Simítás, spline-regresszió, additív modellek

Ferenci Tamás, tamas.ferenci@medstat.hu

2020. június 16.

Tartalom

Tartalom

Tartalom

Tartalom

Előszó

Téma: simítás, spline-regresszió, additív modellek

Minden visszajelzést örömmel veszek a tamas.ferenci@medstat.hu email-címen

Tartalom

Tartalom 2.

Tartalom 3.

A LOESS simító

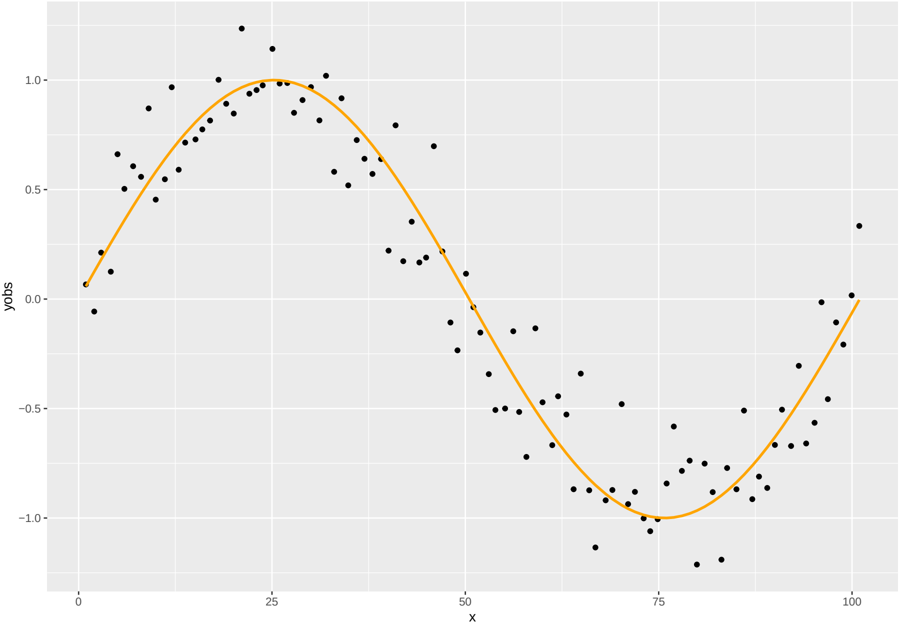
A LOESS simítóról lesz szó

Subsection 1

Motiváció

Motiváció 1.

Motiváció

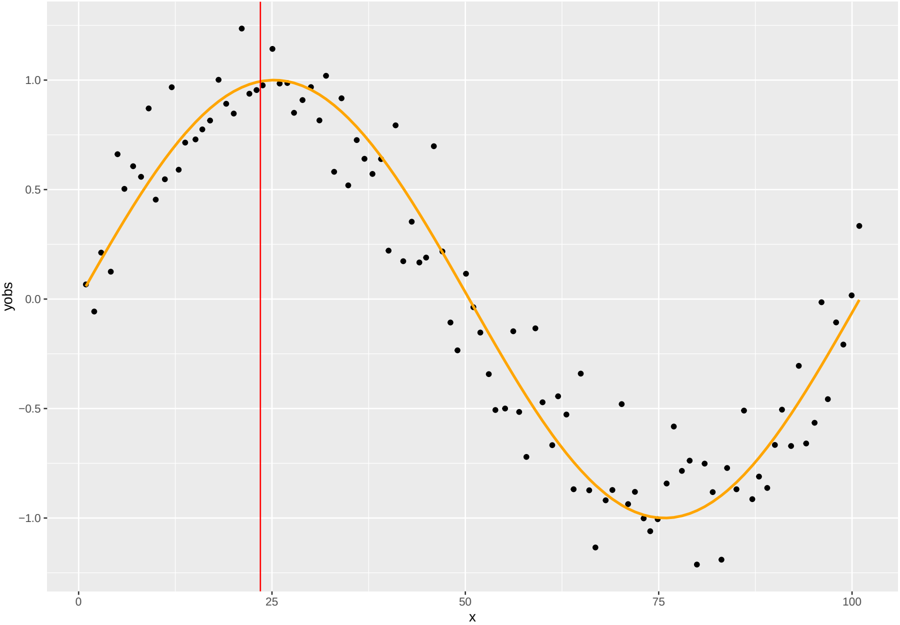


Subsection 2

A LOESS simító alapgondolata

A LOESS simító alapgondolata 1.

A LOESS simító alapgondolata

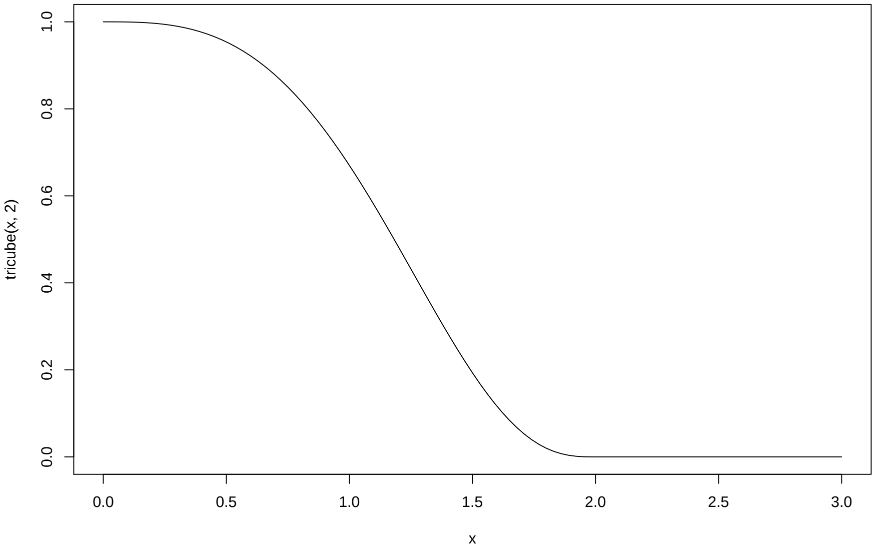


Subsection 3

Lokalitás

Lokalitás 1.

Lokalitás

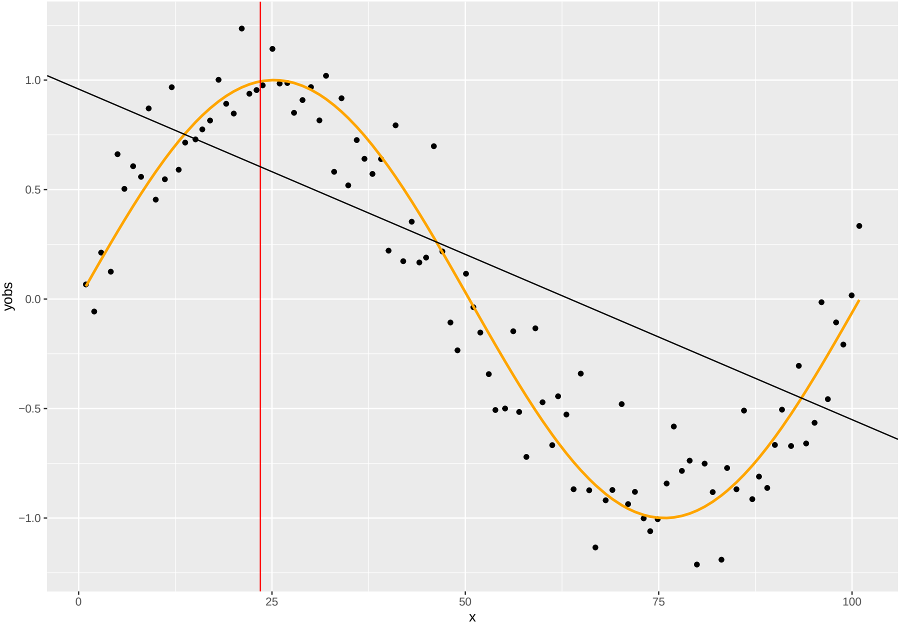


Subsection 4

Polinomiális regresszió

Polinomiális regresszió 1.

