„Shrinkage” módszerek a lineáris regresszióban

Diplomamunka

Írta: Szűcs Katalin

Matematika BSc
Alkalmazott matematikus szakirány

Témavezető:

Csiszár Villő
adjunktus
Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2012

Köszönetnyilvánítás

Ez úton szeretnék köszönetet mondani témavezetőmnek, Csiszár Villőnek, aki kitartó munkájával segítette szakdolgozatom elkészülését.

Köszönettel tartozom szüleimnek és nagyszüleimnek, akik mindvégig segítettek céljaim elérésében.

Bevezetés

A regresszióanalízis során két vagy több véletlen változó között fennálló kapcsolatot modellezzük. A lineáris regresszió ennek egy speciális esete, amikor a mért változót a magyarázó változók lineáris függvényével szeretnénk közelíteni. Ebben az esetben a legkisebb négyzetek módszerét alkalmazhatjuk. Ez az eljárás azonban nem mindig ad kielégítő eredményt. Előfordulhat, hogy a predikciós hiba túl nagy és a modell nehezen értelmezhető. Ilyenkor az összkép érdekében érdemes engedni a részletességből. Ha a közelítés során egy bizonyos korlátot szabunk a magyarázó változók használatára, akkor bár veszítünk a részletességből és a torzítatlanságból, összességében mégis pontosabb becslést és egy könnyebben áttekinthető modellt kapunk.

Szakdolgozatom első felében leírom, milyen különböző módszerek léteznek a korlátozás megtételére. Beszélek a részhalmaz kiválasztó algoritmusokról, amelyek segítségével megtalálhatjuk a magyarázó változók legnagyobb hatással bíró részhalmazát, majd leírom, hogy a „shrinkage", vagyis zsugorító módszerek, a ridge regresszió, a lasso és a legkisebb szög regresszió hogyan zsugorítják a becsült paraméterek együtthatóit, ezzel csökkentve a szórást. Az algoritmusok leírásánál a *The Elements of Statistical Learning* című könyvet és a *Least Angle Regression* című cikket használtam forrásként [1],[2].

Szakdolgozatom második felében egy példa adatsor segítségével mutatom be az eddig tárgyalt algoritmusokat. Az algoritmusok szimulálásához az R program *LARS* programcsomagját használtam [3].

Tartalomjegyzék

1. A lineáris regressziós modell és a legkisebb négyzetek módszere

A regressziós problémában az Y valószínűségi változót (függő változó) szeretnénk az $X^T = (X_1, X_2, \ldots, X_p)$ valószínűségi vektor változó (független változó) függvényével közelíteni legkisebb négyzetes értelemben. Amennyiben ismerjük az $Y, X_1, \ldots, X_p$ véletlen vektor együttes eloszlását, akkor

$$E\left(Y - g(X_1, \ldots, X_p)\right)^2$$

minimumát a p-változós g függvények körében $E(Y|X)$, Y-nak az $X_1, \ldots, X_p$ változók mellett vett feltételes várható értéke szolgáltatja, ezt nevezzük regressziós függvénynek. Mikor ezt a minimumot a lineáris függvények körében keressük, lineáris regresszióról van szó. A lineáris regressziós modell alakja a következő:

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

A regressziós problémában általában adott egy statisztikai minta a függő és független változókra $(x_1, y_1), \ldots, (x_N, y_N)$, ami alapján becsülhetjük a regressziós modellben szereplő, egyelőre ismeretlen β_j paramétereket. A legkisebb négyzetek módszerének alkalmazásakor a $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ együtthatókat úgy választjuk meg, hogy minimalizáljuk az $y_i - f(x_i)$ eltérések négyzetösszegét:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2. \tag{2}$$

(2) minimalizálásához definiáljuk X-et olyan $N \times (p+1)$-es mártrixként, amelynek i. sorában az $(1 : x_i)$ vektor szerepel, y pedig legyen egy N-dimenziós vektor, amelynek koordinátái a statisztikai minta függő változóit tartalmazzák. Ekkor az eltérések négyzetösszege a következő alakba írható:

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

Ezt β szerint differenciálva kapjuk, hogy

$$\begin{aligned} \frac{\partial RSS}{\partial \beta} &= -2X^T(y - X\beta) \\ \frac{\partial^2 RSS}{\partial \beta} &= 2X^T X. \end{aligned}$$

Tegyük fel, hogy X teljes rangú, $X^T X$ pedig pozitív definit mátrix és legyen

$$X^T(y - X\beta) = 0.$$

Ebből

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

Egy x_0 bemeneti vektor alapján a becsült értéket az $\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}$, az illesztett értékeket pedig az

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

összefüggés adja meg, ahol $\hat{y}_i = \hat{f}(x_i)$.

A továbbiakban rögzítsük az x_i változókat, és tegyük fel, hogy az y_i megfigyelések függetlenek és szórásnégyzetük (ezt jelölje σ^2) állandó. Ekkor σ^2 a következő módon becsülhető:

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

A nevezőben N helyett $N - p - 1$ szerepel, így σ^2-re torzítatlan becslést kapunk, vagyis $E(\hat{\sigma}^2) = \sigma^2$.

A további vizsgálatok során feltesszük, hogy (1) megfelelő modell (tehát Y feltételes várható értéke lineáris $X_1, \ldots, X_p$-re nézve), így:

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

alakba írható, ahol ε nulla várható értékű, σ^2 szórásnégyzetű normális eloszlású hiba. Hipotézisvizsgálat segítségével állapíthatjuk meg, hogy a β_j-k között szerepel-e nulla. Ehhez határozzuk meg az együtthatókhoz tartozó Z-pontszámot:

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

ahol v_j az $(X^T X)^{-1}$ mátrix j. diagonális eleme [4]. Ha z_j abszolút értéke nagy, a $\beta_j = 0$ nullhipotézist elvetjük.

Az együtthatókat becsülhetjük konfidenciaintervallumok segítségével is. A β_j-hez tartozó $1 - 2\alpha$ szintű konfidenciaintervallum:

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

ahol $z^{(1-\alpha)}$ a normális eloszlás $1 - \alpha$ értékhez tartozó kvantilise. Hasonló módon készíthetünk egy konfidencia ellipszoidot az egész β paramétervektor becslésére:

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

ahol $\chi_l^{2(1-\alpha)}$ az l-szabadságfokú khi-négyzet eloszlás $1 - \alpha$ értékhez tartozó kvantilise.

2. Részhalmaz kiválasztás

Annak ellenére, hogy a legkisebb négyzetek módszere által nyert becslés torzítása alacsony, a szórása tetszőlegesen nagy lehet. Ezen javíthatunk a magyarázó változók számának csökkentésével, esetleg néhány együttható kinullázásával. Ezáltal csökkentünk a szóráson, és annak ellenére, hogy a torzításon rontunk, összességében mégis egy pontosabb becslést kapunk. Ez az eljárás nem csupán a pontosságon javít. Amikor kevesebb változót vizsgálunk, könnyebben értelmezhető modellt kapunk, mivel eltekintünk a jelentéktelen részletektől. Fontos, hogy megtaláljuk azt a részhalmazt, amelyik a legnagyobb befolyást gyakorolja az összkép szempontjából, és csak olyan változókat zárjunk ki a vizsgálatból, amelyeknek hatása valóban elhanyagolható.
Több különböző stratégia létezik a megfelelő részhalmaz kiválasztására. Ebben a fejezetben ezekből az algoritmusokból ismertetünk néhányat.

2.1. Legjobb részhalmaz

Kisebb elemszámú minta esetén lehetőségünk van minden $k \in \{0, 1, \ldots, p\}$ –re kiválasztani azt a k elemű részhalmazt, amelyik minimalizálja a (2)-ben szereplő reziduális négyzetösszeget. Így minden k-ra megkapjuk a változók egy kívánt halmazát. Az eljárás azonban nem alkalmazható a megfelelő k meghatározására. Ha tekintjük az algoritmus által kiválasztott halmazokra az eltérések négyzetösszegét, k-t növelve csökkenő sorozatot kapunk, így a legjobb közelítést a legnagyobb elemszámú részhalmaz adja. A részhalmaz elemszámának meghatározásakor nem csak a torzítás és a szórás kettősének vizsgálatára kell odafigyelnünk. Fontos szempont a „takarékosság" is, hiszen nem szeretnénk, hogy a modell a túl sok változótól elbonyolódjon.
Több kritérium létezik, amelyek segítségével ezt a szubjektív elvárást precízebben megfogalmazhatjuk (erre jó példa az AIC kritérium).
A következőkben tárgyalt algoritmusok egy része nagyon hasonló. Az adatok felhasználásával egy modellekből álló sorozatot hoznak létre, amelynek tagjai különböző komplexitásúak és egyetlen indexszel vannak paraméterezve.

2.2. Az előre- és a hátralépő regresszió

Az előrelépő regresszió egy mohó algoritmus. Először a β_0 együtthatót illeszti, majd szekvenciálisan, egyesével adja a modellhez azt az elemet, amelyik a legjobban illeszkedik. Ezzel a kiválasztással modellek olyan sorozatát hozhatjuk létre, amelyben a tagok indexe éppen k, a részhalmaz mérete. Az előző eljárással ellentétben most nem kapjuk

meg mindig az optimális megoldást. Ennek ellenére az előrelépő regresszió több okból kifolyólag is előnyösebb a gyakorlati alkalmazások szempontjából. Egyrészt, a legjobb részhalmaz kiválasztással szemben, ez az algoritmus nagyobb elemszámú minta esetén is jól használható, másrészt alacsonyabb szórást ad, mintha az egész modellt tekintenénk.

A hátralépő regresszió az egész modellel indít, majd sorra kizárja azokat az elemeket, amelyek a legkevésbé illeszkednek, vagyis amelyekhez a legkisebb Z-pontszám tartozik. Míg az előrelépő regresszió minden esetben, ez a stratégia csak $N > p$ esetén alkalmazható (a gyakorlatban ez a feltétel általában teljesül). Bizonyos esetekben kevert lépésű algoritmusok is használhatók, amik mind az előre-, mind a hátralépő regressziót magukba foglalják, és a kettő közül minden lépésben a hatékonyabbat alkalmazzák.

