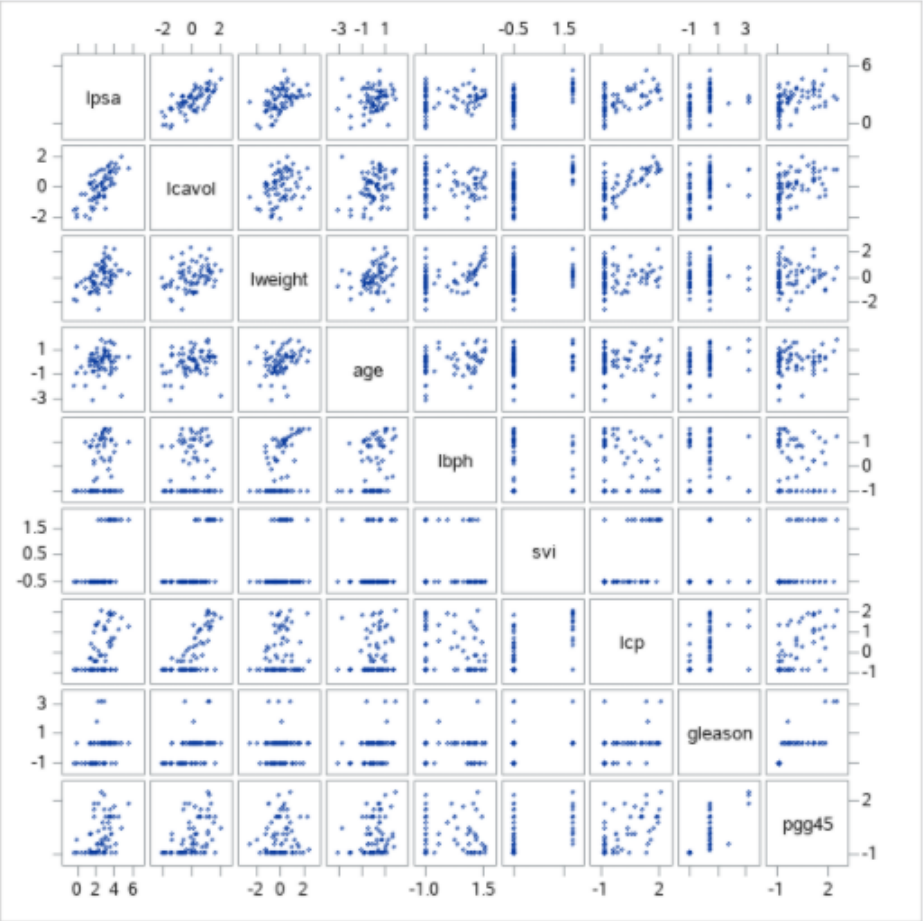
gdje je RSS_1 suma kvadrata reziduala za procjenjene vrijednosti većeg modela sa $p_1 + 1$ parametara, a RSS_0 suma kvadrata reziduala manjeg modela s $p_0 + 1$ parametara. F statistika mjeri promjenu sume kvadrata reziduala po dodanom parametru u većem modelu, a normalizira se procjenom od σ^2[3]. S Gaussovim pretpostavkama i nultom hipotezom koja tvrdi da je manji model dovoljan, F statistika ima $F_{p1-p0,N-p1-1}$ distribuciju. Za velike N kvantili od $F_{p1-p0,N-p1-1}$ teže k $\chi^2_{p1-p0}/(p_1 - p_0)$ kvantilima. Nadalje, $1 - 2\alpha$ pouzdani inteval za paramatar β_j ima oblik:

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

gdje je $z^{1-\alpha}$ $1 - \alpha$ kvantil normalne distribucije.

1.2 Primjer: Rak prostate

Podaci za ovaj primjer potječu iz studije Stamey i suradnici (1989.). Željela se ispitati korelacija između razine antigena specifičnog za prostatu i niza kliničkih mjerenja u muškaraca nad kojima se trebala primijeniti radikalna prostatektomija. Varijable poticaja su logaritam volumena karcinoma (lcavol), logaritam težine prostate(lweight), dob, logaritam količine benigne hiperplazije prostate (lbph), invazija sjemenog mjehurića (svi), logaritam kapsularne penetracije (lcp), Gleasonova vrijednost (gleason) i postotak Gleasonovih vrijednosti 4 ili 5 (pgg45). Iz podataka također vidimo da je svi binarna varijabla, a gleason ordinalna. Prvo standardiziramo prediktore tako da imaju jediničnu varijancu. Na slučajan način podijelimo skup od 97 mjerenja na skup za učenje koji će se sastojati od 67 mjerenja i skup za testiranje od 30 mjerenja. Zatim na skupu za učenje prikažimo matricu raspršenja (eng. scatterplot matrix) te matricu korelacija . Procedure kodiramo u statističkom softveru, SAS-u. U tablici (1.3) kao i na slici (1.2) vidimo nekoliko jakih korelacija između varijabli. Na slici (1.2) prvi redak prikazuje vezu između varijable odaziva lpsa i pojedinih prediktora. Iz prethodnih prikaza uočavamo, na primjer, da postoji snažna veza između varijabli lcavol i lcp te između njih i varijable odaziva. Zatim na skupu za učenje procjenjujemo koeficijente metodom najmanjih kvadrata. Rezultati su prikazani u tablici (1.4). U tablici su još navedene standaradne pogreške te t vrijednosti, odnosno ranije definirani Z-score-ovi koji se dobiju kada procijenjeni koeficijent podijelimo standardnom pogreškom. Standardna pogreška koeficijenta je procjena standardne devijacije koeficijenta. U suštini, ona nam govori o preciznosti dobivenih koeficijenata. Z-score mjeri efekt koji se dobije otpuštanjem pojedine varijable iz modela. Iz zadnjeg stupca tablice vidimo p-vrijednosti kod testiranja nulte hipoteze da su pojedini koeficijenti jednaki 0, za razinu značajnosti uzmemo, npr. 5%, pa za sve p-vrijednosti manje od 0.05 odbacujemo nultu hipotezu.U modelu prediktor lcavol pokazuje najveći utjecaj na varijablu odaziva, dok su učinci varijabli lweight i svi također snažni. Uočimo da lcp nije značajan jednom kad je lcavol u modelu.



Slika 1.2: Matrica raspršenja

1.3 Gauss - Markovljev teorem

Usredotočimo se na procjene bilo koje linearne kombinacije parametara oblika $\theta = \alpha^T\beta$. Procjenitelj za $\alpha^T\beta$ metodom najmanjih kvadrata je $\hat{\theta} = \alpha^T\hat{\beta} = \alpha^T(X^TX)^{-1}X^TY$.

Pretpostavljajući da je matrica X fiksirana, ovo je linearna funkcija vektora odaziva Y oblika c_0^TY. Ako pretpostavimo da je takav linearni model dobar, dobivamo

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

$$= \alpha^T(X^TX)^{-1}X^TX\beta$$

$$= \alpha^T\beta,$$

Pearson Correlation Coefficients, N = 67 Prob > |r| under H0: Rho=0									
	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
lcavol	1.00000	0.30023 0.0136	0.28632 0.0188	0.06317 0.6116	0.59295 <.0001	0.69204 <.0001	0.42641 0.0003	0.48316 <.0001	0.73316 <.0001
lweight	0.30023 0.0136	1.00000	0.31672 0.0090	0.43704 0.0002	0.18105 0.1426	0.15683 0.2050	0.02356 0.8499	0.07417 0.5509	0.48522 <.0001
age	0.28632 0.0188	0.31672 0.0090	1.00000	0.28735 0.0184	0.12890 0.2985	0.17295 0.1616	0.36592 0.0023	0.27581 0.0239	0.22764 0.0639
lbph	0.06317 0.6116	0.43704 0.0002	0.28735 0.0184	1.00000	-0.13915 0.2614	-0.08853 0.4762	0.03299 0.7910	-0.03040 0.8070	0.26294 0.0316
svi	0.59295 <.0001	0.18105 0.1426	0.12890 0.2985	-0.13915 0.2614	1.00000	0.67124 <.0001	0.30688 0.0115	0.48136 <.0001	0.55689 <.0001
lcp	0.69204 <.0001	0.15683 0.2050	0.17295 0.1616	-0.08853 0.4762	0.67124 <.0001	1.00000	0.47644 <.0001	0.66253 <.0001	0.48920 <.0001
gleason	0.42641 0.0003	0.02356 0.8499	0.36592 0.0023	0.03299 0.7910	0.30688 0.0115	0.47644 <.0001	1.00000	0.75706 <.0001	0.34243 0.0046
pgg45	0.48316 <.0001	0.07417 0.5509	0.27581 0.0239	-0.03040 0.8070	0.48136 <.0001	0.66253 <.0001	0.75706 <.0001	1.00000	0.44805 0.0001
lpsa	0.73316 <.0001	0.48522 <.0001	0.22764 0.0639	0.26294 0.0316	0.55689 <.0001	0.48920 <.0001	0.34243 0.0046	0.44805 0.0001	1.00000

Tablica 1.3: Korelacije između varijabli

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.45235	0.08702	28.18	<.0001
lcavol	1	0.71641	0.13350	5.37	<.0001
lweight	1	0.29264	0.10638	2.75	0.0079
age	1	-0.14255	0.10212	-1.40	0.1681
lbph	1	0.21201	0.10312	2.06	0.0443
svi	1	0.30962	0.12539	2.47	0.0165
lcp	1	-0.28901	0.15480	-1.87	0.0670
gleason	1	-0.02091	0.14258	-0.15	0.8839
pgg45	1	0.27735	0.15959	1.74	0.0875

Tablica 1.4: Procjene koeficijenata i ostale statistike

odnosno da je $\alpha^T\hat{\beta}$ nepristran procjenitelj za $\alpha^T\beta$.

