Lineáris regresszió és regularizált változatai

Halász Erik

Témavezető: **Csáji Balázs Csanád**

Matematika BSc

Budapest, 2025

Eötvös Lóránd Tudományegyetem

Tartalomjegyzék

Ábrák jegyzéke

Bevezetés

A statisztikai tanulás napjaink egyik alapvető tudománya, amelyet többek között különféle függvények becslésére használunk. A gépi tanulás egyik leggyakrabban vizsgált területe a *felügyelt tanulás*, ahol az a cél, hogy a rendelkezésre álló bemeneti és kimeneti adatok alapján megtanuljunk egy olyan függvényt, amely jól leírja a bemenet és kimenet közötti kapcsolatot. Az egészségügyben például gyakran szeretnénk előre jelezni bizonyos értékeket – például a vércukorszintet –, vagy megérteni egyes betegségek, például a prosztatarák kialakulásának okait (Hastie, Tibshirani és Friedman 2004). Ezek a problémák jól megfogalmazhatók a felügyelt tanulás keretében, mivel adott bemeneti változók és a hozzájuk tartozó megfigyelt kimenetek alapján keresünk egy olyan modellt, amely a lehető legjobban illeszkedik az adatokhoz.

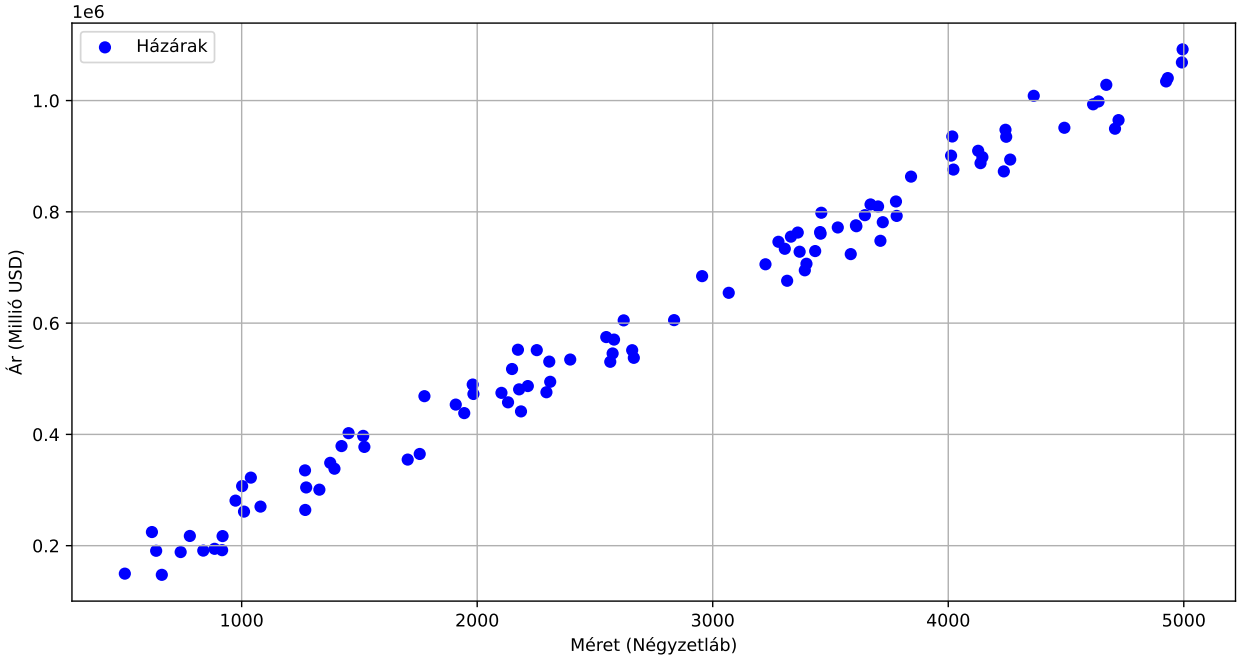
A diplomamunkámban a felügyelt tanulás keretén belül alkalmazott módszerek mélyebb megértésére törekszem, különös tekintettel a függvénybecslésre és a predikciós teljesítmény javítására. Célom olyan megoldások és technikák bemutatása, amelyek segítségével minél jobb általánosító képességű modellt tudunk kialakítani a valós adatok alapján.

1. fejezet

Lineáris regresszió

1.1. Alapvető probléma

Tanulmányaink vagy munkánk során sűrűn előfordulhat velünk, hogy előzetesen mért adataink alapján kell becsülnünk egy új adat értékét. Például, ha szeretnénk eladni az ingatlanunkat, és szeretnénk felmérni, hogy mennyi lehet a reális ára pusztán a mérete alapján.



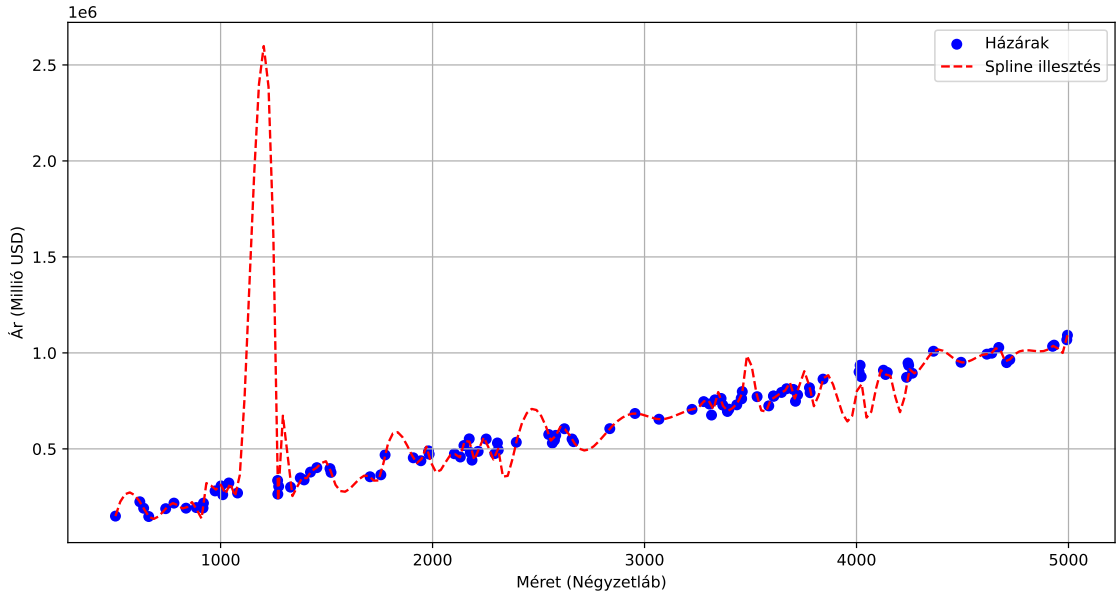
1.1. ábra. Ábra a házárakról a méret függvényében.
Forrás: Kaggle adathalmaz

Az 1000 darabos adathalmazból 100-at választottam ki egyenletesen, majd sorba rendeztem őket. Mint láthatjuk, az adathalmaz azt támasztja alá, amit sejthetünk előzetesen is. Valamifajta lineáris kapcsolat van a házárak és a házméretek között, amit még nem ismerünk.

Természetes ötlet lenne, hogy polinom interpolációt használjunk ahhoz, hogy megtudjuk, mégis mennyibe kerülhet az ingatlanunk. Ehhez használhatunk különböző interpolációs technikákat, például Lagrange-interpolációt. Ennek a módszernek az alapja, hogy olyan polinomot találunk, ami minden ponton áthalad pontosan egyszer, anélkül, hogy a pontokat explicit módon összekapcsolnánk. Magyarán n darab pontra egy $(n-1)$-ed fokú polinomot illesztünk (Faragó és Horváth 2013).

Ezzel az a probléma, hogy bár a modell tökéletes pontossággal megmondja, hogy a már meglévő adataink hogyan helyezkednek el, új adatbevitelkor nem kapunk általában jó választ. Ez azt jelenti, hogy rossz a módszer általánosító képessége, de ez már látszik abból is, hogy ha elveszünk egy pontot a tanuló adatainkból, akkor már egy eggyel kisebb fokú polinomot kapunk.

Ebben az esetben próbálkozzunk meg egy bonyolultabb interpolációs módszerrel. Ami eszünkbe juthat, az a Spline-interpoláció. Ennek a módszernek az alapja, hogy az intervallumunkat kisebb részintervallumokra bontjuk, és az azokban elhelyezkedő pontokra illesztünk kisebb fokú polinomokat.



1.2. ábra. Ábra a kubikusspline interpolációról.

1.2 ábra azt mutatja, hogy a feltételezett közel lineáris kapcsolatot nem látja a modell, egy-egy helyen nagyon kiugrik. Ez többek között azért is van, mert a spline interpolációnál a pontokban a kétoldali deriváltaknak szeretnénk, hogy megegyezzenek„ azért, hogy a függvény folytonosan deriválható legyen.

Bár a módszerünk szofisztikáltabb, sajnos továbbra sem érjük el tökéletesen a kívánt hatást, vagyis, hogy jó legyen a modell általánosító képessége.

1.2. Regresszió

Represszió esetén rendelkezésünkre áll egy (X, Y) véletlen változópár, ahol $X \in \mathbb{R}^d$ és $Y \in \mathbb{R}$. Célunk, hogy feltárjuk a kapcsolatot X és Y között egy olyan $f : \mathbb{R}^d \to \mathbb{R}$ mérhető függvény segítségével, amely $f(X)$ értékével jól közelíti Y-t. A közelítés jóságát egy nemnegatív $\ell(f(X), Y)$ veszteségfüggvénnyel mérjük, amelynek több lehetséges formája létezik; a dolgozatban az L_2-normán alapuló négyzetes veszteségből indulunk ki.

Mivel X és Y valószínűségi változók, ezért a célunk nem csupán a pillanatnyi veszteség minimalizálása, hanem annak várható értékét szeretnénk csökkenteni. Ezt nevezzük *risk*-nek, azaz kockázatnak, amely az L_2-es veszteség esetén a következő formát ölti:

$$R(f) = \mathbb{E}\left[(f(X) - Y)^2\right].$$

A célunk tehát egy olyan f függvény megtalálása, amely minimalizálja ezt a várható négyzetes hibát. (Györfi, Kohler, Krzyzak és Walk 2002).

1.1. Definíció (Négyzetes hiba) *A regresszió várható négyzetes hibájának nevezzük az* $\mathbb{E}[(f(X) - Y)^2]$ *kifejezést.*

Célunk tehát, hogy egy olyan $m^* : \mathbb{R}^d \to \mathbb{R}$ mérhető függvényt találjunk, amire

$$\mathbb{E}[(m^*(X) - Y)^2] = \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}[(f(X) - Y)^2] \tag{1.1}$$

1.2. Definíció (Regressziós függvény) *Az* $m(x) = \mathbb{E}[Y|X = x]$ *függvényt regressziós függvénynek nevezzük.*

1.1. Tétel *A regressziós függvény minimalizálja az L_2 hibát.*

Bizonyítás: Tetszőleges $f : \mathbb{R}^d \to \mathbb{R}$ négyzetesen integrálható függvényre:

$$\begin{aligned}\mathbb{E}[(f(X) - Y)^2] &= \mathbb{E}[(f(X) - m(X) + m(X) - Y)^2] \\ &= \mathbb{E}[(f(X) - m(X)^2] + \mathbb{E}[(m(X) - Y)^2]\end{aligned} \tag{1.2}$$

Itt a toronyszabályt használtuk föl:

$$\begin{aligned}\mathbb{E}[(f(X) - m(X))(m(X) - Y)] &= \mathbb{E}[\mathbb{E}[(f(X) - m(X))(m(X) - Y) \mid X]] \\ &= \mathbb{E}[(f(X) - m(X))\mathbb{E}[m(X) - Y \mid X]] \\ &= \mathbb{E}[(f(X) - m(X))(m(X) - m(X))] \\ &= 0\end{aligned} \tag{1.3}$$

Tehát

$$\mathbb{E}[(f(X) - Y)^2] = \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu(dx) + \mathbb{E}[(m(X) - Y)^2] \tag{1.4}$$

ahol μ az X eloszlása. Ez a kifejezés csak akkor 0, ha $f(x) = m(x)$, különben pozitív. Tehát $m(X)$ optimális becslése Y -nak. □

1.3. Lineáris regresszió

A lineáris regressziós modell felteszi, hogy a regressziós függvény $\mathbb{E}[Y|X]$ a paramétereiben lineáris, vagy könnyen becsülhető úgy. Így, ha a bemeneti vektorunk
$X^{\mathrm{T}} = (X_1, X_2, ..., X_n)$, akkor $f(X, \beta) = \beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_n\beta_n$ a legegyszerűbb lineáris regressziós modell.
Itt a β_i-k ismeretlen együtthatók (amikben a regressziós függvény lineáris), az X_i-k pedig különböző adatok, amiket mérésekkel, illetve azok transzformációival kapunk. Ugyanakkor a linearitás egy szoros megkötése a modellnek. Ennek érdekében kibővíthetjük a függvénycsaládunkat úgy, hogy rögzített nemlineáris függvények lineáris kombinációit vesszük. Az így

kapott modell a következő alakot veszi fel:

$$f(x, \beta) = \beta_0 + \sum_{j=1}^{d} \beta_j \phi_j(x), \tag{1.5}$$

ahol:

- β_0 az affin eltolás (intercept),
- $\{\beta_j\}_{j=1}^{d}$ a modell ismeretlen paraméterei,
- $\{\phi_j(x)\}_{j=1}^{d}$ pedig a rögzített, nemlineáris transzformációk, az úgynevezett bázisfüggvények.