2.3. Szakaszonkénti regresszió

A szakaszonkénti regresszió is első lépésben a β_0 együtthatót illeszti. Kezdetben a centralizált magyarázó változók minden együtthatója nulla. Az algoritmus minden lépésben felismeri az aktuális reziduálissal legjobban korrelált változót. Kiszámítja a reziduális lineáris regressziós együtthatóját azon a változón, és ezt az értéket hozzáadja a változó eredeti együtthatójához. Az eljárás addig tart, míg végül egyetlen változó sem korrelált a reziduálisokkal.

Míg az előrelépő regressziós során nincs szükség további korrekciókra, ha egyszer egy változót hozzáadtunk a modellhez, a szakaszonkénti regresszió esetében a magyarázó változók együtthatóit többször újra kell számítani, mielőtt elérnénk a megfelelő illeszkedést. Ennek következményeként az algoritmus lépésszáma p-nél sokkal nagyobb lehet. Lassúsága miatt ez az eljárás gyakran eredménytelennek bizonyulhat. Magasabb dimenziós problémák esetén azonban versenyképes.

3. ,,Shrinkage" módszerek

A részhalmaz kiválasztó eljárások segítségével egy értelmezhetőbb modellt kapunk, amellyel a becslés hibája várhatóan alacsonyabb, mintha a teljes modellt tekintenénk. Ennek ellenére, mivel ez egy diszkrét eljárás, a szórás továbbra is túl nagy lehet, így a teljes modellt tekintve a predikciós hiba nem csökken megfelelően.

A ,,shrinkage" módszerek (zsugorító módszerek) ennél sokkal folytonosabbak és alacsonyabb szórást adnak.

3.1. Ridge regresszió

A ridge regresszió úgy zsugorítja a regressziós együtthatókat, hogy egy korlátot szab azok méretére. A ridge együtthatók minimalizálják a következő büntető taggal ellátott reziduális négyzetösszeget:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{3}$$

Itt a $\lambda \geq 0$ paraméter szabályozza a zsugorítást, vagyis λ értéke minél nagyobb, az együtthatók annál közelebb kerülnek egymáshoz és a nullához. A ridge probléma egy másik, az előzővel ekvivalens megfogalmazása a következő:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$
$$\text{feltéve, hogy } \sum_{j=1}^{p} \beta_j^2 \leq t. \tag{4}$$

Így az együtthatók méretére szabott korlát expliciten kifejezhető a paraméterrel. Ha egy regressziós modellben túl sok a korrelált változó, akkor ezek együtthatói könnyen válhatnak alulhatározottá, szórásuk pedig túl nagy lehet. Például az egyik változó pozitív együtthatóját eltörölheti egy másik, az előzővel korrelált változó negatív együtthatója. Ha alkalmazzuk a (4)–ben szereplő korlátot az együtthatók méretére, enyhíthetünk ezen a problémán.

Megjegyezzük, hogy a konstans tag nem szerepel a büntető tagban, mivel ez azt jelentené, hogy az eljárás nem független Y origójának választásától.

Mivel a ridge regresszió megoldásai nem függetlenek a magyarázó változók skálázásától, így érdemes azokat centralizálni és normálni az együtthatók kiszámítása előtt, vagyis minden x_{ij}-t helyettesítsünk $\frac{x_{ij}-\bar{x}_j}{\sigma}$-val. Ezentúl mindig feltesszük, hogy a centralizálás megtörtént és az X mátrix már nem $N \times (p+1)$, hanem $N \times p$ dimenziós. A konstans tag β_0 becslése $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, a többi együtthatót pedig ridge regresszióval becsüljük, konstans tag nélkül, a centralizált magyarázó változókat használva.

A (3)-ban szereplő kritériumot mátrixos alakba írva kapjuk, hogy

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

amelyből a ridge regressziós együtthatók könnyen számíthatók:

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

ahol I a $p \times p$ dimenziós egységmátrix. Megjegyezzük, hogy a $\beta^T\beta$ büntető tagot használva, a ridge regressziós együttható y lineáris függvénye, valamint mivel az invertálás előtt egy pozitív konstanst adunk az X^TX mátrix főátlóbeli elemeihez, így a probléma akkor sem válik szinguláris sá, ha X^TX nem teljes rangú.

A centralizált X mátrix szinguláris érték felbontása könnyebben áttekinthetővé teszi a ridge regresszió néhány tulajdonságát. A felbontás alakja a következő:

$$X = UDV^T. \tag{5}$$

Itt U egy $N \times p$ méretű ortogonális mátrix, amelynek oszlopvektorai kifeszítik X oszlopterét, V pedig egy $p \times p$ méretű ortogonális mátrix, amelynek oszlopvektorai X sorterét feszítik ki. D egy $p \times p$ dimenziós diagonális mátrix, amelynek elemei $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ az X mátrix szinguláris értékei. Ha legalább egy $d_j = 0$, akkor az X mátrix szinguláris.

Ha a szinguláris érték felbontás segítségével írjuk fel a legkisebb négyzetek módszerével illesztett értékeket, a következőt kapjuk:

$$\begin{aligned} X\hat{\beta}^{ls} &= X(X^TX)^{-1}X^Ty \\ &= UU^Ty, \end{aligned} \tag{6}$$

ahol U^Ty az y vektor koordinátáit adja, az U oszlopvektoraiból alkotott ortonormált bázissal kifejezve. A ridge regressziós megoldások pedig:

$$\begin{aligned} X\hat{\beta}^{ridge} &= X(X^TX + \lambda I)^{-1}X^Ty \\ &= UD(D^2 + \lambda I)^{-1}DU^Ty \\ &= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^Ty, \end{aligned} \tag{7}$$

ahol u_j az U oszlopait jelöli. Megjegyezzük, hogy $\lambda \geq 0$ esetén, $\frac{d_j^2}{d_j^2+\lambda} \leq 1$. A lineáris regresszióhoz hasonlóan, a ridge regresszió is az U oszlopvektoraiból alkotott ortonormált bázis segítségével fejezi ki y koordinátáit. Az eljárás a koordinátákat $\frac{d_j^2}{d_j^2+\lambda}$ mértéke szerint zsugorítja, vagyis a kisebb d_j^2 értékkel rendelkező bázis vektorokhoz tartozó koordináták zsugorítása nagyobb.

Az X mátrix szinguláris értékek szerinti felbontása arra is lehetőséget ad, hogy kifejezzük vele a mátrix főkomponenseit. A minta kovariancia mátrixa $S = \frac{X^T X}{N}$, és (5) alapján

$$X^T X = V D^2 V^T$$

adja az $X^T X$ mátrix sajátfelbontását. A V mátrix oszlopai az X sajátvektorait tartalmazzák, ezek közül az első sajátvektor, v_1 mutat a legnagyobb szórás irányába. Így $z_1 = Xv_1$ az X mátrix oszlopainak legnagyobb szórással rendelkező normált lineáris kombinációja. Könnyen látható, hogy $Var(z_1) = Var(Xv_1) = \frac{d_1^2}{N}$, és $z_1 = Xv_1 = u_1 d_1$. z_1-et nevezzük X első főkomponensének, u_1 pedig a normált első főkomponens. A további főkomponensek szórása egyre kisebb és a legutolsó főkomponens rendelkezik a legkisebb szórással. Ezért a kisebb szinguláris értékek X oszlopterében az alacsonyabb szórású irányokhoz tartoznak, és a ridge regresszió ezeket az irányokat zsugorítja a legnagyobb mértékben.

A ridge regresszió tényleges szabadságfokát a következő monoton csökkenő függvény adja:

$$\begin{aligned} df(\lambda) &= tr[X(X^T X + \lambda I)^{-1} X^T] \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}. \end{aligned} \tag{8}$$

Általában p változó esetén a lineáris regresszió szabadságfoka p, a szabad paraméterek száma. Ha elengedjük a korlátozást, $\lambda = 0$ választással visszakapjuk ezt az értéket, vagyis ekkor $df(\lambda) = p$, továbbá $\lambda \to \infty$ esetén $df(\lambda) \to 0$. Természetesen egy további szabadságfokkal kell számolnunk a konstans tag miatt, amit előzetesen eltávolítottunk.

3.2. Lasso

A lasso a ridge regresszióhoz hasonló zsugorító módszer. A lasso becslés együtthatóit a következő módon kapjuk:

$$\begin{aligned} \hat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2, \\ \text{feltéve, hogy } \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{9}$$

A ridge regresszióhoz hasonlóan újraparaméterezhetjük a konstans tagot a magyarázó változók centralizálásával. A modellt itt is a konstans tag nélkül illesztjük, β_0 becslése pedig továbbra is $\bar{y}$. A lasso problémát az előzővel ekvivalens Lagrange formában is

felírhatjuk:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{10}$$

Fontos különbség, hogy a ridge regresszióban szereplő $\sum_{j=1}^{p} \beta_j^2$ büntető tagot most $\sum_{j=1}^{p} |\beta_j|$ váltja fel. Ezt a korlátozást használva az együtthatók már nem lineárisak y-ra nézve, így kiszámításuk nem adható meg olyan zárt formulával, mint a ridge regresszió esetében. A lasso együtthatók kiszámítása kvadratikus programozási feladat. Ennek ellenére léteznek olyan algoritmusok, amelyek segítségével hasonló számítási költséggel megtalálható a megoldás, mint a ridge regresszió esetében. Ezeket az algoritmusokat később részletesen tárgyaljuk. Az új büntető tag lehetővé teszi, hogy elegendően kicsi t választással néhány együtthatót nullára állítsunk. Ha t-t $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$-nál nagyobbra választjuk (ahol $\hat{\beta}_j$-k a legkisebb négyzetek módszerével nyert együtthatók), akkor a lasso becslés együtthatói is a $\hat{\beta}_j$-k. Másrészt viszont, ha t-t $t = \frac{t_o}{2}$-nek választjuk, a legkisebb négyzetes együtthatókat átlagosan 50%-kal zsugorítjuk. Ennek ellenére azonban a zsugorítás viselkedése nem mindig egyértelmű.

3.3. A ridge regresszió és a lasso összehasonlítása

Ebben a szakaszban röviden összefoglaljuk és összehasonlítjuk az eddig tárgyalt algoritmusok, a részhalmaz kiválasztás, a ridge regresszió és a lasso tulajdonságait.

Ortogonális X bemeneti mátrix esetén mindhárom eljárás expliciten megoldható. Mindegyik módszer egy egyszerű transzformációt hajt végre a legkisebb négyzetek módszerének együtthatóin. A ridge regresszió arányosan zsugorítja az együtthatókat, míg a lasso a nulla közelében „elvágja” azokat. A legjobb részhalmaz algoritmus csak az első M legnagyobb együtthatóhoz tartozó változót választja be a modellbe, a többit elhagyja.

