(3) Regularizacija linearnog modela [ridge regresija]

Notacija:

- vektore u $\mathbb{R}^k$ ($k \in \mathbb{N}$) shvaćamo kao vektor-stupce

- $$\mathbb{X} = \begin{bmatrix} — & X^{(1)T} & — \\ & \vdots & \\ — & X^{(n)T} & — \end{bmatrix} = \begin{bmatrix} x_{11}, \dots, x_{1p} \\ \vdots \\ x_{n1}, \dots, x_{np} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

  matrica dizajna

- $y = (y_1, \dots, y_n)^T \in \mathbb{R}^{n \times 1}$

- $\mathbb{X}_j = (x_{1j}, \dots, x_{nj})^T \in \mathbb{R}^{n \times 1}$, $j = 1, \dots, p$

  ▷ j-ti stupac od $\mathbb{X}$ (unjedinstiti kovarijate X_j u svim $X^{(i)}$, $i = 1, \dots, n$)

(često u $\mathbb{X}$ dodajemo $\mathbb{X}_0 = (1, \dots, 1)^T \in \mathbb{R}^{n \times 1}$ $x_{10}, \dots, x_{n0}$)

Lin. model: biramo $\hat{f} = f(\tau)$ iz $\mathcal{F} = \{ f_\beta : \beta \in \Theta \}$ uz $\Theta \subseteq \mathbb{R}^{p+1}$ i

$$f_\beta(x) = X^T \cdot \beta, \quad X = (x_1, \dots, x_p) \in \mathbb{R}^{p+1}. \tag{3.1}$$

[dopuštamo (i) $X_2 := f(X_1)$, (ii) interakcije $X_3 := X_1 \cdot X_2$, (ii) kvalitativne kovarijate.]

(3.1) Metoda najmanjih kvadrata → (LS) ["least squares"]

$$\hat{\beta}^{ls} = (\hat{\beta}_0^{ls}, \dots, \hat{\beta}_p^{ls}) := \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - f_\beta(X^{(i)}))^2}_{=: RSS(\beta)}$$

$L_2(f_\beta)$ uz $L = (y - \hat{y})^2$ (3.2)

$$= \underset{\beta}{\operatorname{argmin}} \| y - \mathbb{X}\beta \|_2^2$$

Pretpostavljamo da su $X_0, \dots, X_p \in \mathbb{R}^n$ lin. nezavisni, tj. da je $X^T X \in \mathbb{R}^{(p+1)\times(p+1)}$ poz. definitna (simetrična) matrica.

(mora biti $\boxed{n \geq p+1}$)

- $$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y = \frac{1}{n} \hat{\Sigma}^{-1} X^T y, \qquad (3.3)$$

gdje je $\boxed{\hat{\Sigma} := \frac{1}{n} X^T X.}$

- $$\hat{y} := \left(f_{\hat{\beta}^{OLS}}(x^{(1)}), \dots, f_{\hat{\beta}^{OLS}}(x^{(n)}) \right)^T = X \cdot \hat{\beta} = H y, \qquad (3.4)$$

gdje $H = X (X^T X)^{-1} X^T$ ["hat matrix"],

te je $\hat{y}$ ort. projekcija vektora y na $\operatorname{span}(X_0, \dots, X_p)$.

Napomena U slučaju da $X_0, \dots, X_p$ nisu lin. nezavisni (npr. $n < p+1$ ili $X_j = a \cdot X_i$) $\hat{\beta}^{OLS}$ nije jedinstven. Ipak, $\hat{y}$ je i dalje ortogonalna projekcija od y na $\operatorname{span}(X_0, \dots, X_p)$. ■

(3.2) Statistička svojstva od $\hat{\beta}^{OLS}$

[moramo pretpostaviti model za (X, Y)]

Pretpostavimo da $\boxed{\exists \beta_* \in \mathbb{R}^{p+1}}$ t.d.

$$Y_i = (X^{(i)})^T \beta_* + \varepsilon_i, \quad i = 1, \dots, n \qquad (3.5)$$

pri čemu su [nadi jednostavniji]

- $X^{(i)}$, $i = 1, \dots, n$ neslučajni (tzv. "fiksni dizajn")
- ε_i, $i = 1, \dots, n$, nezavisni t.d. $\boxed{\mathbb{E}[\varepsilon_i] = 0}$, $\boxed{\operatorname{Var}(\varepsilon_i) = \sigma^2}$, $\forall i = 1, \dots, n$ za neki $\sigma^2 > 0$.

Nopr. 1 u (3.5) pretpostavljamo da je lin. model *točan*, tj. da nema *greške aproksimacije* ("bias").