Gyakran kényelmes definiálni egy ún. *álbázisfüggvényt*. Legyen

$$\phi_0(x) = 1. \tag{1.6}$$

Ekkor a függvény az alábbi formában írható fel:

$$f(x, \beta) = \sum_{j=0}^{d} \beta_j \phi_j(x) = \beta^{\mathrm{T}} \phi(x) \tag{1.7}$$

Itt $\beta = (\beta_0, \beta_1, \ldots, \beta_d)^{\mathrm{T}}, \quad \phi(x) = (\phi_0(x), \phi_1(x), \ldots, \phi_d(x))$.

Azáltal, hogy nemlineáris bázisfüggvényeket használunk, el tudjuk érni, hogy a regressziós függvény, az x nemlineáris függvénye legyen. Például ha $\phi_j(x) = x^j$ és csak egyetlen bemeneti vektorunk van x, akkor visszakapjuk az egyszerű polinom regressziót, aminek, mint már említettem, az a limitációja, hogy egyetlen beviteli pont megváltoztatása globálisan megváltoztathatja a függvényt (Bishop 2006).

Szóba jöhetnek más bázisfüggvények is, mint például a Gauss bázisfüggvények, amik a következő alakot veszik fel:

$$\phi_j(x) = \exp\left(\frac{-(x - \mu_j)^2}{2s^2}\right) \tag{1.8}$$

ahol μ_j a várható értékeknek, s pedig a szórásnak felel meg.

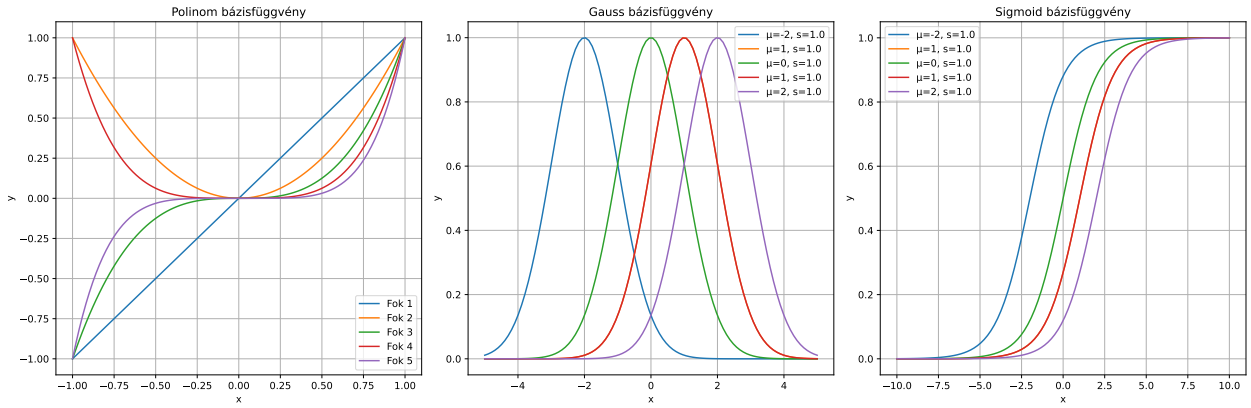
Egy másik gyakori bázisfüggvény a szigmoid bázisfüggvény, aminek az alábbi függvény felel meg:

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right) \tag{1.9}$$

ahol a szigmoid függvény:

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{1.10}$$

Megjegyzés: Ez egy gyakori választás aktivációs függvénynek neurális hálózatokban.



1.3. ábra. Sorban a polinom, Gauss, és sigmoid bázisfüggvények különböző μ-k, és $s = 1$ esetén.

n darab bemeneti vektor esetén jelölje $\Phi = (\phi(x_1), \ldots, \phi(x_n))$ a bemeneti mátrixot, $\beta = (\beta_1, \ldots, \beta_n)$ az ismeretlen együtthatókat, és $y = (y_1, \ldots, y_n)$ a kimeneti vektort. Ekkor átírhatjuk a lineáris regressziót úgy, hogy megpróbáljuk megbecsülni y-t egy lineáris egyenletrendszer segítségével. Tehát:

$$\Phi\beta \approx y \tag{1.11}$$

1.3. Definíció (Reziduális) *Az i. bemeneti vektor reziduálisa az* $\varepsilon_i(\beta) = y_i - \phi_\beta(x_i)$. *Így a reziduálisa a rendszernek az* $\varepsilon(\beta) = y - \Phi\beta$

1.4. Definíció (Legkisebb négyzetek módszere)

$$RSS(\beta) = R(\beta) = ||y - \Phi\beta||_2^2 = (y - \Phi\beta)^{\mathrm{T}}(y - \Phi\beta) \tag{1.12}$$

(RSS: Residual Sum of Squares, vagy más néven OLS: Ordinary Least Squares) (Hastie, Tibshirani és Friedman 2004)

1.1. Megjegyzés A dolgozatban mostantól $||X||$ norma alatt az L_2-es normát értjük.

1.5. Definíció (Átlagos négyzetes hiba)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.13}$$

MSE : Mean squared error

Tegyük fel, hogy Φ teljes rangú ($\mathrm{r}(\Phi) = d$), illetve $n \geq d$. Ekkor célunk tehát találni egy $\hat{\beta}$-t, amire igaz, hogy

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \frac{1}{2}\|y - \Phi\beta\|^2 \tag{1.14}$$

ahol adott $y \in \mathbb{R}^n$ és a $\Phi \in \mathbb{R}^{n \times d}$ mátrix.

1.2. Megjegyzés A feltevéseink miatt egyértelműen létezik $\hat{\beta}$.

1.3. Megjegyzés Az $\frac{1}{2}$-es szorzó könnyítésként van, hogy a gradiensben ne jelenjen meg konstans szorzó.

1.2. Tétel *Ekkor a legkisebb négyzetek módszerének Hesse-mátrixa pozitív definit.*

Bizonyítás: A Hesse-mátrix a következő: $H = \Phi^{\mathrm{T}}\Phi$. $x^{\mathrm{T}}\Phi^{\mathrm{T}}\Phi x = (\Phi x)^{\mathrm{T}}(\Phi x) = ||\Phi x||^2$ Mivel Φ teljes rangú, így minden $x \neq 0$-ra $\Phi x \neq 0$. Tehát $||\Phi x||^2 > 0$ minden $x \neq 0$-ra. Így H pozitív definit. □

1.2.1. Következmény $R(\beta)$ így szigorúan konvex.

Ahhoz, hogy a szélsőérték problémát megoldjuk, kell a legkisebb négyzetek módszerének gradiense β szerint, tehát $\nabla_\beta R(\hat{\beta}) = -\Phi^{\mathrm{T}}(y - \Phi\hat{\beta}) = 0$ -t kell megoldanunk. Ezt kibontva és átrendezve a következő egyenletet kapjuk:

$$\Phi^{\mathrm{T}}y = \Phi^{\mathrm{T}}\Phi\hat{\beta} \tag{1.15}$$

Feltételeink szerint $\hat{\beta}$ egyértelműen létezik (hiszen R szigorúan konvex) és így

$$\hat{\beta} = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}y \tag{1.16}$$

1.3. Tétel (Ortogonális projekció) $\Phi\hat{\beta}$ *a legközelebbi pont* y*-hoz* Φ *oszlopterében.*

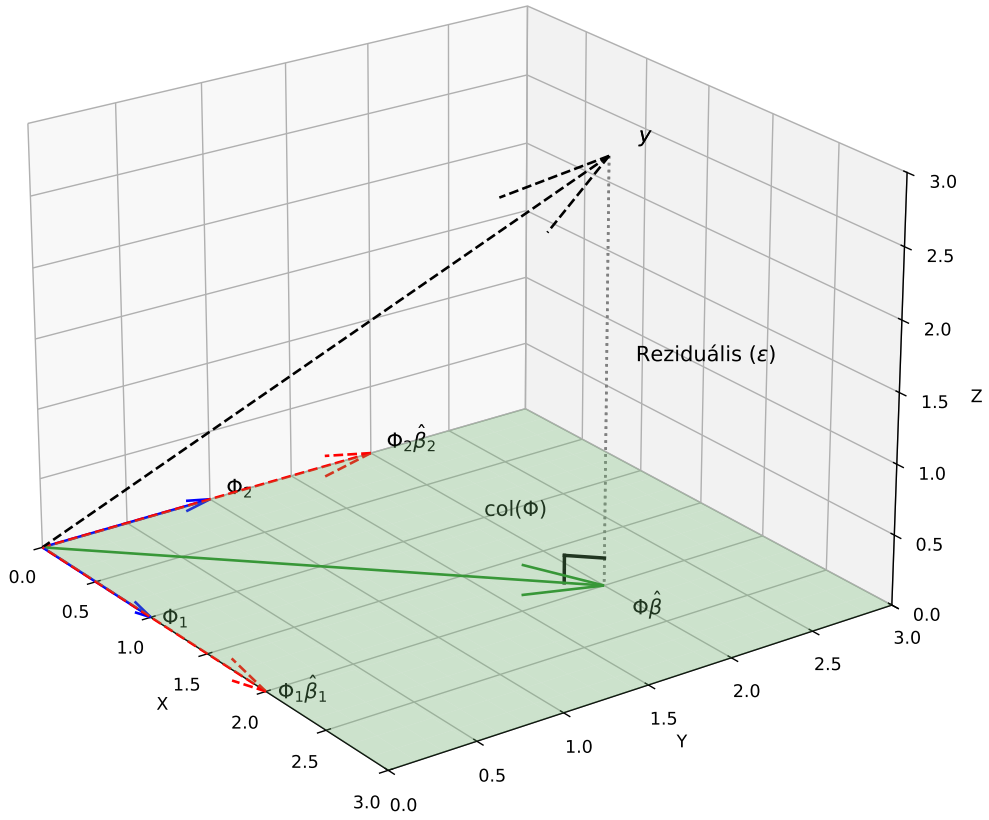
Bizonyítás: Minden $\beta \in \mathbb{R}^d$-re

$$(\Phi\beta)^{\mathrm{T}}\varepsilon(\hat{\beta}) = (\Phi\beta)^{\mathrm{T}}(y - \Phi\hat{\beta}) = \beta^{\mathrm{T}}\Phi^{\mathrm{T}}(y - \Phi\hat{\beta}) = 0 \tag{1.17}$$

Így a reziduális ortogonális Φ oszlopterére. □

1.4. Megjegyzés Az ortogonális projekció mátrixa az 1.16-ből kiszámítható:

$$H = \Phi(\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}} \tag{1.18}$$



1.4. ábra. Ábra arról, hogy a kimeneti vektor y ortogonális projekciója az Φ_1 és Φ_2 által generált altérre az $\Phi\hat{\beta}$.

