Osječki matematički list 25 (2025), 117–136

LASSO Regresija

Nikola Kraljević [*], Igor Velčić[†]

Sažetak

U radu je dana usporedba regresijskih modela s penalizacijom, konkretno Ridge i Lasso regresije. Objašnjen je algoritam koordinatnog spusta koji se koristi za Lasso procjenu parametara. Na simuliranim podacima demonstrirana su svojstva oba modela te je prikazana primjena Bootstrap metode za Lasso model.

Ključne riječi: *linearna regresija, lasso regresija, ridge regresija, rijetki podatci*

LASSO Regression

Abstract

This paper gives a comparison between the Ridge and Lasso regression methods. We also explain the coordinate descent algorithm used for the estimation of the parameters in Lasso regression. The properties of both models are demonstrated on simulated data, as well as the application of the Bootstrap method for the Lasso model.

Keywords: *linear regression, lasso regression, ridge regression, sparse data*

[*]student, Fakultet elektrotehnike i računarstva u Zagrebu, Sveučilište u Zagrebu, email: nikola.kraljevic1c@gmail.com

[†]Zavod za primijenjenu matematiku, Fakultet elektrotehnike i računarstva u Zagrebu, Sveučilište u Zagrebu, email: igor.velcic@fer.hr

1 Uvod

Linearna regresija predstavlja snažan statistički alat s čvrstom teorijskom osnovom, koji omogućuje razumijevanje i kvantificiranje odnosa između ciljne varijable (y) i skupa regresora $(\boldsymbol{x})$. Međutim, u praksi se često susrećemo s korelacijama među opaženim varijablama ili u kojima dimenzionalnost podataka značajno nadmašuje broj opažanja. Kada su temeljne pretpostavke linearne regresije narušene, rješenje je primjena metoda poput Ridge i Lasso regresije.

Ideja koje metode Ridge i Lasso koriste za rješavanje ovih problema jest kompromis između pristranosti i varijance (eng. *bias–variance tradeoff*). Dodavanjem penalizacijskog člana u optimizacijski problem svjesno uvodimo pristranost u procjene, ali time značajno smanjujemo njihovu varijancu. Ovaj pristup daje interpretabilne modele, osobito u uvjetima visoke dimenzionalnosti.

Ridge regresija uvodi penalizaciju u obliku L^2 norme regresijskih koeficijenata, dok Lasso regresija koristi L_1 normu. Postavlja se pitanje: Zašto baš L^1 i L^2 norme?

Odgovor leži u svojstvima optimizacijskog problema. L^1 norma predstavlja najmanju normu za koju problem penalizirane procjene parametara ostaje konveksan, što omogućuje efikasno računanje rješenja. S druge strane, L^2 norma ne samo da zadržava konveksnost problema, već omogućava Ridge regresiji linearno rješenje u zatvorenoj formi, i time procjena parametara nije ništa drugo nego množenje matrica.

2 Linearna Regresija

2.1 Model i pretpostavke

Linearni regresijski model je statistički model koji pretpostavlja linearnu vezu između ciljne (ovisne) varijable i regresora (neovisnih varijabli). Linearna regresija je skup metoda za analizu tog modela, a standardno se regresijski parametri procjenjuju metodom najmanjih kvadrata. Model linearne regresije je:

$$Y_i = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i + \varepsilon_i,$$

gdje je:

- Y_i **ovisna varijabla** (regresand, ciljna varijabla, odziv)

- $\boldsymbol{x}_i \in \mathbb{R}^p$ **neovisne varijable** (regresori, ulazne varijable, prediktori)
- $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ **parametri** (regresijski koeficijenti), oni su nepoznati koeficijenti koji određuju linearnu vezu ovisne varijable s neovisnim varijablama.

  ε_i označava slučajnu grešku: sve utjecaje na ovisnu varijablu Y_i koji nisu objašnjeni neovisnim varijablama modela.
- $p \in \mathbb{N}$ je dimenzija neovisne varijable.
- $i \in \{1, \ldots, N\}$, $N \in \mathbb{N}$ je broj primjera

Primjećujemo da je Y_i slučajna varijabla jer sadrži slučajnu komponentu ε_i. Uzorak na temelju kojeg želimo procijeniti nepoznate parametre zapisujemo kao:

$$U = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^N,$$

gdje je y_i realizacija slučajne varijable Y_i. Procjene parametara β_0 i $\boldsymbol{\beta}$ označavamo sa $\hat{\beta}_0$ i $\hat{\boldsymbol{\beta}}$. **Reziduali** $\hat{\varepsilon}_i$ opisuju razliku između opažene vrijednosti y_i i procijenjene vrijednosti:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i,$$

tj.

$$\hat{\varepsilon}_i = y_i - \hat{y}_i. \tag{1}$$

Stoga vrijedi:

$$y_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\varepsilon}_i.$$

Pretpostavke modela linearne regresije su:

1. **Nezavisnost**: Greške ε_i međusobno su nezavisne.
2. **Homoskedastičnost**: Varijanca grešaka je konstantna za sve vrijednosti $\boldsymbol{x}_i$, tj.
$$\mathrm{Var}(\varepsilon_i) = \sigma^2, \quad \forall i \in \{1, \ldots, N\}$$
3. **Normalnost grešaka**: Greške ε_i su distribuirane prema normalnoj distribuciji s očekivanjem 0 i varijancom σ^2, tj.
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \forall i \in \{1, \ldots, N\}.$$

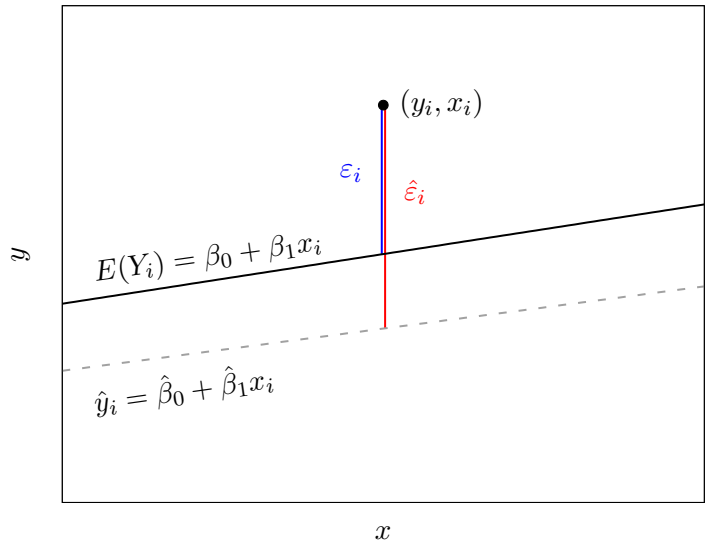
Pod ovim pretpostavkama, klasična teorija linearne regresije predstavlja snažan alat i široko se primjenjuje. Da bismo iskoristili ovaj model u praksi, moramo procijeniti nepoznate parametre β_0 i $\boldsymbol{\beta} \in \mathbb{R}^p$ modela na osnovi dostupnih podataka, konkretno opservacija $y_i \in \mathbb{R}$ i $\boldsymbol{x}_i \in \mathbb{R}^p$.