Polinomiális regresszió

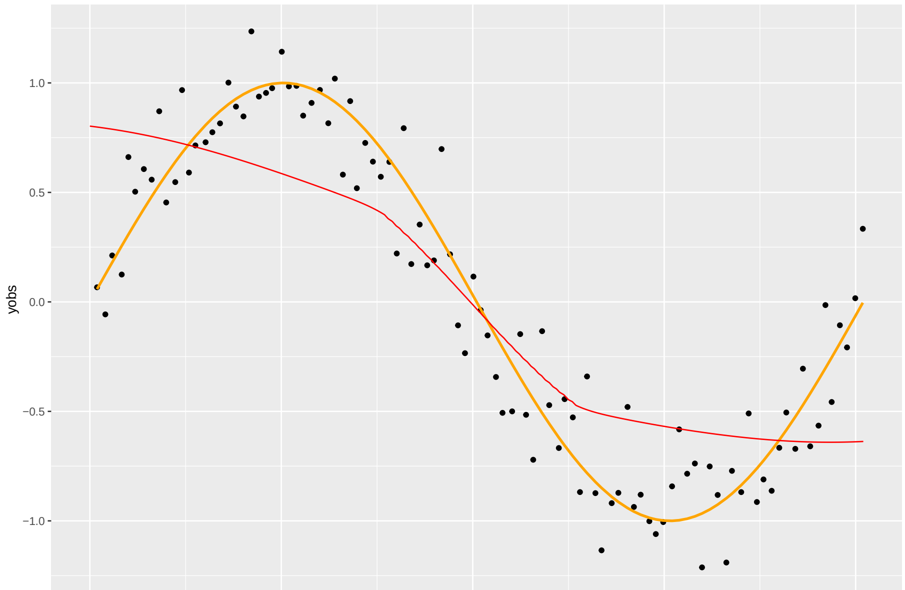


Subsection 5

Összerakva az építőelemeket: lokális polinomiális regressziókkal közelítés

Összerakva az építőelemeket: lokális polinomiális regressziókkal közelítés 1.

Összerakva az építőelemeket: lokális polinomiális regressziókkal közelítés

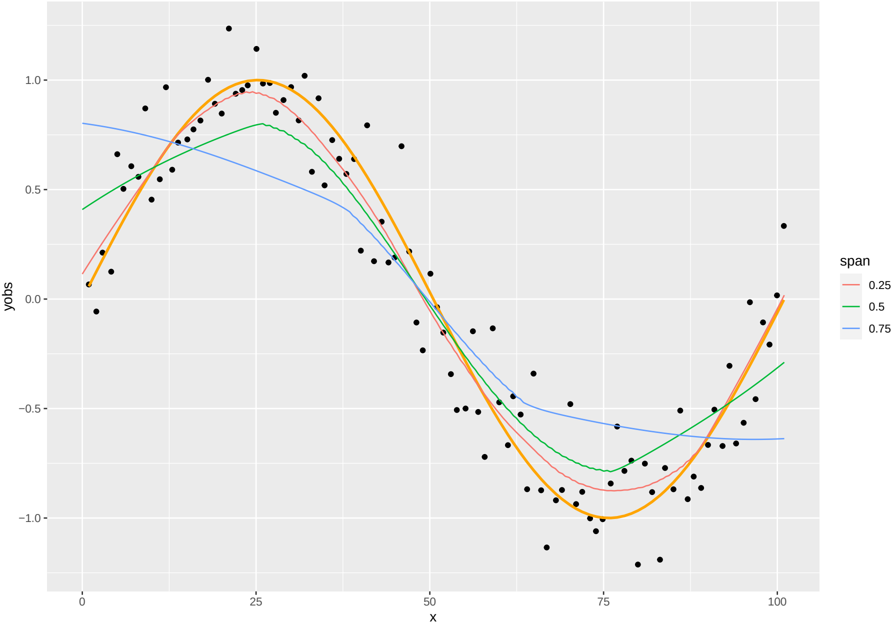


Subsection 6

A paraméterek megválasztásának hatása: lokalitás

A paraméterek megválasztásának hatása: lokalitás 1.

A paraméterek megválasztásának hatása: lokalitás 2.



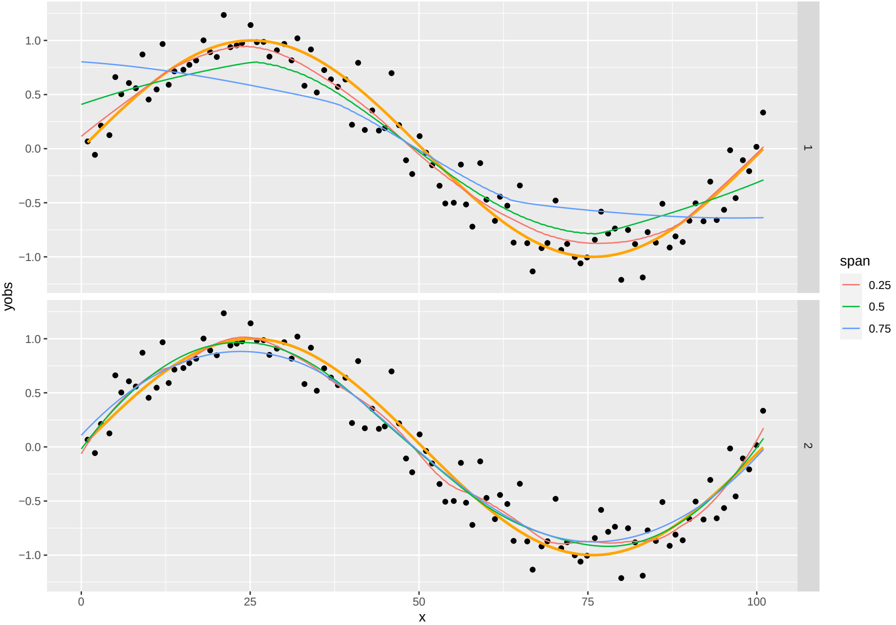
Subsection 7

A paraméterek megválasztásának hatása: a polinom fokszáma

Kitérő: polinomiális regresszió illesztésének szintaktikája R alatt 1.