1.4. Sztochasztikus lineáris regresszió

Célunk tehát, hogy a legjobb becslést kapjuk β^*-ra, adott determinisztikus regressziós mátrix Φ esetén, feltéve, hogy

$$y = \Phi\beta^* + \varepsilon \tag{1.19}$$

ahol

- $\beta^* \in \mathbb{R}^d$ a rögzített helyes paraméter
- Φ determinisztikus és több oszlopa van, mint sora (sovány) és teljes rangú

- ε egy fehérzaj, azaz 0 várható értékű, korrelálatlan és homoszkedasztikus, vagyis $\mathbb{E}\left[\varepsilon\varepsilon^{\mathrm{T}}\right] = \sigma^2 I$, ahol $0 < \sigma^2 < \infty$

(Csáji 2024)

1.4. Tétel *Tegyük fel, hogy az y_i -k korrelálatlanok és konstans σ^2 a szórásnégyzetük, akkor a kovariancia mátrixa a legkisebb négyzetek módszer becslésének megadható a következőképp:*

$$Cov(\hat{\beta}) = (\Phi^{\mathrm{T}}\Phi)^{-1}\sigma^2 \tag{1.20}$$

Bizonyítás: Az 1.16-be $y = \Phi\beta^* + \varepsilon$ -nt behelyettesítve, és felhasználva, hogy $\hat{\beta}$ torzítatlan, mivel

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\left[\beta^* + (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\varepsilon\right] = \beta^* + (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\mathbb{E}[\varepsilon] = \beta^* \tag{1.21}$$

Így

$$\mathrm{Cov}(\hat{\beta}) = \mathbb{E}\left[((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\varepsilon)((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\varepsilon)^{\mathrm{T}}\right]. \tag{1.22}$$

Jelölje $A = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}$, így:

$$\mathrm{Cov}(\hat{\beta}) = \mathbb{E}\left[(A\varepsilon)(A\varepsilon)^{\mathrm{T}}\right] = \mathbb{E}\left[A\varepsilon\varepsilon^{\mathrm{T}}A^{\mathrm{T}}\right] = A\mathbb{E}|\left[\varepsilon\varepsilon^{\mathrm{T}}\right]A^{\mathrm{T}} \tag{1.23}$$

Mivel $\mathbb{E}\left[\varepsilon\varepsilon^{\mathrm{T}}\right] = I\sigma^2$, így

$$\mathrm{Cov}(\hat{\beta}) = A(\sigma^2 I)A^{\mathrm{T}} = \sigma^2 AIA^{\mathrm{T}} = \sigma^2 AA^{\mathrm{T}}. \tag{1.24}$$

Visszahelyettesítve $A = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}$ -t:

$$AA^{\mathrm{T}} = \left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\right)\left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\right)^{\mathrm{T}}. \tag{1.25}$$

Mivel $(\Phi^{\mathrm{T}}\Phi)$ szimmetrikus, így $(\Phi^{\mathrm{T}}\Phi)^{-1}$ is szimmetrikus lesz. Így pedig

$$A^{\mathrm{T}} = \left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\right)^{\mathrm{T}} = \Phi\left((\Phi^{\mathrm{T}}\Phi)^{-1}\right)^{\mathrm{T}} = \Phi(\Phi^{\mathrm{T}}\Phi)^{-1} \tag{1.26}$$

Ezzel továbbszámolva:

$$\begin{aligned} AA^{\mathrm{T}} &= \left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\right)\Phi(\Phi^{\mathrm{T}}\Phi)^{-1} \\ &= (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-1} \\ &= (\Phi^{\mathrm{T}}\Phi)^{-1} \end{aligned} \tag{1.27}$$

Következésképp

$$\mathrm{Cov}(\hat{\beta}) = AA^{\mathrm{T}}\sigma^2 = (\Phi^{\mathrm{T}}\Phi)^{-1}\sigma^2 \tag{1.28}$$

□

Felmerülhet bennünk, hogy miért pont a legkisebb négyzetek módszerével foglalkozunk ennyit. Azon felül, hogy nagyon intuitív, és könnyű vele számolni, van ennek más oka is, ellenben ehhez be kell vezetnünk egy definíciót.

1.6. Definíció (Löwner-részrendezés) *A és B szimmetrikus, pozitív szemidefinit mátrixokra:*

$$A \succeq B \Leftrightarrow A - B \text{ pozitív szemidefinit.}$$

1.5. Megjegyzés Intuitívan ez annyit jelent, hogy A legalább akkora, mint a B mátrix pozitív szemidefinitás szempontjából. Ez részrendezés azért, mert nem minden mátrix hasonlítható így össze.

Most, hogy ezt definiáltuk, kimondhatjuk azt a tételt, ami miatt kulcsfontosságú a legkisebb négyzetek módszere a statisztikában.

1.5. Tétel (Gauss-Markov tétel) *A sztochasztikus lineáris regresszió feltételei mellett, $\forall \beta^* \in \mathbb{R}^d$: $\forall \tilde{\beta}$ lineáris torzítatlan becslésre $Cov(\tilde{\beta}) \succeq Cov(\hat{\beta})$, vagyis a legkisebb négyzetek módszer a legjobb lineáris torzítatlan becslő.*

Bizonyítás: Legyen $\tilde{\beta}$ egy tetszőleges lineáris torzítatlan becslés. Mivel lineáris, ezért felírható $\tilde{\beta} = \hat{\beta} + My$ alakban.
Mivel $\tilde{\beta}$ torzítatlan, ezért $\mathbb{E}[\tilde{\beta}] = \beta^*$ $\forall \beta^*$-ra

$$\begin{aligned} &\mathbb{E}[\tilde{\beta}] = \mathbb{E}[\hat{\beta} + My] = \mathbb{E}[\hat{\beta}] + \mathbb{E}[My] = \beta^* + \mathbb{E}[M(\Phi\beta^* + \varepsilon)] \\ &= \beta^* + M\Phi\beta^* + M\mathbb{E}[\varepsilon] = \beta^* + M\Phi\beta^* \implies M\Phi = 0. \end{aligned} \tag{1.29}$$

Ezt használva $\tilde{\beta}$ kovarianciája pedig felírható a következőképp:

$$\begin{aligned}\mathrm{Cov}(\tilde{\beta}) &= \mathrm{Cov}(\hat{\beta} + My) = \mathrm{Cov}\left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\varepsilon + M\Phi\beta^{*} + M\varepsilon\right) \\ &= \mathrm{Cov}\left(\left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}} + M\right)\varepsilon\right) \\ &= \left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}} + M\right)\mathrm{Cov}(\varepsilon)\left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}} + M\right)^{\mathrm{T}} \\ &= \sigma^2\left((\Phi^{\mathrm{T}}\Phi)^{-1} + MM^{\mathrm{T}}\right) = \mathrm{Cov}(\hat{\beta}) + \sigma^2 MM^{\mathrm{T}} \succeq \mathrm{Cov}(\hat{\beta})\end{aligned} \tag{1.30}$$

□

1.5.1. Következmény Ha feltesszük, hogy a regressziós függvény lineáris, akkor a legjobb, amit tehetünk a paraméterek torzítatlan becslésére az az, hogy a négyzetes hibát vesszük, és aszerint optimalizálunk.

1.5. Modellilleszkedés mérése

1.7. Definíció *Jelölje* $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$. *Ekkor*

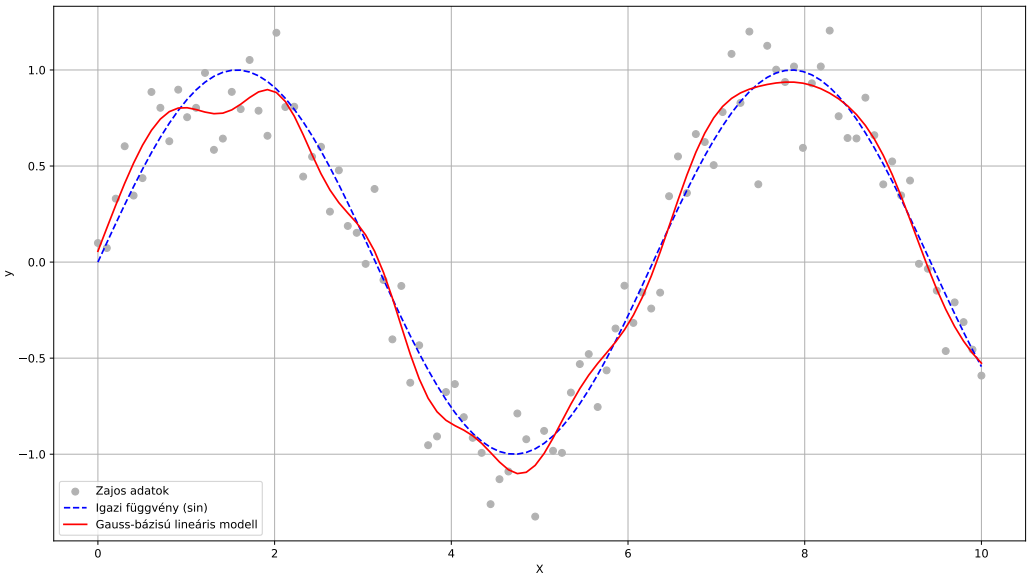
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{1.31}$$

ahol RSS az átlagos négyzetes hiba függvénye (1.4), egy statisztikai mérőszám, amellyel a regressziós függvény illeszkedését mérjük (James, Witten, Hastie és Tibshirani 2013).

1.6. Megjegyzés R^2 értéke 0 és 1 között mozog. Azt mutatja, hogy a modell az adatok varianciájának mekkora részét magyarázza meg. Ha 0, az azt jelenti, hogy a modell teljesen pontatlan, ha 1, akkor a minden adatpont tökéletesen illeszkedik a regressziós függvényre.

1.6. Példa

Az eddigiek alapján szinuszos adathalmaz esetén a kapott függvényünk a következő:



1.5. ábra. Gauss bázisú lineáris regresszió szintetikus adatokra

A 1.5 ábrán Gauss bázisú lineáris regressziót használtam szintetikus adatokon, amelyeket a szinusz függvény + 0,2 szórású normális zajjal generáltam. A gauss bázisfüggvények szórásának 0,5-öt választottam, a várható értékeket pedig egyenletesen az adatok maximuma és minimuma között 20 helyen. A kapott átlagos négyzetes hiba 5-szörös keresztvalidáció után 0.04521 volt, az R^2 pedig 0.94383, ami majdnem tökéletes illeszkedést mutat.

2. fejezet

Torzítatlanság vs Variancia

A gyakorlatban egy modell általánosító képességét úgy mérjük, hogy a bemeneti adatokat tanító- és tesztadathalmazokra bontjuk. A tanítóadathalmazt felhasználva kapjuk meg $\hat{f}$ -ot, majd ennek a függvénynek az átlagos négyzetes hibáját számoljuk ki a tesztadatokon.

2.1. Tétel *Jelölje X a tesztadatok halmazát. Tegyük fel, hogy létezik egy f függvény, hogy minden $(x_0, y_0) \in X$-re $y_0 = f(x_0) + \varepsilon$, ahol ε fehérzaj. Legyen $\hat{f}$, f függvény adott modell szerinti becslése. Ekkor igaz, hogy*

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \left(\mathbb{E}\left[\hat{f}(x_0)\right] - f(x_0)\right)^2 + Var\left(\hat{f}(x_0)\right) + \sigma^2. \tag{2.1}$$

Bizonyítás:

$$\begin{aligned}\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] &= \mathbb{E}\left[\left(f(x_0) + \varepsilon - \hat{f}(x_0)\right)^2\right] \\ &= \mathbb{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] + 2\,\mathbb{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)\varepsilon\right] + \underbrace{\mathbb{E}\left[\varepsilon^2\right]}_{\sigma^2}\end{aligned} \tag{2.2}$$

Mivel ε független 0 várható értékű zaj, így

$$2\,\mathbb{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)\varepsilon\right] = 2\,\mathbb{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)\right]\mathbb{E}[\varepsilon] = 0 \tag{2.3}$$

Végül marad nekünk $\mathbb{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]$, amit még nem ismerünk.

$$\begin{aligned}\mathbb{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] &= \mathbb{E}\left[\left(\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)] + \mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)\right)\right)^2\right] \\ &= \left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)\right)^2\right] \\ &\quad + 2\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\mathbb{E}\left[\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)\right].\end{aligned} \tag{2.4}$$

A második tag tehát

$$\mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)\right)^2\right] = \mathrm{Var}\left(\hat{f}(x_0)\right) \tag{2.5}$$

Az utolsó tag pedig:

$$2\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\underbrace{\mathbb{E}\left[\mathbb{E}[\hat{f}(x_0)] - f(x_0)\right]}_{=\mathbb{E}\left[\hat{f}(x_0)\right]-\mathbb{E}\left[\hat{f}(x_0)\right]=0} = 0 \tag{2.6}$$

És mivel

$$\left(f(x_0) - \mathbb{E}\left[\hat{f}(x_0)\right]\right)^2 = \left(\mathbb{E}\left[\hat{f}(x_0)\right] - f(x_0)\right)^2 \tag{2.7}$$

ami a négyzetes torzítottságot jelenti, így:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \left(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\right)^2 + \mathrm{Var}\left(\hat{f}(x_0)\right) + \sigma^2 \tag{2.8}$$

□

A (2.1) egyenlet tehát megmutatja, hogy ha minimalizálni akarjuk a várható négyzetes hibát a tesztadathalmazon, akkor egy olyan modellt kell választanunk, amire a variancia és a négyzetes torzítás minimális.

2.1. Megjegyzés Tudjuk, hogy a variancia és a négyzetes torzítás mindketten nemnegatív mennyiségek, következésképpen a várható négyzetes hiba sosem lehet kisebb $\mathrm{Var}(\varepsilon) = \sigma^2$ -nél.

A Gauss-Markov tétel szerint a legkisebb négyzetek módszer a legjobb lineáris torzítatlan becslő. De előfordul, hogy a jó átlag mellett meglehetősen nagy a varianciánk. Az a feltétele-

zés, hogy elegendő csupán a jó várható értékre törekednünk, miközben a varianciát figyelmen kívül hagyjuk, lényegében hibás. Vegyünk egy szemléletes példát:

2.1. Példa Tekintsünk egy mérleget, ami 0.5 valószínűséggel 80 kg-ot ír ki, és 0.5 valószínűséggel 160 kg-ot. A becslés várható értéke $0.5 * 80 + 0.5 * 160 = 120$ kg. Viszont a szórásnégyzete:

$$Var(\hat{\theta}) = \mathbb{E}(\hat{\theta}^2) - \mathbb{E}^2(\hat{\theta}) = (0.5 * 80^2 + 0.5 * 160^2) - 120^2 = 1600 \tag{2.9}$$

Tehát a szórás $\sqrt{1600} = 40$ lesz.