(3.5) možemo zapisati kao [tikovi]

$$Y := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \mathbb{X}\beta_* + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} =: \mathbb{X}\beta_* + \mathcal{E}, \qquad (3.6)$$

uz $\mathbb{E}[\mathcal{E}] = 0$, $\mathrm{Var}(\mathcal{E}) = \sigma^2 I_n$.

$$\Rightarrow \quad \hat{\beta}^{LS} = \beta_* + \boxed{\frac{1}{n}\hat{\Sigma}^{-1}\mathbb{X}^T\mathcal{E}} \quad \text{— jedina slučajna}$$

(3.3)

$$\boxed{\begin{aligned} &\bullet\ \mathbb{E}[\hat{\beta}^{LS}] = \beta_* \\ &\bullet\ \mathrm{Var}(\hat{\beta}^{LS}) = \frac{\sigma^2}{n}\hat{\Sigma}^{-1} \end{aligned}} \qquad (3.7)$$

Tipično

$$\hat{\sigma}^2 := \frac{1}{n-p-1}\|Y - \hat{Y}\|_2^2 = \frac{RSS(\hat{\beta})}{n-p-1}. \qquad (3.8)$$

3.2.1 Testna greška

Neka je $\tilde{Y} := \mathbb{X}\beta_* + \tilde{\mathcal{E}}$, pri čemu su $\mathcal{E}$ i $\tilde{\mathcal{E}}$ (tj. Y i $\tilde{Y}$) nezavisni, testni skup.

Promatramo tzv. *testnu grešku unutar uzorka* ("in-sample")

$$L^{in}(f_\beta) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_i - f_\beta(x^{(i)})\right)^2\right]$$

$$= \mathbb{E}\left[\frac{1}{n}\|\tilde{Y} - \mathbb{X}\beta\|_2^2\right], \qquad (3.9)$$

za $\forall \beta \in \mathbb{R}^{p+1}$.

Prop. 3.1 (BVT) Ako je $\hat{\beta} = \hat{\beta}(Y, X)$ procjenitelj za β_*,

$$\mathbb{E}\left[L^{in}(f_{\hat{\beta}})\right] = \sigma^2 + \underbrace{\|\mathbb{E}[\hat{\beta}] - \beta_*\|^2_{\hat{\Sigma}}}_{\text{"pristranost"}} + \underbrace{\mathbb{E}\left[\|\hat{\beta} - \mathbb{E}[\hat{\beta}]\|^2_{\hat{\Sigma}}\right]}_{\text{"varijanca"}}, \tag{3.10}$$

gdje $\boxed{\mathbb{E} = \mathbb{E}_Y}$ te $\boxed{\|\beta\|^2_{\hat{\Sigma}} := \beta^T \hat{\Sigma} \beta}$, $\beta \in \mathbb{R}^{p+1}$.

[kvadratna forma od $\hat{\Sigma}$]

Dokaz:

(1°) Iz (3.6) imamo, $\forall \beta \in \mathbb{R}^{p+1}$, [$\sigma^2$ + reducibilna greška]

$$n \cdot L^{in}(f_\beta) = \mathbb{E}\left[\|Y - X\beta\|_2^2\right] = \mathbb{E}\left[\|X(\beta_* - \beta) + \tilde{\varepsilon}\|_2^2\right]$$

$$= \left[\|a+b\|_2^2 = \langle a+b, a+b\rangle = \|a\|_2^2 + \|b\|_2^2 + 2 \cdot a^T b \;,\; a, b \in \mathbb{R}^{p+1}\right]$$

$$= \mathbb{E}\Big[\|X(\beta_* - \beta)\|_2^2 + \|\tilde{\varepsilon}\|_2^2 + \underbrace{2\left(X(\beta_* - \beta)\right)^T \tilde{\varepsilon}}_{\text{nije slučajno}}\Big]$$

$$= \left[\mathbb{E}\left[\|\tilde{\varepsilon}\|_2^2\right] = n \cdot \sigma^2,\; \mathbb{E}[\tilde{\varepsilon}] = 0 \in \mathbb{R}^n\right]$$

$$= n \cdot \sigma^2 + (\beta_* - \beta)^T X^T X (\beta_* - \beta)$$

$$\Rightarrow \boxed{L^{in}(f_\beta) = \sigma^2 + \|\beta_* - \beta\|^2_{\hat{\Sigma}}}. \tag{3.11}$$

(2°) $\mathbb{E}\left[L^{in}(f_{\hat{\beta}})\right] \overset{(3.11)}{=} \sigma^2 + \mathbb{E}\left[\|\beta_* - \hat{\beta}\|^2_{\hat{\Sigma}}\right]$ ← ovisi o Y

↳ $\mathbb{E}\left[\|\beta_* - \hat{\beta}\|^2_{\hat{\Sigma}}\right] = \mathbb{E}\left[\|(\beta_* - \mathbb{E}[\hat{\beta}]) + (\mathbb{E}[\hat{\beta}] - \hat{\beta})\|^2_{\hat{\Sigma}}\right]$

$$= \left[\|a+b\|^2_{\hat{\Sigma}} \overset{?}{=} \|a\|^2_{\hat{\Sigma}} + \|b\|^2_{\hat{\Sigma}} + 2\, a^T \hat{\Sigma} b\right]$$

$$= \mathbb{E}\left[\|\beta_* - \mathbb{E}[\hat{\beta}]\|^2_{\hat{\Sigma}}\right] + \mathbb{E}\left[\|\mathbb{E}[\hat{\beta}] - \hat{\beta}\|^2_{\hat{\Sigma}}\right] +$$

$$+ 2\, \mathbb{E}\Big[\underbrace{(\beta_* - \mathbb{E}[\hat\beta])^T \hat\Sigma}_{\text{nije slučajno}} (\mathbb{E}[\hat\beta] - \hat\beta) \Big]$$

$$= (\beta_* - \mathbb{E}[\hat\beta])^T \hat\Sigma \underbrace{\mathbb{E}\big[\mathbb{E}[\hat\beta] - \hat\beta\big]}_{=0}$$