Gauss-Markovljev teorem tvrdi da ako imamo bilo koji drugi linearni procjenitelj $\tilde{\theta} =$

$c^T Y$ koje je nepristran za $\alpha^T\beta$, tada vrijedi

$$Var(\alpha^T\hat{\beta}) \leq Var(c^T Y),$$

odnosno procjenitelj za β dobiven metodom najmanjih kvadrata ima najmanju varijancu od svih linearnih nepristranih procjenitelja. Dokaz teorema se može pronaći u knjizi [3].

Pogledajmo sada srednje kvadratnu pogrešku (eng. mean squared error) procjenitelja $\tilde{\theta}$ kod procjene θ:

$$MSE(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2$$

$$= Var(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2,$$

gdje je drugi član u drugoj nejednakosti kvadratna pristranost. Gauss-Markovljev teorem implicira da procjenitelj dobiven metodom najmanjih kvadrata ima najmanju srednje kvadratnu pogrešku od svih linearnih procjenitelja bez pristranosti. Međutim, možda postoji pristrani procjenitelj s manjom srednjom kvadratnom pogreškom. Takav procjenitelj zamijenio bi malo pristranosti za veće smanjenje varijance[3]. Obično se koriste pristrane procjene. Bilo koja metoda u kojoj se smanji ili postavi na nulu neki od koeficijenata dobivenih metodom najmanjih kvadrata može rezultirati pristranom procjenom. Takve metode pojavljuju se u 2. i 3. poglavlju.

1.4 Višestruka regresija

Višestruki linearni regresijski model je linearni model s $p > 1$ ulaznih varijabli. Krećemo sa jednovarijatnim modelom ($p = 1$) bez slobodnog člana, tj.

$$Y = X\beta + \varepsilon.$$

Procjena metodom najmanjih kvadrata i reziduali tada izgledaju:

$$\hat{\beta} = \frac{\sum_1^N x_i y_i}{\sum_1^N x_i^2},$$

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

Ako označimo s $y = (y_1, ..., y_N)^T$ i $x = (x_1, ..., x_N)^T$ i definiramo skalarni produkt između x i y kao

$$\langle x, y\rangle = \sum_{i=1}^{N} x_i y_i$$

$$= x^T y,$$

tada možemo pisati:

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

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

Vidjeti ćemo, da ova jednostavna jednovarijatna regresija daje temelj za višestruku linearnu regresiju. Pretpostavimo da su stupci matrice X $x_1, ..., x_p$ ortogonalni, tj. vrijedi $< x_j, x_k >= 0$ za svaki $j \neq k$. Tada se lako vidi da su pojedine procjene $\hat{\beta}$ jednake $\langle x_j, y\rangle/\langle x_j, x_j\rangle$, odnosno kada imamo ortogonalnost, ulazne varijable nemaju utjecaj na međusobne parametre procjene. No, takve varijable se skoro nikada ne pojavljuju u primjeni, pa ćemo ih morati ortogonalizirati kako bismo mogli sprovesti ovu ideju. Pretpostavimo da imamo samo jedan ulaz x. Tada procijenjeni koeficijent pridružen x ima oblik

$$\hat{\beta}_1 = \frac{\langle x - \bar{x}1, y\rangle}{\langle x - \bar{x}1, x - \bar{x}1\rangle},$$

gdje je $\bar{x} = \sum_i x_i/N$ te $1 = x_0$, vektor od N jedinica. Prethodni procjenitelj dobijemo koristeći jednostavnu regresiju u dva koraka:

1. provedemo regresiju $x - a$ na 1, stvarajući ostatak $z = x - 1\bar{x}$

2. provedemo regresiju $y - a$ na z da bi dobili koeficijent $\hat{\beta}_1$,

gdje regresija b na a znači običnu jednovarijatnu regresiju od b s obzirom na a, bez slobodnog člana, stvarajući koeficijent $\hat{\gamma} = \langle a, b\rangle/\langle a, a\rangle$ te rezidual $b - \hat{\gamma}a$. Slika 1.5 prikazuje prethodni postupak za dva općenita ulaza x_1 i x_2.

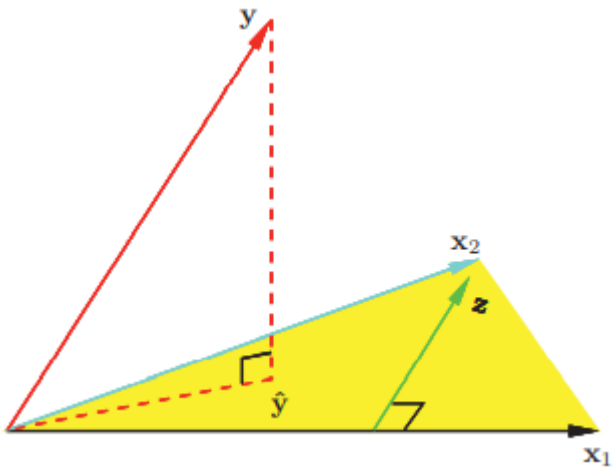
Sada dajemo generalizirani algoritam s p ulaza.

Algorithm 1: Regresija uzastopnom ortogonalizacijom

1: Inicijaliziramo $z_0 = x_0 = 1$.
2: Za $j = 1, 2, ..., p$
Napravimo regresiju x_j na $z_0, z_1, ..., z_{j-1}$ kako bismo dobili koeficijente $\hat{\gamma}_{lj} = \langle z_l, x_j\rangle/\langle z_l, z_l\rangle, l = 0, 1, ..., j-1$ te vektor reziduala $z_j = x_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} z_k$.
3: Napravimo regresiju y-a na rezidual z_p da dobijemo procjenu $\hat{\beta}_p$

Rezultat algoritma je procjenitelj:

$$\hat{\beta}_p = \frac{\langle z_p, y\rangle}{\langle z_p, z_p\rangle}.$$



Slika 1.5: Regresija metodom najmanjih kvadrata ortogonalizacijom ulaza. Provedemo regresiju vektora x_2 na vektor x_1. Time dobivamo rezidualni vektor z. Zatim regresija od y na z daje višestruki regresijski koeficijent od x_2. Zbrajajući zajedno projekcije od y na x_1 te z dobivamo procjenitelj najmanjih kvadrata $\hat{y}$

Ulazi $z_0, z_1, ..., z_{j-1}$ u 2. koraku su ortogonalni. Prethodni algoritam je poznat kao Gram-Schmidt-ov postupak za višestruku regresiju. Iz rezultata algoritma također možemo dobiti formulu za varijancu procjenitelja, tj.

$$Var(\hat{\beta}_p) = \frac{\sigma^2}{\|z_p\|^2},$$

koja predstavlja veličinu neobjašnjenosti x_p s preostalim x_k-ovima. Drugi korak algoritma može se prikazati u matričnom obliku na sljedeće načine:

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

gdje matrica Z sadrži redom z_j kao stupce, Γ je gornje trokutasta matrica s koeficijentima $\hat{\gamma}_{kj}$, D je dijagonalna matrica t.d. $D_{jj} = \|z_j\|$. Zadnji izraz predstavlja QR dekompoziciju matrice X, gdje je Q $N \times (p+1)$ ortogonalna matrica te R $(p+1) \times (p+1)$ gornje trokutasta matrica. Tada su rješenja metodom najmanjih kvadrata dana s:

$$\hat{\beta} = R^{-1}Q^T y,$$

$$\hat{y} = QQ^T y.$$

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:

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \varepsilon_k$$

$$= f_k(X) + \varepsilon_k.$$

U matričnoj notaciji ovo možemo zapisati kao

$$Y = XB + E,$$

gdje je za N pokusnih promatranja, Y $N \times K$ matrica odaziva s elementima y_{i_k}, 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 jednovarijantne funkcije gubitka je dana s

$$RSS(B) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2$$

$$= tr[(Y - XB)^T (Y - XB)].$$

Matrica procjenitelja ima istu formu kao i prije:

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

Poglavlje 2

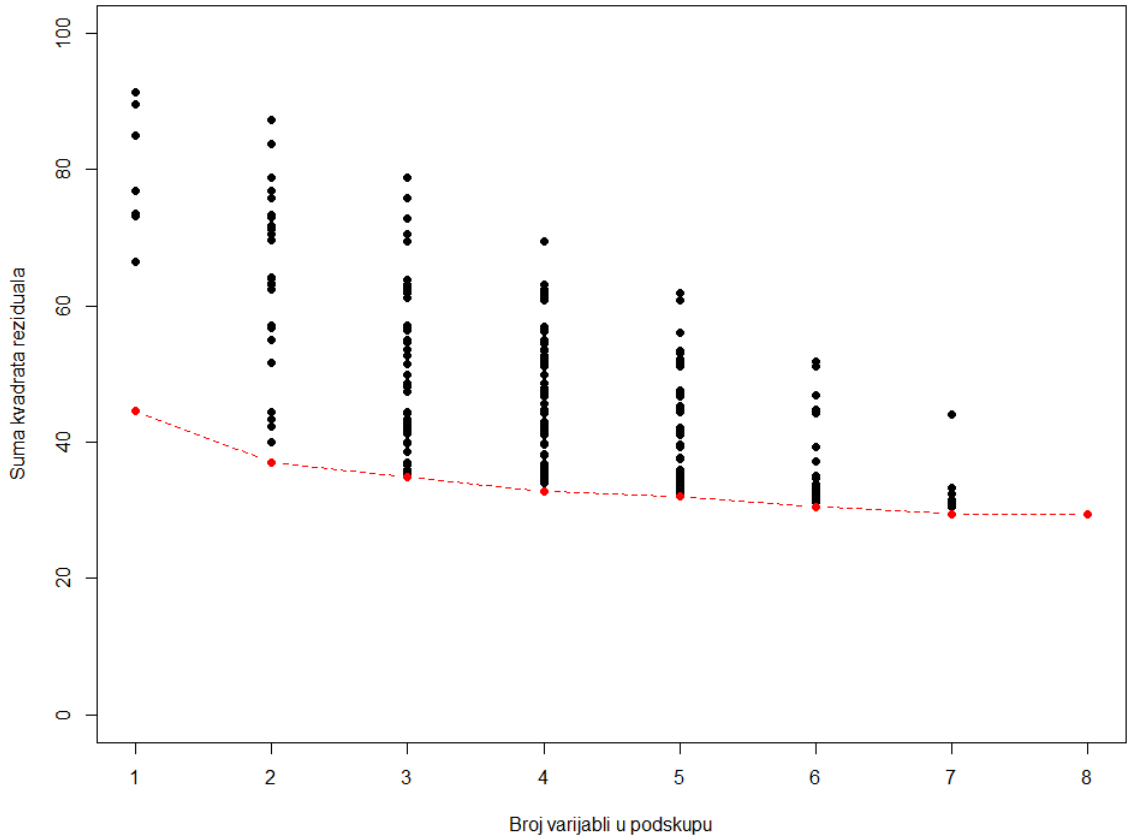
Metode odabira podskupa

Općenito nam nisu potrebne sve dostupne ulazne varijable, odnosno prediktori. Kako bismo povećali preciznost procjene varijable odaziva, smanjujemo broj varijabli koje sudjeluju u modelu, žrtvujući malo pristranosti u zamjenu za smanjenje varijabilnosti procijenjenih vrijednosti. Cilj nam je pronaći podskup prediktora koji će imati najveći utjecaj u modelu. U nastavku se obrađuju upravo takve metode reduciranja skupa prediktora.

2.1 Metoda odabira najboljeg podskupa

Kod konstrukcije regresijskog modela, uklanjanje irelevantnih varijabli učinit će model lakšim za interpretaciju i manje sklonim prekomjernom prilagođavanju podacima, samim time više generaliziranim.

Metoda odabira najboljeg podskupa (eng. Best-Subset Selection method) ima za cilj 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 $k \in \{1, 2, ..., p\}$. Tada pronalazi najbolji model sa jednom varijablom, dvije varijable i tako dalje. Pod pojmom "najbolji" smatramo model s najmanjom sumom kvadrata reziduala. U statističkom softveru R, kodiramo sljedeću sliku. Na slici 2.1 prikazujemo sve modele podskupova za primjer raka prostate. Donje crvene točke predstavljaju modele koji su kvalificirani za odabir putem metode odabira najboljeg podskupa. Crvena krivulja najboljeg odabira strogo pada, dakle što je veći broj k varijabli u modelu podskupa to je suma kvadrata reziduala manja, što nije dobar pokazatelj kvalitete modela pa nam to ne može biti validan kriterij odabira najbolje veličine podskupa. Odgovor na pitanje kako odabrati k uključuje kompromis između pristranosti i varijance, zajedno sa subjektivnom željom za štednjom. Postoje brojni kriteriji koje možemo koristiti, no obično odabiremo najmanji model koji minimizira procjenu očekivane pogreške predviđanja.



Slika 2.1: Svi mogući modeli podskupa za primjer raka prostate. Za svaku veličinu podskupa prikazujemo zbroj kvadrata reziduala za svaki model te veličine.

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. Treba imati na umu da automatizirani odabir varijabli ponekad ne bi trebao zamijeniti stručno mišljenje. Zapravo, važne varijable procijenjene prema pozadinskom znanju i dalje bi trebale biti unesene u model, čak i ako nisu statistički značajne.

Mana ovog algoritma je da broj modela koji se moraju uzeti u obzir raste eksponencijalno s brojem prediktora koji se razmatraju.

Ostali pristupi o kojima raspravljamo u ovom poglavlju funkcioniraju na sličan način.

2.2 Stepenasta selekcija unaprijed i unazad

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

Stepenasta selekcija (eng. Stepwise selection) se pojavljuje u dva oblika: selekcija 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 postiglo, potrebna je neka vrsta pravila odlučivanja o tome mijenja li se R do željenog stupnja dodavanjem 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 $[-1, 1]$. Većina programskih jezika koji imaju sposobnost raditi stepenastu regresiju daju korisnicima određenu kontrola nad kriterijem koji se koristi za odlučivanje hoće li se dodati ili ukloniti određena varijabla.

Metoda stepenaste selekcija unaprijed (eng. Forward-stepwise selection method) funkcionira tako da 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, tj. $|r_{YX_j}|$. Zatim od preostalih $k-1$ varijabli metoda pronalazi varijablu (P_2) koja najviše povećava R kada se doda u model koji sadrži samo P_1. Metoda nastavlja istim postupkom dok ne dobijemo model sa svih k prediktora. Dakle, očito metoda generira k modela, a analitičar tada može odabrati model koji balansira između velike vrijednosti od R te malog broja prediktora. Varijable koje su dodane kasnije u model najčešće ne doprinose toliko povećanju od R, pa se te varijable obično ne smatraju vrijednim zadržavanja u predikcijskom modelu, osobito ako se radi o varijablama čije je podatke zahtjevno sakupiti, a potrebni su za buduće primjene tog modela.

Upravo 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[2]. U prvom koraku metode, varijabla P_1 se bira samo u slučaju da je korelirana s Y pomoću statistički značajnog kriterija, kao što je na primjer 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. Ali ako se pronađe takav, metoda dodaje P_2 modelu otkrivenom u ovom koraku. 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 selekcije 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 stepenaste regresije unaprijed je zapravo kombinacija prethodno opisane stepenaste selekcije unaprijed i selekcije unatrag koju opisujemo u nastavku.

Dok selekcija unaprijed počinje bez prediktora, a zatim gradi model dodavanjem prediktora jedan po jedan, stepenasta selekcija unatrag (eng. Backward-stepwise selection) započinje modelom sa svih k prediktora i zatim ih uklanja jednog po jednog koristeći neki kriterij za uklanjanje[2]. U stepenastoj selekciji unaprijed, varijabla se dodaje u trenutnom koraku ako najviše povećava R. U selekciji unazad, u svakom koraku se uklanja varijabla koja snižava R najmanje u odnosu na ostale prediktore u modelu.

U praksi se često koristi test statističke značajnosti za utvrđivanje je li neku varijablu treba ukloniti u određenom koraku ili možda zaustaviti proces. Za razliku od selekcije unaprijed, varijabla se u modelu čuva ako bi njeno uklanjanje značajno smanjilo R (što je ekvivalentno sljedećem: dodavanjem varijable u model koji je do sada nije sadržavao, povećao bi se R do statistički značajnog stupnja)[2]. Ako više od jedne varijable zadovoljava ovaj test, onaj za kojeg se najmanje smanjuje R je odabir za uklanjanje u tom koraku. Proces se završava kada više ne postoje varijable koje se mogu ukloniti bez značajnog smanjenja R.

2.3 Etapna regresija unaprijed

Etapna regresija unaprijed (eng. Forward-stagewise regression) još je više ograničena od stepenaste regresije unaprijed[3]. Počinje kao i ona, sa slobodnim članom jednakim $\bar{y}$, i centriranim prediktorima 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 te je dodaje trenutnom koeficijentu odabrane varijable. Postupak se nastavlja sve dok nijedna varijabla nema značajne korelacije s rezidualima.

Poglavlje 3

Metode sažimanja

Zadržavanjem određenog podskupa od skupa svih prediktora i odbacivanjem ostatka, metoda odabira podskupa konstruira model koji je interpretabilan i koji vjerojatno ima manju pogrešku predviđanja od punoga modela, odnosno modela sa svim uključenim prediktorima. Međutim, budući da se radi o diskretnom procesu u kojem se varijable ili zadržavaju ili odbacuju, vrlo često je podložan velikoj varijanci pa samim time ne smanjuje pogrešku predviđanja punog modela. Metode sažimanja (Metode skupljanja) (eng. Shrinkage methods) su kontinuiranije i ne pate toliko od visoke varijabilnost[3].

3.1 Regresija grebenom

Regresija grebenom (eng. Ridge regression) smanjuje koeficijente regresije dodjeljivanjem kazne na temelju njihove veličine. Koeficijenti regresije grebenom minimiziraju penalizirajuću sumu kvadrata reziduala:

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

za vrijednosti $\lambda \geq 0$. Varijabla λ u ovoj formuli predstavlja parametar kompleksnosti koji kontrolira količinu sažimanja: što je veći λ, to je veća količina sažimanja. Koeficijenti se smanjuju prema nuli (i jedni prema drugima). Ekvivalentan način zapisivanja problema grebena je sljedeći,

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \tag{3.2}$$

za

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

što izričito ograničava veličinu koeficijenata. Postoji 1-1 korespodencija između varijable λ iz (3.1) i varijable t iz (3.2).