Most tekintsük azt az esetet, amikor az X mátrix nem ortogonális. Ha csupán két változó együtthatóit, β_1-et és β_2-t vizsgáljuk egy koordináta rendszerben, a reziduálisok négyzetösszegének szintvonalai olyan ellipsziseket írnak le, amelyek középpontja a legkisebb négyzetek módszerével kapott értékeket adja. A ridge regresszió esetében az együtthatók méretére a $\beta_1^2 + \beta_2^2 \leq t$ feltételt szabjuk. Ez a korlátozás egy körlapot határoz meg a síkon. A lasso esetében ugyanez a $|\beta_1| + |\beta_2| \leq t$ rombusz. Mindkét módszer megoldása az a pont, ahol az ellipszis érinti a korlátozott terület határát. Jól látszik, hogy a lasso esetében előfordulhat, hogy a rombusz egyik csúcsát kapjuk megoldásként, és ekkor valamelyik együttható nulla. $p > 2$ esetén rombusz helyett egy romboidot kapunk, amelynek több csúcsa, éle és lapja van, így még nagyobb az esélye annak, hogy valamelyik becsült paraméterre nullát kapunk.

Az algoritmusokat tekinthetjük a következő általános alakban:

$$\tilde{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \tag{11}$$

ahol $q \geq 0$. Itt $q = 0$ választással a részhalmaz kiválasztás együtthatóit kapjuk, hiszen ekkor a büntető tag a nem nulla paraméterek számát adja. A $q = 1$ eset a lassohoz, míg a $q = 2$ eset ridge regresszióhoz tartozik. A $q \in (1, 2)$ választás kompromisszumot eredményez a lasso és a ridge regresszió között. Ekkor $|\beta_j|^q$ differenciálható a nullában, így nincs meg az a lehetőség, hogy néhány együtthatót kinullázzunk. Részben ebből az okból kifolyólag, részben pedig a kezelhetőbb számítási tulajdonságai miatt vezette be 2005-ben Zou és Hastie az elasztikus háló elnevezésű büntető tagot, amelynek alakja a következő:

$$\lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1 - \alpha)|\beta_j|). \tag{12}$$

Ez egy másfajta kompromisszum a lasso és a ridge regresszió között. Az elasztikus háló a lassohoz hasonlóan választja ki a megfelelő változókat, és úgy zsugorítja a korrelált magyarázó változók együtthatóit, mint a ridge regresszió. Ezen felül kedvező számítási tulajdonságokkal bír.

3.4. Legkisebb szög regresszió

A legkisebb szög regresszió (LSR) egy viszonylag új eljárás (2004). Tekinthető úgy, mint az előrelépő regresszió egy változata. Látni fogjuk, hogy az LSR hatékony algoritmust biztosít a lasso együtthatóinak kiszámítására.

Az előrelépő regresszió szekvenciálisan építi fel a modellt. Minden lépésben egyetlen elemet ad az aktuális halmazhoz, azt a változót, amelyik a legjobban illeszkedik. Ezután újra kiszámolja a legkisebb négyzetes együtthatókat az új halmaz figyelembevételével.

Az LSR is hasonló stratégiát alkalmaz, avval a különbséggel, hogy „csak annyit enged egy magyarázó változónak, amennyit az megérdemel". Az első lépésben felismeri azt a változót, amelyik a legjobban korrelált a mért változóval. Ahelyett, hogy illesztené ezt a változót, folytonosan mozgatni kezdi az együtthatóját a legkisebb négyzetes érték felé (ezzel csökkentve a reziduálissal vett korrelációjának abszolút értékét). Amint egy másik változó korrelációja a reziduálissal meghaladja az előző változóval számított értéket, az eljárás megáll. A második változót is hozzáveszi az aktív halmazhoz. Ezután a beválasztott két elem együtthatóit együtt változtatja úgy, hogy csökkenjen azok korrelációja a reziduálissal. Ez az eljárás addig folytatódik, amíg végül az összes változó bekerül az aktív halmazba. A végeredmény a legkisebb négyzetes illesztés. Az algoritmus lépései részletesen:

A legkisebb szög regresszió algoritmusa

1. Standardizáljuk a magyarázó változókat, hogy normájuk egységnyi, várható értékük pedig nulla legyen. Kezdetben a reziduális: $r = y - \bar{y}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.

2. Keressük meg azt az x_j változót, amelyik a legjobban korrelált a reziduálissal.

3. Közelítsük β_j értékét x_j legkisebb négyzetes együtthatója, $\langle x_j, r \rangle$ felé addig, amíg valamelyik másik magyarázó változó, x_k korrelációja a reziduálissal el nem éri az x_j-vel számított értéket.

4. Mozgassuk β_j-t és β_k-t az x_j és x_k által meghatározott legkisebb négyzetes együtthatók felé addig, amíg egy következő x_l magyarázó változó korrelációs értéke a reziduálissal el nem éri az előzőekkel számított értéket.

5. Addig folytassuk az eljárást, amíg mind a p változó be nem kerül az aktív halmazba. $\min(N-1, p)$ lépés után elérjük a legkisebb négyzetes illesztést.

A k. lépés elején a változók aktív halmazát jelölje $\mathcal{A}_k$, a változókhoz tartozó együtthatóvektor pedig legyen $\beta_{\mathcal{A}_k}$. Ez utóbbinak $k-1$ darab nemnulla értéke van, és egy darab nulla, ami a legutoljára beválasztott változóhoz tartozik. Ha az aktuális reziduális $r_k = y - X_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$, akkor ehhez a lépéshez tartozó irány

$$\delta_k = (X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1} X_{\mathcal{A}_k}^T r_k.$$

Az együtthatóvektor pedig a következőképpen változik: $\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha\delta_k$. Ha a lépés elején az illesztett vektor $\hat{f}_k$, akkor a lépés után a következő módon változik az értéke: $\hat{f}_k(\alpha) = \hat{f}_k + \alpha u_k$, ahol $u_k = X_{\mathcal{A}_k}\delta_k$, az új illesztési irány. A legkisebb szög elnevezés onnan ered, hogy u_k zárja be a legkisebb szöget az aktív halmazban szereplő magyarázó változók mindegyikével.

Az eljárás folyamán az együtthatók szakaszonként lineárisan változnak. Megjegyezzük, hogy emiatt a tulajdonság miatt nem kell az algoritmus harmadik lépésében apránként megállni és újraellenőrizni a korrelációt, hanem előre meghatározhatjuk a következő lépés pontos hosszát.

A legkisebb szög regresszióhoz hasonlóan a lasso együtthatók is szakaszonként lineárisan változnak. Ha ugyanazon az adatsoron összevetjük, hogy hogyan alakulnak az együtthatók a két eljárás során, láthatjuk, hogy szinte teljesen megegyeznek. Eltérést csak akkor tapasztalunk, amikor egy aktív változó együtthatója áthalad a nullán. Ez az észrevétel motiválta azt az egyszerű módosítást az LSR negyedik lépésében, hogy ha egy

addig nemnulla együttható eléri a nullát, akkor a hozzá tartozó változót vegyük ki az aktív halmazból és számítsuk újra a legkisebb négyzetes irányt. Ez a módosított változat már kiválóan alkalmas a pontos lasso értékek kiszámítására. Műveletigénye ugyanannyi, mint a legkisebb négyzetes illesztésé p változó esetén. A kívánt együtthatókat p lépésben megadja, míg a lassonak esetenként p lépésnél többre van szüksége.

A következőkben adunk némi magyarázatot arra, miért is olyan hasonló ez a két eljárás. Annak ellenére, hogy az LSR algoritmus a korrelációk meghatározására épül, ha a bemenő adatok standardizáltak, egyszerűbb inkább belső szorzatokkal dolgozni. Ez ekvivalens avval, mikor a korrelációkkal számolunk. Tegyük fel, hogy az algoritmus egy állapotában $\mathcal{A}$ az aktív változók halmaza. Az $\mathcal{A}$ halmaz elemeivel és az aktuális reziduálissal számított belső szorzatok abszolút értékét a következő módon írhatjuk le:

$$x_j^T(y - X\beta) = \gamma s_j, \forall j \in \mathcal{A} \tag{13}$$

ahol $y - X\beta$ az aktuális reziduális, $s_j \in \{-1, 1\}$ a belső szorzat előjele, γ pedig a közös érték. Ekkor $|x_k^T(y - X\beta)| \leq \gamma, \forall k \notin \mathcal{A}$ is teljesül. Tekintsük a lasso problémát (10) vektoros alakban:

$$R(\beta) = \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1. \tag{14}$$

Legyen $\mathcal{B}$ a változók aktív halmaza az algoritmus végén, adott λ értékre. Ezekre a változókra $R(\beta)$ differenciálható és a stacionárius állapotokat

$$x_j^T(y - X\beta) = \lambda \text{sign}(\beta_j), \forall j \in \mathcal{B} \tag{15}$$

adja. Ha (13)-at és (15)-öt összehasonlítjuk, láthatjuk, hogy csak akkor térnek el, ha β_j előjele eltér a belső szorzat előjelétől. Ez az oka annak, hogy az LSR és a lasso akkor kezd el eltéréseket mutatni, amikor egy aktív változó együtthatója áthalad a nullán. Ilyen esetben ez a változó nem tesz eleget a (15)-ben szereplő feltételnek, így az kikerül az aktív halmazból.

A lasso és az LSR szabadságfoka

Most arra a kérdésre keressük a választ, hogy hány paramétert, vagy szabadságfokot használunk az egyes algoritmusok alkalmazása során.