□

Prop. 3.2

$$\mathbb{E}\big[L^{in}(\hat\beta^{ols}) \big] = \sigma^2 + 0 + \boxed{\frac{\sigma^2(p+1)}{n}} \qquad (3.12)$$

Dokaz. Slijedi iz Prop. 3.1 jer $\mathbb{E}[\hat\beta^{ols}] = \beta_*$ te

$$\mathbb{E}\big[\| \hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}] \|_{\hat\Sigma}^2 \big] = \mathbb{E}\Big[\underbrace{(\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}])^T \hat\Sigma (\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}])}_{\in \mathbb{R}} \Big]$$

$$= \mathbb{E}[\, tr(\,\cdot\,) \,] = \mathbb{E}\Big[tr\big((\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}]) (\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}])^T \hat\Sigma \big) \Big]$$

($\hat\Sigma$: nije slučajno)

$$= tr\Big(\underbrace{Var(\hat\beta^{ols})}_{= \frac{\sigma^2}{n} \hat\Sigma^{-1}} \cdot \hat\Sigma \Big) = \frac{\sigma^2}{n}\, tr(I_{p+1}) = \frac{\sigma^2(p+1)}{n}$$

□

Alternativno,

$$\mathbb{E}\big[\| \underbrace{\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}]}_{= \hat\beta^{ols} - \beta_*} \|_{\hat\Sigma}^2 \big] = \mathbb{E}\big[\| (X^2 X)^{-1} X^T \varepsilon \|_{\hat\Sigma}^2 \big]$$

$$= \dots = \frac{1}{n}\, \mathbb{E}\big[\underbrace{\varepsilon^T H \varepsilon}_{= tr(\varepsilon^T H \varepsilon)} \big] = \frac{1}{n}\, \mathbb{E}\big[tr(\varepsilon \varepsilon^T H) \big] = \frac{\sigma^2}{n}(p+1)$$

($\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 I_n$, $+ tr(H) = p+1$)

□

Nap. 3.3 |

(i) Iz $\hat{y} = Hy$ te Prop. 2.7 slijedi da

$$df(LS) = tr(H) \underset{\uparrow}{=} p+1 = \#\text{ procijenjenih parametara.}$$

H ort. projektor na $\underbrace{\text{span}\{x_0, \dots, x_p\}}_{\text{lin. nez.}}$

Nadalje, iz Prop. 2.5 imamo $\left(\mathbb{E}[L_T(\hat{f})] = \mathbb{E}_y[L^{in}(\hat{f})]\right)$,

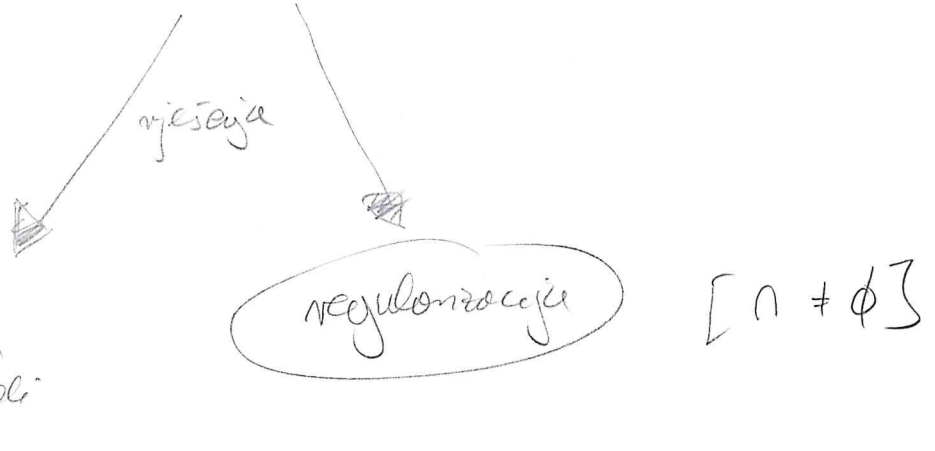
$$\mathbb{E}\left[L_T(f_{\hat{\beta}^{LS}})\right] = \sigma^2 \oplus \frac{\sigma^2(p+1)}{n}$$

$$= \sigma^2 \frac{n-p-1}{n} \quad \left[\xrightarrow{p+1 \to n} 0\right]$$

[Specijalno, za $\hat{\sigma}^2$ iz (3.8) imamo $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.]

(ii) Dakle, čak i kad je lin. model točan (pa je pristranost = 0), višjenca (a samim time i ukupna greška) je velika (u odnosu na σ^2) ako je $\frac{p+1}{n}$ velik!

Uočimo, svaka dodatna kovarijata X_j povećava grešku za $\frac{\sigma^2}{n}$ bez obzira je li $\beta_j \approx 0$ i/ili $X_j \approx X_i$ za $i \neq j$.