Polinom fokszámának változtatása 1.

Polinom fokszámának változtatása

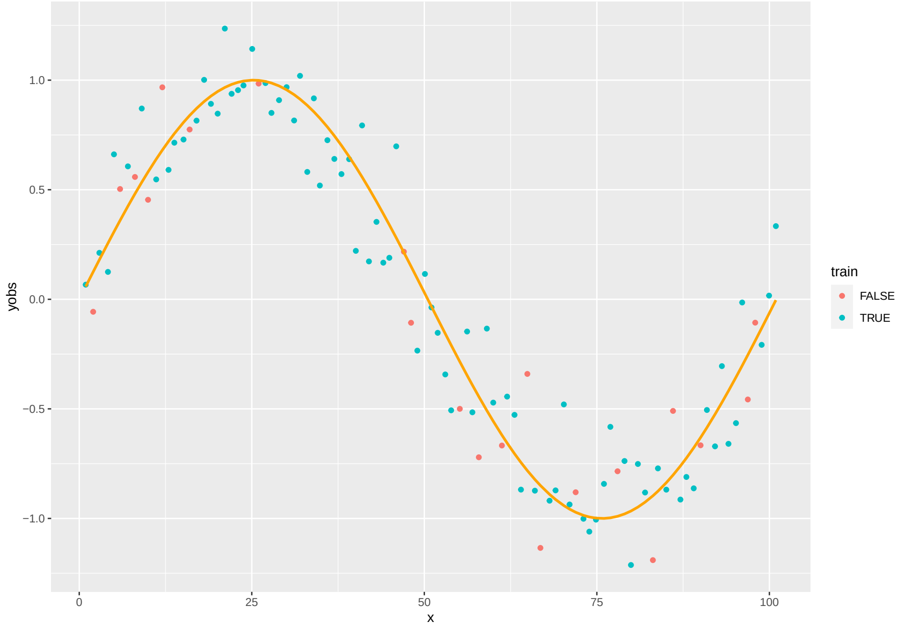


Subsection 8

A paraméterek megválasztása

A paraméterek megválasztása

A paraméterek megválasztása



Tartalom

Tartalom

Tartalom

Subsection 1

A regresszió

A regresszió

A regresszió legtöbb alkalmazott statisztikai terület talán legfontosabb eszköze

Regresszió: változók közti kapcsolat (illetve annak becslése minta alapján)

„Kapcsolat" formalizálása: függvény a matematikai fogalmával, tehát keressük az

$$Y = f\left(X_1, X_2, \ldots, X_p\right) + \varepsilon = f(\mathbf{X})$$

függvényt

(Y eredményváltozó, X_i-k a magyarázó változók)

Subsection 2

Regresszió becslése mintából

Regresszió becslése mintából 1.

Paraméteres regresszió: ha *a priori* feltételezzük, hogy az f függvény valamilyen – paraméterek erejéig meghatározott – függvényformájú (az „alakja" ismert), és így a feladat e paraméterek becslésére redukálódik

Tipikus példa a **lineáris regresszió**:
$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = \mathbf{X}^T\boldsymbol{\beta}$, így $Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$

Ha rendelkezésre állnak az $\left\{y_i, \mathbf{x}_i\right\}_{i=1}^n$ megfigyeléseink a háttéreloszlásra, akkor e mintából megbecsülhetjük a paramétereket például **hagyományos legkisebb négyzetek** (OLS) módszerével:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}} \sum_{i=1}^n \left[Y_i - \mathbf{X}_i^T\mathbf{b}\right]^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$$

Itt tehát $\mathbf{X}$ az a mátrix, amiben a magyarázó változók elé egy csupa 1 oszlopot szúrtunk, a neve **modellmátrix** vagy design mátrix

Subsection 3

Paraméteres és nem-paraméteres regresszió

De cserében mindig ott lebeg felettünk a kérdés, hogy a függvényformára *jó feltételezést* tettünk-e (hiszen ez nem az adatokból következik, ezt „ráerőszakoljuk” az adatokra)

(Persze ezért van a modelldiagnosztika)

A nem-paraméteres regresszió *flexibilis*, olyan értelemben, hogy minden a priori megkötés nélkül követi azt, ami az adatokból következik (a valóság ritkán lineáris?)

Cserében nehezebb becsülni, és nem kapunk analitikus – jó esetben valamire hasznosítható – regressziós függvényt, nem lehet értelmesen interpolálni és extrapolálni („fordul a kocka” a paraméteres esethez képest)

Subsection 4

A lineáris regresszió kibővítése, nemlinearitások

A lineáris regresszió kibővítése, nemlinearitások 1.

Maradva a paraméteres keretben, arra azért mód van, hogy a függvényformát kibővítsük (és így flexibilisebbé tegyük)

Ezzel a különféle **nemlineáris regressziókhoz** jutunk el

E nemlinearitásoknak két alaptípusa van

- Változójában nemlineáris modell (pl. $\beta_0 + \beta_1 x + \beta_2 x^2$): csak a szó „matematikai értelmében” nemlineáris, ugyanúgy becsülhető OLS-sel
- Paraméterében nemlineáris modell (pl. $\beta_0 x_1^{\beta_1} x_2^{\beta_2}$): felrúgja a lineáris struktúrát, így érdemileg más, csak linearizálás után, vagy NLS-sel becsülhető

Mi most az első esettel fogunk foglalkozni

Az itt látott „polinomiális regresszió” valóban nagyon gyakori módszer a flexibilitás növelésére

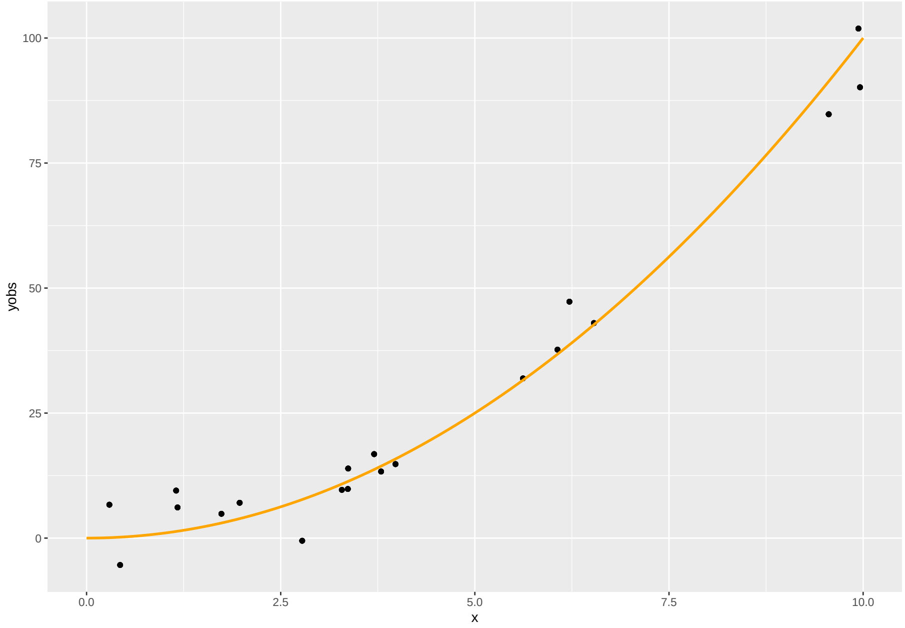
Subsection 5

Egy példa

Egy példa

Tekintsünk most egy másik példát, egy zajos másodfokú függvényt, kevesebb pontból:

Egy példa

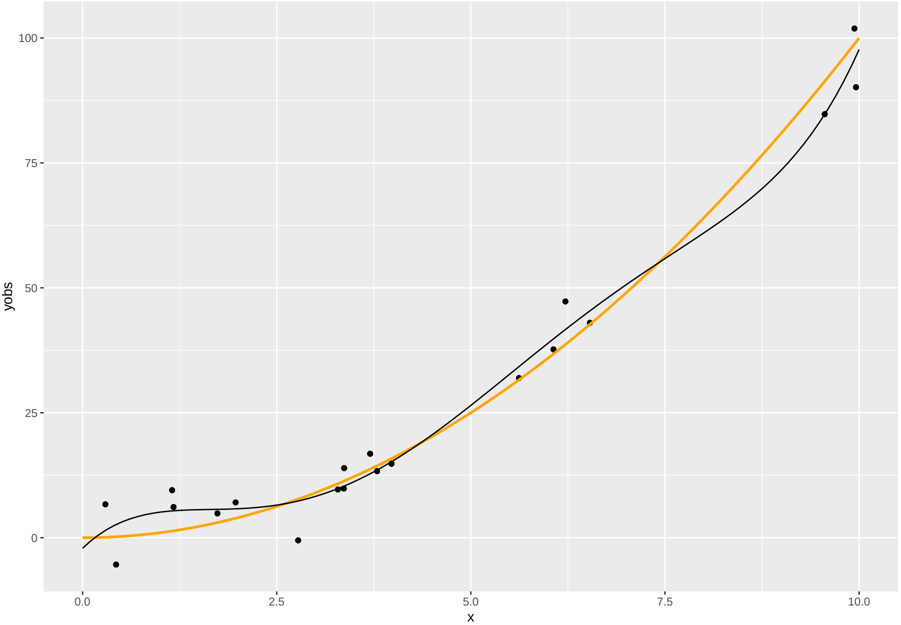


Subsection 6

Regresszió ötödfokú polinommal

Regresszió ötödfokú polinommal 1.

Regresszió ötödfokú polinommal



Subsection 7

Módosítás

Módosítás

Mondjuk, hogy nagyobb flexibilitásra vágyunk

- Például figyelembe akarjuk venni, hogy ez nem tűnik teljesen lineárisnak, vagy meg akarjuk ragadni a finomabb tendenciákat is

Emeljük a polinom fokszámát (ez nyilván növeli a flexibilitást, hiszen a kisebb fokszám nyilván speciális eset lesz), például 10-re

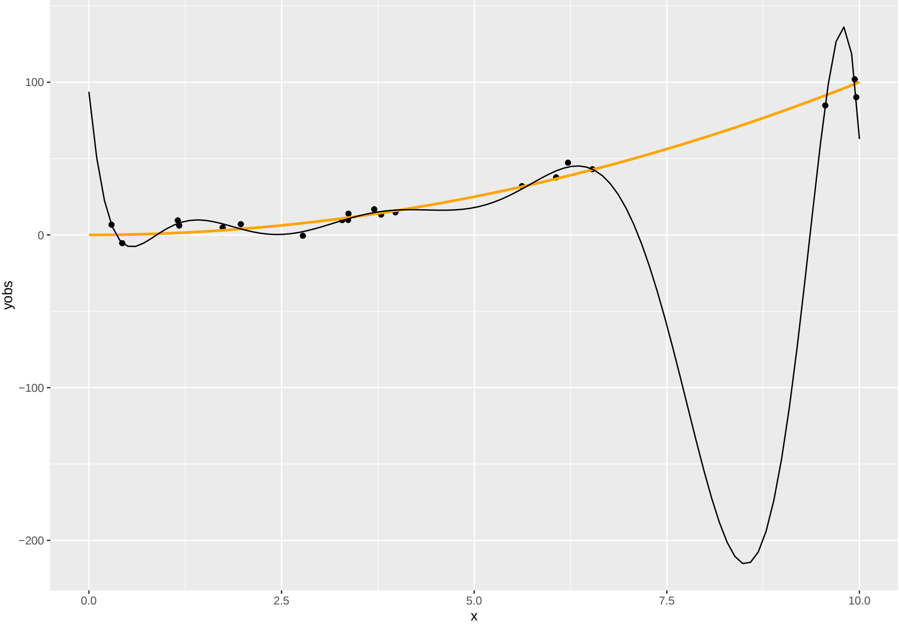
Szokás azt mondani, hogy a rang 5 illetve 10 (a polinom fokszáma, a becsülendő paraméterek száma nyilván egyezik a modellmátrix rangjával, de ez a fogalom később, amikor nem is polinomunk van, akkor is használható)

Subsection 8

Regresszió tizedfokú polinommal

Regresszió tizedfokú polinommal 1.

Regresszió tizedfokú polinommal



Subsection 9

Mi a jelenség oka?

Mi a jelenség oka? 1.

Szokás azt mondani, hogy *túlilleszkedés*, ami persze igaz is, de itt többről van szó

A polinomok elsősorban *lokálisan* tudnak jól közelíteni (a Taylor-sorfejtéses érvelés miatt), de nekünk arra lenne szükségünk, hogy *globálisan* jól viselkedő függvényformát találjunk

Pedig a polinomokat amúgy szeretjük, többek között azért is, mert szép sima görbét írnak le (matematikai értelemben véve a simaságot: végtelenszer folytonosan deriválhatóak, C^∞-beliek)

Mi lehet akkor a megoldás?

Subsection 10

Mi lehet a megoldás?

Mi lehet a megoldás? 1.

Egy lehetséges megközelítés: „összerakjuk a globálisat több lokálisból”

Azaz szakaszokra bontjuk a teljes intervallumot, és mindegyiket *külön-külön* polinommal igyekszünk modellezni

Így próbáljuk kombinálni a két módszer előnyeit

Persze a szakaszosan definiált polinomok önmagában még nem jók: a szakaszhatárokon találkozniuk kell (e találkozópontok neve: **knot**, „csomópont”, a számukat $q-2$-val jelöljük, a pozíciójukat x_i^*-vel)

Sőt, ha a simasági tulajdonságokat is át akarjuk vinni, akkor az érintkezési pontokban a deriváltaknak (magasabbrendűeknek is) is egyezniük kell

Ha p-edfokú polinomokat használunk, akkor az első $p-1$ derivált – és persze a függvényérték – egyezését kell kikötnünk a knot-okban (és esetleg még valamit a végpontokra)

Mi lehet a megoldás? 2.