Bár az átlag megegyezik a várható értékkel, a nagy szórás miatt a mérleg mérései rendkívül ingadozóak, és kevés információt adnak a valós értékről.

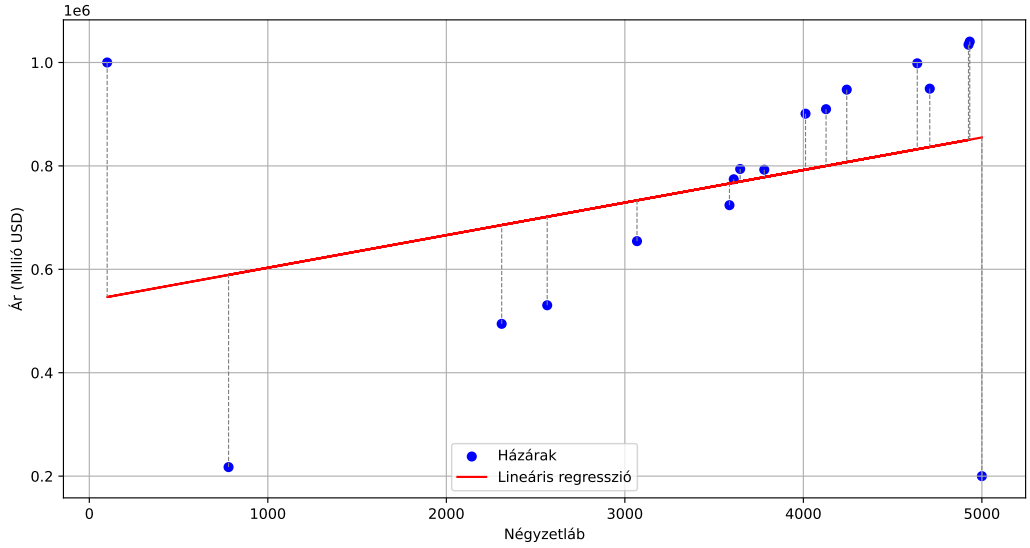
De mit is jelent pontosan egy modell varianciája és torzítottsága? Egy statisztikai tanulási módszer varianciája azt fejezi ki, hogy mennyire változna $\hat{f}$, ha más tanulóadathalmazból becsülnénk meg. Ideálisan, f becslésének nem kellene függnie nagyban a tanulóadathalmaztól, de ha a módszerünknek nagy a varianciája, akkor kis változtatások a tanulóhalmazon is nagy változással járhatnak $\hat{f}$-ban. Így, ha a modellünk nagy hibával teljesít a tesztadathalmazon, akkor azt mondjuk, hogy magas a varianciája.

A modell torzítottsága alatt pedig azt értjük, hogy mekkora a hiba, ha esetleg egy szimplább módszert alkalmazunk egy bonyolultabb problémára. Ezt a jelenséget nevezzük alultanulásnak. Például, ha feltesszük egy adathalmazról, hogy lineáris a kapcsolat az adatok között, akkor az egy nagyon erős megkötést jelent.

Ezzel szemben, ha a modellünk túl összetett az adathalmazunkhoz képest, előfordulhat, hogy tanításkor a tanítóadatokon nagyon magas pontosságot érünk el. Ez úgy fordulhat elő, ha a modellünk rátanul az adatainkban levő zajra. Például modelltanításkor kaphatunk olyan eredményt, hogy bár a tanulóadathalmazunkon 100%-os pontosságot értünk el, a tesztadathalmazon csak 70%-ot. Az a 30%-os különbség arra utalhat, hogy a modellünk rátanult a tanítóadathalmaz zajára. Ezt a jelenséget túltanulásnak nevezzük. Általánosságban, ha a modellünk a tanítóadathalmazon meglehetősen jobban teljesít, mint a tesztadathalmazon, az azt jelenti, hogy a modell veszített az általánosító képességéből, azaz túltanult.

2.2. Megjegyzés Ezt a jelenséget megfigyelhetjük a 1.5 példa esetén.

2.2. Példa A 1.1-es házáras ábrán láttuk, hogy közel lineáris kapcsolat lehet az ingatlanok mérete és ára között. Tegyük fel, hogy létezik még két lakás, egy, ami kicsi, ellenben nagyon drága, és egy másik, ami nagy, de nagyon olcsó. Ha ezekre a házáras adatokra számoljuk ki a lineáris regresszió által adott egyenest, akkor azt várhatjuk, hogy a két kiugró adat miatt kisebb meredekségű lesz.

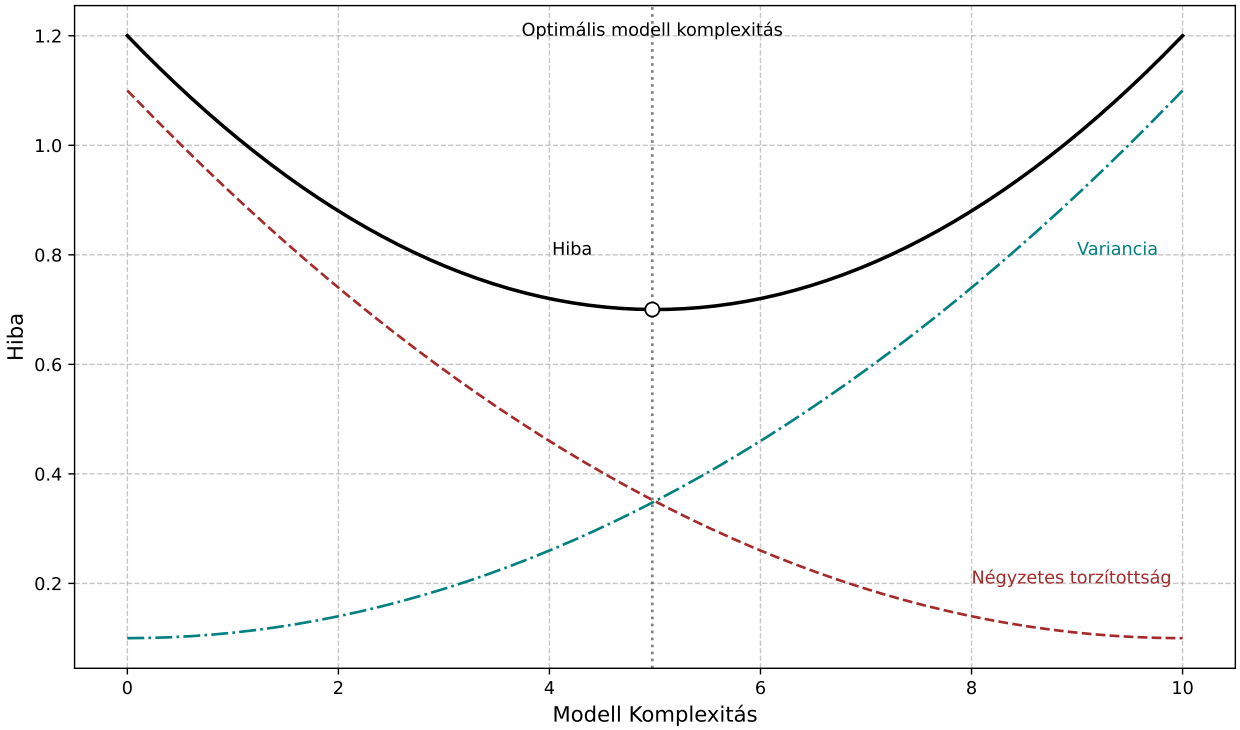


2.1. ábra. Lineáris regresszió kiugró adatokkal

A 2.1 ábrán az egyszerű lineáris regressziós egyenest illesztettem 17 darab pontra, amik közül 15-öt egyenletesen választottam ki a házáras adataink közül, és 2 kiugró pontot manuálisan adtam meg. Az egyenes egyenlete az $y = 2.99x + 540071.17$, az $R^2 = 0.1120$-tel, $\sigma = 242683.7126$ szórással, illetve $MSE = 58895384384.8405$ átlagos négyzetes hibával. Láthatjuk, hogy az egyenes nem illeszkedik jól az adatokhoz, mint ahogy azt első látásra elvárnánk a két kiugró pont miatt. Ha a két pont még kiugróbb lenne, az egyenes akár negatív meredekségűvé is válhatna, ami ugyan megőrizné a jó átlagot, de még nagyobb szórással járna és teljesen félrevezető képet adna az adatok közötti kapcsolat természetéről. Modellválasztáskor tehát fontos, hogy ne legyen érzékeny a kiugró adatokra.

Az alábbi képen láthatjuk tehát, hogy hogyan alakul a torzítottság és a variancia, mo-

dellkomplexitástól függően.



2.2. ábra. Torzítottság, variancia és hiba a modell komplexitás függvényében.

Az 2.2 ábra azt mutatja, hogy ha a modellünk összetettsége túl alacsony, akkor magas torzítottságra és alacsony varianciára számíthatunk, ami alultanuláshoz vezethet, tehát a tanító- és tesztadathalmazon is alacsony pontosságra, azaz nagy hibára számíthatunk. Ugyanakkor, ha túl nagy a modell komplexitása, akkor $\hat{f}$ túlságosan szépen simul rá a tanítóadathalmazbeli pontokra, ugyanakkor a tesztadathalmazon nagy varianciával jár, ami túltanuláshoz vezet. Tehát az optimális modellt a minimális összhibánál találjuk, a megfelelő torzítottsággal és varianciával.

Egyes adathalmazok esetén célunk tehát egy olyan modell kiválasztása statisztikai tanuláskor, ami által adott $\hat{f}$ nem túlságosan torzított, viszont nem is olyan nagy a varianciája.

3. fejezet

Regularizáció

Az előző fejezetben kitárgyaltuk a modellhibát jellemző két kulcstényezőt - a torzítottságot és a varianciát. Láttuk, hogy a modell komplexitásának csökkentésével alacsonyabb lesz a variancia, ami alultanuláshoz vezet, ugyanakkor a magas modell összetettség következtében nő a variancia, ami pedig túltanulást eredményezhet.

Túltanulás kezelésére sok módszer létezik, például feltételezhetjük, hogy csak a legfontosabb jellemzői érdekelnek minket az adatnak, és minden mást kinullázhatunk, így csökkentve a modell komplexitását. Ezt hívják változókiválasztásnak (angolul: subset selection). Ellenben, mivel ez egy diszkrét folyamat, (változókat vagy kiválasztjuk, vagy kinullázzuk) gyakran magas varianciával számolhatunk, így magas marad a hiba a tesztadatokon.

Egy alternatív megközelítés, ha nem nullázzuk ki teljesen a kevésbé fontos változókat, hanem büntetjük a modellben szereplő paraméterek nagyságát, így ösztönözve a kisebb értékekre. Ezt nevezzük zsugorító módszereknek (shrinkage methods), amelyek a modell súlyait „összehúzzák", és ezáltal szabályozzák a modell összetettségét. Ez finomabb módon szabályozza a modell komplexitását, és gyakran jobb általánosító képességet eredményez.

A regularizáció a túltanulás csökkentésén, és az általánosítás növelésén kívül több problémát is segít megoldani. Például, a legkisebb négyzetek módszerénél a megoldás nem egyértelmű, ha $n < d$, hiszen egy pontra több egyenest is tudunk illeszteni. A regularizáció segítségével ilyen esetekben is egyértelmű megoldás lehet. Egy másik előnye a regularizációnak, hogy rosszul kondicionált feladatokat jól kondícionálttá teheti. Tehát, ha a bemeneti mátrixunk kondíciószáma nagyon nagy ($||\Phi|| \cdot ||\Phi^{-1}|| >> 1$), akkor a módosított mátrix már

kondícionáltabb lehet.

Ebben a fejezetben a három fő zsugorítómódszerrel, a Ridge regresszióval, a LASSO-val és az elasztikus hálóval foglalkozunk.

3.1. Ridge regresszió

A ridge regresszió összezsugorítja a modell együtthatóit úgy, hogy az L_2 -es normájukkal együtt minimalizál.

3.1. Definíció *Adott $\lambda \geq 0$ esetén a ridge regresszió függvénye:*

$$\nu(\beta) = ||y - \Phi\beta||_2^2 + \lambda||\beta||_2^2 = R(\beta) + \lambda||\beta||_2^2 \tag{3.1}$$

Egy ezzel ekvivalens definíció:

$$\nu(\beta) = ||y - \Phi\beta||_2^2 \ \textit{ahol} \ ||\beta||^2 \leq t \tag{3.2}$$

3.1. Megjegyzés Egyértelmű megfeleltetés létezik λ és t között.

A $\lambda \geq 0$ a modell paramétere, mely kontrollálja a zsugorítás mértékét. Egyszerű látni, hogy $\lambda = 0$ esetén a négyzetes hibát kapjuk vissza. A ridge regresszió λ-jának növelésével az együtthatók közelednek a nullához (és egymáshoz).