U sljedećem poglavlju razmatramo kako se procjena parametara može formulirati kao problem maksimiziranja vjerodostojnosti parametara, te pokazujemo da je takva formulacija problema ekvivalentna metodi najmanjih kvadrata.

2.2 Procjena parametara

U prijašnjem poglavlju definirali smo reziduale (1), koji predstavljaju opažene greške modela, tj. odstupanja predviđenih vrijednosti $\hat{y}_i$ od opaženih vrijednosti y_i. Slika 1 ilustrira razliku između stvarne greške ε_i i reziduala $\hat{\varepsilon}_i$. Pri procjeni parametara cilj nam je pronaći parametre takve da su reziduali što manji. U literaturi se najčešće samo navodi da se procjena parametara radi minimiziranjem sume kvadrata reziduala:

$$\min_{\hat{\beta}_0, \hat{\boldsymbol{\beta}}} \left\{ \sum_{i=1}^{N} \hat{\varepsilon}_i^2 \right\}.$$



Slika 1. Slika prikazuje razliku između reziduala ε_i i greške modela $\hat{\varepsilon}_i$

Carl Friedrich Gauss (1777–1855) njemački matematičar i fizičar

Ruđer Josip Bošković (1711–1787) hrvatski fizičar, astronom i matematičar

Ovu metodu je uveo Carl Friedrich Gauss oko 1795. godine. Iako ova metoda intuitivno jasna, postavlja se pitanje, zašto kvadriramo reziduale? Zašto ne minimiziramo, primjerice, njihove apsolutne vrijednosti? Naravno, kvadriranje vodi do linearne zadaće za procjenu parametara, što bitno olakšava račun, ali pitanje je postoji li i neko drugo opravdanje za kvadriranje pogrešaka.

Zanimljivo je spomenuti da je još 1757. godine Ruđer Bošković predložio pristup minimiziranjem sume apsolutnih vrijednosti pogrešaka. Njegova metoda, poznata kao *Least Absolute Deviations*, temelji se na minimiziranju sume apsolutnih vrijednosti reziduala. U nastavku ćemo predstaviti formalniji pristup procjeni parametara koji se može primijeniti na gotovo svaki statistički model, metodu maksimalne vjerodostojnosti.

Uz dani uzorak predstavljen vektorom odziva $\boldsymbol{y} \in \mathbb{R}^N$ i matricom dizajna $\mathbf{X} \in \mathbb{R}^{N \times p}$, te vektorom parametara $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$, funkcija vjerodostojnosti uz dani uzorak je:

$$\mathcal{L}(\hat{\boldsymbol{\theta}}|U) = f_{\boldsymbol{Y}}(\boldsymbol{y} \mid \mathbf{X}, \hat{\boldsymbol{\theta}}) = \prod_{i=1}^{N} f_Y(y_i|x_i, \hat{\boldsymbol{\theta}}).$$

Ovdje je $f_{\boldsymbol{Y}}$ funkcija gustoće slučajnog vektora $Y = (Y_1, \ldots, Y_N)$, koja je zbog pretpostavke nezavisnosti jednaka umnošku gustoća $f_Y(y_i|x_i, \hat{\boldsymbol{\theta}})$ slučajnih varijabli Y_i, $i \in \{1, \ldots n\}$.

Procjena maksimalne vjerodostojnosti (eng. *Maximum Likelihood Estimate*) definira se kao vektor parametara koji maksimizira funkciju vjerodostojnosti:

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta \mid U).$$

Problem maksimizacije obično se transformira u ekvivalentan problem minimizacije negativnog logaritma funkcije vjerodostojnosti:

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\theta}{\operatorname{argmin}} \left\{ -\ln \mathcal{L}(\theta \mid U) \right\}.$$

Na taj način, produkt prelazi u sumu, što značajno pojednostavljuje račun. U slučaju linearne regresije, pretpostavlja se da su greške normalno distribuirane, pa gustoća ima oblik normalne razdiobe:

$$f_Y(y_i \mid \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left\{ \frac{\hat{\varepsilon}_i^2}{2\hat{\sigma}^2} \right\}.$$

Problem linearne regresije možemo zapisati u matričnom obliku:

$$\underset{N\times 1}{\boldsymbol{y}} = \underset{N\times(p+1)}{\mathbf{X}} \cdot \underset{(p+1)\times 1}{\hat{\boldsymbol{\beta}}} + \underset{N\times 1}{\hat{\boldsymbol{\varepsilon}}},$$

gdje smo proširili matricu dizajna $\mathbf{X}$ s jediničnim vektor stupcem, i vektor $\hat{\boldsymbol{\beta}}$ smo proširili s β_0.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} + \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_N \end{bmatrix}.$$

Sada imamo osnovu za pokazati da MLE procjena vodi do metode najmanjih kvadrata:

$$\underset{\hat{\beta}}{\operatorname{argmin}}\left\{ -\ln \mathcal{L}(\hat{\boldsymbol{\beta}}, \sigma^2 | \mathbf{X}, \boldsymbol{y}) \right\} = \underset{\hat{\beta}}{\operatorname{argmin}}\left\{ -\ln \prod_{i=1}^{N} f_Y(y_i | \boldsymbol{x}_i; \hat{\boldsymbol{\beta}}, \sigma^2) \right\} =$$

$$\underset{\hat{\beta}}{\operatorname{argmin}}\left\{ -\sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\hat{\varepsilon}_i)^2/2\sigma^2} \right\} =$$

$$\underset{\hat{\beta}}{\operatorname{argmin}}\left\{ -\sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{N} \ln e^{-(\hat{\varepsilon}_i)^2/2\sigma^2} \right\} =$$

$$\underset{\hat{\beta}}{\operatorname{argmin}}\left\{ \sum_{i=1}^{N} \hat{\varepsilon}_i^2 \right\} = \underset{\hat{\beta}}{\operatorname{argmin}}\left\{ \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \right\} = \underset{\hat{\beta}}{\operatorname{argmin}}\left\{ (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \right\}.$$