Ez így már jó konstrukció lesz, a neve: **spline**

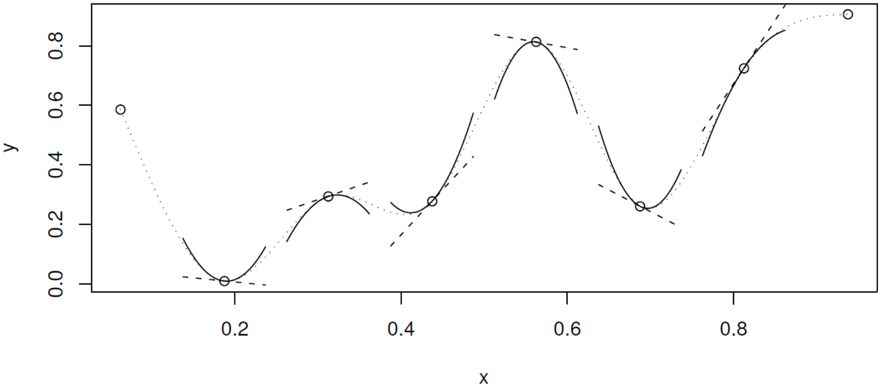
Subsection 11

Természetes köbös spline

Természetes köbös spline 1.

(Azért köbös, mert harmadfokúak a polinomok, és azért természetes, mert azt kötöttük ki, hogy a végpontokban nulla legyen a második derivált)

Természetes köbös spline 2.



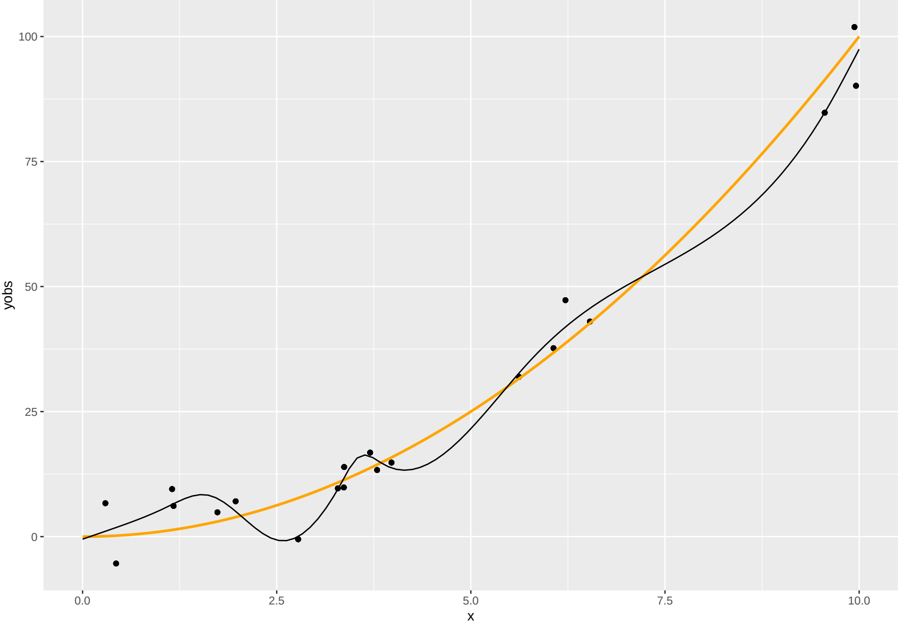
ábra 1: Természetes köbös spline

Subsection 12

A példa regressziója természetes köbös spline-nal

A példa regressziója természetes köbös spline-nal

A példa regressziója természetes köbös spline-nal

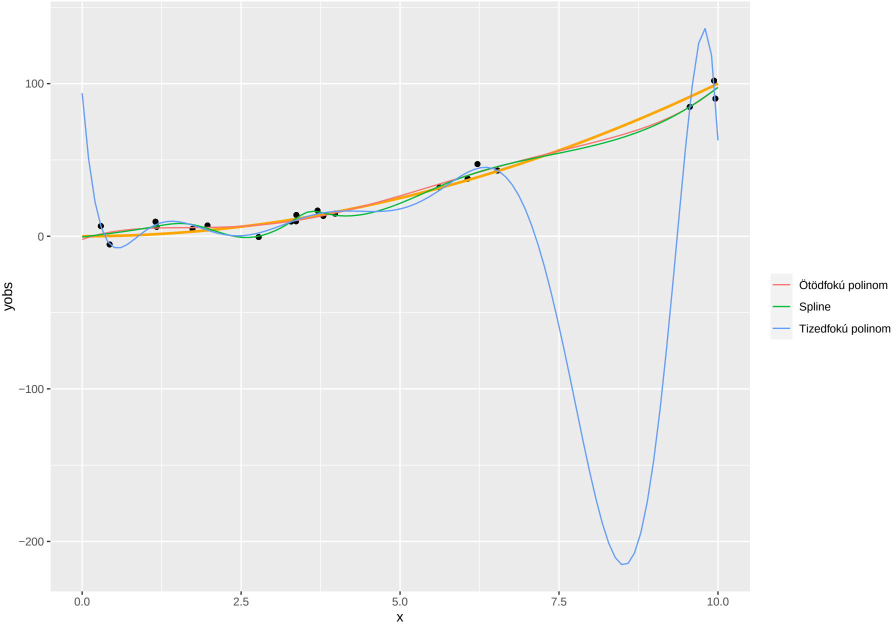


Subsection 13

Mi az előbbiben a fantasztikus?

Mi az előbbiben a fantasztikus?

Mi az előbbiben a fantasztikus?



Subsection 14

A spline-regresszió ereje

A spline-regresszió ereje

Nem csak az a jó, hogy szépen illeszkedik (tulajdonképpen még annál is jobban, mint a tizedfokú polinom, még ott is, ahol az jól illeszkedik amúgy)

…hanem, hogy – most már elárulhatom – *ez is ugyanúgy 10 rangú* mint a tizedfokú polinom!

Mégis: nyoma nincs túlilleszkedésnek

Tartalom

Tartalom 2.

Tartalom 3.

Subsection 1

Bázisfüggvényekkel felírás

Hogyan becsüljük meg a spline-regressziót?

Amiről nem beszéltünk eddig: ez mind szép, de hogyan tudunk ténylegesen is megbecsülni egy ilyen spline-regressziót?

Ehhez visszalépünk pár lépést, és bevezetünk egy első kicsit absztraktnak tűnő, de később rendkívül jó szolgálatot tevő megközelítést

Bár a célunk a spline-regresszió becslésének a megoldása, de a dolog – értelemszerűen – alkalmazható polinomiális regresszióra is (legfeljebb nincs sok értelme, mert az hagyományos módszerekkel is jól kézbentartható), úgyhogy először azon fogjuk illusztrálni

Polinomok tere mint függvénytér

A másodfokú polinomok – mint függvények – összessége **függvényteret** alkot

Ez egy olyan *vektortér*, aminek az elemei a függvények, a skalárok a valós számok, a két művelet pedig

- Skalárral szorzás: $\left(cf\right)(x) = cf(x)$
- Vektorok (azaz függvények) összeadása: $\left(f+g\right)(x) = f(x) + g(x)$, tehát pontonkénti összeadás

Belátható, hogy ez teljesíti a vektortéraxiómákat, mert zárt a két műveletre (másodfokú polinomok összege másodfokú polinom és másodfokú polinom konstansszorosa másodfokú polinom), és a többi követelményt is teljesíti

Polinomok terének bázisa

Szuper, de mindez mire jó?