Tegyük fel, hogy k elemű részhalmaz felhasználásával szeretnénk lineáris modellt illeszteni. Ha ez a részhalmaz a mért változóktól függetlenül előre meghatározott, akkor definíció szerint az illesztés szabadságfoka k. A klasszikus statisztikában a lineárisan független paraméterek számát nevezzük szabadságfoknak. Most tegyük fel, hogy az illesztés előtt a legjobb részhalmaz algoritmus segítségével határozzuk meg a változók optimális k elemű részhalmazát. Ekkor annak ellenére, hogy a modellben k paraméter szerepel, bizonyos értelemben több mint k szabadságfokot használtunk fel. Ebben az esetben a tényleges

szabadságfok meghatározására egy általánosabb definícióra van szükségünk. A következő módon határozzuk meg egy illesztett $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ vektor szabadságfokát:

$$df(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i). \tag{16}$$

Itt $\mathrm{Cov}(\hat{y}_i, y_i)$ a minta kovariancia a becsült érték és a mért változó között. A (16) definíció intuitívan annyit jelent, hogy minél jobban illeszkednek az adatok, annál nagyobb a kovariancia, és így $df(\hat{y})$ értéke. A szabadságfok kiszámításának ez a fajta megközelítése más modellekben is kiválóan alkalmazható, beleértve azokat a modelleket is, amelyeket adaptívan illesztünk a mért változókra. Könnyen látható, hogy ha k előre rögzített változón végzünk lineáris regressziót, a tényleges szabadságfok $df(\hat{y}) = k$. Hasonlóképpen a ridge regresszió esetén visszakapjuk a (8)-as összefüggést. Mindkét esetben (16) kiértékelése egyszerű, hiszen $\hat{y}$ a mért változó lineáris függvénye. Ha az előző példában vizsgáljuk $df(\hat{y})$ értékét, mikor az illesztés előtt a legjobb részhalmaz algoritmust használtuk, látható, hogy $df(\hat{y}) \geq k$. A legjobb részhalmaz algoritmus tényleges szabadságfokának kiszámítására nincs zárt formulás eljárás.

A LSR és a lasso sokkal simább módon illeszkednek, mint a legjobb részhalmaz algoritmus, így a tényleges szabadságfokuk kiszámítása is sokkal jobban kezelhető. Speciálisan az is megmutatható, hogy az LSR algoritmus k. lépése után az illesztett vektor tényleges szabadságfoka éppen k. A lasso esetében viszont a módosított LSR eljárás gyakran igényel p-nél több lépést, mivel ott a változók ki is kerülhetnek az aktív halmazból. Így ebben az esetben $df(\hat{y})$ a modellben szereplő aktív változók számával egyezik meg.

4. Az algoritmusok illusztrálása egy példa adatsoron

4.1. A diabétesz adatsor

Ebben a fejezetben egy példa adatsor segítségével szemléltetjük az eddig tárgyalt módszereket. Az adatsor 442 diabéteszes beteg adatait tartalmazza. A betegeket a következő szempontok szerint vizsgálták: életkor, nem, testtömeg index (BMI), vérnyomás (BP), valamint megmérték a vérszérum hat összetevőjének szintjét. Ez a tíz adat alkotja a magyarázó változókat, amelyeket majd standardizálunk, hogy várható értékük nulla, normájuk pedig egységnyi legyen. Egy év elteltével az orvosok újabb vizsgálatot tettek és feljegyezték, hogy mennyit romlott az alanyok állapota az első vizsgálat óta. Ezeknek az adatoknak a számokban kifejezett értéke alkotja a magyarázott változót.

					. . . . . A vérszérum adatai.						
	Életkor	Nem	BMI	BP	tc	ldl	hdl	tch	ltg	glu	Célváltozó
Betegek	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

1. táblázat. A minta adatsor egy részlete.

A 1. táblázat a példa adatsor egy részletét tartalmazza még a standardizálás előtt. A 2. táblázat a magyarázó változók korrelációit mutatja. Láthatjuk, hogy a legjobban korrelált változók például a koleszterin (tc) és az alacsony sűrűségű lipoprotein (ldl). Ha az adatsoron a legkisebb négyzetek módszerét alkalmazzuk, az együtthatók a 3. táblázat szerint alakulnak.

Az együtthatókhoz tartozó Z-pontszám kiszámítása után azt kapjuk, hogy a diabétesz betegség kialakulása szoros kapcsolatban áll a vérben lévő koleszterin (tc) és lamotrigin (ltg) szinttel. Ezen kívül a nem és a testtömeg index is meghatározó szerepet játszik.

Ekkor az illesztett modell reziduálisait az 1. ábra mutatja. Látható, hogy a hiba szórása itt még relatíve nagy.

	Életkor	Nem	BMI	BP	tc	ldl	hdl	tch	ltg
Nem	0.174								
BMI	0.185	0.088							
BP	0.335	0.241	0.395						
tc	0.260	0.035	0.249	0.242					
ldl	0.219	0.142	0.261	0.185	0.896				
hdl	-0.075	-0.379	-0.366	-0.178	0.051	-0.196			
tch	0.203	0.332	0.413	0.257	0.542	0.659	-0.738		
ltg	0.270	0.149	0.446	0.393	0.515	0.318	-0.398	0.617	
glu	0.301	0.208	0.388	0.390	0.325	0.290	-0.273	0.417	0.464

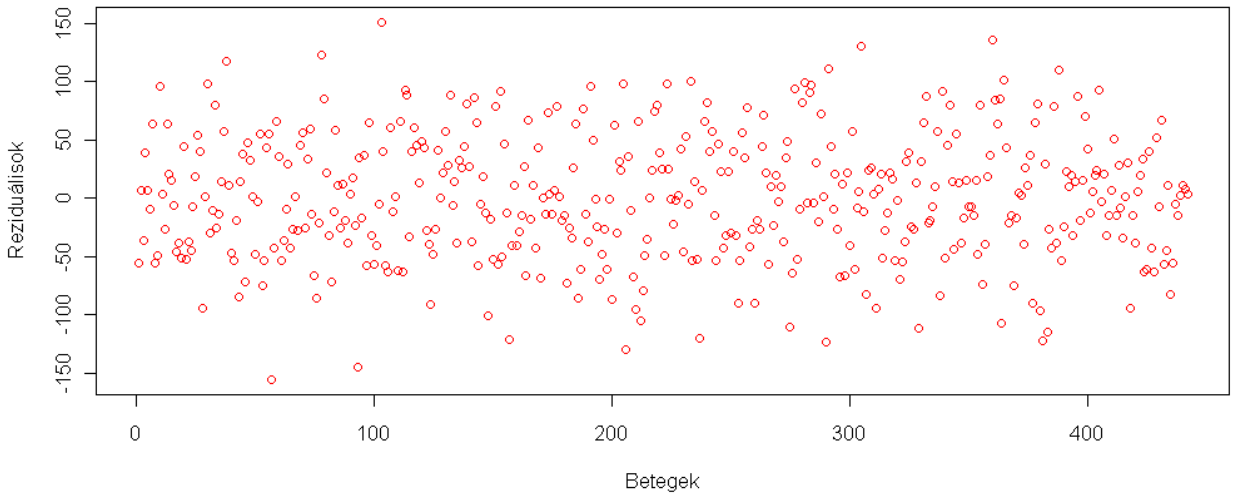
2. táblázat. A változók korreláció mátrixa.

Változó	Életkor	Nem	BMI	BP	tc	ldl	hdl	tch	ltg	glu
Együttható	-10.0	-239.8	519.8	324.3	-792.1	476.7	101.0	177.0	751.2	67.6

3. táblázat. A legkisebb négyzetes együtthatók.

Változó	Életkor	Nem	BMI	BP	tc	ldl	hdl	tch	ltg	glu
Z-pontszám	-0.338	-0.866	0.877	0.428	-2.133	0.778	-0.083	0.090	1.408	-0.160

4. táblázat. A változókhoz tartozó Z-pontszám.



1. ábra. A reziduálisok értéke a legkisebb négyzetek módszerénél.

4.2. A Mallows-féle Cp statisztika és a keresztvalidáció

Eddig a részhalmaz kiválasztó algoritmusok és a zsugorító módszerek esetén modellek egy sorozatát adtuk meg, és a paraméterek becslésének egész útvonalát. Az alkalmazások során egy alkalmas modellre és a paraméterek egyetlen $\hat{\beta}$ becslésére van szükségünk. Hogyan dönthetjük el, hogy melyik modellt válasszuk? Erre a kérdésre az egyik lehetséges választ a Mallows-féle Cp statisztika adja. Ennek segítségével értékelhetjük a különböző regressziós modellek illeszkedését. Ha a változók K elemű halmazának egy $P < K$ elemű részhalmazát választjuk ki, a következő módon számíthatjuk ki a Cp statisztika értékét ezen a részhalmazon:

$$Cp = \frac{\sum_{i=1}^{N}(Y_i - Y_{Pi})^2}{S^2} - N + 2P,$$

ahol Y_{Pi} a célváltozó i. koordinátájának becsült értéke a változók P elemű részhalmazának segítségével, S^2 a teljes, K elemű halmazon végzett regresszió során a reziduálisok négyzetének átlaga, N pedig a minta mérete [5]. A modellek sorozatából választhatjuk azt az elemet, amelyik a legkisebb Cp értéket adja.

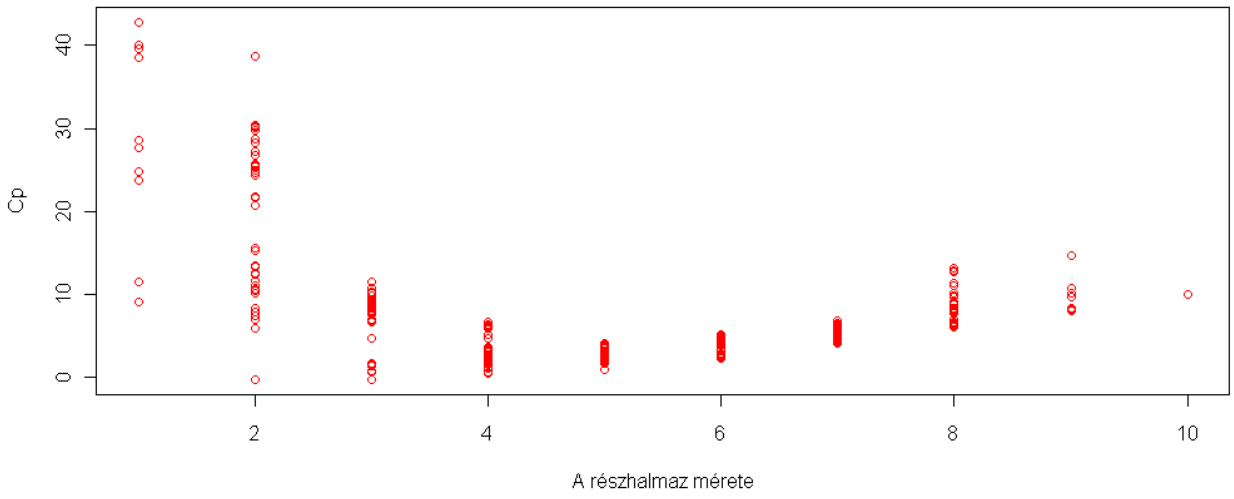
Egy másik eljárás a modellek illeszkedésének tesztelésére a k-szoros keresztvalidáció (CV). Ennek lényege, hogy az adathalmazt véletlenszerűen felosztjuk k darab, nagyjából egyenlő méretű részhalmazra. Az első $k-1$ halmazon megépítjük a modellünket, a k-adikon pedig kiértékeljük a pontosságát. Ezt nevezzük egy ismétlésnek. Ezután ugyanezt végrehajtjuk úgy, hogy az első $k-2$ és a k. halmazon építjük a modellünket és a $k-1$-ediken értékeljük ki. Ugyanezt ismételve k modellépítés után a k darab pontossági érték átlaga lesz a becslésünk [6]. A módszer egyik előnye, hogy az eljárásban minden megfigyelést