Kad u linearnom regresijskom modelu postoji mnogo koreliranih varijabli, njihovi koeficijenti mogu biti loše određeni i mogu rezultirati velikim varijancama. Izrazito veliki pozitivni koeficijent na jednoj varijabli može se poništiti pomoću približno velikog negativnog koeficijenta na njegovom koreliranom paru. Nametanjem ograničenja na veličine koeficijenta, kao u (3.2), postižemo ublažavanje problema. Rješenja regresije grebenom ovise o skaliranju ulaznih podataka, tako da se oni normalno standardiziraju prije rješavanja (3.1). Uočimo da je slobodan član β_0 izostavljen iz penaliziranja. Penaliziranje slobodnog člana imalo bi za posljedicu to da bi postupak ovisio o podrijetlu izabranom za Y; odnosno dodavanje konstante c svakom od y_i jednostavno ne bi rezultiralo pomakom predviđanja za isti iznos c.

Rješenje od (3.1) se može razdvojiti u dva dijela, a nakon reparametrizacije koristeći centrirane ulaze: svaki x_{ij} se zamjenjuje sa $x_{ij} - \bar{x}_j$. Tada β_0 procjenjujemo s $\bar{y} = \frac{1}{N}\sum_1^N y_i$. Preostali koeficijenti se procjenjuju regresijom grebenom bez slobodnog člana, koristeći centrirane x_{ij}[3]. Od sada pretpostavljamo da su x_{ij} centrirane, pa da matrica ulaza X ima p (umjesto $p + 1$) stupaca. Napisan u matričnoj normi, kriterij (3.1) ima sljedeci oblik:

$$RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta,$$

pa su rješenja regresije grebenom dana s:

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^Ty, \tag{3.3}$$

gdje je I $p \times p$ jedinična matrica.

Primjetimo da smo odabirom kvadratnog penaliziranja $\beta^T\beta$ doibli rješenje regresije grebenom koje je ponovo linearna funkcija od y. Prije invertiranja, rješenje dodaje pozitivnu konstantu dijagonali matrice X^TX te to čini problem nesingularnim, čak i ako X^TX nije punog ranga. To je bio glavni motiv za regresiju grebenom kada je prvi put uvedena u statistiku (1970.).Tradicionalni opisi regresije grebenom započinju s definicijom (3.3)[3]. Motiviramo je preko (3.1) i (3.2), jer pružaju uvid u to kako regresija grebenom funkcionira. U slučaju ortogonalnih ulaznih podataka, procjene $\hat{\beta}^{ridge}$ su zapravo skalirane verzije procjene najmanjih kvadrata, odnosno $\hat{\beta}^{ridge} = \hat{\beta}/(1 + \lambda)$.

Regresija grebenom također se može izvesti kao srednja vrijednost ili mod 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 („novi dokazi“). Drugim riječima, aposteriorna distribucija sažima ono što znamo nakon što su podaci opaženi[5]. Dakle, pretpostavimo $Y_i \sim N(\beta_0 + x_i^T\beta, \sigma^T)$, a koeficijenti β_j

su apriorno distribuirani kao $N(0, \tau^2)$, nezavisno jedan od drugog. Zatim (negativna) log-aposteriorna gustoća od β, pretpostavljajući da su τ^2 i σ^2 poznati, jednaka je izrazu u vitičastim zagradama u (3.1), za $\lambda = \sigma^2/\tau^2$. Pa je tako $\hat{\beta}^{ridge}$ mod aposteriorne distribucije. Budući da je distribucija Gaussova, to je ujedno i aposteriorna srednja vrijednost.

Dekompozicija singularne vrijednosti (SVD, eng. singular value decomposition) centrirane matrice ulaza X daje nam dodatni uvid u prirodu regresije grebenom[3]. SVD $N \times p$ matrice X ima oblik

$$X = UDV^T. \tag{3.4}$$

Ovdje su U i V $N \times p$ i $p \times p$ dimenzionalne. U je matrica s ortonormiranim stupcima, a V je ortogonalna matrica, gdje stupci od U razapinju prostor stupaca od X, te stupci od V razapinju prostor redaka od X. D je dijagonalna $p \times p$ matrica, s dijagonalnim elementima $d_1 \geq d_2 \geq ... \geq d_p \geq 0$ koje nazivamo singularnim vrijednostima matrice X budući da, ako je jedan ili više $d_j = 0$, matrica X je singularna.

Koristeći dekompoziciju singularne vrijednosti možemo zapisati vektor prilagodbe s obzirom na koeficijente dobivene metodom najmanjih kvadrata kao

$$\begin{aligned} X\hat{\beta}^{ls} &= X(X^TX)^{-1}X^Ty \\ &= UU^Ty. \end{aligned}$$

U^Ty su koordinate od y s obzirom na ortonormiranu bazu U. Sada su rješenja regresije grebenom dana s

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

gdje su u_j stupci od U. Uočimo da je $d_j^2/(d_j^2 + \lambda) \leq 1$ budući da je $\lambda \geq 0$. Regresija grebenom, poput linearne regresije, računa koordinate od y s obzirom na ortonormiranu bazu U te zatim reducira dobivene koordinate faktorom $d_j^2/(d_j^2 + \lambda)$. To znači da su na koordinate vektora baze s manjim d_j^2 primijenjene veće količine sažimanja.

No, što mala vrijednost od d_j^2 zapravo znači? SVD centrirane matrice X možemo shvatiti kao drugi način iskazivanja glavnih komponenti (eng. principal components) varijabli iz X. Uzoračka kovarijacijska matrica je dana s

$$S = X^TX/N,$$

pa iz (3.4) imamo

$$X^TX = VD^2V^T,$$

što predstavlja svojstvenu dekompozicija od X^TX (i od S, do na faktor N). Svojstvene vektore v_j (stupce matrice V) još nazivamo smjerovi glavnih komponenti od X. v_1, odnosno smjer prve glave komponente, ima svojstvo da $Z_1 = Xv_1$ ima najveću uzoračku varijancu od svih normiranih linearnih kombinacija stupaca iz X. Uzoračka varijanca od Z_1 je jednaka

$$Var(Z_1) = Var(Xv_1) = \frac{d_1^2}{N},$$

također možemo pisati $Z_1 = Xv_1 = U_1d_1$. Dobivena varijabla Z_1 se naziva prva glavna komponenta od X. U_1 je tada normirana prva glavna komponenta.

Slijed glavnih komponenti Z_j koje imaju redom najveće varijance d_j^2/N, je ortogonalan na prijašnje glavne komponente. Očito, posljednja glavna komponenta ima minimalnu varijancu. Stoga male singularne vrijednosti d_j odgovaraju smjerovima u prostoru stupaca od X koji imaju male varijance, a regresija grebenom najviše smanjuje upravo takve smjerove.

Definirajmo funkciju efektivnih stupnjeva slobode prilagodbe regresijom grebenom:

$$\begin{aligned} df(\lambda) &= tr[X(X^TX + \lambda I)^{-1}X^T], \\ &= tr(H_\lambda) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}. \end{aligned} \tag{3.5}$$

Ovo je monotono 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 $df(\lambda) \to 0$ kad $\lambda \to \infty$. Imamo još jedan dodatan stupanj slobode za slobodan član, no on je bio apriorno uklonjen.

3.2 Laso

Laso (eng. Lasso) je također metoda sažimanja s time da se od prethodne metode razlikuje u jednom sulptilnom, ali bitnom detalju. Procjenitelj metodom laso definiran je s

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

za

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

Baš kao i kod regresije grebenom, možemo reparametrizirati konstantu β_0 standardiziranjem prediktora. Rješenje za β_0 je $\bar{y}$, te stoga prilagođavamo model bez slobodnog člana. Problem procjene metodom laso možemo zapisati u takozvanoj Lagrange-ovoj formi

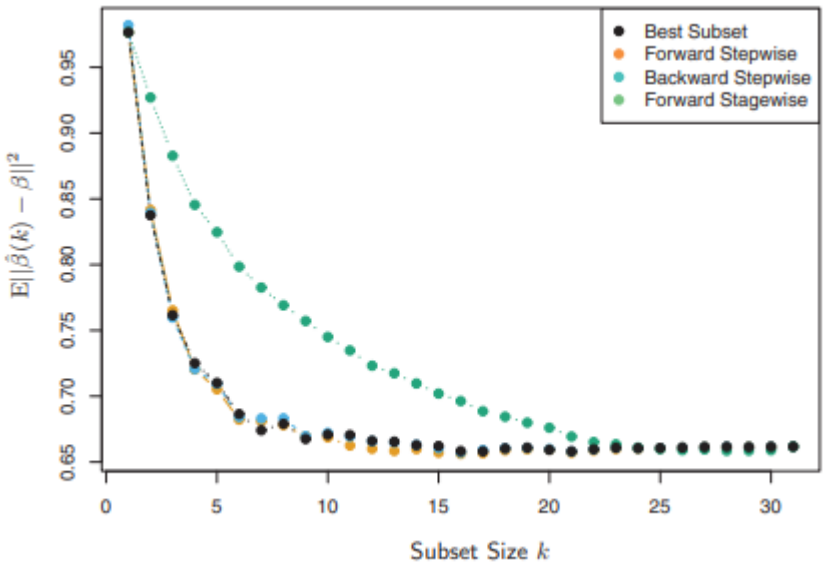
$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.6}$$

Uočimo sličnost s regresijom grebenom (3.1) ili (3.2): penaliziranje izraza $\sum_1^p \beta_j^2$ u metodi grebenom je zamijenjeno penaliziranjem $\sum_1^p |\beta_j|$ u metodi laso. Ovo potonje ograničenje čini rješenja nelinearnima u y_i, te ovdje ne postoji izraz zatvorenog oblika za procjene koeficijenata kao u regresiji grebenom. Računanje rješenja metodom laso je problem kvadratnog programiranja. Zbog prirode ograničenja, za t dovoljno mali neki će koeficijenti biti točno nula, pa možemo reći da metoda laso čini neku vrstu kontinuiranog odabira podskupa. Ako je odabrani t veći od $t_0 = \sum_1^p |\hat{\beta}_j|$, gdje su $\hat{\beta}_j = \hat{\beta}_j^{ls}$ procjenitelji dobiveni metodom najmanjih kvadrata, tada su procjenitelji metodom laso $\hat{\beta}_j$, $j = 1, 2, \ldots, p$. S druge strane, za, na primjer $t = t_0/2$, koeficijenti najmanjih kvadrata su u prosjeku smanjeni za oko 50% [3]. Međutim, priroda sažimanja nije uvijek očita, pa je dalje istražujemo u odjeljku 3.4. Kao i kod odabira veličine podskupa u metodi odabira podskupa ili parametra penaliziranja u regresiji grebenom, t treba biti adaptivno odabran kako bi minimalizirali procjenu očekivane greške predviđanja.