Uvest ćemo pomoćnu varijablu $\boldsymbol{b} \in \mathbb{R}^p$ umjesto $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, te ćemo derivirati izraz po $\boldsymbol{b}$ i izjednačiti s nulom kako bismo dobili traženi procjenitelj u zatvorenoj formi:

$$\begin{aligned} &\nabla_{\boldsymbol{b}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{b})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{b}) = 0 \\ &\nabla_{\boldsymbol{b}} \left(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{b}^T\mathbf{X}^T\boldsymbol{y} - \boldsymbol{y}^T\mathbf{X}\boldsymbol{b} + \boldsymbol{b}^T\mathbf{X}^T\mathbf{X}\boldsymbol{b} \right) = 0 \\ &\nabla_{\boldsymbol{b}} \left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\mathbf{X}\boldsymbol{b} + \boldsymbol{b}^T\mathbf{X}^T\mathbf{X}\boldsymbol{b} \right) = 0 \\ &-2\mathbf{X}^T\boldsymbol{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{b} = 0 \\ &\mathbf{X}^T\mathbf{X}\boldsymbol{b} = \mathbf{X}^T\boldsymbol{y} \\ &\boldsymbol{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}, \end{aligned}$$

gdje smo koristili činjenicu da su u drugom retku svi izrazi *skalari*[1] i pravila *matričnog deriviranja.* [2] Može se pokazati da je tako dobiveni procjenitelj nepristran, vidi npr. [4].

Dobili smo elegantno rješenje za MLE procjenu parametara. Naravno, treba primijetiti da se pri procjeni parametara računa inverz takozvane Gramove matrice $\boldsymbol{X^T X}$, pa se moramo zapitati u kojim slučajevima inverz ne postoji. Može se pokazati da je rang matrice $\boldsymbol{X}$ jednak rangu matrice $\boldsymbol{X^T X}$, i uz pretpostavku $N > p + 1$, inverz neće postojati ako imamo linearno zavisne stupce u matrici $\boldsymbol{X}$. U praksi stupci matrice $\boldsymbol{X}$ rijetko budu linearno zavisni, ali često su numerički blizu linearne zavisnosti. Takva situacija dovodi do tzv. približno singularne Gramove matrice. Numerički, to znači da iako inverz matrice postoji, vrlo je osjetljiv na male promjene u podacima.

Jedan pristup za ublažavanje tog problema jest uvođenje regularizacije u postupak procjene parametara. Na taj način uvodimo određenu pristranost kako bismo smanjili varijancu parametara. U idućim poglavljima ćemo vidjeti kako uvođenjem ograničenja na regresijske koeficijente dobijemo mogućnost kontroliranja kompleksnosti modela.

3 Ridge Regresija

Ridge regresija je karakterizirana time da uvodi kaznu (penalizaciju) ili ograničenje u optimizacijski problem u obliku L^2 normi koeficijenata. Optimizacijski problem Ridge regresije je:

$$\begin{aligned} \min_{\boldsymbol{b}} \quad & (\boldsymbol{y} - \boldsymbol{Xb})^T(\boldsymbol{y} - \boldsymbol{Xb}) \\ \text{uz ograničenje} \quad & \sum_{j=1}^{p} b_j^2 < t, \quad t \geq 0. \end{aligned}$$

Možemo interpretirati da nam parametar t nam daje određeni budžet koji možemo "potrošiti" na koeficijente. Padajuće vrijednosti parametra t uvode jače pristranosti procjena prema nuli, kako nam se "budžet" smanjuje. Ovdje je bitno napomenuti da sumacija L^2 norme počinje od $j = 1$. To ima smisla jer ne želimo uvesti pristranost parametra β_0 prema 0. Teorija konveksne optimizacije nam govori da je problem optimizacije uz ograničenje ekvivalentan minimizaciji Lagrangeove dualne funkcije:

[1] $\boldsymbol{b^T X^T y} = (\boldsymbol{b^T X^T y})^T = \boldsymbol{y^T Xb}$

[2] $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{Ax} = \boldsymbol{A}$ i $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x^T Ax} = \boldsymbol{x}^T(\boldsymbol{A} + \boldsymbol{A}^T)$ za $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, $\boldsymbol{x} \in \mathbb{R}^{n \times 1}$

$$\max_{\lambda \geq 0} \min_{\boldsymbol{b}} \quad \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) + \lambda \left(\sum_{j=1}^{p} b_j^2 - t \right) \right\}.$$

S obzirom na to da ćemo mi birati t iz podataka, to na neki način odgovara biranju λ. Također za fiksni λ član $-\lambda t$ u gornjem izrazu ne utječe na vrijednost $\boldsymbol{b}$ za koju se postiže minimum. To nas vodi na sljedeću zadaću:

$$\min_{\boldsymbol{b}} \quad \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) + \lambda \sum_{j=1}^{p} b_j^2 \right\},$$

gdje $\lambda \geq 0$ treba biti izabran. Dakle, parametar t je implicitno sadržan u ciljnoj funkciju u obliku Lagrangeovog multiplikatora λ. Rješenje ovog problema je također dano u zatvorenoj formi, te je račun potpuno analogan iskazu iz prošlog poglavlja, uz napomenu da se L^2 norma može zapisati kao:

$$\lambda \sum_{j=1}^{p} b_j^2 = \lambda \, \boldsymbol{b}^T \boldsymbol{I}^* \boldsymbol{b}$$

$$\text{gdje je} \quad \boldsymbol{I}^* = \operatorname{diag}(0, 1, \ldots, 1) \in \mathbb{R}^{(p+1)\times(p+1)}.$$

Rješenje za Ridge regresiju u zatvorenoj formi je dano sa:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}^*)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Vidimo da je jedina razlika u tome što dodajemo vrijednost λ na dijagonalu Gramove matrice, time smanjujemo linearne zavisnosti u i ona postaje stabilnija pri invertiranju.

Parametar λ kontrolira kompleksnost modela, različite vrijednosti daju različite modele. Za $\lambda = 0$ dobijemo metodu najmanjih kvadrata, dok velike vrijednosti λ vode prema rješenju $\hat{\boldsymbol{\beta}} = \boldsymbol{0}$. O odabiru parametra λ govorimo u poglavlju 5.