Ha vektortér, akkor létezik **bázisa**, azaz olyan vektorok halmaza, melyekből lineáris kombinációval minden vektor – egyértelműen – előállítható (bázis: lineárisan független generátorrendszer)

A bázis nem feltétlenül egyértelmű, de az elemszáma igen, ez a vektortér **dimenziója**

Például a másodfokú polinomok jó bázisa $\left\{1, x, x^2\right\}$, nyilvánvaló, hogy ebből tényleg minden $ax^2 + bx + c$ másodfokú polinom előállítható lineáris kombinációval (triviálisan, a súlyok c, b és a)

Függvényterek esetében a bázis elemeit **bázisfüggvényeknek** is szokás nevezni, az $\left\{1, x, x^2\right\}$ tehát a másodfokú polinomok bázisfüggvényei

A polinomok terének dimenziója

Mivel mutattunk egy konkrét bázist, így a dimenzió nyilván 3, de a későbbiek szempontjából jól jön egy másik módszer is

Azzal, hogy az $ax^2 + bx + c$ polinomot megfeleltettük az (a, b, c) valós számhármasnak, a polinomok tere és a valós számhármasok tere (az $\mathbb{R}^3$) között létesítettünk egy izomorfizmust (a leképezés művelettartó és kölcsönösen egyértelmű)

Emiatt a polinomok terének ugyanaz a dimenziója, mint az $\mathbb{R}^3$-nak, ami viszont természetesen 3

Ez a módszer általában is használható: a dimenzió a felíráshoz szükséges paraméterek száma (feltéve, hogy ezek valós számok, valamint mindegyikhez tartozik egy polinom és viszont)

Spline-ok függvénytere

Mindez a spline-okra is igaz!

Érthető: minden pontban két polinomot adunk össze, vagy polinomot szorzunk skalárral, az eredmény polinom (már láttuk) – így tud spline adott pontja lenni!

Azaz: spline-okat is elő tudunk állítani bázisfüggvények lineáris kombinációjaként!

Hány dimenziós a spline-ok tere? 1.

Mielőtt megkeressük a spline-ok terének egy bázisát (azaz a konkrét bázisfüggvényeket), tisztázni kellene, hogy hány bázisfüggvényt keresünk egyáltalán, azaz hány dimenziós a spline-ok függvénytere

Naiv ötlet (köbös spline-okat használva példaként): van $q-1$ szakasz ($q-2$ knot, ami meghatároz $q-3$ szakaszt meg a két vége; úgy is felfogható, hogy a két végével együtt q knot van, ami meghatároz $q-1$ szakaszt) és mindegyiken egy harmadfokú polinom (aminek 4 paramétere van), akkor az $4q-4$ paraméter

Igen ám, de vannak megkötések: a knotokban a függvényérték és az első két derivált egyezik

Minden megkötés minden pontban 1 egyenlet, az 1-gyel csökkenti a paraméterek számát: van $q-2$ knot és 3 megkötés, az $3q-6$ csökkentés, marad $q+2$ paraméter

Hány dimenziós a spline-ok tere? 2.

De mivel természetes, így a végpontokban is van 1-1 megkötés: marad q paraméter, azaz q dimenziós a természetes köbös spline-ok tere (ezért neveztük a knot-ok számát $q-2$-nek!)

Mik a spline-ok bázisfüggvényei? 1.

Természetesen itt is igaz, hogy adott, rögzített spline-osztályra (pl. természetes köbös) is végtelen sok bázis van

Köztük célszerűség alapján választhatunk

A részletek nélkül két példa:

- $b_1(x) = 1, b_2(x) = x, b_i(x) = \left|x - x^*_{i-2}\right|^3$ $(i = 3, 4, \ldots, q)$
- $b_1(x) = 1, b_2(x) = x, b_i(x) = R\left(x, x^*_{i-2}\right)$ $(i = 3, 4, \ldots, q)$, ahol R egy nevezetes – elég hosszú, bár nem túl bonyolult – függvény (hamar látni fogjuk, hogy ez miért előnyös), annyi fontos, hogy x a $\left[0, 1\right]$ intervallumban essen (egyszerű átskálázssal mindig elérhető)

Most már csak a regresszió kivitelezését kell kitalálnunk

Subsection 2

Modellmátrix előállítása

A bázisfüggvények használatának ereje

A bázisfüggvények használatának két hatalmas előnye van:

- A probléma visszavezethető velük a sima lineáris regresszióra
- Sőt, ehhez a modellmátrix is könnyen előállítható

Bázisfüggvények használata másodfokú polinomnál

Legyen $b_1(x) = 1$, $b_2(x) = x$ és $b_3(x) = x^2$ a bázisunk

Az eredeti regresszió:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$$

Átírva bázisokra (lényegében transzformált magyarázó változók):

$$y_i = \beta_1 b_1\left(x_i\right) + \beta_2 b_2\left(x_i\right) + \beta_3 b_3\left(x_i\right) + \varepsilon_i$$

Ez már tiszta lineáris regresszió

Bázisfüggvények használatának előnye

Ez úgy tűnik, hogy csak egy nagyon nyakatekert felírás egy amúgy egyszerű problémára

Valójában viszont egy elképesztően erőteljes dolgot nyertünk: *minden* olyan függvény, legyen bármilyen komplikált is, ami felírható bázisfüggvényekkel (azaz az osztálya függvényosztályt alkot), az berakható egy *kutyaközönséges* regresszióba (azaz lehet ő a regrssziós függvény) a fenti transzformációval, tehát

$$\sum_{i=1}^{q} \beta_i b_i(x)$$

alakban

(Azaz minden függvény, ami egy függvénytér eleme)

A bázisfüggvények ereje, 1. felvonás

Még egyszer: *minden* függvény, ami felírható bázisfüggvényekkel

Azaz: *minden*

…és az összesnek *pontosan ugyanúgy* az lesz az alakja, hogy

$$\sum_{i=1}^{q} \beta_i b_i(x),$$

egyedül a bázisfüggvényt kell az adott esetnek megfelelően megválasztani

Tehát a spline is mehet ugyanígy (csak megfelelő b_i-kkel)!