(3.3) Ridge regresija

Pr.1 (Motivacija)

Pretp. da $z_1, \dots, z_n$ njd t.d. $z_i \sim N(\boxed{\theta}, \sigma^2)$.

- $MSE(\bar{z}) := \mathbb{E}\left[(\bar{z} - \theta)^2\right] = Var(\bar{z}) + 0 = \frac{\sigma^2}{n} + 0$

- $MSE(\bar{z}/2) := \mathbb{E}\left[\left(\frac{\bar{z}}{2} - \theta\right)^2\right] \underset{\text{Dz}}{=} Var(\bar{z}/2) + (\mathbb{E}[\bar{z}/2] - \theta)^2 = \frac{\sigma^2}{4n} + \boxed{\frac{\theta^2}{4}}$

$\hookrightarrow MSE(\bar{z}) \leq MSE(\bar{z}/2)$

$\Leftrightarrow$

$\boxed{n \geq \frac{3\sigma^2}{\theta^2}}$ ← $\boxed{\text{velike ako } \theta \approx 0.}$ ∎

"L_2 "regularizacija"

Za $\lambda \geq 0$ defin.

$$\hat{\beta}^r_\lambda = \hat{\beta}^r = (\hat{\beta}^r_0, \dots, \hat{\beta}^r_p) = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\beta(x_{i\cdot}))^2 + \boxed{\lambda \sum_{j=1}^{p} \beta_j^2} \right\}$$

$$= \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ RSS(f_\beta) + n\lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

(3.13)

ne uključuje β_0

Nap.3.4 $\forall \lambda \geq 0$, $\forall \beta_1, \dots, \beta_p \in \mathbb{R}$,

$$\underset{\beta_0 \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + n\lambda \sum_{j=1}^{p} \beta_j^2 \right\} = \bar{y} - \sum_{j=1}^{p} \beta_j \bar{x}_j,$$

gdje je $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $j = 1, \dots, p$.

$$\Rightarrow (\hat{\beta}^r_1, \dots, \hat{\beta}^r_p) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \bar{y} - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j \right)^2 + n\lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

centriranje ∎

Zbog Pop. 3.4 u nastavku prepostavljamo da je

$$\bar{x}_j = 0, \ \forall j = 1, \dots, p, \quad \bar{y} = 0, \qquad (3.14)$$

te rješavamo

$$\hat{\beta}^r = (\beta_1^r, \dots, \beta_p^r) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \| y - X\beta \|_2^2 + n\lambda \|\beta\|_2^2 \right\} \qquad (3.15)$$

$X \in \mathbb{R}^{n \times p}$

(vidi (3.24))

Pop. 3.5 | Uz centriranje, kovarijate tipično još i skalirane t.d. imaju uzoračku varijancu 1 jer je kazna $\|\beta\|_2^2$ "nepoštena" ako kovarijate nisu u istim mjernim jedinicama.

$$\left(\left(\frac{Y - \bar{y}}{s_y} \right) = \hat{\beta}_1^r \cdot \left(\frac{x_1 - \bar{x}_1}{s_1} \right) + \dots + \hat{\beta}_p^r \cdot \left(\frac{x_p - \bar{x}_p}{s_p} \right) \Rightarrow Y = \hat{\beta}_0^r + \hat{\beta}_1^r \frac{s_y}{s_1} x_1 + \dots + \hat{\beta}_p^r \frac{s_y}{s_p} x_p \right) \blacksquare$$

$\hat{\beta}_0^r = \bar{y} - \sum_{j=1}^{p} \bar{x}_j \frac{s_y}{s_j} \cdot \hat{\beta}_j^r$

Prp 3.6 | $\forall \lambda \geq 0$,

$$\hat{\beta}^r = \frac{1}{n} \left(\hat{\Sigma} + \lambda I \right)^{-1} X^T y, \qquad (3.16)$$

Dokaz | (DZ). → analogno kao za $\hat{\beta}^{ols}$.

Nap. | (i) za $\lambda > 0$, $\hat{\beta}^r$ je uvijek dobro definirano (čak, i u slučaju $p > n$!)

(ii)
- $\hat{\beta}^r = \hat{\beta}^{ols}$ za $\lambda = 0$,
- $\hat{\beta}^r \xrightarrow{\lambda \to +\infty} 0$.

} hiperparametar λ kontrolira efekt smanjenja (engl. shrinkage) koeficijenata prema 0.

Prp. 3.7 | Pod prepostavkama u (3.5), $\forall \lambda \geq 0$,

$$\mathbb{E}\left[L^{in}\left(f_{\hat{\beta}_\lambda^r} \right) \right] = \sigma^2 + \underbrace{\lambda^2 \beta_*^T \left(\hat{\Sigma} + \lambda I_p \right)^{-2} \hat{\Sigma} \beta_*}_{\text{bias}^2} + \underbrace{\frac{\sigma^2}{n} \operatorname{tr}\left[\hat{\Sigma}^2 \left(\hat{\Sigma} + \lambda I \right)^{-2} \right]}_{\text{varijanca}}. \qquad (3.17)$$

$$\hat{\beta}^r_\lambda = \frac{1}{n}\left(\hat{\Sigma} + \lambda I_p\right)^{-1} X^T X \beta_* + \frac{1}{n}\left(\hat{\Sigma} + \lambda I_p\right)^{-1} X^T \varepsilon \quad (*)$$