4 Lasso Regresija

Lasso regresija uvodi kaznu (penalizaciju) u obliku L^1 norme koeficijenata. Optimizacijski problem Lasso regresije je:

$$\min_{\boldsymbol{b}} \quad (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \tag{2}$$

$$\text{uz ograničenje} \quad \sum_{j=1}^{p} |b_j| < t, \quad t \in \mathbb{R}.$$

Na isti način kao u prethodnom poglavlju dolazimo do problema:

$$\min_{\boldsymbol{b}} \quad \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) + \lambda \sum_{j=1}^{p} |b_j| \right\}, \quad \lambda \geq 0. \tag{3}$$

S obzirom na to da je problem striktno konveksan, postoji jedinstveno rješenje. Iako ga je moguće napisati preko podgradijenta, prikazat ćemo kako ga numerički (iterativno) računamo.

4.1 Cikličan Koordinatni Spust

Pogledajmo prvo slučaj s jednim regresorom

$$\min_{b} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - z_i b)^2 + \lambda |b| \right\}. \tag{4}$$

i pretpostavimo da su podatci prethodno transformirani tako da vrijedi:

$$\frac{1}{N} \sum_{i=1}^{N} y_i = 0, \quad \frac{1}{N} \sum_{i=1}^{N} z_i = 0, \quad \frac{1}{N} \sum_{i=1}^{N} z_i^2 = 1. \tag{5}$$

Ovo se uvijek može postići oduzimanjem jedne konstante od svih y_i, druge konstante od svih x_i te potom skaliranjem svih transformiranih x_i (što mijenja λ). Te centrirajuće konstante upravo su aritmetički prosjeci pripadnih vektora, tj. $\frac{1}{N}\sum_{i=1}^{N} y_i$ i $\frac{1}{N}\sum_{i=1}^{N} x_i$. Standardizacija stupaca matrice dizajna dovodi ih na istu skalu, što olakšava usporedbu regresijskih koeficijenata; po potrebi, koeficijenti se mogu vratiti na izvornu skalu. Centriranje vektora odziva i vektore prediktora omogućuje zanemarivanje β_0 pri procjeni tj. uzimanje $\beta_0 = 0$. Vraćanjem na originalnu skalu dobivamo $\beta_0 \neq 0$.

Jednadžbu 4 ne možemo direktno derivirati jer funkcija apsolutne vrijednosti nije derivabilna, ali je konveksna. Poopćenje derivacije funkcije za funkcije koje su konveksne i nisu nužno derivabilne je koncept **podgradijenta**. Za konveksnu funkciju $f : \mathbb{R}^p \to \mathbb{R}$, kažemo da je $\boldsymbol{z} \in \mathbb{R}$ podgradijent funkcije f u $(\boldsymbol{x_0})$ ako vrijedi:

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x_0}) + \boldsymbol{z}^T(\boldsymbol{x} - \boldsymbol{x_0}) \quad , \forall \boldsymbol{x} \in \mathbb{R}^p.$$

Skup svih podgradijenata funkcije u točki $\boldsymbol{x}_0$, označen s $\partial f(\boldsymbol{x}_0)$, naziva se **poddiferencijal** funkcije u točki $\boldsymbol{x}_0$. Može se pokazati da konveksna funkcija f ima globalni minimum u točki x_0 ako i samo ako $0 \in \partial f(\boldsymbol{x}_0)$.

Podgradijent funkcije $f(x) = |x|$ je:

$$\partial f = \begin{cases} \{+1\}, & x > 0 \\ [-1, 1], & x = 0 \\ \{-1\}, & x < 0. \end{cases}$$

Primijetimo da funkcija $f(x) = |x|$ ima ekstrem za $x_0 = 0$ i da $0 \in \partial f(0)$.

Nužan i dovoljan uvjet optimalnog rješenja jednadžbe 4 problema izražavamo pomoću podgradijenta ($s \in \partial |\cdot|$):

$$\begin{aligned} -\frac{1}{N}\sum_{i=1}^{N}(y_i - z_i\beta)z_i + \lambda s = 0 \\ -\frac{1}{N}\sum_i y_i z_i - \frac{1}{N}\sum_i z_i^2\beta + \lambda s = 0 \\ -\frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle - \beta + \lambda s = 0 \\ \beta = \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle - \lambda s, \end{aligned}$$

gdje je $s = sign(\beta)$ kada je $\beta \neq 0$ odnosno $s \in [-1, 1]$ kada je $\beta = 0$. Uočimo da smo koristili uvjete 5. Kako su s i β međuovisni, imamo slučajeve:

$$\widehat{\beta} = \begin{cases} \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle - \lambda, & \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle > \lambda \\ 0, & \frac{1}{N}|\langle \mathbf{z}, \mathbf{y}\rangle| \leq \lambda \\ \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle + \lambda, & \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle < -\lambda. \end{cases} \tag{6}$$

Dakle, rješenje za slučaj jednog regresora je zapravo u zatvorenoj formi, i preslikavanje 6 je poznato kao operator mekog praga (*Soft-threshold operator*) u oznaci:

$$\widehat{\beta} = \mathcal{S}_\lambda\left(\frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle\right), \tag{7}$$

gdje je

$$\mathcal{S}_\lambda(x) = \begin{cases} x - \lambda, & x > \lambda \\ 0, & |x| \leq \lambda \\ x + \lambda, & x < -\lambda. \end{cases}$$

Algoritam cikličnog koordinatnog spusta u slučaju više regresora direktno koristi rezultat 7. Optimizacijski problem više regresora ako fiksiramo sve koeficijente osim β_j glasi:

$$\min_{\beta_j}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j|\right\}.$$

Ovdje opet koristimo pretpostavku da su podaci transformirani tj. da vrijedi

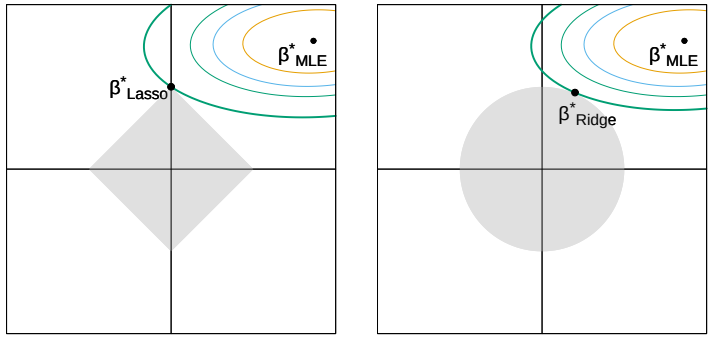
$$\frac{1}{N}\sum_{i=1}^{N} y_i = 0, \quad \frac{1}{N}\sum_{i=1}^{N} x_{ij} = 0, \forall j, \quad \frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1, \forall j.$$