És ha ez az alak megvan, akkor onnantól természetesen *sima lineáris regresszióval* elintézhető

A bázisfüggvények ereje, 2. felvonás

Ráadásul az $\mathbf{X}$ modellmátrix (design mátrix) előállítása is nagyon könnyű lesz: az i-edik sora

$$\left[b_1\left(x_i\right), b_2\left(x_i\right), \ldots, b_q\left(x_i\right)\right]$$

Így maga a mátrix az $\mathbf{x}$ és az $\left[1, 2, \ldots, q\right]$ vektor *külső szorzata* (tenzorszorzata), ha a művelet alatt az oszlopban szereplő érték által meghatározott bázisfüggvény sorbeli elemre történő alkalmazását értjük, tehát $i \otimes j$ 6= $b_j\left(x_i\right)$, és így

$$\begin{array}{cc} & \begin{pmatrix} 1 & 2 & \cdots & q \end{pmatrix} \\ \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} & \begin{bmatrix} b_1\left(x_1\right) & b_2\left(x_1\right) & \cdots & b_q\left(x_1\right) \\ b_1\left(x_2\right) & b_2\left(x_2\right) & \cdots & b_q\left(x_2\right) \\ \vdots & \vdots & \ddots & \vdots \\ b_1\left(x_n\right) & b_2\left(x_n\right) & \cdots & b_q\left(x_n\right) \end{bmatrix} \end{array}$$

A bázisfüggvények ereje, 2. felvonás

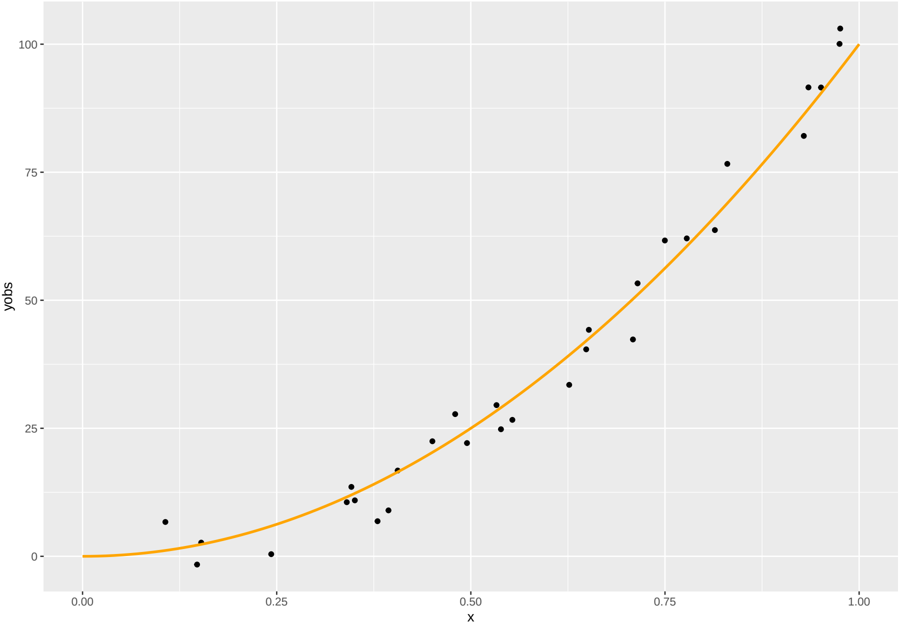
Így, a teljes modellmátrix egy lépésben megkapható…

… majd közvetlenül rakható is bele a sima lineáris regresszióba (ld. 1. előny):

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Megvalósítás R alatt 1.

Megvalósítás R alatt 2.



Subsection 3

Penalizálás

Dimenzió meghatározása 1.

A q dimenzió tehát az illeszkedés szabadságát határozza meg

Valahogy ezt is meg kellene határozni

Jön a fő kérdéskör: a túlilleszkedés elleni védekezés

Milyen legyen a „simítás foka”?

Simítás fokának meghatározása 1.

Tehát q-t kellene valahogy jól belőni

Egyszerű *modellszelekció*?

- Vagy nem beágyazott modellek szelekciója, vagy nem ekvidisztáns knot-ok, egyik sem túl szerencsés

Alternatív ötlet: q legyen inkább rögzített (elég nagy értéken, kicsit a várható fölé lőve), de a függvényformát nem engedjük teljesen szabadon alakulni

Hogyan? Büntetjük a túl „zizegős” függvényt!

Ez épp a **penalizált regresszió** alapötlete

És ami rendkívül fontos: így már jellemzően sem q pontos megválasztása, sem a knot-ok pontos helye nem bír nagy jelentőséggel (választhatjuk például egyenletesen)!

Penalizált regresszió

Klasszikus megoldás: a második derivált jelzi adott pontban a „zizegősséget”, ezt kiintegrálva kapunk egy összesített mértéket az egész függvényre

Valamilyen súllyal ezt vegyük figyelembe:

$$\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 + \lambda \int_0^1 \left[f''(x)\right]^2 \mathrm{d}x$$

A λ a *simítási paraméter*, ez határozza meg a trade-off-ot a jó illeszkedés és a simaság között

- $\lambda = 0$: penalizálatlan becslés, $\lambda \to \infty$: egyenes regressziós függvény

A simasági büntetőtag meghatározása

A regressziós függvény alakja: $f(x) = \sum_{i=1}^{q} \beta_i b_i(x)$

Kétszer deriválva: $f''(x) = \sum_{i=1}^{q} \beta_i b_i''(x)$

Négyzetre emelve: $\left[f''(x)\right]^2 = \sum_{i=1}^{q} \sum_{j=1}^{q} \beta_i b_i''(x) b_j''(x) \beta_j$

Kiintegrálva: $\int_0^1 \left[f''(x)\right]^2 \mathrm{d}x = \sum_{i=1}^{q} \sum_{j=1}^{q} \beta_i \left(\int_0^1 b_i''(x) b_j''(x) \mathrm{d}x \right) \beta_j$

De hát ez épp egy *kvadratikus alak*! ($\sum_{i=1}^{q} \sum_{j=1}^{q} x_i a_{ij} x_j = \mathbf{x}^T \mathbf{A} \mathbf{x}$)

Legyen $S_{ij} = \int_0^1 b_i''(x) b_j''(x) \mathrm{d}x$ és $\mathbf{S}$ az ezekből alkotott mátrix, akkor tehát a simítási büntetőtag:

$$\lambda \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}$$

Az előbb definiált R-rel $\mathbf{S}$ alakja nagyon egyszerű lesz: $S_{i+2,j+2} = R\left(x_i^*, x_j^*\right)$, az első két oszlop és sor pedig csupa nulla

Megvalósítás R alatt

A simítási büntetőtag beépítése a regressziós célfüggvénybe

Kényelmes lenne, ha $\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|^2+\lambda\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}$ helyett írhatnánk egyetlen normát célfüggvényként

Ez nem nehéz, ha a második tagot át tudjuk normává alakítani, hiszen (innentől némi blokkmátrix műveletekre szükség lesz)

$$\|\mathbf{a}\|^2+\|\mathbf{b}\|^2=\left\|\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right\|^2$$

Legyen $\mathbf{B}$ olyan, hogy $\mathbf{B}^T\mathbf{B}=\mathbf{S}$ (pl. spektrális dekompozícióval, vagy Cholesky-dekompozícióval megtalálható a mátrix ilyen „négyzetgyöke”), ekkor

$$\lambda\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}=\lambda\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}=\lambda\left(\mathbf{B}\boldsymbol{\beta}\right)^T\mathbf{B}\boldsymbol{\beta}=\left(\sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\right)^T\left(\sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\right)$$

A simítási büntetőtag beépítése a regressziós célfüggvénybe 2.