$Y = X\beta_* + \varepsilon$

Iz Prop. 3.1, $\mathbb{E}\left[L^{in}(t\,\hat{\beta}^r_\lambda)\right] - \sigma^2 = B + V$, gdje

- $B = \| \mathbb{E}[\hat{\beta}^r_\lambda] - \beta_* \|^2_{\hat{\Sigma}}$

$$\mathbb{E}[\hat{\beta}^r_\lambda] \overset{\mathbb{E}[\varepsilon]=0}{=} \frac{1}{n}\left(\hat{\Sigma} + \lambda I_p\right)^{-1} X^T X \beta_* = \left(\hat{\Sigma} + \lambda I_p\right)^{-1}\left[\hat{\Sigma}\beta_* \pm \lambda I_p \beta_*\right]$$

$$= \beta_* - \lambda\left(\hat{\Sigma} + \lambda I_p\right)^{-1}\beta_*$$

$$\Rightarrow B = \| \lambda\left(\hat{\Sigma} + \lambda I_p\right)^{-1}\beta_* \|^2_{\hat{\Sigma}}$$

$$= \lambda^2 \beta_*^T \left(\hat{\Sigma} + \lambda I_p\right)^{-1} \hat{\Sigma} \left(\hat{\Sigma} + \lambda I_p\right)^{-1} \beta_*$$

$\hat{\Sigma}^T = \hat{\Sigma}$; $\hat{\Sigma}\left(\hat{\Sigma} + \lambda I_p\right)^{-1} = \left(\hat{\Sigma} + \lambda I_p\right)^{-1}\hat{\Sigma}$ (Iz)

$$= \lambda^2 \beta_*^T \left(\hat{\Sigma} + \lambda I_p\right)^{-2} \hat{\Sigma}\beta_*.$$

- $V = \mathbb{E}\left[\| \hat{\beta}^r_\lambda - \mathbb{E}[\hat{\beta}^r_\lambda] \|^2_{\hat{\Sigma}}\right]$

$$= \mathbb{E}\left[\| \frac{1}{n}\left(\hat{\Sigma} + \lambda I_p\right)^{-1} X^T \varepsilon \|^2_{\hat{\Sigma}}\right]$$

$$= \mathbb{E}\left[tr\left(\frac{1}{n} \varepsilon^T X \left(\hat{\Sigma} + \lambda I_p\right)^{-1} \hat{\Sigma} \left(\hat{\Sigma} + \lambda I_p\right)^{-1} X^T \varepsilon \right)\right]$$

$$= \frac{\sigma^2}{n} tr\left(\hat{\Sigma}\left(\hat{\Sigma} + \lambda I_p\right)^{-1} \hat{\Sigma} \left(\hat{\Sigma} + \lambda I_p\right)^{-1} \right)$$

($\mathbb{E}[\varepsilon \varepsilon^T] = \sigma^2 I_n$; kruženje)

$$= \frac{\sigma^2}{n} tr\left(\hat{\Sigma}^2 \left(\hat{\Sigma} + \lambda I_p\right)^{-2} \right)$$

∎

Nap. 1 (i) Pokazat ćemo da je

$$\operatorname{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right) = \sum_{j=1}^{p} \frac{\lambda_j^2}{(\lambda_j+\lambda)^2}$$

gdje su $\lambda_1, \dots, \lambda_p \geq 0$ svojstvene vrijednosti od $\hat{\Sigma}$.

$\Rightarrow$ $\forall \lambda > 0$

varijanca $< \frac{d^2}{n} \cdot p$, te

varijanca $\searrow 0$, za $\lambda \nearrow +\infty$.

(ii) pristranost $= \beta_*^T \left(\frac{\hat{\Sigma}}{\lambda} + I_p\right)^{-2} \hat{\Sigma} \beta_* \nearrow \beta_*^T \hat{\Sigma} \beta_* = \|\beta_*\|_{\hat{\Sigma}}^2$,

$(\neq 0!)$

kada $\lambda \nearrow +\infty$. ∎

(iii) Može se pokazati da postoji $\lambda_* > 0$ t.d.

$$\mathbb{E}\left[L^{in}\left(t_{\hat{\beta}^{ols}}\right)\right] \geq \mathbb{E}\left[L^{in}\left(t_{\hat{\beta}^{r}_{\lambda_*}}\right)\right].$$

↪ u praksi, λ odabiremo CV metodom.

[λ_* ne možemo eksplicitno izračunati jer ovisi o nepoznatim veličinama.]

(3.3.1) Kako ridge smanjuje koeficijente?

SVD od $\mathbb{X} \in \mathbb{R}^{n \times p}$ je

$$\underset{n\times p}{\mathbb{X}} = \underset{n\times p}{U} \cdot \underset{p\times p}{D} \cdot \underset{p\times p}{V^T}, \qquad (3.18)$$

pri čemu

- $U^T U = I_p$
- $VV^T = V^T V = I_p$

} stupci $u_1, \dots, u_p \in \mathbb{R}^{n\times 1}$ od U i $v_1, \dots, v_p \in \mathbb{R}^{p\times 1}$ su ortonormirani.