Uz supstituciju parcijalnog reziduala $r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\beta_k$, rješenje je:

$$\widehat{\beta}_j = \mathcal{S}_\lambda\left(\frac{1}{N}\langle \mathbf{x_j}, \mathbf{r}^{(j)}\rangle\right). \tag{8}$$

Za fiksnu vrijednost λ, algoritam koordinatnog spusta iterativno primjenjuje pravilo ažuriranja (8) za svaki regresijski koeficijent β_j, prolazeći redom kroz sve koordinate $j \in \{1, \ldots, p\}$. Kreće se s vektorom $\boldsymbol{\beta} = 0$ te se ažurira β_1 (uzimajući da su ostale koeficijenti nula), zatim se uz novo izračunati β_1 ažurira β_2 (uzimajući da je $\beta_3 = \beta_4 = \cdots = \beta_p = 0$) itd. Nakon ažuriranja pojedine koordinate, potrebno je ažurirati i vektor reziduala $\boldsymbol{r}$, jer on ovisi o regresijskim koeficijentima. Dakle, jedan prolaz algoritma odgovara ažuriranju svih koeficijenata, a algoritam se zaustavlja kada razlika između vektora koeficijenata u dva uzastopna prolaza postane manja od unaprijed zadane tolerancije.

Povoljno svojstvo Lasso regresije jest to da rezultira modelima u kojima je velik broj koeficijenata točno jednaki nuli, i time poboljšavamo interpretabilnost. Taj efekt je jasno vidljiv na slici 2, koja prikazuje prostor koeficijenata te razlike u Lasso i Ridge procjenama.



(a) L^1 norma (b) L^2 norma

Slika 2. Razlika u procjenama gdje su zadovoljena ograničenja

Slika 2 obrazlaže heuristiku da se, kada se ograničenje zada u L^1 normi, minimum postiže na šiljku tj. kada su neki od koeficijenata nula.

5 K-struka Unakrsna Validacija

U prijašnjem poglavlju prikazali smo algoritam cikličnog koordinatnog spusta i parametar $\lambda \in \mathbb{R}$ koji kontrolira složenost modela. Za $\lambda = 0$ regularizacija je ugašena, *Soft-threshold* operator S_0 ponaša se kao identitet, a rješenje je približno jednako metodi najmanjih kvadrata. Za λ_{MAX} vrijedi $\hat{\boldsymbol{\beta}} = \mathbf{0}$, dok za svaki $\lambda < \lambda_{\text{MAX}}$ barem jedan koeficijent $\beta_j \neq 0$. Optimalan $\lambda_{\text{CV}} \in \langle 0, \lambda_{\text{MAX}} \rangle$ određujemo K-strukom unakrsnom validacijom (npr. $K = 5$ do $K = 10$). Skup podataka dijelimo na K podskupova, a za svaki λ_t iz guste mreže u $\langle 0, \lambda_{\text{MAX}} \rangle$ procijenimo regresijske koeficijente na uniji $K - 1$ podskupova, dok MSE računamo na preostalom. Prosjek K takvih procjena daje $MSE_{\text{CV}}(\lambda_t)$.

Odabir K ovisi o nekoliko čimbenika. Prvenstveno, ovisi o veličini uzorka: veći uzorci dopuštaju veće vrijednosti K, dok smo na manjim uzorcima primorani birati manje vrijednosti K. Veća vrijednost K dovodi do nešto manje preciznih procjena MSE-a u pojedinom koraku, ali rezultira stabilnijim procijenjenim parametrima, dok manji K povećava varijabilnost procjene parametara, ali poboljšava kvalitetu procjena MSE-a. Ovaj kompromis između stabilnosti procjene parametara i preciznosti procjene MSE-a treba imati na umu pri odabiru broja podskupova. Ekstremni slučaj na manjim uzorcima je $K = N$, što se naziva *Leave One Out Cross Validation* (LOOCV), gdje u svakom koraku evaluiramo MSE na jednoj opservaciji, dok procjenjujemo parametre na preostalih $N - 1$ opservacija.

Može se pokazati [3] da je najmanji λ za kojeg koordinatni spust vraća opet nul vektor jednak:

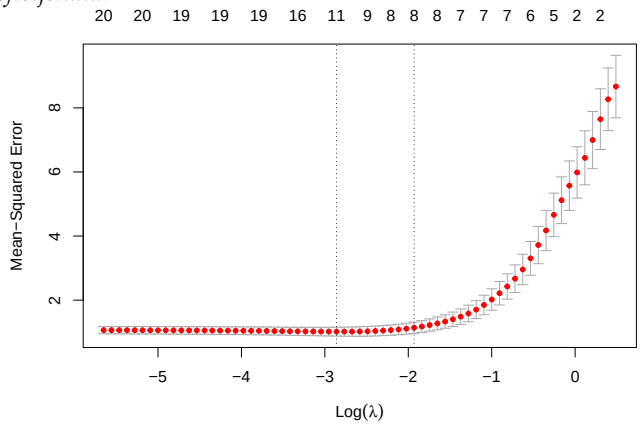
$$\lambda_{MAX} = \max_j \{\frac{1}{N} |\langle \boldsymbol{x_j}, \boldsymbol{y} \rangle|\},$$

gdje je $\boldsymbol{x_j} = (x_{ij})_{i=1}^N$.

Primjer 5.1. *Prikazat ćemo rezultat algoritma unakrsne validacije na testnim podacima koji su dostupni u biblioteci* [2, *glmnet*]. *Na slici 3 prikazana je krivulja unakrsne validacije za lasso regresiju. Postupak unakrsne validacije započinje s desne strane grafa, gdje vidimo da modeli s dva regresijska koeficijenta imaju visoku vrijednost MSE_{CV}. Kako se vrijednost λ smanjuje, smanjuje se i MSE_{CV}, sve do točke $\lambda = 0.057$ (odnosno $\ln(\lambda) = -2.86$), nakon koje krivulja počinje*

[3] Potrebno je prvo pokazati da je pravilo ažuriranja 8 ekvivalentno $S_\lambda(\beta_j + \frac{1}{N}\langle \boldsymbol{x_j}, \boldsymbol{r} \rangle)$, pa onda raspisati izraz kada je $\boldsymbol{\beta} = \mathbf{0}$ (**NB**: $r_i = y_i - \sum_{k=1}^p x_{ik}\beta_k$)

blago rasti. Optimalna vrijednost Lagrangeovog multiplikatora je ona koja minimizira krivulju unakrsne validacije, označena kao λ_{CV}, što odgovara modelu s devet koeficijenata.