3.2. Megjegyzés Az L_2 regularizáció nemcsak a lineáris regresszió esetén hasznos. Széles körben alkalmazzák például neurális hálók tanításánál, ahol L_2-es büntetés néven vagy *weight decay*-ként szerepel, és a túltanulás elleni egyik legismertebb technika.

3.1. Tétel *A ridge regresszió függvényét, $\nu(\beta)$, négyzetes hiba alakba írható át.*

Bizonyítás: Legyen

$$\tilde{\Phi} = \begin{bmatrix} \Phi \\ \sqrt{\lambda}I \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} y \\ \mathbf{0} \end{bmatrix} \tag{3.3}$$

Ekkor a ridge regresszió célfüggvénye az alábbi módon írható fel:

$$\nu(\beta) = \|y - \Phi\beta\|^2 + \lambda\|\beta\|^2 \tag{3.4}$$

átírható a következő alakra:

$$\nu(\beta) = \left\|\tilde{y} - \tilde{\Phi}\beta\right\|^2. \tag{3.5}$$

Ugyanis,

$$\tilde{y} - \tilde{\Phi}\beta = \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} \Phi \\ \sqrt{\lambda}I \end{bmatrix} \beta = \begin{bmatrix} y - \Phi\beta \\ -\sqrt{\lambda}I\beta \end{bmatrix} \tag{3.6}$$

és ennek a normanégyzete a következő:

$$\left\|\tilde{y} - \tilde{\Phi}\beta\right\|^2 = \left\| \begin{bmatrix} y - \Phi\beta \\ -\sqrt{\lambda}I\beta \end{bmatrix} \right\|^2 = \|y - \Phi\beta\|^2 + \left\|\sqrt{\lambda}\beta\right\|^2 = \|y - \Phi\beta\|^2 + \lambda\|\beta\|^2 \tag{3.7}$$

□

Célunk tehát megtalálni $\hat{\beta}$ -ot, amire igaz, hogy

$$\hat{\beta}_\lambda = \underset{\beta\in\mathbb{R}^d}{\arg\min} \frac{1}{2}\|y - \Phi\beta\|^2 + \frac{\lambda}{2}||\beta||^2 \tag{3.8}$$

Ha ezt a kifejezést β szerint deriváljuk, és nullával egyenlővé tesszük:

$$-\Phi(y - \Phi\beta) + \lambda\beta = 0 \tag{3.9}$$

Majd ezt kibontva, és β -ra rendezve megkapjuk az optimális $\hat{\beta}_\lambda$-t:

$$\hat{\beta}_\lambda = (\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\mathbf{y} \tag{3.10}$$

3.3. Megjegyzés Ha $\lambda > 0$, akkor $(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}$ mindig létezik, mivel $\Phi^{\mathrm{T}}\Phi + \lambda I$ sajátértékei a $\Phi^{\mathrm{T}}\Phi$ pozitív szemidefinit mátrix sajátértékeinek és λ-nak az összegei, ezek pedig minden esetben pozitívak. (Mohri, Rostamizadeh és Talwalkar 2018)

3.1.1. Tikhonov-regularizáció

Sok esetben szeretnénk beépíteni egyfajta relevanciát a paramétereinkbe. Például a házáras példánál lehet, hogy nem annyira fontosak már a több, mint 10 éves adatok hiszen az infláció és az általános ingatlanár-változás jelentősen változtatja ennyi idő alatt az árakat. Egy másik példa, hogy egy munkavégzésénél szeretnénk a modellnek megadni valahogyan, hogy a szerszámaink mennyire régiek, hiszen ez valamelyest befolyásolja az elvégzés időtartamát. Ekkor beszélhetünk a Tikhonov-regularizációról, amelynek a célfüggvénye a következő:

$$\nu(\beta) = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\|y - \Phi\beta\|^2 + ||\Gamma\beta||^2 \tag{3.11}$$

A Γ az úgynevezett *Tikhonov-mátrix*, ami általában szigorúan pozitív definit és sokszor λI-nek választják. Észrevehető, hogy ha $\Gamma = \sqrt{\lambda} I$, akkor a ridge regressziót (3.1) kapjuk vissza. Ugyanakkor ha $0 < \lambda < 1$, akkor Γ-t választhatjuk $diag(\lambda^d, \lambda^{d-1} \ldots, \lambda^1, 1)$-nek, és ebben az esetben valóban egyfajta változórelevanciát adunk meg.

3.4. Megjegyzés Tikhonov-regularizációnál nem írunk ki külön λ-t, mivel az már implicit módon szerepel a Γ mátrixban.

A 3.11 egyenletet átírhatjuk a 3.1 szerint négyzetes hiba alakba, vagyis

$$\|y - \Phi\beta\|_2^2 + \|\Gamma\beta\|_2^2 = \left\| \begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} \Phi \\ \Gamma \end{pmatrix} \beta \right\|_2^2 \tag{3.12}$$

Ha erre alkalmazzuk a megoldóképletet, akkor az optimális $\hat{\beta}$ kifejezhető a következőképpen:

$$\hat{\beta} = (\Phi^{\mathrm{T}}\Phi + \Gamma^{\mathrm{T}}\Gamma)^{-1}\Phi^{\mathrm{T}} y \tag{3.13}$$

3.1.2. Ridge regresszió torzítottsága és varianciája

Torzítottság

A (3.10)-ból kiindulva felírhatjuk $\mathbb{E}\left[\hat{\beta}_\lambda\right]$-ot (Sun 2022).

$$\mathbb{E}\left[\hat{\beta}_\lambda\right] = \mathbb{E}\left[(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\mathbf{y}\right] = (\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\mathbb{E}\left[y\right], \tag{3.14}$$

ahol felhasználtuk, hogy Φ rögzített. Mivel tudjuk, hogy $y = \Phi\beta^* + \varepsilon$ és $\mathbb{E}[\varepsilon] = 0$ megkapjuk azt, hogy:

$$\begin{aligned}\mathbb{E}[\hat{\beta}_\lambda] &= \left(\Phi^{\mathrm{T}}\Phi + \lambda I\right)^{-1}\Phi^{\mathrm{T}}\Phi\beta^* = \left(\Phi^{\mathrm{T}}\Phi + \lambda I\right)^{-1}\left(\Phi^{\mathrm{T}}\Phi + \lambda I - \lambda I\right)\beta^* \\ &= \left(\Phi^{\mathrm{T}}\Phi + \lambda I\right)^{-1}\left(\Phi^{\mathrm{T}}\Phi + \lambda I\right)\beta^* - \lambda\left(\Phi^{\mathrm{T}}\Phi + \lambda I\right)^{-1}\beta^* \\ &= \beta^* - \lambda\left(\Phi^{\mathrm{T}}\Phi + \lambda I\right)^{-1}\beta^*.\end{aligned} \tag{3.15}$$

Jelölje $T(\lambda)$ a torzítottságot λ függvényében. Ekkor ez felírható úgy, hogy

$$T(\lambda) := \sum_{i=1}^{n}\left(\mathbb{E}[\hat{\beta}]^{\mathrm{T}}\phi(x_i) - (\beta^*)^{\mathrm{T}}\phi(x_i)\right)^2 = \sum_{i=1}^{n}\left((\mathbb{E}[\hat{\beta}] - \beta^*)^{\mathrm{T}}\phi(x_i)\right)^2 \tag{3.16}$$

Ezt behelyettesítve, majd kibontva a már megkapott $\mathbb{E}[\hat{\beta}_\lambda]$-t:

$$\begin{aligned}T(\lambda) &= \sum_{i=1}^{n}\lambda^2\left((\beta^*)^{\mathrm{T}}\left(\Phi^{\mathrm{T}}\Phi + \lambda I\right)^{-1}\phi(x_i)\right)^2 \\ &= \lambda^2\sum_{i=1}^{n}(\beta^*)^{\mathrm{T}}(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\phi(x_i)^{\mathrm{T}}\phi(x_i)(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}(\beta^*) \\ &= \lambda^2(\beta^*)^{\mathrm{T}}(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\sum_{i=1}^{n}\left[\phi(x_i)^{\mathrm{T}}\phi(x_i)\right](\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}(\beta^*) \\ &= \lambda^2(\beta^*)^{\mathrm{T}}(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}(\beta^*)\end{aligned} \tag{3.17}$$

$\Phi^{\mathrm{T}}\Phi$ sajátértékfelbontását felírva $\Phi^{\mathrm{T}}\Phi = U\Sigma U^{\mathrm{T}}$, ahol $\Sigma = diag(\sigma_1, \ldots, \sigma_d)$ úgy, hogy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, és U ortonormális mátrix.

Mivel $\Phi^{\mathrm{T}}\Phi + \lambda I$ sajátvektorjai az U oszlopai és a sajátértékei a $\sigma_i + \lambda$ minden $i \in [1, \ldots, d]$-re, így $\Phi^{\mathrm{T}}\Phi + \lambda I = U(\Sigma + \lambda I)U^{\mathrm{T}}$. Ezt a sajátértékfelbontást felhasználva, felírhatjuk a torzítottságot sajátértékek segítségével:

$$T(\lambda) = \lambda^2(\beta^*)^{\mathrm{T}}U(\Sigma + \lambda I)^{-1}U^{\mathrm{T}}U\Sigma U^{\mathrm{T}}U(\Sigma + \lambda I)^{-1}U^{\mathrm{T}}\beta^* \tag{3.18}$$

Felhasználva, hogy U ortonormált, azaz $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I$

$$= \lambda^2(\beta^*)^{\mathrm{T}}U(\Sigma + \lambda I)^{-1}\Sigma(\Sigma + \lambda I)^{-1}U^{\mathrm{T}}\beta^* \tag{3.19}$$

$$= \lambda^2 (\beta^*)^{\mathrm{T}} U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1+\lambda)^2} & 0 & \cdots & 0 \\ 0 & \frac{\sigma_2}{(\sigma_2+\lambda)^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sigma_d}{(\sigma_d+\lambda)^2} \end{bmatrix} U^{\mathrm{T}} \beta^* \tag{3.20}$$

$$= (\beta^*)^{\mathrm{T}} U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1/\lambda+1)^2} & 0 & \cdots & 0 \\ 0 & \frac{\sigma_2}{(\sigma_2/\lambda+1)^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sigma_d}{(\sigma_d/\lambda+1)^2} \end{bmatrix} U^{\mathrm{T}} \beta^* \tag{3.21}$$

3.1.1. Következmény Ha $\lambda \to 0$, akkor a diagonális mátrix elemei $\frac{\sigma_i}{(\sigma_i/\lambda+1)^2} \to 0$. Tehát a torzítottság is tartani fog a nullához. Egyértelműen, ha $\lambda = 0$, akkor torzítatlan.
Ha $\lambda \to +\infty$, akkor $\frac{\sigma_i}{(\sigma_i/\lambda+1)^2} \to \sigma_i$. Vagyis

$$\lim_{\lambda\to\infty} T(\lambda) = (\beta^*)^{\mathrm{T}} U \Sigma U^{\mathrm{T}} \beta^* = (\beta^*)^{\mathrm{T}} \Phi^{\mathrm{T}} \Phi \beta^* = \sum_{i=1}^{n} (\phi(x_i)^{\mathrm{T}} \beta^*)^2$$

Mivel, ha $\lambda \to +\infty$, akkor a ridge regresszióban $\hat{\beta} \to 0$, vagyis mindig nullát fogunk prediktálni, amiből következik, hogy $\mathbb{E}[\hat{\beta}] \to 0$, vagyis ilyenkor erősen torzított a becslésünk.