- $D = \operatorname{diag}(d_1, \dots, d_p)$ za $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$

["singularne vrijednosti"]

Zbog pretp. $\bar{x}_j = 0, \forall j = 1, \dots, p$,

$\hat{\Sigma} = \frac{1}{n} \mathbb{X}^T \mathbb{X} \in \mathbb{R}^{p\times p}$ = uzoračka kovarijacijska matrica od $x^{(1)}, \dots, x^{(n)}$.

$$\hat{\Sigma} \overset{(3.18)}{=} V \frac{D^2}{n} V^T,$$

tj. $\vartheta_j := \frac{d_j^2}{n}, j = 1, \dots, p$ su svojstv. vrijednosti od $\hat{\Sigma}$, a

$v_j, j = 1, \dots, p$ odgovarajući svojstv. vektori.

tzv. smjerovi glavnih komponenti (od $\mathbb{X}$)

↳ pretp. da $\operatorname{rang}(\mathbb{X}) = p$ (te $n > p$) $\Rightarrow$ $d_p > 0$.

Nadalje, uchtne

$$Z_j := \mathbb{X} v_j = \begin{bmatrix} v_j^T x^{(1)} \\ \vdots \\ v_j^T x^{(n)} \end{bmatrix} \in \mathbb{R}^{n \times 1}, \quad j = 1, \dots, p \qquad (3.10)$$

zovemo glavne komponente (od $\mathbb{X}$).

($\text{Proj}_{v_j}(x) = (v_j^T \cdot x) \cdot v_j$, za $x \in \mathbb{R}^p$)

Nap.: 1) Vrijedi, $\forall j = 1, \dots, p$,

$$v_j = \underset{\substack{v \in \mathbb{R}^p, \\ \|v\|_2 = 1}}{\arg\max} \widehat{\text{Var}}(\mathbb{X} v),$$

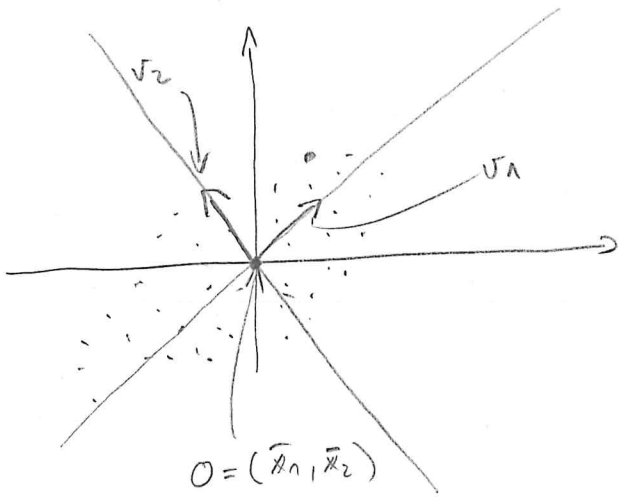
↖ uzoračka varijanca od $v^T x^{(i)}$, $i = 1, \dots, n$

uz uvjet

$$v_j^T v_\ell = 0, \quad \forall \ell = 1, \dots, j-1.$$

[bez dokaza → Primjerjena statistika]

npr. $p = 2$



[v_1 je smjer v koji maksimizira varijancu projekcija od $x^{(1)}, \dots, x^{(n)}$ na v; v_2 je smjer $v \perp v_1$ koji maksimizira

——— " ——— , itd.]

Neka je

$$\mathcal{Z} := \mathcal{X} V = \begin{bmatrix} | & & | \\ z_1 & \cdots & z_p \\ | & & | \end{bmatrix} \in \mathbb{R}^{n \times p}, \qquad (3.20)$$

zapravo početaka u bazi $v_1, \dots, v_p$.

(3.18) $\Rightarrow$ $\mathcal{Z} = U \cdot D$, tj.

$$\boxed{z_j = d_j \cdot u_j, \; j = 1, \dots, p} \qquad (3.21)$$

Specijalno,

- $\|z_j\|_2^2 = z_j^T z_j = d_j^2, \; \forall j,$

te

$$\widehat{Var}(z_j) = \frac{d_j^2}{n} = \lambda_j, \; \forall j. \qquad (3.22)$$

$$\left[\bar{z}_j = (\bar{\mathcal{X}})^T \cdot u_j = 0, \; \forall j = 1, \dots, p \right]$$

- $\boxed{z_j^T z_k = 0 = \widehat{Cov}(z_j, z_k), \; \forall j \neq k}$ (3.23)

$\Rightarrow \operatorname{span}(x_1, \dots, x_p) = \operatorname{span}(z_1, \dots, z_p).$

↳ Ridge regresija i glavne komponente

Prop. da je $\boxed{\widehat{Cov}(x_j, x_k) = 0 = x_j^T x_k, \; \forall j \neq k.}$

$\Rightarrow$ $X^T X = n \hat{\Sigma} = \operatorname{diag}(\|x_1\|_2^2, \dots, \|x_p\|_2^2)$

te

- $\hat{\beta}_j^{ls} = \dfrac{x_j^T y}{\|x_j\|_2^2}, \; j = 1, \dots, p$ ($\hat{y} = \operatorname{Proj}_{x_1, \dots, x_p}(y)$)

Isto kao kod regresije od y samo na x_j!