Slika 3. Krivulja unakrsne validacije. Pri vrhu slike je broj regresijskih koeficijenata koji su različiti od nule. $\lambda_{\text{CV}} = 0.057$

6 Simulacije

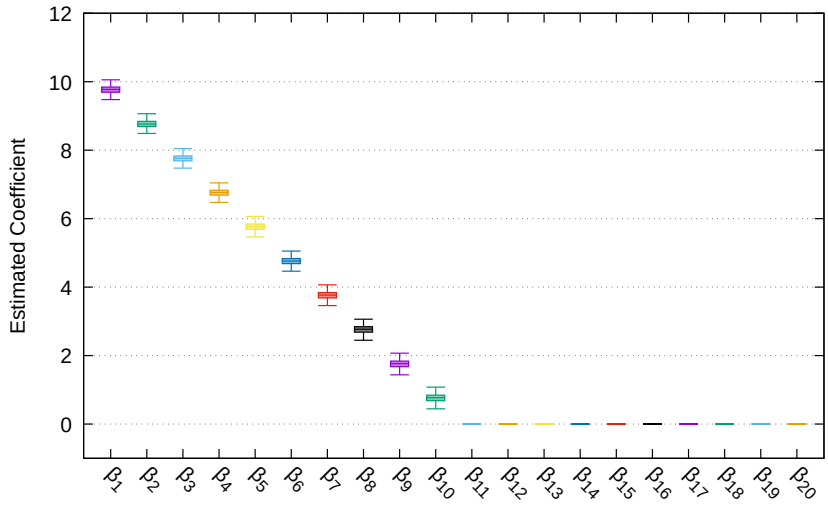
6.1 Slučaj rijetkih podataka

Simulirat ćemo uzorak s populacijskim parametrima:

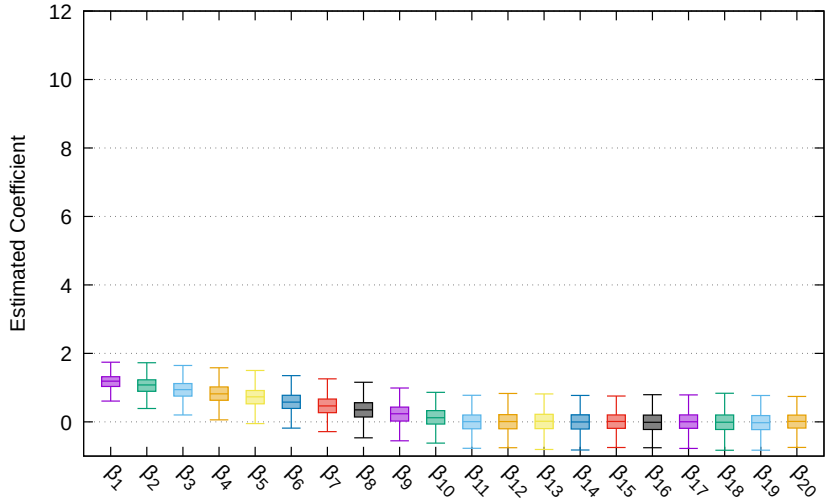
$$\boldsymbol{\beta} = [10, 9, 8, \ldots, 2, 1, 0, 0, \ldots, 0]^T_{100\times 1} \quad ; \quad \sigma = 0.1 \quad ; \quad N = 50.$$

Vektor $\mathbf{y}_{N\times 1}$ ćemo dobiti na način da na $\boldsymbol{X\beta}$ dodamo slučajan Gaussov šum. Potom ćemo procijeniti parametre Lasso i Ridge regresije (s optimalnim λ_{CV}), ponavljat ćemo navedene korake ukupno 1000 puta i pogledati empirijske distribucije procijenjenih parametara. Parametar β_0 ćemo izostaviti iz razmatranja.

Iz slika 4(a) i 4(b) vidimo da se Lasso ponaša puno bolje od Ridge regresije u $p \gg N$ slučaju. Također vidimo pristranost koju uvodi Lasso, parametri β_1 do β_{10} su nam blago podcijenjeni. Moramo spomenuti da je ovo prilično pogodan slučaj za Lasso, populacijski parametar koji modelira šum u podacima je relativno malen, $\sigma = 0.1$. To je pozitivno utjecalo na stabilnost procjene parametara. Iako je riječ o slučaju s niskim šumom, koeficijenti β_{11} do β_{100} bili su procijenjeni različito od nule u otprilike 5% situacija. To se vidi u tablici 1, 4808/100000.



(a) Lasso parametri β_1 do β_{20}. Parametri β_1 do β_{10} su blago podcijenjeni, dok su parametri β_{11} do β_{20} često procijenjeni točno na nulu (oznake za kvartile nisu ni vidljive)



(b) Ridge parametri β_1 do β_{20}. Za razliku od 4(a), procjene β_1 do β_{10} su puno dalje od stvarnih vrijednosti, dakle ridge se ne ponaša dobro u $p \gg N$ slučaju. Parametri β_{11} do β_{100} su često procijenjeni na ne-nul vrijednosti.

Slika 4. Empirijske distribucije za prvu simulaciju ($\sigma = 0.1$).
Stršeće vrijednosti su izostavljene iz boxplotova.

Matrica zabune (eng. *confusion matrix*) u tablici 1 prikazuje kako se ponašaju Lasso procjene.

Tablica 1. Matrica zabune ($\sigma = 0.1$).
Precision $= 0.675$, *Recall* $= 0.99$

		Populacijski parametri		
		$\neq 0$	$= 0$	Σ
Lasso	$\neq 0$	9999	4808	14 807
	$= 0$	1	85 192	85 193
	Σ	10 000	90 000	100 000

Slučajevi od interesa u matrici zabune su:

- **True Positive (TP)** označava broj koeficijenata koji su u populaciji različiti od nule i koje je Lasso ispravno procijenio kao različito od nula (ovdje 9999).

- **True Negative (TN)** predstavlja koeficijente koji su u populaciji jednaki nuli i koje je Lasso ispravno procijenio kao nula (ovdje 85 192).

- **False Positive (FP)** su koeficijenti koji su u populaciji zapravo jednaki nuli, ali ih je Lasso pogrešno označio kao različite od nula (ovdje 4808).

- **False Negative (FN)** su koeficijenti koji su u populaciji različiti od nula, a Lasso ih je pogrešno procijenio kao nula (ovdje 1).