3.3 Usporedba metoda na primjeru

Pogledajmo prvo usporedbu opisanih metoda odabira podskupa nad simuliranim podacima na slici 3.1. Vidimo da se sve navedene metode ponašaju slično bez obzira na veličinu odabranog podskupa, što je općenito vrlo čest slučaj. Uočavamo blago odudaranje etapne regresije unaprijed kod manjih veličina podskupa prediktora. Kod našeg primjera raka prostate sve četiri metode odabira podskupa daju isti podskup od samo 2 prediktora (kod etapne i stepenastih metoda uzmemo razinu značajnosti od 5% za ulazak/izlazak pojedinih varijabli u/iz model/a.) tako da u nastavku uspoređujemo samo metodu odabira najboljeg podskupa s metodama sažimanja.

Tablica 3.1 prikazuje koeficijente dobivene metodom odabira najboljeg podskupa, regresijom grebenom i metodom laso zajedno s prethodno izračunatim koeficijentima dobivenih metodom najmanjih kvadrata. Koeficijente smo dobili primjenom odgovarajućih funkcija u R-u. Svaka metoda ima parametar kompleksnosti koji je odabran da minimizira procjenu predikcijske greške na temelju unakrsne validacije s deseterostrukim preklapanjem (eng. tenfold cross-validation). Unakrsna validacija tehnika je za procjenu prediktivnih modela particioniranjem originalnog uzorka na skup za učenje modela i testni skup



Slika 3.1: Usporedba 4 navedene metode na simuliranom linearnom regresijskom modelu $Y = X^T\beta + \varepsilon$. Simulacija je provedena sa $N = 300$ opažanja na $p = 31$ standardnih Gaussovih varijabli s međusobnim korelacijama jednakim 0.85. Za 10 varijabli koeficijenti se izabiru nasumično iz $N(0, 0.4)$ distribucije, a ostali su nula. Distribucija slučajne greške je $\varepsilon \sim N(0, 6.25)$. Rezultati su uprosječeni nad 50 simulacija. Na y-osi grafa prikazana je srednje kvadratna greška procijenjenih koeficijenata $\hat{\beta}(k)$ od stvarnih koeficijenata β za svaku veličinu podskupa k.

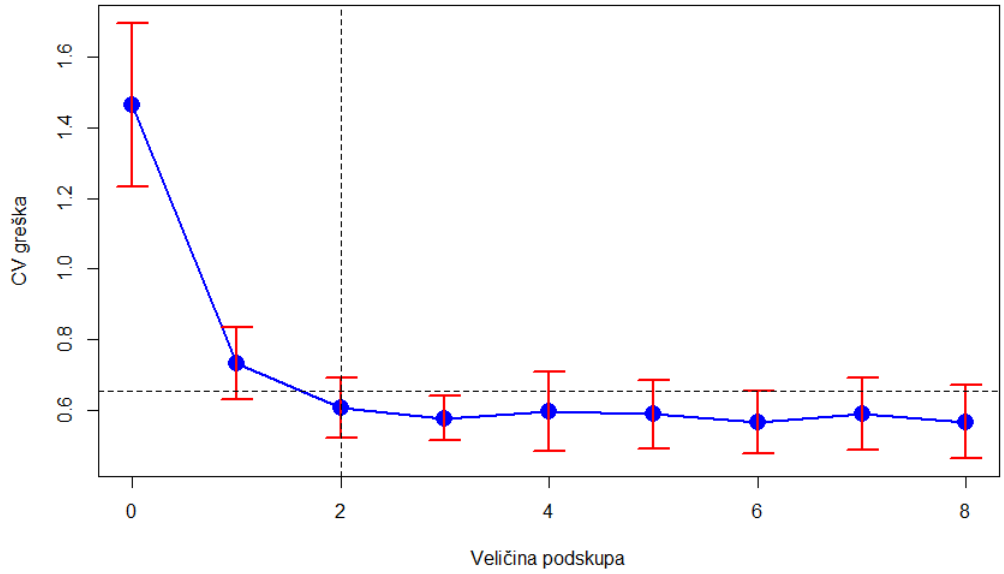
za njegovu procjenu. U k-strukoj unakrsnoj validaciji, originalni uzorak nasumično je podijeljen na k poduzorka jednake veličine. Od k poduzorka, jedan poduzorak se zadržava i predstavlja validacijske podatke za testiranje modela, a preostalih $k - 1$ poduzoraka koristi se za učenje modela. Postupak unakrsne validacije tada se ponavlja k puta (preklapanja), pri čemu se svaki od k poduzorka koristi točno jednom za validacijski proces. k rezultata iz preklapanja tada se mogu uprosječiti (ili kombinirati na drugi način) kako bi se dobila jedna procjena. Prednost ove metode je što se sva opažanja koriste i za učenje i za provjeru valjanosti, a svako se opažanje koristi za validaciju točno jednom. Naš skup podatak za rak prostate je već otprije bio podijeljen na skup za učenje veličine 60 i skup za testiranje veličine 30 tako da unakrsnu validaciju primjenjujemo na skupu za učenje, budući da je odabir parametra sažimanja dio procesa učenja. Na temelju zadnjeg retka tablice zaključujemo da metoda laso daje najbolji prediktivni model s obzirom na testne podatke, dok, očekivano, puni model dobiven jednostavnom linearnom regresijom daje najlošije testne rezultate.

Na slici 3.2 je prikazana krivulja procijenjenih predikcijskih grešaka (CV grešaka) s obzirom na parametar kompleksnosti modela (ovdje, kod primjera metode odabira najboljeg podskupa, to je veličina podskupa prediktora). Prikazujemo je zajedno s pripadnim rasponom procijenjenih standardnih pogrešaka, izračunatih unakrsnom validacijom. Koristi se pravilo "jedna-standardna-pogreška" (eng. "one-standarad-error") prema kojem se bira onaj model koji ima najmanje prediktora, a čija se CV greška nalazi unutar jedne standardne pogreške od podmodela s najmanjom CV greškom.

	MNK	MONP	RG	ML
slobodni član	2.452	2.452	2.452	2.452
lcavol	0.716	0.780	0.355	0.571
lweight	0.293	0.352	0.225	0.216
age	-0.143		-0.016	0.071
lbph	0.212		0.146	
svi	0.310		0.208	0.146
lcp	-0.289		0.052	
gleason	-0.021		0.050	
pgg45	0.277		0.117	0.043
Test MSE	0.549	0.548	0.506	0.446

Tablica 3.1: Procijenjeni koeficijenti te srednja kvadratna pogreška za metode: MNK- metoda najmanjih kvadrata, MONP - metoda odabira najboljeg podskupa, RG - regresija grebenom, ML - metoda laso. Prazni unosi odgovaraju izostavljenim varijablama. Zadnji redak prikazuje aritmetičku sredinu kvadarata reziduala (eng. mean squared error, MSE) nad testnim podacima.

Slike 3.3 i 3.4 prikazuju grafove ovisnosti srednje kvadratne pogreške potencijalnog modela s obzirom na povećanje parametra sažimanja λ kod primjera raka prostate. Prva slika se odnosi na regresiju grebenom, a druga na metodu laso. Također, iznad grafa možemo vidjeti odgovarajući broj potencijalnih prediktora za odgovarajuću vrijednost $\log(\lambda)$. Uočimo da su dvije vrijednosti od λ ($\log(\lambda)$) odabrane kao optimalne. Lijeva, odnosno manja vrijednost predstavlja λ_{min} koji daje najmanju vrijednost srednje unakrsno validacijske greške (eng. mean cross-validation error, CV error). Veći λ predstavlja λ_{1se}, odnosno vrijednost parametra sažimanja koji daje model s najmanje prediktora, ali da se njegova greška