Veće smanjenje ako je $\|x_j\|_2^2 = n \cdot \widehat{Var}(x_j)$ manja!

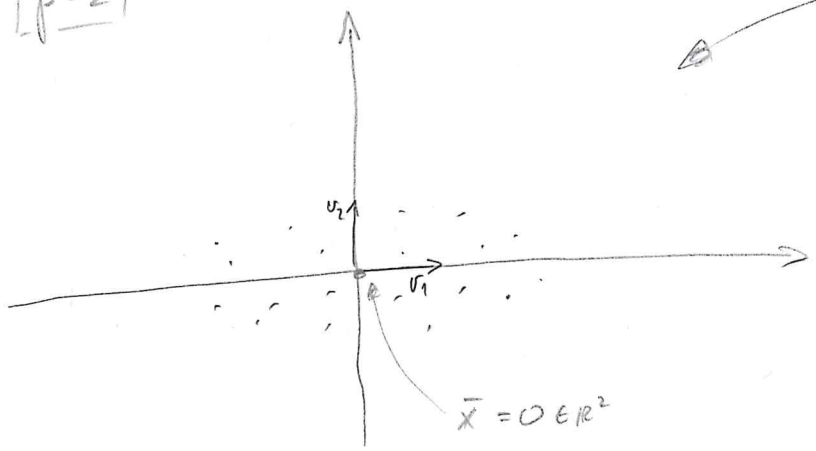
(3.24)

- $\hat{\beta}_j^{r} = \dfrac{\|x_j\|_2^2}{\|x_j\|_2^2 + n\lambda} \cdot \hat{\beta}_j^{ls}, \; j = 1, \dots, p$

(P2) → (3.16)

< 1, za $\lambda > 0$!

mpr. 1 [p=2]



$Var(x_1) > Var(x_2)$ te x_1 puno veći

- Intuitivno, ako prilagodavamo lin. plohu (Y-os je tada koordinata), $Var(\hat{\beta}^{LS})$ (a samim time i varijanca u $\mathbb{E}[L^{in}(\hat{f}_{\hat{\beta}^{LS}})]$) bit će velika jer procjena nagiba u smjeru v_2 jako ovisi o T.

- Ridge regresija smanjuje tu varijancu t.d. smanjuje procjene nagiba u smjerovima v_1 i v_2, a više u smjeru v_2.

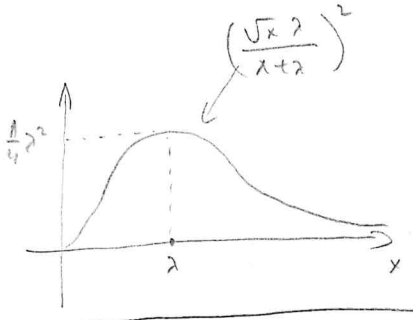
→ pristranost ovisi o λ te o tome je li stvarni nagib u smjeru v_2 (tj. β_2) velik ili ne!

↳ Iz $\hat{\Sigma} = diag(\underbrace{\widehat{Var}(x_j)}_{=s_j} : j=1,\dots,n)$ imamo da je pristranost u $\mathbb{E}[L^{in}(\hat{f}_{\hat{\beta}^R_\lambda})]$ jednaka (Dt)

$$\sum_{j=1}^{p} \left(\frac{\sqrt{s_j}\cdot\lambda}{s_j+\lambda}\right)^2 (\beta_j^*)^2$$

$s_j \to +\infty$, $s_j \to 0$ → 0

max za $s_j = \lambda$.



→ ridge - simulacijski primjer.pdf

Opčenito, $\forall \beta \in \mathbb{R}^p$

$X = U D V^T = Z V^T$

$$\|y - X\beta\|_2^2 + \lambda n \|\beta\|_2^2 = \|y - Z\cdot\alpha\| + \lambda n \|\alpha\|_2^2)$$

gdje je $\alpha = V^T \beta$ ($V^2V = I \Rightarrow \|\alpha\|_2 = \|\beta\|_2$).

$\alpha \overset{1-1}{\leftrightarrow} \beta$

Iz $Z^T Z = \operatorname{diag}(d_1^2, \dots, d_p^2)$,

(3.22), (3.23)

$$\Rightarrow \quad \hat{\alpha}_j^r = \frac{d_j^2}{d_j^2 + n\lambda} \cdot \hat{\alpha}_j^{ls}, \quad j = 1, \dots, p$$

(3.24)

($\hat{\beta}^r = V \cdot \hat{\alpha}^r$)

opčenito, ridge smanjuje ls projekcije, ali mahom projekcije podataka na $v_1, \dots, v_p$, te je smanjuje više za $\hat{\alpha}_j^{ls}$ za koje je $\widehat{Var}(z_j) = \frac{d_j^2}{n} = \lambda_j$ manje.

Nap. 3.9 | df(ridge) = ?