használunk mind a modellépítés, mind a kiértékelés során. A kiértékelésekhez minden adatot pontosan egyszer. A gyakorlatban az ismétlések száma általában tíz, az illeszkedés mértékét pedig az átlagos négyzetes hiba (MSE) nagyságával jellemezzük, ami a következő összeg:

$$MSE(\hat{y}) = \sum_{i=1}^{N} \frac{1}{N}(\hat{y}_i - y_i)^2.$$

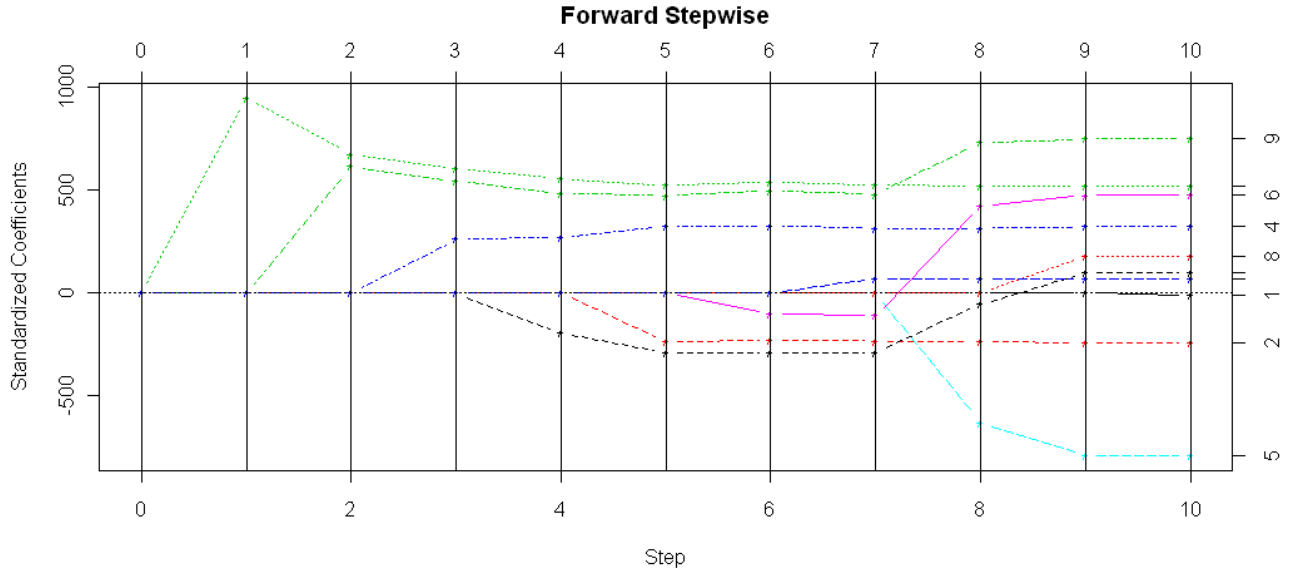
4.3. A részhalmaz kiválasztó algoritmusok alkalmazása

A 2.1 fejezetben láttuk, hogy ha a legjobb részhalmaz algoritmust a reziduálisok négyzetösszegének minimalizálására használjuk, annál jobb illeszkedést kapunk, minél több változót veszünk be az aktív halmazba. Így ez a módszer nem alkalmas a változók legnagyobb hatással bíró k elemű részhalmazának megtalálására. A reziduálisok négyzetösszege helyett tekintsük a Mallows-féle Cp statisztikát. Ahogy az 1. ábra is mutatja, Cp értéke már nem csökken automatikusan attól, ha egy új változót veszünk be a modellbe.



2. ábra. A legjobb részhalmaz algoritmus a Mallows-féle Cp statisztika szempontjából. A legkisebb Cp értéket a BMI, BP, ltg változókat tartalmazó háromelemű részhalmaz adja.

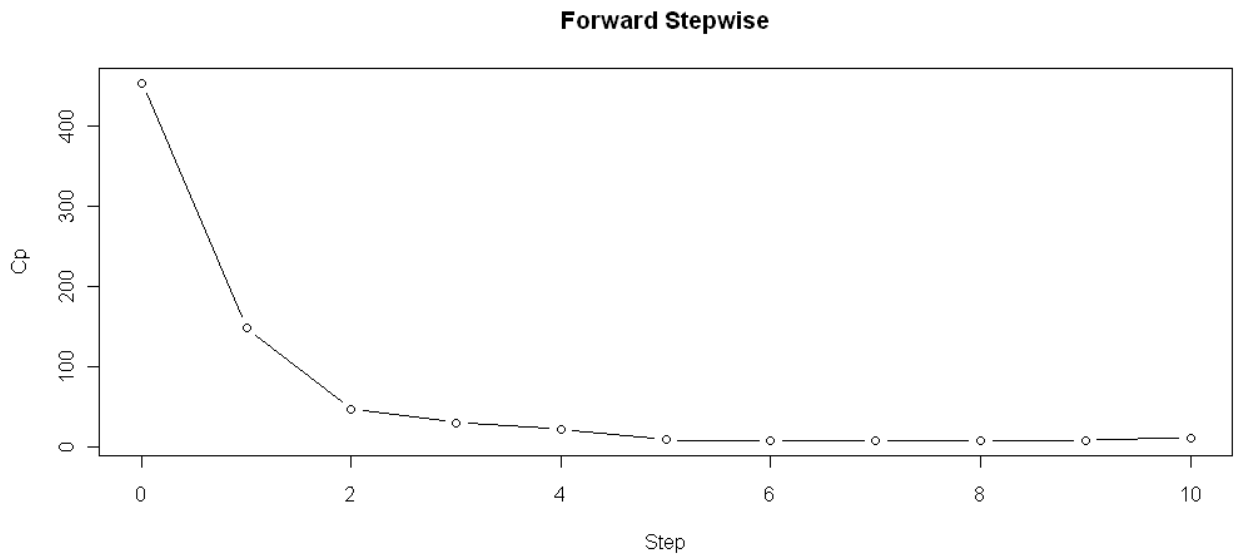
Az előrelépő regresszió esetében is a Cp statisztikát használjuk a megfelelő modell kiválasztásához. A 3. ábra azt mutatja, hogyan változnak az előrelépő regresszió során az együtthatók, míg végül a 10. lépésben megkapjuk a legkisebb négyzetes értékeket. Az 4. ábrán jól látható, hogy az algoritmus ötödik lépése után a Cp statisztika szempontjából már nem javul jelentős mértékben az illeszkedés attól, ha új változót adunk a modellhez. Így e szerint az eljárás szerint a legjobb modellt akkor kapjuk, ha a változók közül a testtömeg indexet (BMI), a vérnyomást (BP), a nemet, a magas sűrűségű lipoproteint (hdl) és a lamotrigint (ltg) tartalmazó ötelemű részhalmazt tartjuk meg, a 5. táblázatban szereplő együtthatókkal. A 5. ábra azt mutatja, hogyan alakulnak a reziduálisok, ha ezt a modellt alkalmazzuk. Látható, hogy a szórás nem csökkent jelentős mértékben.



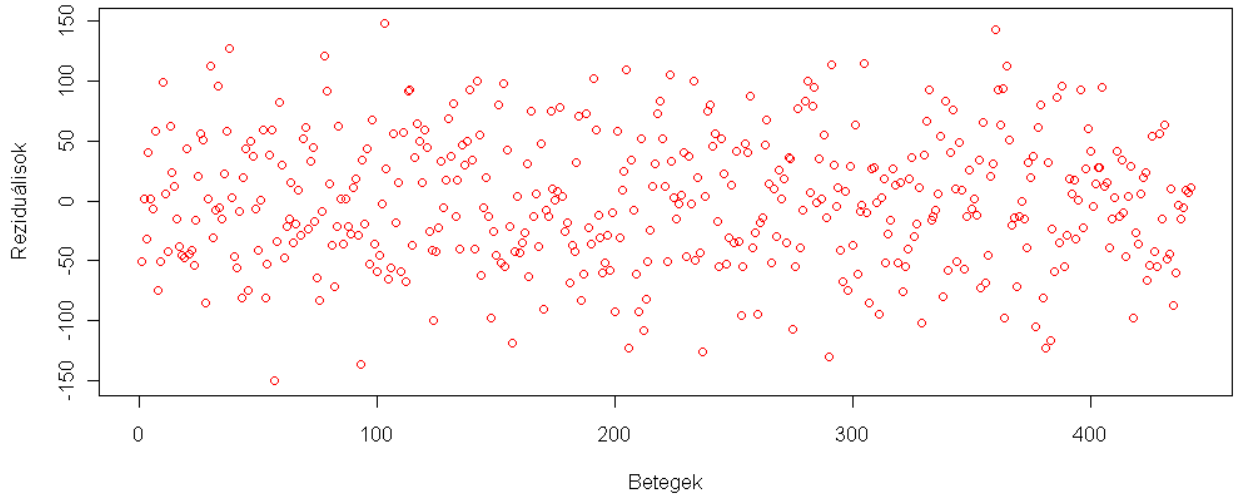
3. ábra. Az együtthatók útja az előrelépő regresszió során.

Változó	Életkor	Nem	BMI	BP	tc	ldl	hdl	tch	ltg	glu
Együttható	0	-235.77	523.56	326.23	0	0	-289.11	0	474.29	0

5. táblázat. Az előrelépő regresszió együtthatói.