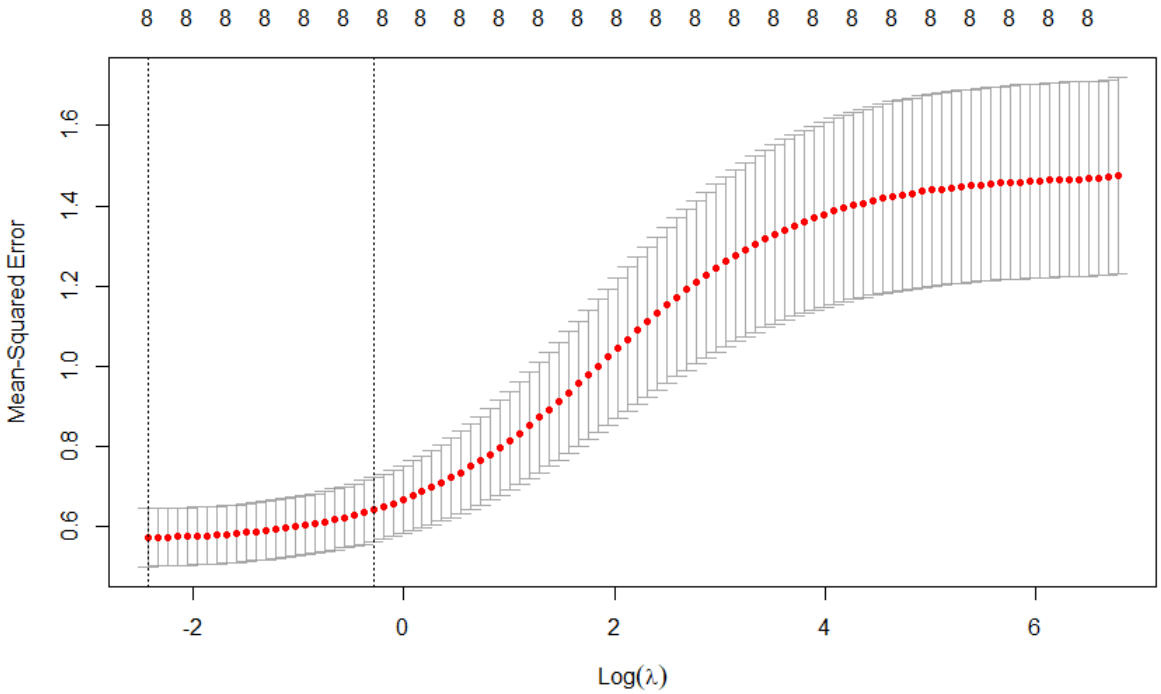
Slika 3.2: Krivulja procjenjenih predikcijskih grešaka sa standardnim pogreškama za metodu odabira najboljeg podskupa. S isprekidanom vertikalnom linijom odabran je model prema pravilu "jedna-standardna-pogreška".

unakrsne validacije nalazi unutar jedne standardne greške od minimalne greške unakrsne validacije. Mi odabiremo drugu vrijednost kao optimalnu. Iz prve slike vidimo da će za odabrani optimalni λ u naš model dobiven regresijom grebena biti uključeno svih osam prediktora, dok će (druga slika) model dobiven metodom laso zadržati pet prediktora. Preostali koeficijenti odgovarajućih prediktora biti će jednaki 0 kao što vidimo u tablici 3.1.

Tablica 3.2 prikazuje kako metoda odabira podskupa, regresija grebenom i metoda laso primjenjuju jednostavnu transformaciju procjenitelja $\hat{\beta}_j$ dobivenih metodom najmanjih kvadrata u slučaju ortonormirane ulazne matrice X.

Regresija grebenom radi proporcionalno sažimanje, metoda laso 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 sljedećom slikom. Slika 3.6 prikazuje metodu laso (lijevo) i regresiju grebenom (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 regresije grebenom je disk $\beta_1^2 + \beta_2^2 \leq t$, dok je to za metodu laso dijamant $|\beta_1| + |\beta_2| \leq t$. Obje metode daju rješenje prikazanog optimizacijskog problema tamo gdje

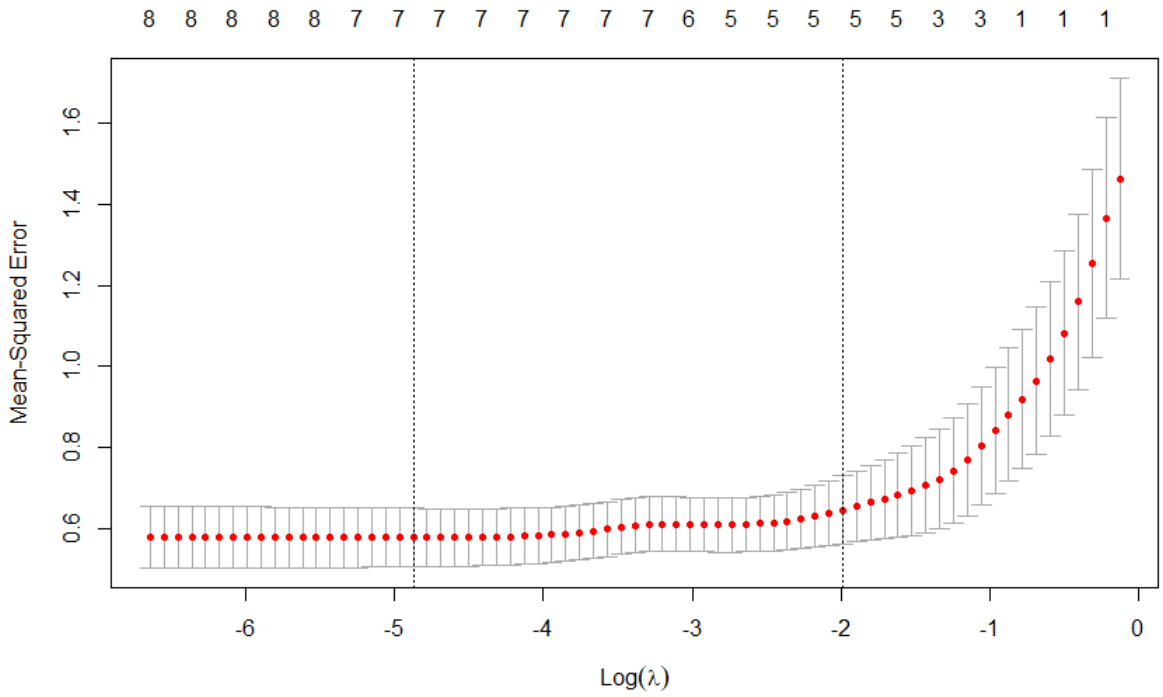


Slika 3.3: Graf ovisnosti srednje kvadratne pogreške o logaritamskoj vrijednosti od λ kod regresije grebenom. Vertikalne isprekidane linije predstavalju logaritam optimalnih λ dobivenih unakrsnom validacijom skupa za učenje.

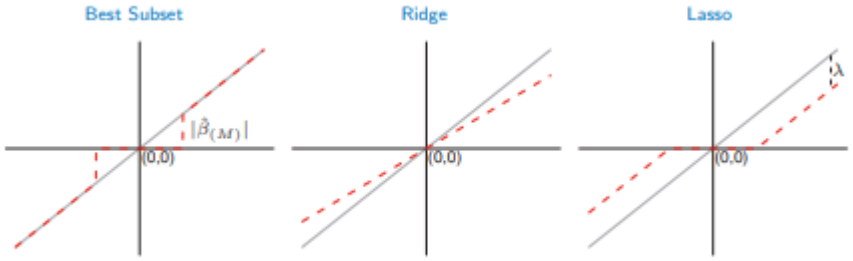
Procjenitelj za	Formula
Najbolji podksup (veličine M)	$\hat{\beta}_j \cdot \mathbb{1}(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Metoda grebenom	$\hat{\beta}_j/(1+\lambda)$
Metoda laso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

Tablica 3.2: Procjenitelji za koeficijente β_j u slučaju ortonormiranih stupaca matrice X. M i λ su konstante odabrane odgovarajućim tehinikama. sign označava predznak argumenta, x_+ označava pozitivni dio od x, a $\mathbb{1}$ označava karakterističnu funkciju navedenog skupa.

eliptične konture pogađaju ograničena područja. Za razliku od diska, dijamant ima uglove. Ako se rješenje dogodi na uglu, tada je jedan parametar β_j jednak nuli. Kad je $p > 2$, dija-



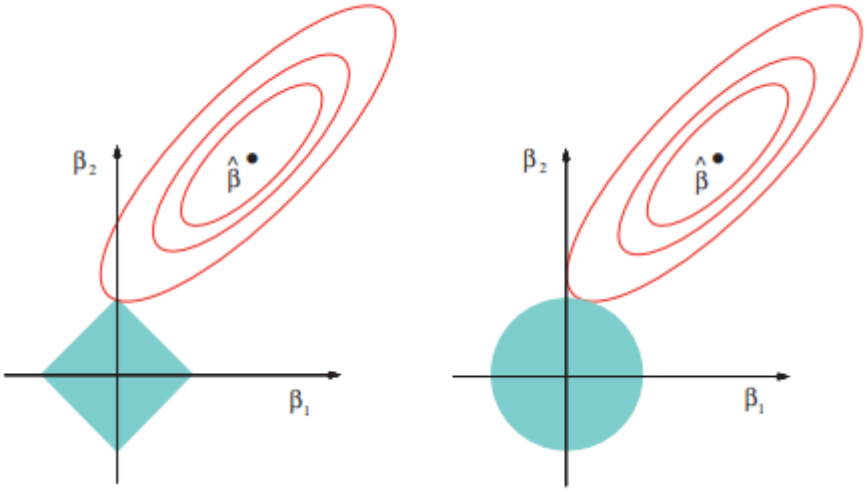
Slika 3.4: Graf ovisnosti srednje kvadratne pogreške o logaritamskoj vrijednosti od λ kod metode laso. Vertikalne isprekidane linije predstavalju logaritam optimalnih λ dobivenih unakrsnom validacijom skupa za učenje.



Slika 3.5: Primjer efekta procjenitelja iz tablice (crvene isprekidane linije) 3.2 s obzirom na nerestringirane procjenitelje (siva linija) za sve tri navede metode.

mant postaje romboid koji ima mnogo stranica, ravnih rubova i uglova. Stoga, procijenjeni

parametri imaju puno više mogućnosti za postizanje nule.



Slika 3.6: Procjene za regresiju grebenom (lijevo) i regresiju laso (desno). Plava područja su ograničena područja pripadnih metoda, dok su crvene elipse konture funkcije greške procjenitelja dobivenih metodom najmanjih kvadarata.