Vidimo da je mjera *Recall* $= 99\%$ $(TP/(TP + FN))$, odnosno u ovom slučaju s malim šumom Lasso dobro procijeni koeficijente koji su zapravo različiti od nule. Povećavanjem parametra σ, omjer signala i šuma u podacima postaje manji. Kada je to slučaj, možemo očekivati da će Lasso morati uvesti jaču regularizaciju kako bi izdvojio signal iz podataka. Ponovit ćemo simulacije opisane na početku poglavlja, ali s populacijskim parametrom $\sigma = 1$. Možemo očekivati da će se mjera *Recall* malo smanjiti, zato što jačom regularizacijom uvodimo veću pristranost svih koeficijenata prema nuli. Specifično, povećavanje šuma najviše će utjecati na populacijski parametar $\beta_{10} = 1$, Lasso će taj parametar početi privlačiti prema nuli zbog veće pristranosti, odnosno regularizacije. Matrica zabune ponovljene simulacije s većim šumom dana je u tablici 2.

Tablica 2. Matrica zabune, za prvu simulaciju ($\sigma = 1$).
Precision $= 0.329$, *Recall* $= 0.998$

		Populacijski parametri		
		$\neq 0$	$= 0$	Σ
Lasso	$\neq 0$	9977	20 350	30 327
	$= 0$	23	69 650	69 673
	Σ	10 000	90 000	100 000

Možemo vidjeti da je *Recall* još uvijek dobar, ali nam je *Precision* ($TP/TP+FP$) dvostruko manji. Učinak ovisnosti λ_{CV} o parametru σ prikazuje tablica 3. Vidimo da s porastom populacijskog parametra σ raste i vrijednost λ, što dovodi do izraženije pristranosti regresijskih koeficijenata prema nuli.

Tablica 3. Lasso procjene za prvu simulaciju za 1000 ponavljanja za različite parametre σ ($\overline{\beta_i} \pm$ s.d.)

σ	λ_{CV}	β_1	β_9	β_{10}	β_{11}
1	0.12 ± 0.02	9.69 ± 0.27	1.68 ± 0.27	0.68 ± 0.27	0.00 ± 0.07
2	0.19 ± 0.09	9.44 ± 0.52	1.42 ± 0.53	0.48 ± 0.42	0.00 ± 0.17
3	0.29 ± 0.15	9.15 ± 0.80	1.18 ± 0.70	0.38 ± 0.49	-0.01 ± 0.26
4	0.40 ± 0.23	8.85 ± 1.01	1.00 ± 0.85	0.33 ± 0.54	0.00 ± 0.33
5	0.55 ± 0.32	8.51 ± 1.29	0.84 ± 0.89	0.30 ± 0.58	-0.01 ± 0.38

6.2 Bootstrap metoda za lasso

U prijašnjoj simulaciji smo vidjeli da je Lasso robustan što se tiče dohvata signala u podatcima. Pod signalom ovdje podrazumijevamo stvarni utjecaj prediktora na odziv, tj. one parametre regresijskog modela koji su različiti od nule. Parametre koji su bili različiti od nule Lasso je i procijenio kao različite od nule, uz određenu pristranost prema nuli, što se odrazilo u visokoj vrijednosti *Precision*-a. Čak i kada je prisutan značajniji šum od $\sigma = 1$ u podatcima, *Precision* je ostao vrlo visok, ali je *Recall* znatno opao. S većim šumom svi parametri su privučeni više prema nuli, te se znalo dogoditi da je $\hat{\beta}_{10} \approx 0.40$, a da su neki parametri koji su zapravo nula procijenjeni na sličnu vrijednost.

U ovoj simulaciji prikazat ćemo bootstrap metodu za Lasso regresiju. Ideja je da poduzorkovanjem i opetovanim procjenama na poduzorcima dobijemo histogram na kojem možemo vidjeti koliko je puta pojedini koeficijent bio pritegnut točno na nulu. Iako nećemo dobiti točne intervale

pouzdanosti, imat ćemo puno bolji uvid u značajnost pojedinih procjena. Vidjet ćemo kako se bootstrap procjene mijenjaju s povećanjem šuma.

Simulirat ćemo dva nezavisna uzorka podataka generirana prema modelu linearne regresije kao u prvoj simulaciji. Parametri modela su:

$$\boldsymbol{\beta} = [10, \ldots, 1, 0, \ldots, 0]^T_{100\times 1} \quad ; \quad \sigma \in \{1, 4\} \quad ; \quad N = 50.$$

Bootstrap postupak uključuje iduće korake:

1. Dohvati slučajan podskup od $k < N$ opservacija $(k = 40)$.

2. Unakrsnom validacijom izračunaj λ_{CV}.

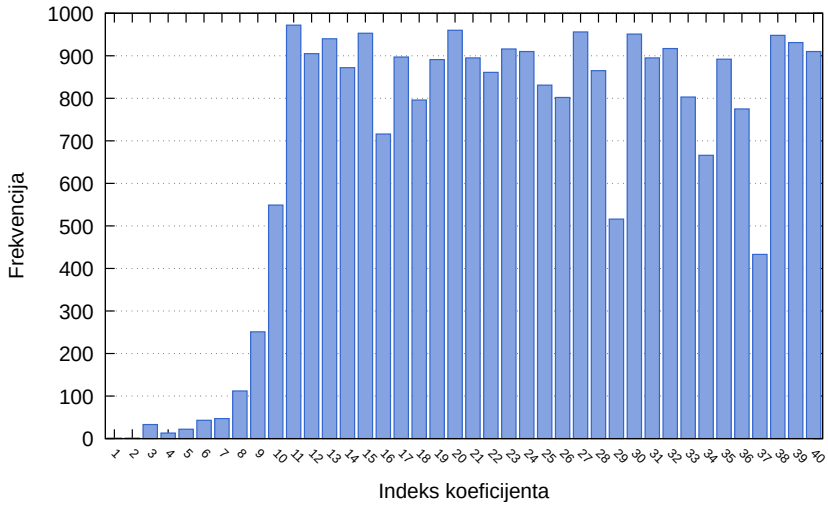
3. Za λ_{CV} procijeni parametre $\boldsymbol{\beta}$ i zapamti procijenjene parametre.