Imamo $\forall \lambda \geq 0$,

$$\hat{y}_\lambda^r = X \hat{\beta}_\lambda^r = Z \cdot \hat{\alpha}_\lambda^r \overset{(3.16)}{=} Z (Z^T Z + n\lambda I_p)^{-1} Z^T y$$

$\operatorname{diag}(d_1^2, \dots, d_p^2)$

$$= U \cdot \operatorname{diag}\left(\frac{d_1^2}{d_1^2 + n\lambda}, \dots, \frac{d_p^2}{d_p^2 + n\lambda}\right) U^T \cdot y$$

$(n \times p)$ $(p \times n)$

$Z = UD$

$=: S_\lambda$

$$\Rightarrow df(\text{ridge}_\lambda) = \operatorname{tr}(S_\lambda) = \operatorname{tr}\left(\underbrace{U^T U}_{=I_p} \operatorname{diag}(\cdot)\right)$$

$$= \boxed{\sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + n\lambda}} \quad \left(= \sum_{j=1}^{p} \frac{\gamma_j}{\gamma_j + \lambda}\right) \qquad (3.25)$$

$\gamma_j = \frac{d_j^2}{n}$

↳ $df(\text{ridge}_\lambda) < p$, za $\lambda > 0$, te

$df(\text{ridge}_\lambda) \searrow 0$, za $\lambda \to +\infty$.

[ideja projekcijom p parametara]

(Dz) Pokažite da je $\operatorname{tr}\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I)^{-2}\right) = \sum_{j=1}^{p} \frac{\gamma_j^2}{(\gamma_j + \lambda)^2}$, $\forall \lambda \geq 0$.

(vidi Prop. 3.7). ∎

(3.3.2) Multikolinearnost

Iz modela (3.5),

$\hat{\Sigma}$ = kovarijacijska matrica od $(X^{(1)}, \dots, X^{(n)}) \in \mathbb{R}^{p \times p}$; $\bar{X}_j = 0, \forall j$

$$\operatorname{Var}\left(\hat{\beta}_j^{ls}\right) = \frac{\sigma^2}{n} \left(\hat{\Sigma}^{-1}\right)_{jj}.$$

$\Rightarrow$ ako je $\hat{\Sigma} = \operatorname{diag}(\gamma_1, \dots, \gamma_p)$, $\operatorname{Var}(\hat{\beta}_j^{ls}) = \frac{\sigma^2}{n} \cdot (\gamma_j)^{-1}$.

$\widehat{\operatorname{Var}}(X_1)$

Općenito, uvjedi [uvjetne varijance]

$$\operatorname{Var}\left(\hat{\beta}_j^{ls}\right) = \frac{\sigma^2}{n} \boxed{\widehat{\operatorname{Var}}\left(X_j - \hat{X}_{j|-j}\right)^{-1}} \qquad (3.26)$$

gdje je $\hat{X}_{j|-j} := \operatorname{Proj}_{\text{span}(X_k ; k \neq j)}(X_j)$.

$\Rightarrow$ ako $X_j \approx \sum_{k \in I} a_k X_k$ za neke $I \subseteq \{1, \dots, p\} \setminus \{j\}$ te $a_k \in \mathbb{R}$,

$\operatorname{Var}(\hat{\beta}_j^{ls})$ će biti veliko. — "problem multikolinearnosti"

S druge strane, ako $p << n$, testna greška neće biti velika (ako je lin. model točan, naravno).

3°)

npr.) Prep. da $X_1 \approx X_2$ (dakle, $\rho(X_1, X_2) \approx 1$) te vrijedi

$$Y = \underbrace{X_1 + X_2}_{\approx c X_1 + (2-c) X_2, \ (\forall c \in \mathbb{R}!)} + \varepsilon \qquad (\text{dakle, } \beta_* = (1,1))$$

Intuitivno

$\Rightarrow$ $RSS((c, 2-c)) \approx RSS((1,1))$, $\forall c \in \mathbb{R}$, a $\operatorname{argmin}_{\beta \in \mathbb{R}^2} RSS(\beta)$ jako ovisi o $\mathbb{X}$ i y!

↳ ridge regresija penalizira više vektore $(c, 2-c)$ za $|c| >> 0$, pa $Var(\hat{\beta}_j^r) \overset{(<<)}{<} Var(\hat{\beta}_j^{ls})$. ◻

→ ridge - kolinearnost.R

Kako prepoznati multikolinearnost

1°) pogledati korelacije među koreacijetima

2°) Variance Inflation Factor:

$$VIF(\beta_j) := \frac{(\hat{\Sigma}^{-1})_{jj}}{s_j^{-1}} \qquad (3.27)$$

$$= \frac{\text{varijanca od } \hat{\beta}_j^{ls}}{\text{—— || —— u slučaju } x_j^T x_k = 0, \forall k \neq j}$$

↳ gledamo je li $VIF(\beta_j) \geq 5$ ili 10. (heuristika)

3°) Ako λ_k = svojstv. vrijedn. od $\hat{\Sigma} \approx 0$, [k-ta po veličini]

$$\Rightarrow 0 \approx \lambda_k = \widehat{Var}(\underset{= \mathbb{X} \cdot v_k}{z_k}) = \widehat{Var}\left(\sum_{j=1}^{p} x_j \cdot V_{jk}\right)$$

$$\Rightarrow \sum_{j=1}^{p} X_j V_{jk} \approx \text{const!}$$

◻