3.5. Megjegyzés Észrevehetjük, hogy a torzítottság, $T(\lambda)$, monoton nő, ahogy λ nő.

Variancia

A (3.14)-et felhasználva írjuk fel a varianciát a következőként. Jelölje $V(\lambda)$ a varianciát λ függvényében. Ekkor:

$$V(\lambda) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \left(\mathbb{E}[\hat{\beta}] - \hat{\beta}\right)^{\mathrm{T}} \phi(x_i)\right)^2\right] = \mathbb{E}\left[\left(\mathbb{E}[\hat{\beta}] - \hat{\beta}\right)^{\mathrm{T}} \Phi\Phi^{\mathrm{T}} \left(\mathbb{E}[\hat{\beta}] - \hat{\beta}\right)^{\mathrm{T}}\right] \tag{3.22}$$

Jelölje $tr(A)$ az A mátrix nyomát. Általánosságban $\mathbb{E}[\varepsilon A\varepsilon^{\mathrm{T}}] = \mathbb{E}[\mathrm{tr}(A\varepsilon\varepsilon^{\mathrm{T}})]$. Mivel már megkaptuk a formulát $\hat{\beta}$ -ra (3.10) és $\mathbb{E}[\hat{\beta}]$-ra (3.14), így felírhatjuk a két elem különbségét:

$$\begin{aligned}\mathbb{E}[\hat{\beta}] - \hat{\beta} &= (\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi\beta^* - (\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}(\Phi\beta^* + \varepsilon) \\ &= (\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi\beta^* - (\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi\beta^* - (\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\varepsilon \\ &= -(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\varepsilon\end{aligned} \tag{3.23}$$

$$\begin{aligned}V(\lambda) &= \mathbb{E}\left[\varepsilon^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\varepsilon\right] \\ &= \mathbb{E}\,\mathrm{tr}\left(\varepsilon^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\varepsilon\right) \\ &= \mathbb{E}\,\mathrm{tr}\left(\varepsilon\varepsilon^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\right) \\ &= \mathrm{tr}\left(\mathbb{E}[\varepsilon\varepsilon^{\mathrm{T}}]\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\right) \\ &= \mathrm{tr}\left(\sigma^2 I \cdot \Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\right) \\ &= \sigma^2\,\mathrm{tr}\left(\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\right) \\ &= \sigma^2\,\mathrm{tr}\left(\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\right),\end{aligned} \tag{3.24}$$

ahol felhasználtuk, hogy $tr(AB) = tr(BA)$ és $\mathbb{E}\left[\varepsilon\varepsilon^{\mathrm{T}}\right] = \sigma^2 I$.

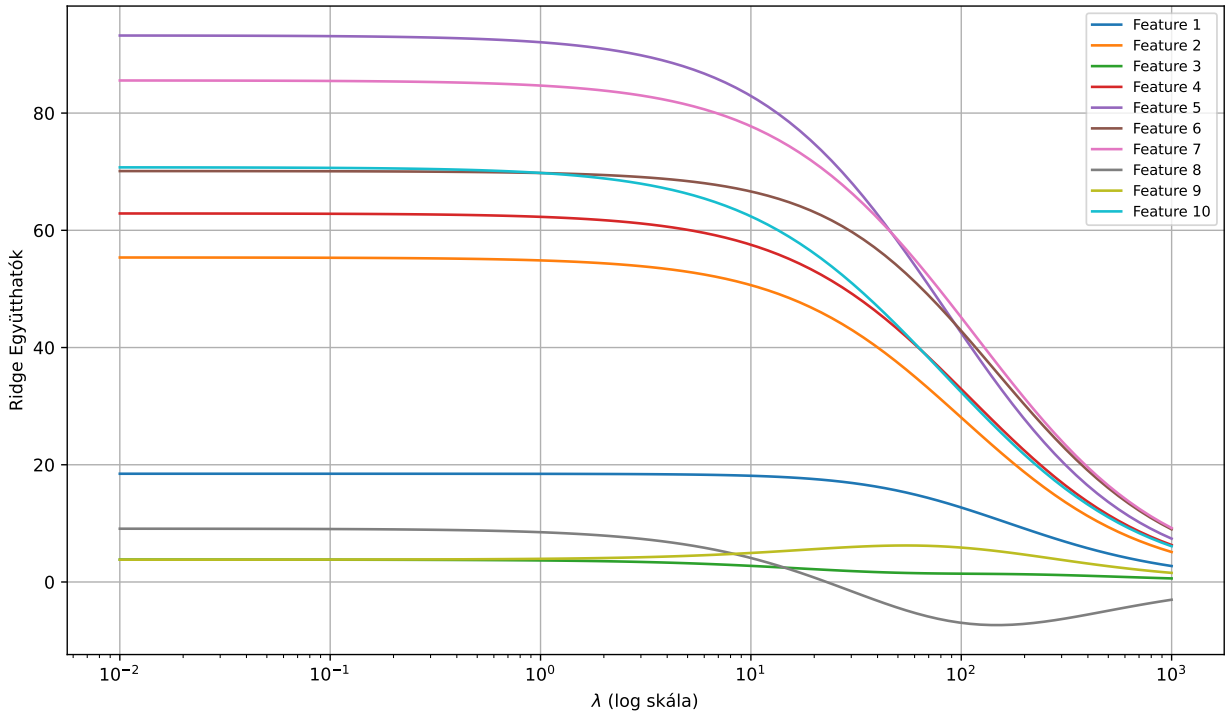
Ha megint használjuk $\Phi^{\mathrm{T}}\Phi = U\Sigma U^{\mathrm{T}}$ sajátértékfelbontást ($\Phi^{\mathrm{T}}\Phi + \lambda I$-re is), akkor megkapjuk, hogy:

$$\begin{aligned}V(\lambda) &= \sigma^2\,\mathrm{tr}\left(U\Sigma U^{\mathrm{T}}U(\Sigma + \lambda I)^{-1}U^{\mathrm{T}}U\Sigma U^{\mathrm{T}}U(\Sigma + \lambda I)^{-1}U^{\mathrm{T}}\right) \\ &= \sigma^2\,\mathrm{tr}\left(U\Sigma(\Sigma + \lambda I)^{-1}\Sigma(\Sigma + \lambda I)^{-1}U^{\mathrm{T}}\right) \\ &= \sigma^2\,\mathrm{tr}\left(\Sigma(\Sigma + \lambda I)^{-1}\Sigma(\Sigma + \lambda I)^{-1}\right) \\ &= \sigma^2\sum_{i=1}^{d}\frac{\sigma_i^2}{(\sigma_i + \lambda)^2}\end{aligned} \tag{3.25}$$

Itt az $U^{\mathrm{T}}U = I$ azonnosságot használtuk fel, illetve az utolsó egyenlőségnél $\Sigma(\Sigma + \lambda I)^{-1}\Sigma(\Sigma + \lambda I)^{-1}$ egy diagonális mátrix $\frac{\sigma_i^2}{(\sigma_i+\lambda)^2}$ főátlóbeli elemekkel.

3.1.2. Következmény Ha $\lambda \to +\infty$, akkor $\sigma^2\frac{\sigma_i^2}{(\sigma_i+\lambda)^2} \to 0$, vagyis a variancia, $V(\lambda) \to 0$. Ez annak is a következménye, hogy ha $\lambda \to +\infty$, akkor $\hat{\beta} \to 0$.
Ha $\lambda \to 0$, akkor $\frac{\sigma_i^2}{(\sigma_i+\lambda)^2} \to 1$, vagyis $V(\lambda) = \sigma^2\sum_{i=1}^{d} 1 \to d\sigma^2$

3.6. Megjegyzés Észrevehetjük, hogy a variancia, $V(\lambda)$, csökken, ahogy λ nő.



3.1. ábra. Ridge regresszió együtthatói különböző λ-k esetén

A 3.1 ábrán 10 darab tulajdonságra generáltam egyenként 100 darab adatot szintetikusan, 10 zajjal. Az ábrán látható, hogy $\lambda \to \infty$ esetén a modell együtthatói tartanak a nullához, ellenben nem érik el azt.

3.2. LASSO

A ridge regressziónak egyik főbb tulajdonsága, hogy nem nulláz ki együtthatókat. Tehát a végső modellben az összes paraméter benne van (kivéve, ha $\lambda = \infty$), mivel a $\lambda||\beta||^2$ csak csökkenti az együtthatókat nulla felé. Ez persze nem jelent problémát a modell pontossága esetén, ellenben nehezebben lesz értelmezhető, ha sok paraméter van. A LASSO (least absolute shrinkage and selection operator) egy olyan zsugorító módszer, ami megoldást jelent erre a problémára úgy, hogy a ridge regresszióban levő L_2-es norma helyett L_1-es normájú büntetőtagot minimalizálunk. Tehát:

3.2. Definíció *Adott $\lambda \geq 0$ esetén a LASSO célfüggvénye*

$$\nu(\beta) = ||y - \Phi\beta||_2^2 + \lambda||\beta||_1 = R(\beta) + \lambda||\beta||_1 \tag{3.26}$$

Egy ezzel ekvivalens definíció:

$$\nu(\beta) = ||y - \Phi\beta||_2^2 \ ahol \ ||\beta||_1 \leq t \tag{3.27}$$

Vagyis a célunk megtalálni $\hat{\beta}$-ot, amire igaz, hogy

$$\hat{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\|y - \Phi\beta\|^2 + \lambda|\beta||_1 \tag{3.28}$$

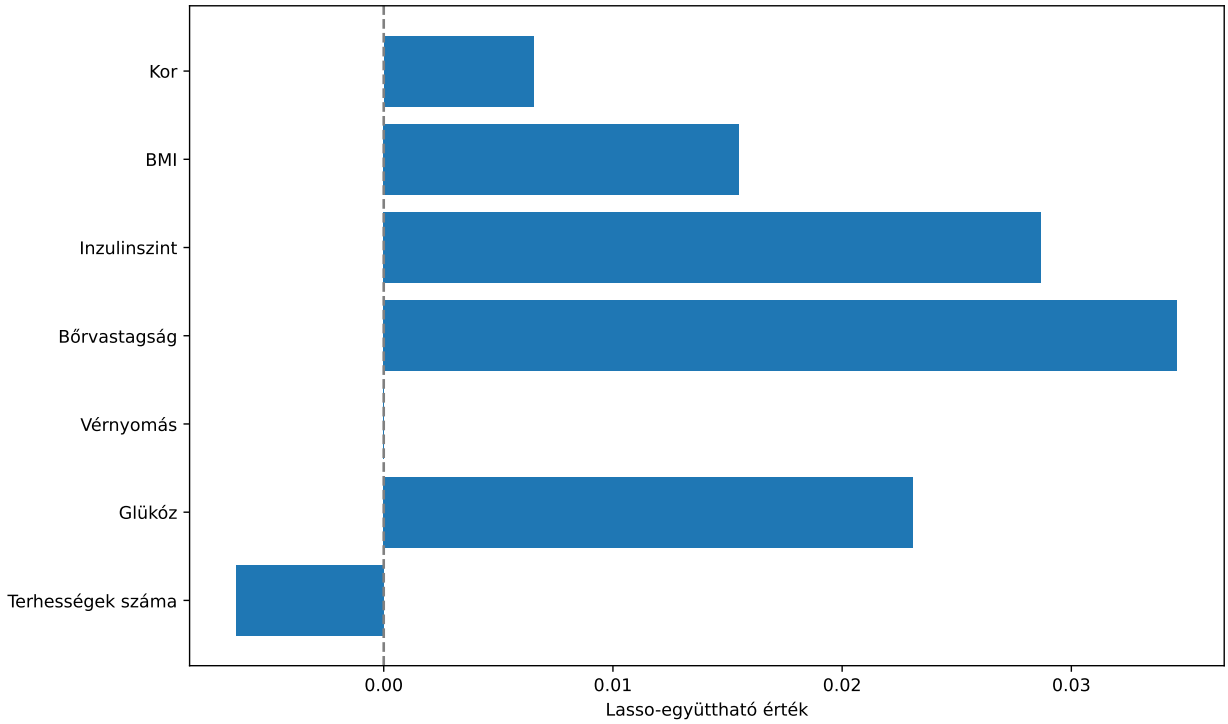
Avagy:

$$\hat{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\|y - \Phi\beta\|^2 \text{ ahol } ||\beta||_1 \leq t \tag{3.29}$$

3.7. Megjegyzés A ridge regresszióval való hasonlóság ellenére, a LASSO esetén az L_1 norma miatt a megoldás nem lesz lineáris y-ban, és nincs is zárt formula rá. A becsült $\hat{\beta}_\lambda$ kiszámítása kvadratikus programozási feladat, amivel nem foglalkozunk a diplomamunkában.