Nadalje, možemo generalizirati regresiju grebenom i metodu laso u obliku Bayesovih procjenitelja. Promatramo kriterij:

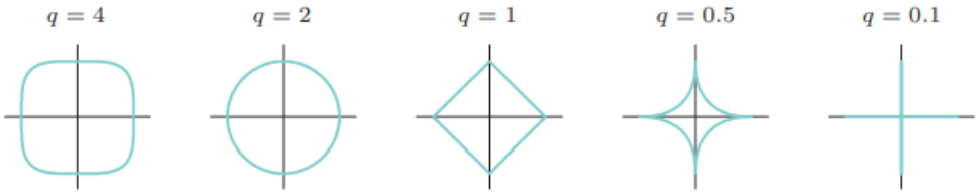
$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \tag{3.7}$$

za $q \geq 0$. Konture za konstantne vrijednosti od $\sum_j |\beta_j|^q$ za slučaj $p = 2$ prikazane su na slici 3.7

Na $|\beta_j|^q$ možemo gledati kao na log-apriornu gustoću od β_j. Vrijednost $q = 0$ odgovara odabiru podskupa varijabli, budući da penaliziranje tada daje broj ne-nul koeficijenata. $q = 1$ odgovara metodi laso, a $q = 2$ regresiji grebenom. Uočimo da za $q \leq 1$ apriorna distribucija nema uniformni smjer, već koncentrira više mase u smjeru koordinatnih osi. Apriorna distribucija koja odgovara slučaju $q = 1$ je nezavisna dvostruko eksponencijalna (ili Laplaceova) distribucija za svaki ulaz, s gustoćom:

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

za $\tau = 1/\lambda$. Također vidimo da je $q = 1$ (metoda laso) najmanji q za koju je ograničeno područje konveksno. Nekonveksna ograničena područja čine optimizacijski problem puno težim.

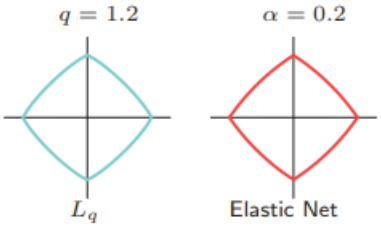


Slika 3.7: Konture za konstantne vrijednosti od $\sum_j |\beta_j|^q$ za pripadne vrijednosti q.

Promatrajući kriterij (3.7), mogli bismo pokušati upotrijebiti neke druge vrijednosti od q osim 0, 1 ili 2. Iako bi se moglo razmisliti o procjeni q-a iz podataka, iskustvo kaže da se ne isplati truditi za višak nastale varijance. Nadalje, iako bi vrijednosti od $q \in (1, 2)$ sugerirale kompromis između regresije grebenom i metode laso, za $q > 1$, $|\beta_j|^q$ je diferencijabilna u 0, pa nema sposobnost postavljanja koeficijenata na nulu, što je slučaj kod metode laso ($q = 1$). Djelomično zbog toga razloga, uvedeno je penaliziranje elastičnom mrežom (eng. elastic - net penalty):

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

drugačiji kompromis između metode laso i regresije grebenom. Slika 3.8 uspoređuje L_q penaliziranje za $q = 1.2$ te penaliziranje elastičnom mrežom za $\alpha = 0.2$. Naizgled, teško je uočiti bilo kakvu razliku, no elastična mreža ima oštre, nediferencijabilne uglove, dok $q = 1.2$ penaliziranje nema. Elastična mreža odabire varijable kao metoda laso te sažima koeficijente koreliranih prediktora kao regresija grebenom[3].



Slika 3.8: 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).

3.4 Regresija najmanjim kutom

Regresija najmanjim kutom (LAR, eng. Least angle regression) relativno je nova metoda (2004.), kao što ćemo vidjeti, blisko je povezana s lasom, a zapravo pruža iznimno učinkovit algoritam za računanje cijele putanje, odnosno razvoja koeficijenata i dobivanja rješenja metode laso.

Algoritam je sličan stepenastoj regresiji unaprijed, ali umjesto identificiranja i uključivanja najefektivnije nove varijable u model u svakom koraku, procijenjeni parametri se povećavaju u određenom smjeru s obzirom na korelaciju svake od njih s rezidualom, odnosno unosi samo "onoliko" od prediktora koliko zaslužuje. LAR metoda radi sa centriranim podacima.

U početku su svi koeficijenti postavljeni na nulu. Identificira se prediktor koji ima najveću korelaciju s trenutnim rezidualom, a čini ga samo odaziv u ovom koraku. Zatim se poduzima korak u smjeru ovog prediktora. Duljina ovog koraka, koja odgovara koeficijentu ovog prediktora, odabire se tako da neki drugi prediktor (tj. drugi prediktor koji ulazi u model) i trenutni predviđeni odaziv imaju istu korelaciju s trenutnim rezidualom. Zatim se, predviđeni odaziv kreće u smjeru koji se nalazi jednakokutno (zatvara isti kut) između ova dva prediktora. Krećući se u ovom zajedničkom smjeru osigurava se da će ova dva prediktora i dalje imati zajedničku korelaciju s trenutnim rezidualom. Predviđeni odaziv kreće se u ovom smjeru sve dok treći prediktor nema istu korelaciju s trenutnim rezidualom kao i prethodna dva prediktora koji su već u modelu. Određuje se novi zajednički smjer koji je ponovno pozicioniran jednakokutno između ova tri prediktora te se predviđeni odaziv kreće u tom smjeru sve dok se četvrti prediktor, koji ima istu korelaciju s trenutnim rezidualom, ne pridruži aktivnom skupu. Ovaj proces se nastavlja sve dok svi prediktori ne uđu u model. Sljedeći algoritam prikazuje pojedinosti.

Pretpostavimo sada da je A_k aktivni skup prediktora na početku k-tog koraka. Neka je β_{A_k} vektor koeficijenata za varijable iz A_k u tom koraku. $k-1$ koeficijenata će biti različito od nule, a jedan koji je u tom koraku ušao će biti jednak nuli. Ako je $r_k = Y - X_{A_k}\beta_{A_k}$ trenutni rezidual, tada je smjer za taj korak jednak

$$\delta_k = (X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T r_k.$$

Profil koeficijenata tada se razvija kao $\beta_{A_k}(\alpha) = \beta_{A_k} + \alpha \cdot \delta_k$. Ovako definiran smjer radi ono što metoda tvrdi: zadržava korelacije vezane i padajuće.

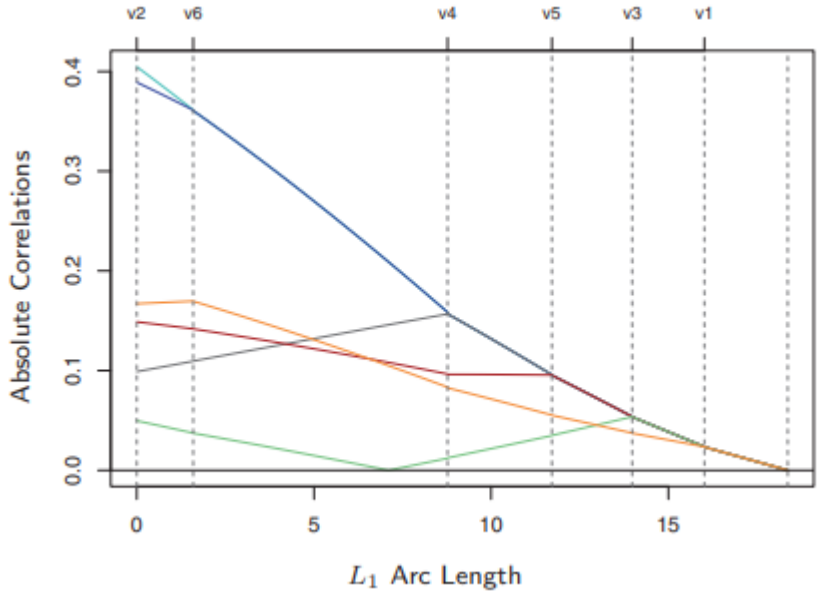
Naziv "najmanji kut" dolazi iz geometrijske interpretacije metode: u svakom koraku pronalazimo najmanji (i jednak) kut sa svakim od prediktora iz A_k. Napomenimo da ne moramo poduzimati male korake i ponovno provjeravati korelacije u koraku 3. Koristeći znanje o kovarijancama prediktora i linearnosti algoritma po dijelovima, možemo odredite točnu duljinu koraka na početku svakoga koraka.

Slika 3.9 prikazuje funcioniranje LAR algoritma, odnosno smanjivanje apsolutnih korelacija u svakom koraku algoritma na simuliranim podacima sa šest prediktora. Korela-

Algorithm 2: Regresija najmanjim kutom

1: Standardizirajte prediktore tako da imaju srednju vrijednost 0, te jediničnu normu. Započinjemo s rezidualom $r = Y - \bar{Y}, \beta_1, \beta_2, ..., \beta_p = 0$.
2: Nađite prediktor X_j najviše koreliran s r.
3: Pomaknite β_j s 0 prema njegovom koeficijentu dobivenim metodom najmanjih kvadrata $< X_j, r >$, sve dok ne naiđemo na drugi prediktor X_k koji je jednako ili više koreliran od X_j s trenutnim rezidualom.
4: Pomaknite β_j i β_k u smjeru definiranom s njihovim zajedničkim koeficijentom dobivenim metodom najmanjih kvadrata regresijom trenutnog reziduala s obzirom na (X_j, X_k), sve dok ne naiđemo na prediktor X_l koji je jednako ili više koreliran od prethodnih prediktora s trenutnim rezidualom.
5: Nastavite postupak dok svih p prediktora nije ubačeno u model. Nakon $\min(N-1, p)$ koraka, dolazimo do potpunog rješenja najmanjih kvadrata.

cije su prikazane s obzirom na duljine koraka algoritma mjerenih u duljinama lukova u L_1 normi.