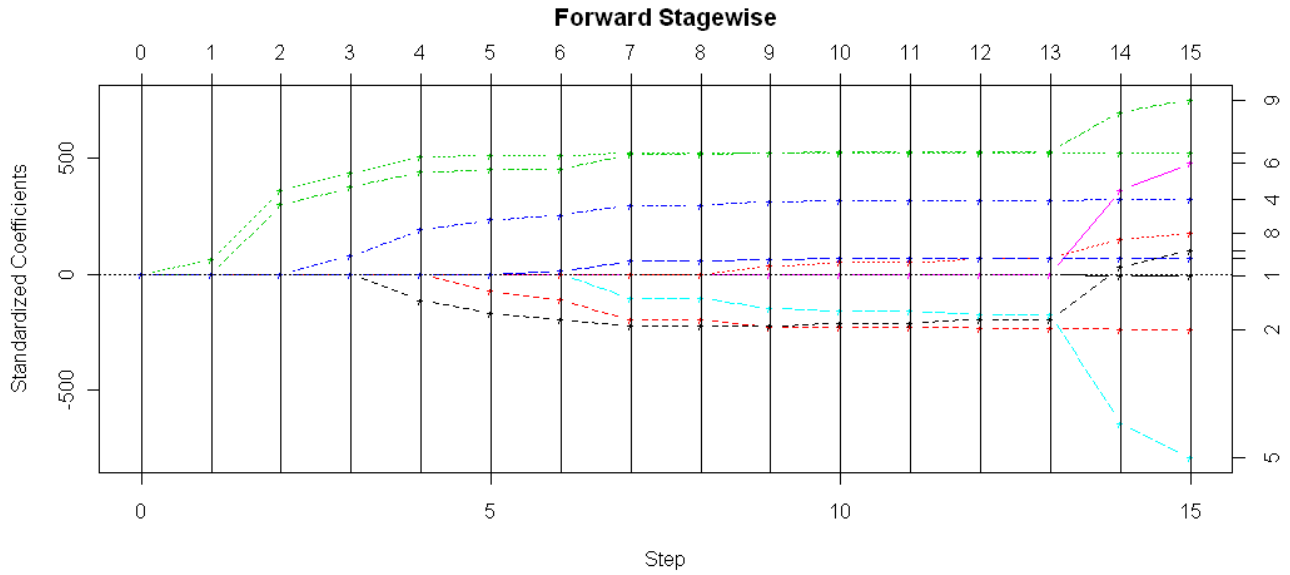


4. ábra. A Cp statisztika értéke az előrelépő regresszió során kapott modellsorozat elemein.



5. ábra. A reziduálisok értéke az előrelépő regresszióval meghatározott modell esetében.

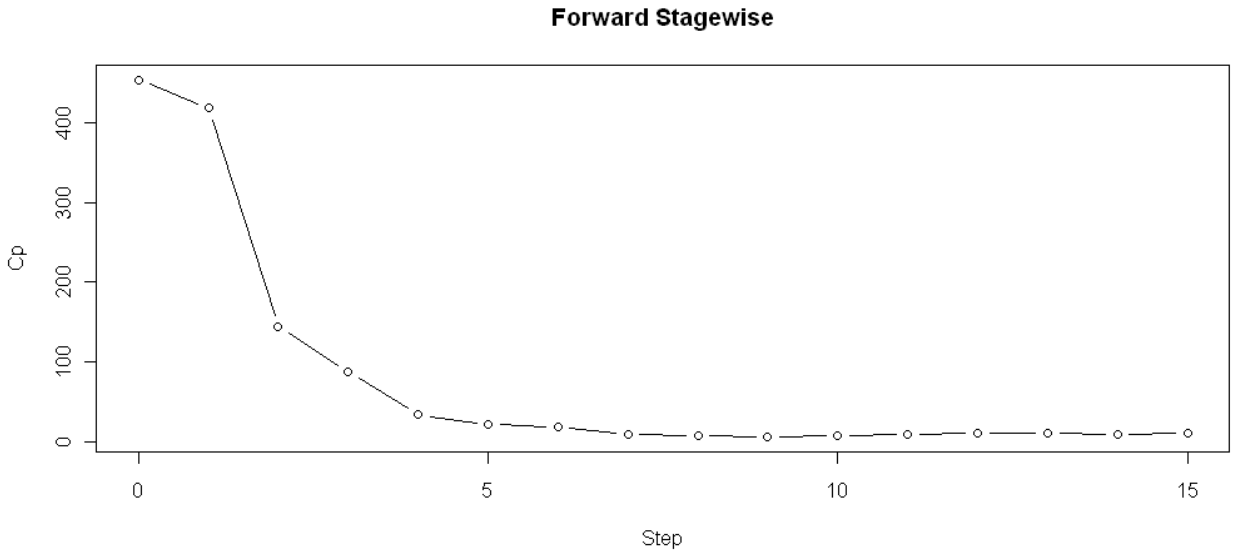
A szakaszonkénti regresszióban is az előzőekhez hasonlóan járunk el. Távolról szemlélve az együtthatók útja is több hasonlóságot mutat (6. ábra). Ennek az algoritmusnak azonban már több lépésre van szüksége leállásig. A 7. ábráról azt olvashatjuk le, hogy az algoritmus hetedik lépése után a Cp statisztika értéke már nem csökken jelentős mértékben. Ebben a lépésben a modell a Nem, BMI, BP, tc, hdl, ltg, glu változókat tartalmazza, a 6. táblázatban szereplő együtthatókkal.



6. ábra. A szakaszonkénti regresszió a diabétesz adatsor esetében 15 lépés után éri el a legkisebb négyzetes együtthatókat.

Változó	Életkor	Nem	BMI	BP	tc	ldl	hdl	tch	ltg	glu
Együttható	0	-197.75	522.26	297.15	-103.94	0	-223.92	0	514.74	54.76

6. táblázat. A szakaszonkénti regresszió együtthatói.



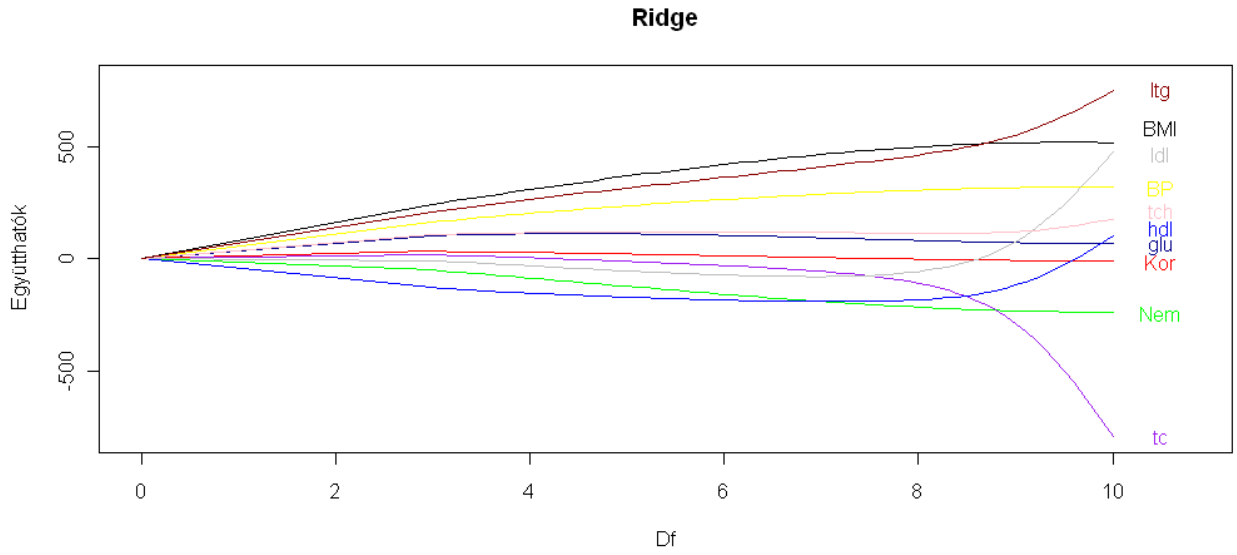
7. ábra. A szakaszonkénti regresszió során a Cp statisztika értéke.

4.4. A „Shrinkage” módszerek alkalmazása

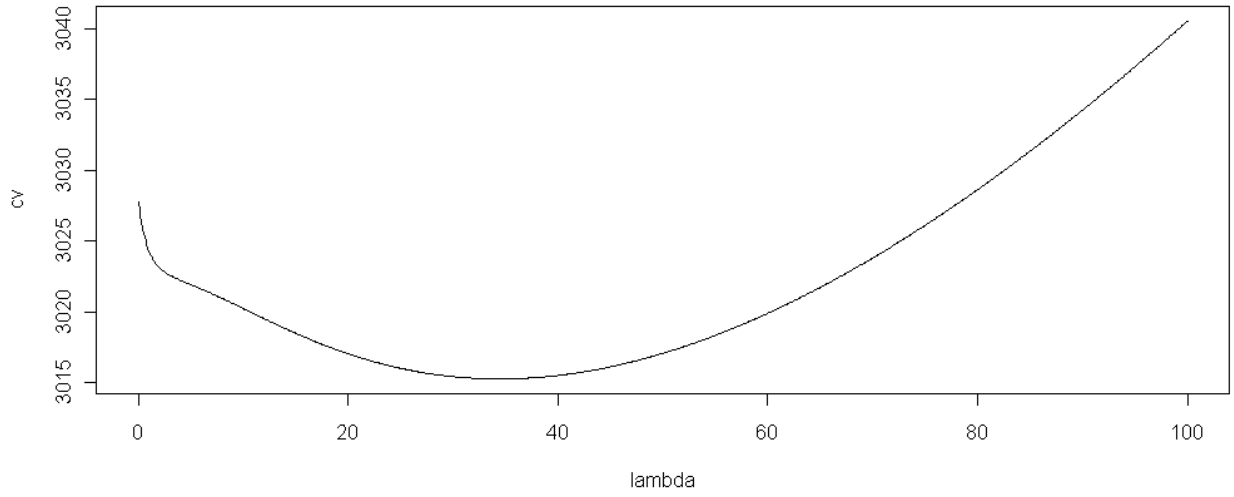
A ridge regresszió merőben különbözik az eddigi eljárásoktól. Ez a módszer lehetőséget ad arra, hogy nagyobb mértékben csökkentsük a szórást. A paraméter értékét és az együtthatókat most keresztvalidáció segítségével határozzuk meg. A kapott modellben most minden változó nullától különböző együtthatóval szerepel.