Ha t elég kicsi, akkor egyes együtthatói a modellnek kinullázódnak, így egyfajta folytonos változókiválasztási módszert is érthetünk alatta. Ha $t \geq t_0$, ahol $t_0 = ||\hat{\beta}||_1$ (ahol $\hat{\beta}$ a legkisebb négyzetek módszer becslése), akkor a $\hat{\beta}_\lambda$ megegyezik a legkisebb négyzetek módszer becslésével. Ugyanakkor, ha $t = \frac{t_0}{2}$, akkor a legkisebb négyzetek módszer által becsült paraméterek átlagosan feleződnek. Ellenben a zsugorítás hatása a paraméterekre nem ennyire egyértelmű. (Hastie, Tibshirani és Friedman 2004)

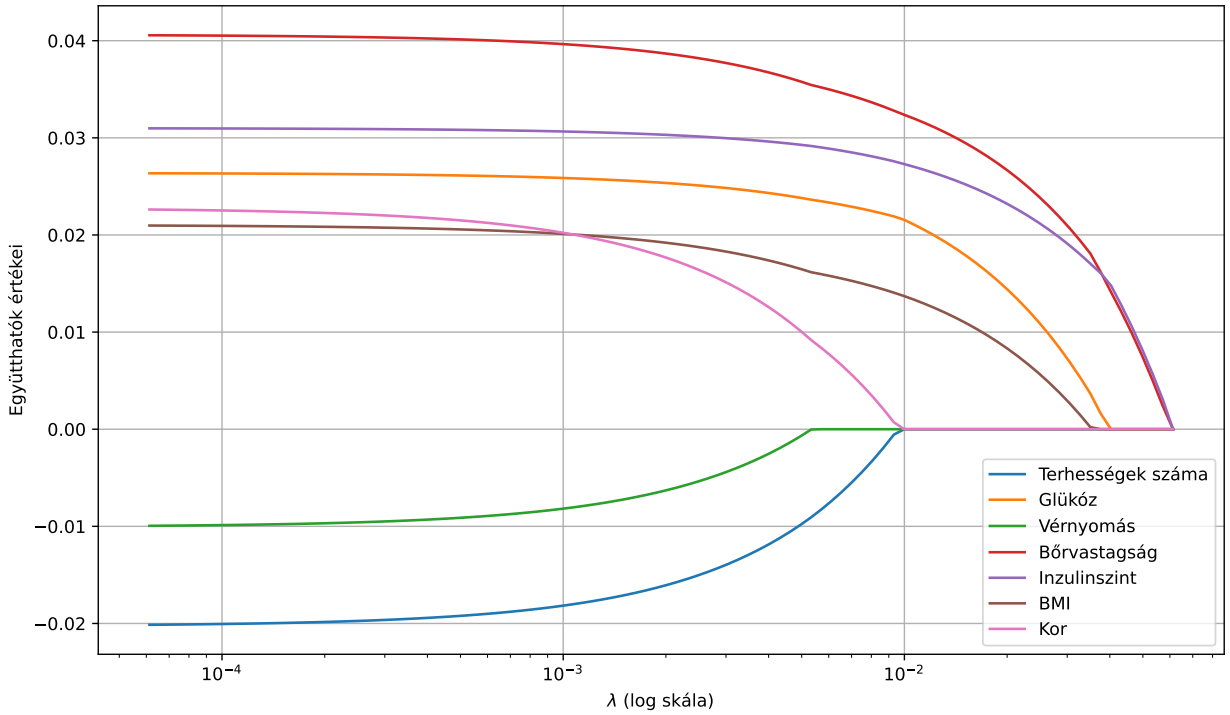
3.1. Példa Tekintsünk egy olyan adathalmazt (Forrás: Kaggle adathalmaz), amiben a célváltozó az úgynevezett *DiabetesPedigreeFunction*, egy olyan érték, ami azt méri, hogy adott illető mekkora valószínűséggel cukorbeteg. Az adathalmazunkban 7 darab változó van, ezek a következők: terhességek száma, glükóz, vérnyomás, bőrvastagság, inzulinszint, BMI és a kor. Meg szeretnénk tudni, hogy ezek közül melyeknek van szignifikáns hatása a cukorbetegségre. Ha LASSO regressziót alkalmazunk ezen az adathalmazon, akkor megkapjuk a következőt:



3.2. ábra. LASSO együtthatók

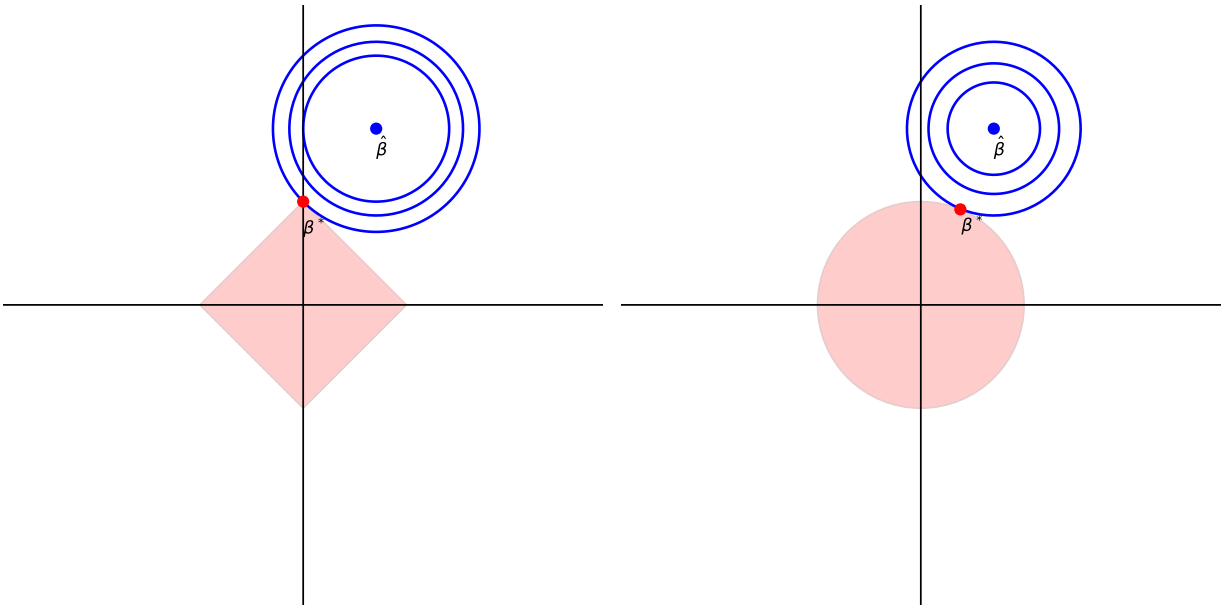
A 3.2 ábra azt mutatja, hogy a legnagyobb hatással a cukorbetegségre, a bőrvastagság, a glükóz, és az inzulinszint voltak, melyek köztudottan szoroskapcsolatban állnak vele. Ugyanakkor észrevehetjük, hogy a vérnyomás teljesen kinullázódott a modellben, sőt a terhességek száma negatív hatással van rá. A modellben az optimális λ a 0.00657 volt, melyet ötszörös keresztvalidációval kaptam meg. Az átlagos négyzetes hiba (MSE) 0.1059 volt.

Megkaphatjuk azt is, hogy különböző lambdákra, hogyan változnak a modell együtthatói.



3.3. ábra. Változók nagysága különböző λ-k esetén

A 3.3 ábrán láthatjuk, hogy minél nagyobb lambda esetén, a változók tartanak a nullához, és elég nagy lambda esetén ki is nullázódnak..



3.4. ábra. A LASSO és a ridge regresszió megoldásai

A 3.4 ábrán a regularizálatlan hiba függvényt láthatjuk (kék), a LASSO regularizációval a bal oldalon, és a jobboldalon pedig a ridge regularizációval kiegészítve. Ha t elég nagy, akkor a megkötési tartomány magába fogja foglalni $\hat{\beta}$-t, és ekkor a ridge és LASSO becslések megegyeznek (egy ilyen nagy t a $\lambda = 0$-nak felel meg). Az ellipszisek egyes körvonalain ugyanazokat az RSS értékeket veszik fel, és ahogy növekszik az ellipszis el $\hat{\beta}$-tól, úgy nő. A 3.2 és (3.27) egyenletek azt súgalják, hogy a LASSO és ridge együttható becslések ott vannak, ahol először érintkezik az ellipszis a megkötési tartománnyal. Mivel a ridge regressziónak a büntetőtagja egy gömb, így az érintkezés nem egy tengely mentén lesz, és így a becslés sosem lesz nulla. Ellenben a LASSO büntetőtagjának vannak sarkai minden tengely mentén, és így az ellipszis gyakran fogja azt érinteni. Ekkor az adott β_j kinullázódhat. Magasabb dimenziókban egyszerre több β_j is lehet nulla.

3.3. Általánosított regularizáció

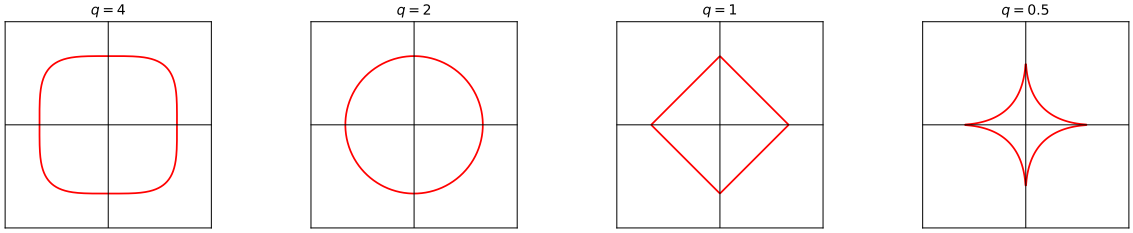
Alkalmazhatunk általánosabb büntetőtagot is adott $q \geq 0$-ra is:

$$\nu(\beta) = ||y - \Phi\beta||^2 + \lambda||\beta||_q^q = R(\beta) + \lambda||\beta||_q^q \tag{3.30}$$

Ha a $||\beta||_q^q$ kifejezést a β_j paraméter log-prior sűrűségének tekintjük, akkor ezek egyben a paraméterek prior eloszlásának szintvonalai is (Hastie, Tibshirani és Friedman 2004). Ha $q = 0$, akkor ez a változókiválasztásnak felel meg, mivel ekkor a büntetés, a paraméterek darabszámát számolja. A $q = 1$ a LASSO-nak, a $q = 2$ a ridge regressziónak felel meg. Ha $q \leq 1$, akkor a prior eloszlás tömege a koordinátairányok mentén koncentrálódik. A $q = 1$-hez tartozó prior eloszlás a Laplace eloszlás minten β_j paraméterre, aminek a sűrűségfüggvénye a következő:

$$f(x) = \frac{1}{2\tau} \exp\left(-\frac{|\beta|}{\tau}\right), \quad \text{ahol } \tau = \frac{1}{\lambda} \tag{3.31}$$

Megfigyelhetjük, hogy a $q = 1$ eset a legkisebb olyan q, hogy a megkötési tartomány konvex. A nem konvex megkötési tartományok nehezebbé teszik az optimalizálási problémát.



3.5. ábra. L_q normák

3.4. Elasztikus háló

A LASSO regularizációnak van néhány nemkívánatos tulajdonsága. Például, ha $n < d$, akkor maximum n darab nem nulla együttható lesz a modellben. Ez a ridge regresszió esetén nem fordulhat elő. Másodszor, a LASSO gyakran alulteljesít, amikor β^*-nak sok elég kicsi, de nem nulla eleme van. Annak érdekében, hogy ezeket a problémákat megelőzzük, jöhet szóba az elasztikus háló becslés, amellyel először Trevor Hastie és Hui Zou foglalkozott 2005-ben (Zou

és Hastie 2005). Ez a következő minimalizálás problémát mutatja be:

$$\hat{\beta}(\lambda_1, \lambda_2) = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\|y - \Phi\beta\|^2 + \lambda_2||\beta||_2^2 + \lambda_1|\beta||_1, \tag{3.32}$$

ahol $\lambda_1, \lambda_2 \geq 0$.

Egyértelműen $\lambda_1 = 0$ esetén visszakapjuk a ridge regressziót, $\lambda_2 = 0$ esetén pedig a LASSO-t. Itt a büntetőtagot felfoghatjuk mint a skálázott súlyozott átlagát az L_1 és L_2 hibának.

$$P_{\lambda_1,\lambda_2}(\beta) = (\lambda_1 + \lambda_2)\left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\|\beta\|_2^2 + \frac{\lambda_1}{\lambda_1 + \lambda_2}\|\beta\|_1\right) \tag{3.33}$$

Így pedig hihetőnek tűnik, hogy megfelelő (λ_1, λ_2) választásával, az elasztikus háló el tud érni egyfajta egyensúlyt a LASSO és a ridge regresszió előnyei között, a hátrányaik elkerülésével.