Slika 3.9: Kretanje apsolutnih korelacija prediktora metode LAR po koracima, odnosno ulascima pojedinih prediktora u aktivni skup prediktora (oznake $v_1, \ldots, v_6$). Koraci metode su mjereni duljinama lukova u L_1 normi.

Algorithm 3: Regresija najmanjim kutom: Laso modifikacija

Ako ne-nul koeficijent postigne nulu, izbaciti pripadnu varijablu iz aktivnog skupa prediktora te ponovno izračunati trenutni zajednički smjer najmanjih kvadrata.

LAR algoritam iznimno je učinkovit i zahtijeva isti redoslijed izračunavanja kao i metoda najmanjih kvadrata s p prediktora. Regresiji najmanjim kutom uvijek je potrebno p koraka da dođe do potpunih procjenitelja najmanjih kvadrata[3]. Putanja metode laso može imati više od p koraka.

Sada dajemo heuristički argument zašto su LAR i metoda laso toliko slične. Iako je LAR algoritam naveden u terminima korelacija, ako su ulazi standardizirani, lakše je i ekvivalentno raditi sa skalarnim produktima. Pretpostavimo da je A aktivni skup varijabli u nekoj fazi algoritma, vezan s njihovim apsolutnim skalarnim produktom s trenutnim rezidualima $Y - X\beta$. To možemo izraziti kao

$$x_j^T(Y - X\beta) = \gamma \cdot s_j, \forall j \in A \tag{3.8}$$

gdje $s_j \in \{-1, 1\}$ označava predznak skalarnog produkta, a γ je zajednička vrijednost. Također vrijedi $|x_k^T(Y - X\beta)| \leq \gamma, \forall k \notin A$.

Pogledajmo sada laso kriterij (3.6) u vektorskoj formi

$$R(\beta) = \frac{1}{2} \|(Y - X\beta)\|_2^2 + \lambda \|\beta\|_1 .$$

Neka je sada B aktivni skup varijabli u rješenju za danu vrijednost λ. Za ove varijable $R(\beta)$ je diferencijabilna funkcija po β, a uvjeti stacionarnosti daju

$$x_j^T(Y - X\beta) = \lambda \cdot \text{sign}(\beta_j), \forall j \in B. \tag{3.9}$$

Uspoređujući (3.8) s (3.9) vidimo da su oni identični jedino ako predznak od β_j odgovara predznaku skalarnog produkta. Upravo zbog toga se LAR i metoda laso počinju razlikovati kada aktivni koeficijent mijenja predznak. Tada je uvjet (3.9) prekršen za pripadnu varijablu, te se ona izbacuje iz aktivnog skupa B. Uvjeti stacionarnosti za neaktivne varijable zahtijevaju

$$|x_j^T(Y - X\beta)| \leq \lambda, \forall k \notin B,$$

što je opet u skladu s LAR algoritmom[3].

Sada detaljnije diskutiramo formulu stupnjeva slobode za LAR i metodu laso. Pretpostavimo da smo procijenili linearni model pomoću regresije najmanjim kutom, zaustavljajući se na nekom broju koraka $k < p$, ili ekvivalentno, koristeći laso granicu t koja kreira ograničenu verziju prilagodbe dobivene pomoću metode najmanjih kvadrata. Zanima nas koliko parametara ili "stupnjeva slobode" smo pritom koristili.

Razmotrimo prvo linearnu regresiju koristeći podskup od k prediktora. Ako je ovaj podskup unaprijed određen bez pozivanja na podatke za treniranje, tada su stupnjevi slobode korišteni u modelu postavljeni na k. Doista, u klasičnoj statistici, broj nezavisnih parametara je ono što se misli pod pojmom "stupnjevi slobode". Alternativno, pretpostavimo da provodimo metodu odabira najboljeg podskupa kako bismo odredili optimalni skup od k prediktora. Tada rezultirajući model ima k parametara, ali smo u nekom smislu koristili i više od k stupnjeva slobode pa nam je stoga potrebna općenitija definicija stupnjeva slobode. Definiramo stupnjeve slobode procijenjenog vektora $\hat{Y} = (\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_N)$ kao

$$df(\hat{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} Cov(\hat{Y}_i, Y_i). \tag{3.10}$$

Ovdje se $Cov(\hat{Y}_i, Y_i)$ odnosi na uzoračku kovarijancu između procijenjene vrijednosti $\hat{Y}_i$ te pripadne varijable odaziva Y_i. Ovo također ima smisla na intuitivnoj razini: što teže učimo iz podataka, veća će biti kovarijanca, a time i veća vrijednost od $df(\hat{Y})$. Izraz (3.10) koristan je predstavnik stupnjeva slobode, jer je onaj koji se može primijeniti na predviđanje $\hat{Y}$ bilo kojeg modela.

Sada je za linearnu regresiju s k fiksnih prediktora lako pokazati da je $df(\hat{Y}) = k$. Slično za regresiju grebenom, ova definicija vodi do izraza zatvorenog oblika (3.5): $df(\hat{Y}) = tr(S_\lambda)$. U oba ova slučaja, (3.10) je jednostavno procijeniti jer je $\hat{Y} = H_Y Y$ linearno u Y. Razmišljamo li o definiciji (3.10) u kontekstu metode odabira najboljeg podskupa veličine k, čini se jasnim da će $df(\hat{Y})$ biti veće od k, a to može biti provjereno procjenom $Cov(\hat{Y}_i, Y_i)/\sigma^2$ simulacijom. Međutim, ne postoji metoda zatvorenog oblika za procjenu $df(\hat{Y})$ kod ove metode.

Kod LAR i metode laso događa se nešto impresivno. Ove tehnike su prilagodljive na elegantniji način od metode odabira najboljeg podskupa, te je stoga i procjena stupnjeva slobode više prilagodljiva. Posebno, može se pokazati da nakon k-tog koraka LAR procedure, efektivni stupanj slobode je točno k. Za laso, (modificirani) LAR postupak, često je potrebno više od p koraka, budući da prediktori mogu iskočiti van iz aktivnog skupa prediktora. Stoga je definicija malo drugačija. Za metodu laso, u bilo kojoj fazi, $df(\hat{Y})$ je približno jednak broju prediktora u modelu. Dok ova aproksimacija radi poprilično dobro bilo gdje na laso putanji, za svaki k radi najbolje na posljednjem modelu u slijedu koji sadrži k prediktora[3].

Bibliografija

[1] E. A. Pack, D. C.Montgomery i G. Geoffrey Vining, *Introduction to Linear Regression Analysis, 5th Edition*, Wiley, 2012.

[2] Richard B. Darlington i Andrew F. Hayes, *Regression analysis and linear models, Concepts, Applications, and Implementation*, The Guilford press, 1990.

[3] J. Friedman, T. Hastie i R. Tibshirani, *The Elements of Statistical Learning, Second Edition*, Springer, 2009.

[4] X. Gang Su i X. Yan, *Linear Regression Analysis, Theory and Computing*, World Scientific, 2009.

[5] Stephanie Glen, *"Posterior Probability the Posterior Distribution" From StatisticsHowTo.com: Elementary Statistics for the rest of us!*, `https://www.statisticshowto.com/posterior-distribution-probability/`.

[6] T. Hastie, G. James, R. Tibshirani i D. Witten, *An Introduction to Statistical Learning: With Applications in R, 1st ed.*, Springer, 2013.

Sažetak

U ovome radu upoznali smo se sa principom i svrhom korištenja linearne regresije u različitim formama. Vidjeli smo kako funkcioniraju neke od osnovnih metoda za pronalazak koeficijenata (i mnogih drugih statistika) linearnog regresijskog modela, odnosno kako ponekad možemo izmijeniti (metode sažimanja) prvobitno dobivene koeficijente ili neke prediktore u potpunosti izbaciti (metode odabira podskupa) u svrhu poboljšanja snage predikcije. Zaključujemo da bez obzira na prednost neke metode s obzirom na druge u određenom problemu, ne postoji savršeni model koji bi u svakome problemu predikcije davao optimalno rješenje, već moramo svaki puta uzeti u obzir više metoda kako bismo pronašli model koji nam u tom trenutku najviše odgovara gledajući kriterij koji nas najviše zanima.

Summary

In this paper, we are introduced to the principle and purpose of using linear regression in different forms. We have seen how some of the basic methods for finding the coefficients (and many other statistics) of a linear regression model work, that is, how we can sometimes modify (shrinkage methods) originally obtained coefficients or eliminate some predictors completely (subset selection methods) to improve prediction power. We conclude that regardless of the advantage of some methods over others in a certain problem, there is no perfect model that would predict the optimal solution in each problem, instead each time we must consider several methods to find the model that suits us best at that time by looking at the criterion that interests us the most.

Životopis

Rođen sam 20. siječnja 1998. godine u Zaboku. Pohađao sam Osnovnu školu Matije Gupca u Gornjoj Stubici. Srednjoškolsko obrazovanje stekao sam u prirodoslovno-matematičkoj V. gimnaziji u Zagrebu. Preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu upisujem 2016. godine. Sa zvanjem sveučilišnog prvostupnika 2019. godine upisujem diplomski sveučilišni studij Matematička statistika.