Változó	Életkor	Nem	BMI	BP	tc	ldl	hdl	tch	ltg	glu
Együttható	-0.3	-213	497	305	-100	-61	-185	114	458	83

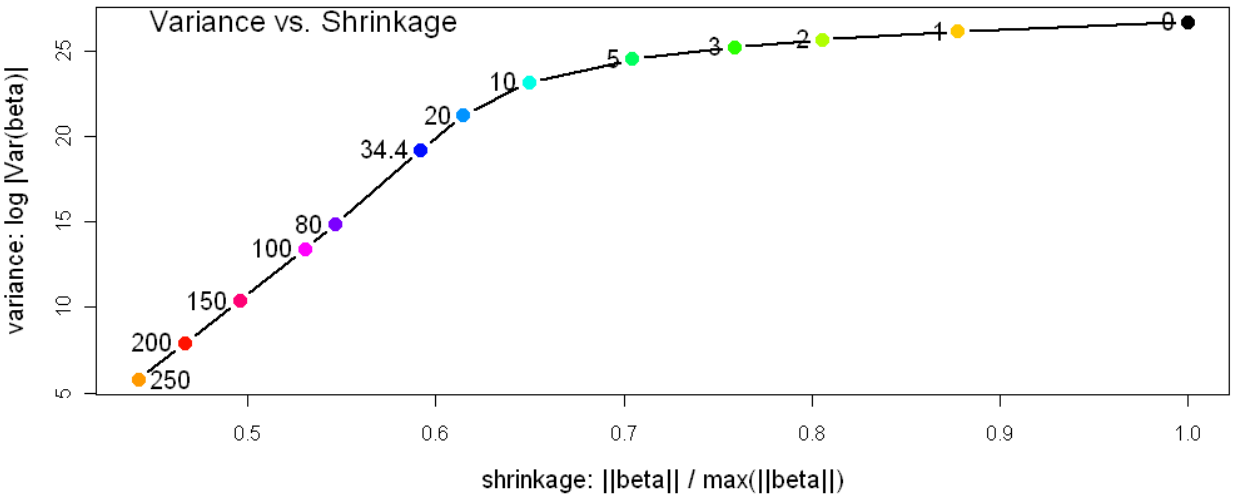
7. táblázat. A ridge regresszió során a kiválasztott modellben szereplő változók együtthatói.



8. ábra. Az ábra a ridge regresszió együtthatóinak útját mutatja, miközben változtatjuk a szabadságfokot.



9. ábra. A keresztvalidáció értékei a ridge regresszió esetén. Az optimális lambda értéknek $\lambda = 34.4$-et kaptuk.

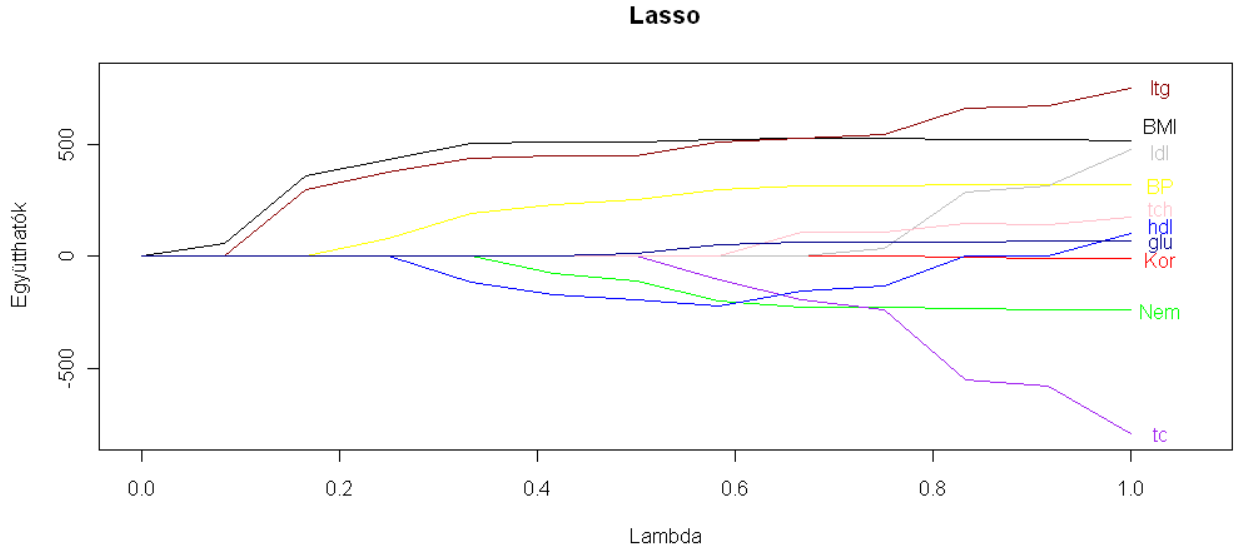


10. ábra. Az ábrán szereplő számok a λ értékét mutatják, az x tengelyről pedig a hozzájuk tartozó zsugorítás mértékét olvashatjuk le. Az y tengelyen az együtthatók szórását láthatjuk. Jól látszik, hogy minél nagyobb λ értéket választunk, annál nagyobb a zsugorítás mértéke és annál kisebb szórást kapunk.

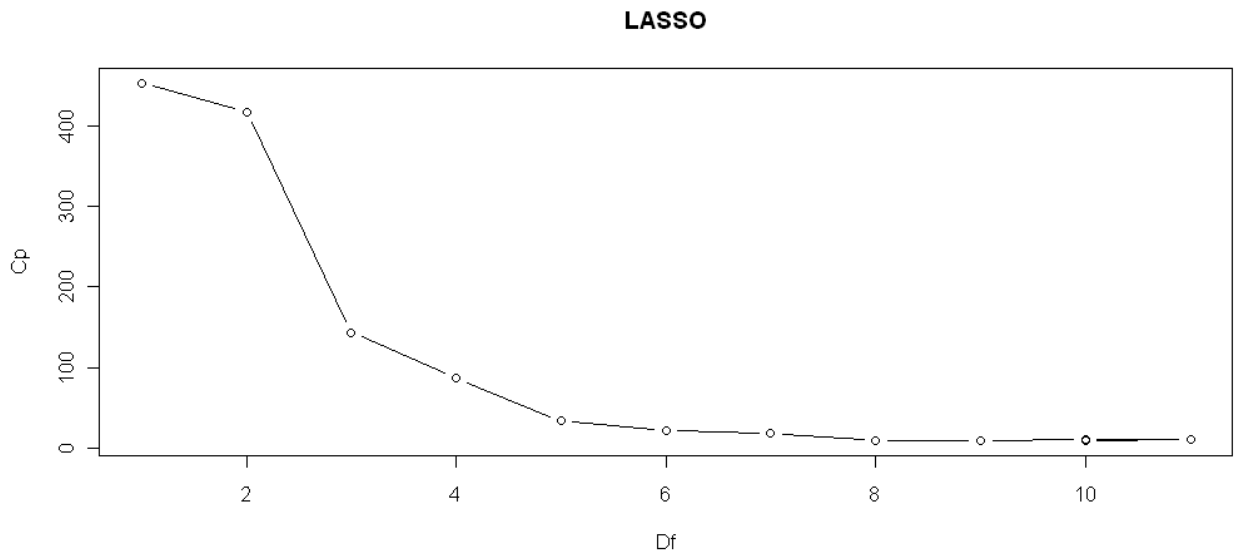
A lasso algoritmus esetében a megfelelő modell kiválasztásánál hasonló eredményre jutunk mind a keresztvalidáció, mind a Cp statisztika értékének vizsgálata esetén. Mindkét esetben azt tapasztaljuk, hogy körülbelül a $\lambda = 0.4$ után már nem tudjuk jelentős mértékben csökkenteni sem a Cp statisztika értékét, sem az átlagos négyzetes hiba értékét a keresztvalidáció esetében. Ennek következtében a modellsorozatból válasszuk a $\lambda = 0.4$ értékhez tartozó elemet. Ekkor a változók együtthatói a 8. táblázatban szereplő értékek alapján alakulnak. Láthatjuk, hogy a ridge regressziótól eltérően most néhány változó éppen nulla együtthatóval szerepel a modellben.

Változó	Életkor	Nem	BMI	BP	tc	ldl	hdl	tch	ltg	glu
Együttható	0	-74.9	511.3	234.1	0	0	-169.7	0	450.6	0

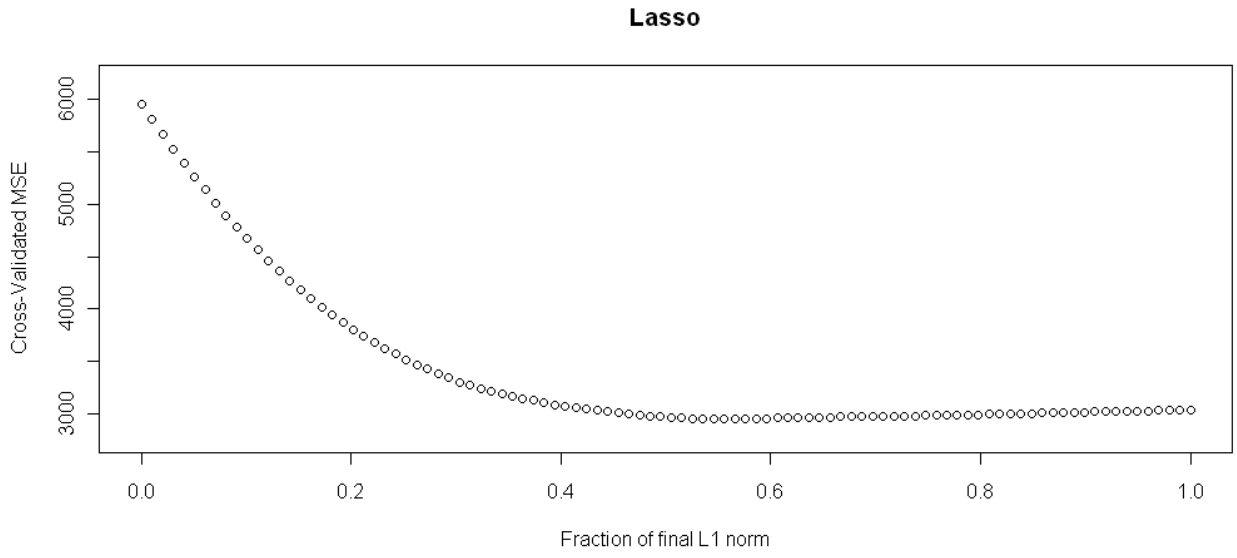
8. táblázat. A lasso eljárás során a kiválasztott modellben szereplő változók együtthatói.