Ezzel meg is vagyunk, hiszen a norma egyszerűen $\|\mathbf{a}\|^2 = \mathbf{a}^T\mathbf{a}$, így

$$\lambda\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} = \left\|\sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\right\|^2$$

ahonnan

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \left\|\sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\right\|^2$$

és így, az előzőek szerint

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \left\|\sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\right\|^2 = \left\|\begin{pmatrix}\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \\ \sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\end{pmatrix}\right\|^2$$

Jó lenne $\boldsymbol{\beta}$-t kiemelni; ez nem is túl nehéz, hiszen $\mathbf{a}$ és $-\mathbf{a}$ normája ugyanaz:

$$\left\|\begin{pmatrix}\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \\ \sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\end{pmatrix}\right\|^2 = \left\|\begin{pmatrix}\mathbf{y} \\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{X} \\ \sqrt{\lambda}\mathbf{B}\end{pmatrix}\boldsymbol{\beta}\right\|^2$$

Regresszió megoldása a penalizálással 1.

Innentől a regresszió játszi könnyedséggel (értsd: a szokványos, nem is penalizált eszköztárral) megoldható, csak $\mathbf{X}$ szerepét $\begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{B} \end{pmatrix}$, $\mathbf{y}$ szerepét $\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$ játssza

Így az „$\mathbf{X}^T\mathbf{X}$” épp $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{B}^T\mathbf{B} = \mathbf{X}^T\mathbf{X} + \lambda\mathbf{S}$ lesz

Az „$\mathbf{X}^T\mathbf{y}$” pedig $\mathbf{X}^T\mathbf{y}$ (a kiegészített eredményváltozóban lévő nullák épp a magyarázó változók kiegészítését ütik ki)

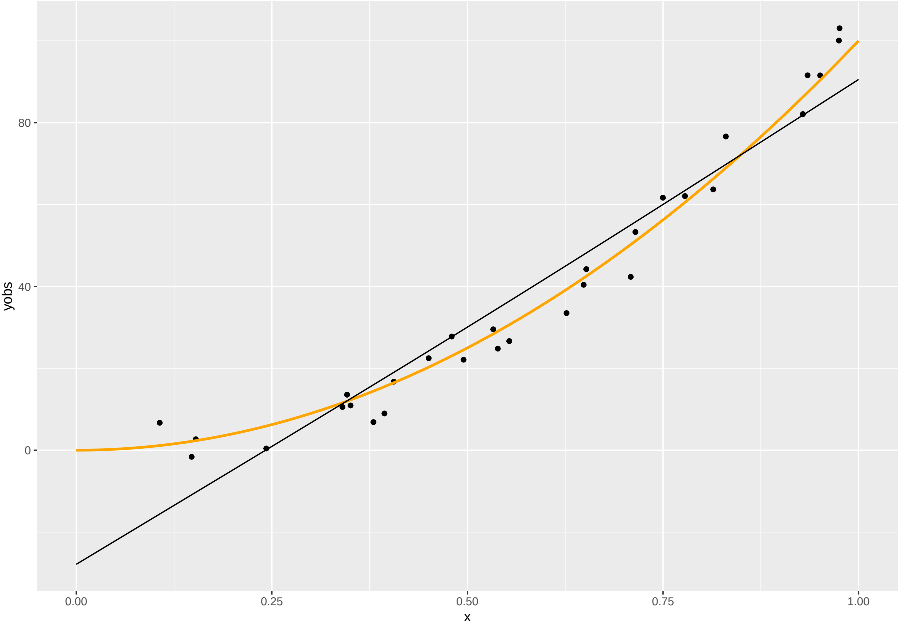
Így az OLS megoldás:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

(Persze a gyakorlatban ennek közvetlen számítása helyett célszerűbb az augmentált eredmény- és magyarázóváltozókat berakni egy hatékonyabb lineáris regressziót megoldó módszerbe)

Megvalósítás R alatt

Megvalósítás R alatt



Subsection 4

Simítási paraméter meghatározása

A simítási paraméter meghatározása

Kérdés még a λ értéke

Sima OLS-jellegű eljárással, tehát a reziduális négyzetösszeg minimalizálást tűzve ki célul nyilván nem határozható meg (hiszen az mindig 0-t adna)

Épp az a lényeg, hogy a túlilleszkedésre is tekintettel legyünk

Ötlet: keresztvalidáció

Keresztvalidációs módszerek: OCV

Mindig egy pontot hagyunk ki, és így számolunk hibát: OCV

(Szokták egy-kihagyásos keresztvalidációnak, LOOCV-nek is nevezni)

Tehát:

$$E_{OCV} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_i^{[-i]} - y_i \right)^2$$

Szerencsére nem kell ténylegesen n-szer lefuttatni a regressziót mert belátható, hogy

$$E_{OCV} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_i \right)^2 / \left(1 - A_{ii} \right)^2 ,$$

ahol $\mathbf{A}$ az influence mátrix

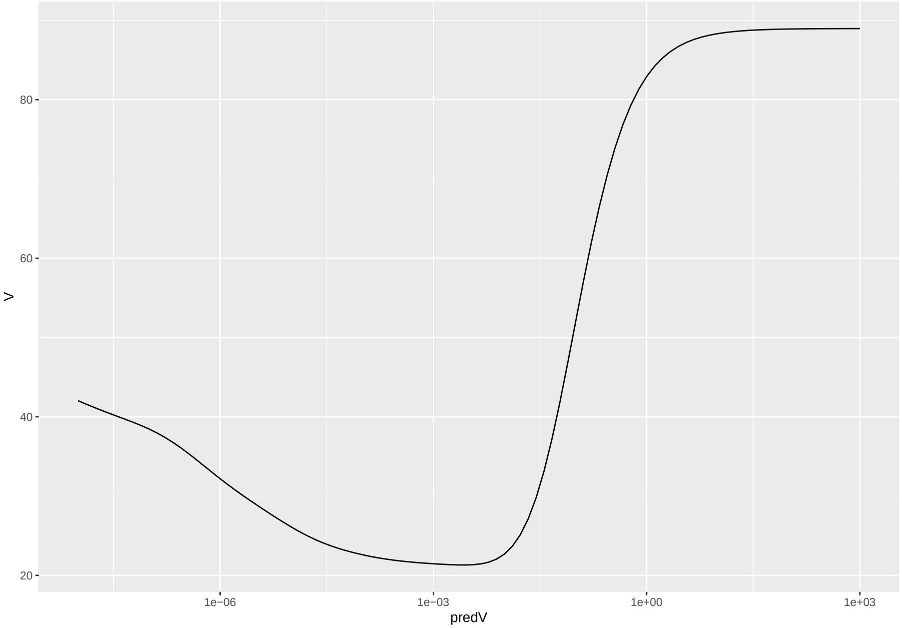
Ha az A_{ii}-ket az átlagukkal helyettesítjük, akkor az általánosított keresztvalidációhoz jutunk (GCV)

Tehát:

$$E_{GCV} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_i \right)^2 / \left[\mathrm{tr}\left(\mathbf{I} - \mathbf{A} \right) \right]^2$$

Megvalósítás R alatt

Megvalósítás R alatt



Tartalom

Tartalom

Tartalom

Subsection 1

Több magyarázó változó

Több magyarázó változó

Eddig egy magyarázó változó esetével foglalkoztunk