4. Ponavljati korake 1 do 3 ukupno 1000 puta.

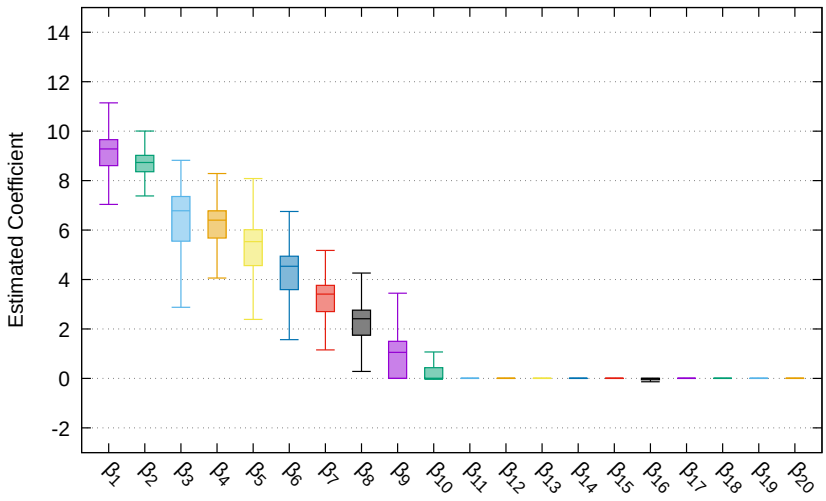
Na kraju ćemo imati 1000 bootstrap procjena parametara [4] i možemo vidjeti koliko puta su parametri bili privučeni na 0.

Iz slika možemo vidjeti kako veći šum rezultira većom varijabilnošću u procjenama (usporedi slike 5(b) i 6(b)), ali čak i u prisutnosti visokog šuma $(\sigma = 4)$ bootstrap metoda daje prilično pogodne rezultate. Iz slike 6(b) vidimo da nam je šum prevelik i da je Lasso privukao β_{10} na nulu. Iz histograma 6(a) vidimo da su parametri s indeksima 1, 2, 4 i 7 za simulirani uzorak značajniji od ostalih parametara. Na parametre veće magnitude visoki šum nije toliko utjecao, ali je dosta povećana varijanca procjena.

[4] Dobivene bootstrap distribucije parametara nisu klasični intervali pouzdanosti. Autori biblioteke [2, glmnet] namjerno izostavljaju bootstrap procjene kako ne bi došlo do zabune s intervalima pouzdanosti.

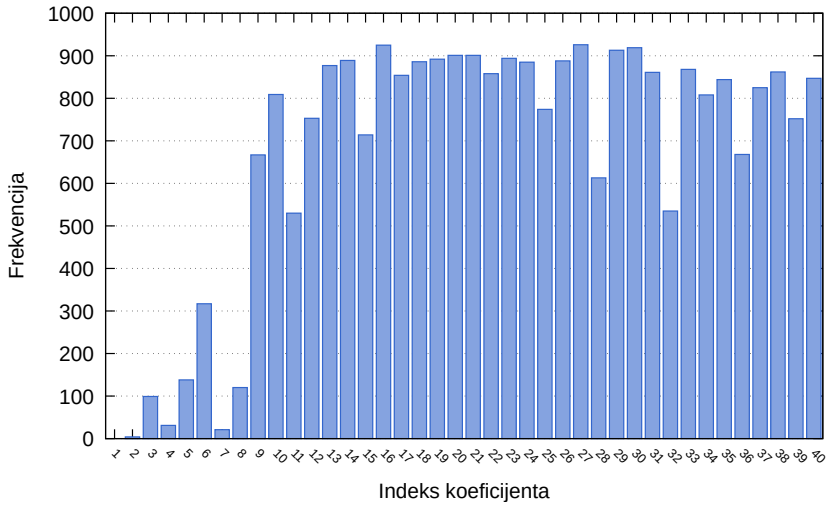


(a) Histogram bootstrap realizacija. Frekvencija označava koliko je pojedini parametar procijenjen točno na nulu. Parametri s indeksima 1 do 10 su nam bili različiti od nule i imali su opadajuće vrijednosti. Vidimo da su parametri veće magnitude manji broj puta procijenjeni na nulu.

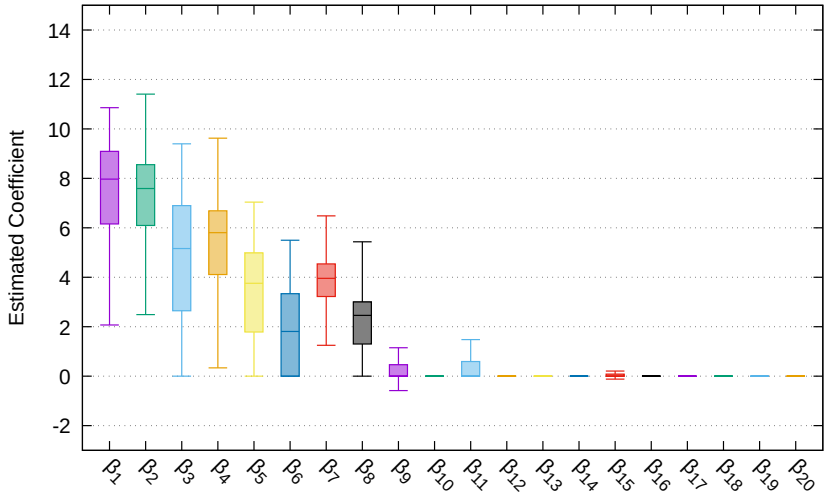


(b) Bootstrap distribucije koeficijenata. Vidimo da su parametri β_1 do β_{10} podcijenjeni i imaju relativno velik varijabilitet, dobili smo dobru informaciju koji parametri imaju utjecaj na y.

Slika 5. Rezultat bootstrap metode za $\sigma = 1$



(a) Histogram bootstrap realizacija. Frekvencija označava koliko je pojedini parametar procijenjen točno na nulu. U slučaju s većim šumom je puno teže zaključiti da je $\beta_6 \neq 0$. Parametre s indeksima $1, 2, 4$ i 7 bi lagano uključili u model.



(b) Bootstrap distribucije koeficijenata. Vidimo da smo dobili takav uzorak gdje β_9 i β_{11} imaju slične distribucije. Dakle kako šum raste lakše nam je napraviti grešku tako da izbacimo iz modela parametre koji imaju relativno malenu magnitudu.

Slika 6. Rezultat bootstrap metode za $\sigma = 4$

Literatura

[1] T. Hastie, R. Tibshirani i M. Wainwright , *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press Taylor & Francis group, 2016.

[2] T. Hastie, J. Qian, K. Tay, Biblioteka za R: *glmnet*, `https://glmnet.stanford.edu/articles/glmnet.html`.

[3] R-*project*, `https://cran.r-project.org/web/packages/rlang/index.html`.

[4] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.