11. ábra. Az együtthatók útja a lasso algoritmus során.

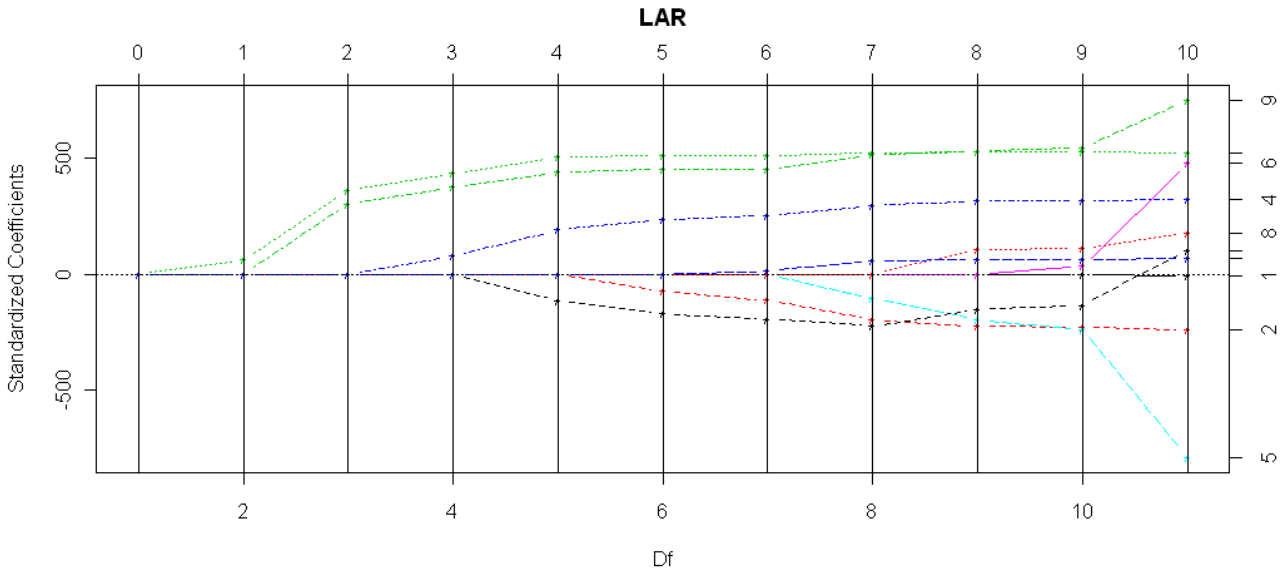


12. ábra. A Cp statisztika értékének változása a szabadságfok függvényében.

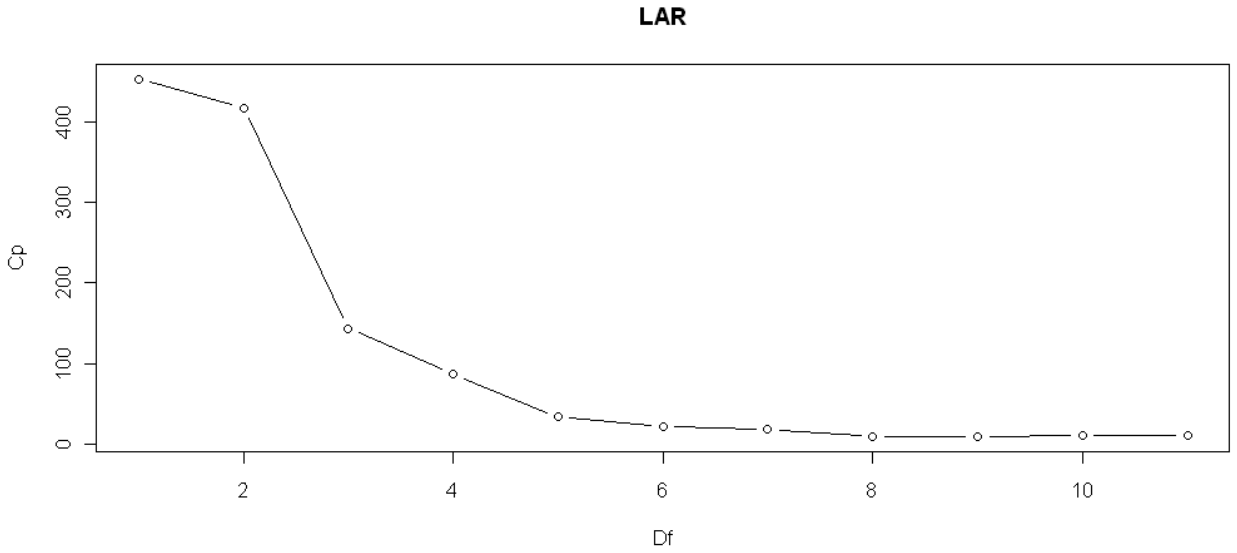


13. ábra. A keresztvalidációs eljárás értékei a zsugorítás mértékének függvényében.

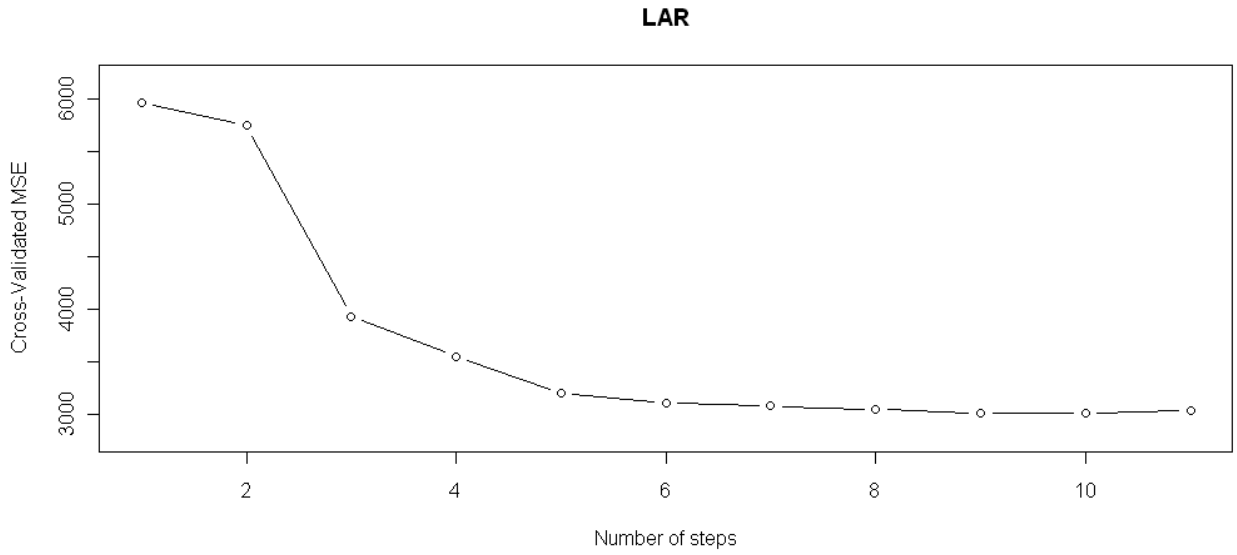
A legkisebb szög regresszió esetében az előzőekkel tökéletesen megegyező eredményekhez jutunk. Ennek az az oka, hogy az eljárás szinte ugyanaz, mint az előbb. Eltérés csak azután tapasztalhatunk, ha valamelyik együttható átlépi a nullát. Ahogy a 15. és a 16. ábra is mutatja, mind a keresztvalidációs eljárás, mind a Cp statisztikával történő vizsgálat esetén ugyanazt a modellt érdemes választani, mint a lasso esetében.



14. ábra. A legkisebb szög regresszió együtthatói.



15. ábra. A Cp statisztika értéke a legkisebb szög regresszió esetén.



16. ábra. A keresztvalidációs eljárás értékei az algoritmus lépései során.

5. Összefoglalás

Az algoritmusok példa adatsoron való alkalmazása után elmondhatjuk, hogy a várakozásainknak megfelelő eredményeket kaptunk.

A részhalmaz kiválasztó algoritmusok esetében azt tapasztaltuk, hogy az előrelépő regresszionak a szakaszonkénti regresszióhoz képest kevesebb számítási költségre volt szüksége és ennek ellenére egy jobban illeszkedő modellt talált.

A zsugorító módszerek alkalmazásakor sikerült csökkentenünk a teljes modell szórását. Míg a ridge regresszió során a kapott modellben minden változó nullától különböző együtthatóval szerepel, a lasso eljárás segítségével sikerült néhány együtthatót kinullázni. A lasso és a legkisebb szög regresszió pontosan ugyanazt az eredményt adta. Az általuk meghatározott modellben a változóknak ugyanaz az ötelemű részhalmaza szerepel, mint amit az előrelépő regresszió esetében kaptunk. A különbség csupán annyi, hogy a változókhoz tartozó együtthatók abszolút értéke a lasso esetében kisebb. Ezen a példán jól megfigyelhető a különbség a részhalmaz kiválasztó algoritmusok és a zsugorításos módszerek között. Minden szempontot figyelembe véve a célváltozó közelítésére a lasso eljárással létrehozott modell felel meg a legjobban.

Hivatkozások

[1] Jerome Friedman T.Hastie, R. Tibshirani. *The Elements of Statistical Learning*. Springer, 2008.

[2] Iain Johnstone Bradley Efron, Trevor Hastie and Robert Tibshirani. Least angle regression. `http://www.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf`, 2003.

[3] Trevor Hastie and Brad Efron. Least angle regression, lasso and forward stagewise. `http://cran.r-project.org/web/packages/lars/lars.pdf`, 2012.

[4] Standard score. `http://en.wikipedia.org/wiki/Standard_score`.

[5] Mallows' cp. `http://en.wikipedia.org/wiki/Mallows'_Cp`.

[6] Cross-validation. `http://en.wikipedia.org/wiki/Cross-validation_(statistics)`.