3.2. Tétel *Adott (X, Y) adathalmaz és (λ_1, λ_2) paraméterek esetén készítsük el az (X^*, Y^*) adathalmazt a következőképp:*

$$X^* = \frac{1}{\sqrt{1 + \lambda_2}}\begin{bmatrix} X \\ \sqrt{\lambda_2} I_d \end{bmatrix}, \quad Y^* = \begin{bmatrix} Y \\ 0_d \end{bmatrix},$$

ahol I_d egy $d \times d$ egységmátrix, és $0_d \in \mathbb{R}^d$ egy nullvektor. Legyen $\gamma = \lambda_1/\sqrt{1 + \lambda_2}$ és jelölje $\beta^* = \sqrt{1 + \lambda_2}\beta$. Ekkor az elasztikus háló átírható LASSO problémára a módosított adatokkal:

$$L(\gamma, \beta^*) = \|y^* - X^*\beta^*\|_2^2 + \gamma\|\beta^*\|_1 \tag{3.34}$$

Sőt, jelölje:

$$\hat{\beta}^* = \arg\min_{\beta^*} L(\gamma, \beta^*) \tag{3.35}$$

Ekkor az elasztikus háló megoldása a következő:

$$\hat{\beta} = \frac{1}{\sqrt{1 + \lambda_2}}\hat{\beta}^* \tag{3.36}$$

Bizonyítás: Az elasztikus hálóból (3.32) kiindulva:

$$\begin{aligned} L(\lambda_1, \lambda_2, \beta) &= \|y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1 \\ &= y^\top y - 2y^\top X\beta + \beta^\top X^\top X\beta + \lambda_2\beta^\top\beta + \lambda_1\|\beta\|_1 \\ &= y^\top y - 2y^\top X\beta + \beta^\top(X^\top X + \lambda_2 I_d)\beta + \lambda_1\|\beta\|_1 \end{aligned} \tag{3.37}$$

Most vegyük a felnagyított $(n + d) \times d$-s mátrixot:

$$\tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda_2} I_d \end{bmatrix}$$

Ekkor

$$\tilde{X}^\top \tilde{X} = X^\top X + \lambda_2 I_d \tag{3.38}$$

Ellenben $\tilde{X}$ nem normális minden $j \in \{1, \ldots, p\}$-re:

$$\tilde{X}_j^\top \tilde{X}_j = X_j^\top X_j + \lambda_2 \tag{3.39}$$

Így vehetjük a normálissá tett alakját:

$$X^* = \frac{1}{\sqrt{1 + \lambda_2}} \tilde{X} = \frac{1}{\sqrt{1 + \lambda_2}} \begin{bmatrix} X \\ \sqrt{\lambda_2} I_d \end{bmatrix}$$

Definiáljuk az $y^* = (y, 0_d)^{\mathrm{T}}$ vektort. Mivel y centrálva van (az elemei átlaga nulla), így y^* is centrálva van. Így pedig:

$$y^{*\top} y^* = y^\top y, \quad y^{*\top} X^* = \frac{1}{\sqrt{1 + \lambda_2}} y^\top X \tag{3.40}$$

Most átírhatjuk a veszteséget:

$$\begin{aligned} L(\lambda_1, \lambda_2, \beta) &= y^\top y - 2y^\top X\beta + \beta^\top (X^\top X + \lambda_2 I_d)\beta + \lambda_1 \|\beta\|_1 \\ &= y^{*\top} y^* - 2 \cdot \frac{1}{\sqrt{1+\lambda_2}} y^{*\top} X^* \beta + (1+\lambda_2)\beta^\top X^{*\top} X^* \beta + \lambda_1 \|\beta\|_1 \end{aligned} \tag{3.41}$$

Helyettesítsünk $\beta^* = \sqrt{1+\lambda_2}\beta$, és definiáljuk $\gamma = \lambda_1/\sqrt{1+\lambda_2}$. Ekkor:

$$\begin{aligned} L(\lambda_1, \lambda_2, \beta) &= y^{*\top} y^* - 2y^{*\top} X^* \beta^* + \beta^{*\top} X^{*\top} X^* \beta^* + \gamma \|\beta^*\|_1 \\ &= \|y^* - X^* \beta^*\|_2^2 + \gamma \|\beta^*\|_1 \end{aligned} \tag{3.42}$$

Ez a LASSO definíciója. Így, ha:

$$\hat{\beta}^* = \arg\min_{\beta^*} \left(\|y^* - X^* \beta^*\|_2^2 + \gamma \|\beta^*\|_1 \right), \tag{3.43}$$

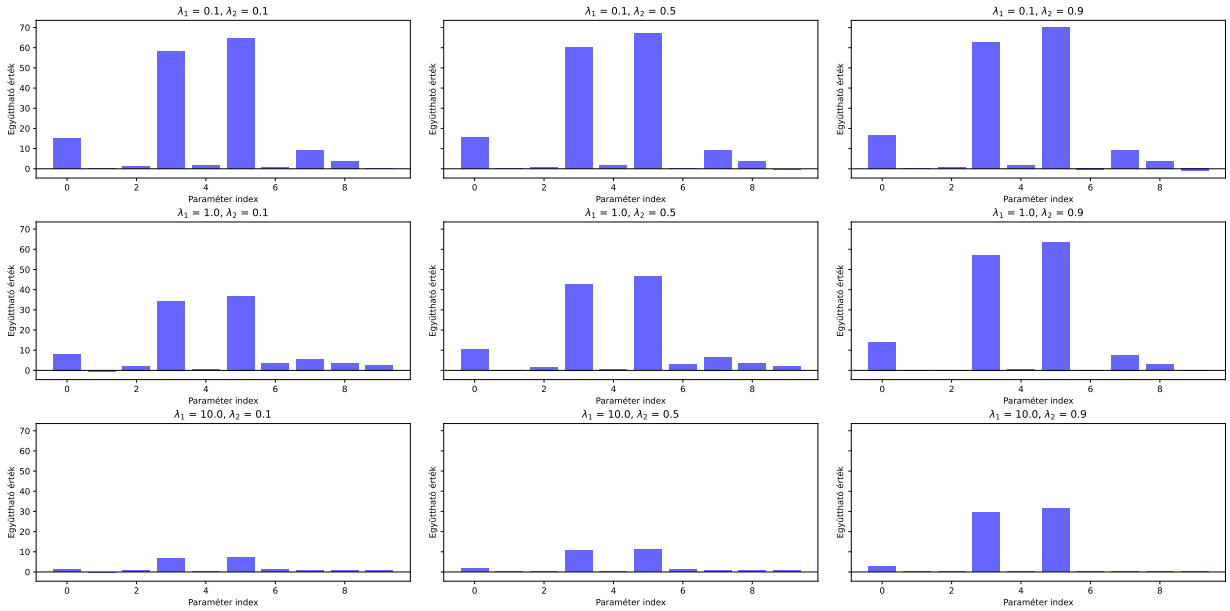
akkor az elasztikus háló megoldását megkapjuk a következőképp:

$$\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}} \hat{\beta}^* \tag{3.44}$$

□

3.8. Megjegyzés Így, hogy visszavezettük a LASSO regresszióra, ugyancsak kvadratikus programozási probléma a megoldás.

3.2. Példa Nézzük meg, hogyan viselkedik az elasztikus háló egy példán keresztül:



3.6. ábra. Elasztikus háló együtthatók különböző λ_1, λ_2 értékekre

A 3.6 ábrán szintetikusan generáltam 10 különböző paramétert, mindegyikből 100-at, 10 Gauss eloszlásbeli szórással. Ezután elasztikus háló modellt használtam $\lambda_1 = (0.1, 1, 10)$ és $\lambda_2 = (0.1, 0.5, 0.9)$ esetekben. A modell a legjobban a $\lambda_1 = 0.1$ és $\lambda_2 = 0.9$ esetén illeszkedett, $MSE = 99.08$-al. Az ábrán több dolgot is észre lehet venni. Például, ha λ_1 elég kicsi, de λ_2 ellenben nagy, akkor a LASSO paraméter kinullázó hatása elmarad, az együtthatóink nem lesznek nullák. Ellenben, ha pont fordítva van az eset, vagyis az L_2-es büntetőtag λ-ja kicsi, és az L_1-esé nagy, akkor a legtöbb változó kinullázódik, ezzel is megmutatva a LASSO regresszió egyik fő tulajdonságát.

Összegzés

Dolgozatom elsősorban olyan függvények kereséséről szól, amik jól leírnak egy valóvilágbeli jelenséget, például szeretnénk megtudni egy ingatlan reális árát úgy, hogy csak a méretét, szobáinak a számát, és elhelyezkedését ismerjük.

Diplomamunkámban elsőként a polinominterpoláció segítségével motiváltuk a lineáris regresszió alkalmazását. Megmutattuk, hogy önmagában nem elegendő olyan függvényt konstruálni, amely minden tanítópontra illeszkedik, mivel az ilyen modellek általánosító képessége gyenge lehet, vagyis új bemenetek esetén nem biztosítanak jó predikciót.

Ezt követően bevezettük a regressziós problémát, melynek célja megtalálni az X és Y közötti kapcsolatot egy mérhető f függvény segítségével. Ha feltételezzük, hogy a regressziós függvény $m(x)$ a paramétereiben lineáris, akkor lineáris regresszióról beszélhetünk. A klasszikus linearitás azonban gyakran túl szűk feltétel, ezért általánosítottuk a fogalmat: a függvény akkor is lehet lineáris, ha rögzített, nemlineáris bázisfüggvények lineáris kombinációjaként írható fel. Ezután bemutattuk a legkisebb négyzetek módszerét, és elemeztük annak tulajdonságait. Megmutattam, hogy ha a bemeneti mátrix Φ teljes rangú és sovány, akkor a becsült $\hat{\beta}$ paramétervektor létezik és egyértelmű. Azt is beláttuk, hogy a reziduálisa a rendszernek merőleges Φ oszlopterére. Ezt követően tárgyaltuk a sztochasztikus lineáris regressziót, amely megfelelő feltételek mellett hasonló tulajdonságokat mutat, mint a determinisztikus eset, illetve megmutattuk, hogy a lineáris és torzítatlan becslések között a legjobb.

A következő fejezetben a legkisebb négyzetek módszerének korlátaira hívtuk fel a figyelmet. Bemutattuk, hogy a várható négyzetes hiba három komponensre bontható: a torzítottság négyzetére, a varianciára és az adatokat jellemző zaj varianciájára. Ebből következik, hogy a várható négyzetes hiba alsó korlátja a zaj varianciája. Különböző példákon keresztül szemléltettük, hogy egy torzítatlan modell sem feltétlenül optimális, mivel magas varianciával járhat.

Ismertettük az alultanulás és túltanulás jelenségét is, és példákon keresztül érzékeltettük a gyakorlati hatásukat.

A regularizációval foglalkozó fejezetben megoldási lehetőségeket mutattunk arra, hogyan módosítható a legkisebb négyzetek módszere kis mértékű torzítás bevezetésével annak érdekében, hogy csökkenjen a modell varianciája és javuljon az általánosítóképessége, illetve egyéb előnyöket is elérhessünk, mint például a jól kondicionáltság. A zsugorítómódszerek közül elsőként a ridge regressziót mutattuk be, amely egy L_2 normán alapuló büntetőtaggal egészíti ki a célfüggvényt. Megmutattuk, hogy a regularizációs paraméter λ növelésével a modell torzítottabb lesz, viszont a varianciája csökken, és $\lambda \to \infty$ esetén a becsült paraméterek nullához tartanak. Kitértünk a Tikhonov-regularizációra is, ahol a büntetőtag súlyozását egy tetszőleges pozitív definit mátrixszal (Tikhonov-mátrix) lehet végezni.

Ezután bemutattuk a LASSO-módszert, amely L_1 normát használ büntetőtagként. Láttuk, hogy nagy λ esetén a paraméterek nullává válhatnak, így a módszer alkalmas változószelekcióra is. Végül ismertettük az általánosított regularizációt, amelyben tetszőleges L_q normákat alkalmazhatunk. A dolgozatot az elasztikus háló módszerével zártuk, amely a ridge és a LASSO kombinációja. Megmutattam, hogy ez a módszer visszavezethető a LASSO megoldására, és egy konkrét példán keresztül illusztráltuk, hogyan változik a modell viselkedése különböző λ_1 és λ_2 értékek esetén.

Megjegyzés: A dolgozatban szereplő valamennyi ábrát saját magam készítettem. Ezeket, valamint az ábrák előállításához használt Python-kódokat feltöltöttem a következő GitHub-repozitóriumba: Kódok, ábrák repozitóriuma.

Irodalomjegyzék

Aster, Richard Craig, Clifford H Thurber és Brian Borchers (2005). *Parameter Estimation and Inverse Problems*. Elsevier Academic Press.

Bishop, Christopher M (2006). *Pattern Recognition and Machine Learning*. Springer.

Csáji, Balázs Csanád (2024). *Data Mining and Machine Learning, ELTE órai jegyzet*.

Faragó, István és Róbert Horváth (2013). *Numerikus Módszerek*. Második, javított kiadás. Typotex.

Györfi, László, Michael Kohler, Adam Krzyzak és Harro Walk (2002. aug.). *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media.

Hastie, Trevor, Robert Tibshirani és J H Friedman (2004). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

James, Gareth, Daniela Witten, Trevor Hastie és Robert Tibshirani (2013). *An Introduction to Statistical Learning : with Applications in Python*. Springer.

Mohri, Mehryar, Afshin Rostamizadeh és Ameet Talwalkar (2018). *Foundations of Machine Learning*. The MIT Press.

Sun, Wen (2022). *Bias-Variance Decomposition in Ridge Linear Regression, Cornell Lecture Notes*. [elérhető itt].

Zou, Hui és Trevor Hastie (2005. ápr.). „Regularization and Variable Selection via the Elastic Net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 301–320. old.

MI alapú eszközök használatáról szóló nyilatkozat

Alulírott Halász Erik nyilatkozom, hogy szakdolgozatom elkészítése során az alább felsorolt feladatok elvégzésére a megadott MI alapú eszközöket alkalmaztam:

Feladat	Felhasznált eszköz	Felhasználás helye
Irodalomkeresés	GPT-4	Teljes dolgozat
Nyelvhelyesség ellenőrzése	Writefull	Teljes dolgozat

A felsoroltakon túl más MI alapú eszközt nem használtam.