A lineáris regressziós modell becslése mintából, vektoros-mátrixos formalizmus, az OLS-becslő

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 12.

Tartalom

Tartalomjegyzék

1. Az OLS-elv

Előkészületek az OLS-becsléshez

- Nem kell hozzá semmilyen regresszió, a legközönségesebb következtető statisztikai példán is elmondható
- Például: sokasági várható érték becslése normalitás esetén (legyen a szórás is ismert)
- Ami fontos: bár egy alap következtető statisztika kurzuson nem szokták mondani, de lényegében itt is az a helyzet, hogy egy *modellt* feltételezünk a sokaságra
- Jelesül $Y \sim \mathcal{N}\left(\mu, \sigma_0^2\right)$, amit nem mellesleg úgy is írhatnánk, hogy $Y = \mu + \varepsilon$, ahol $\varepsilon \sim \mathcal{N}\left(0, \sigma_0^2\right)$
- A másik ami fontos: a modellből következik egy *becsült érték* minden mintabeli elemhez

- Jelen esetben, ha m egy feltételezett érték az ismeretlen sokasági várható értékre:

$$\widehat{y_i} = m$$

Az OLS-elv

- OLS-elvű becslés: az ismeretlen sokasági paraméterre az a becsült érték, amely mellett a tényleges mintabeli értékek, és az adott paraméter melletti, modellből származó becsült értékek közti eltérések négyzetének összege a legkisebb:

$$\widehat{\mu} = \arg\min_m \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \arg\min_m \sum_{i=1}^{n} (y_i - m)^2$$

- (Aminek a megoldása természetesen $\widehat{\mu} = \overline{y}$)

A mintavétel a lineáris regressziós feladatban

- Tételezzük fel, hogy az $(Y, X_1, X_2, \ldots, X_k)$ változóinkra veszünk egy n elemű mintát
- Az i-edik mintaelem: $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$
- Feltételezzük azt is, hogy a mintavétel fae (független, azonos eloszlású)

A fae feltevés keresztmetszeti esetben sokszor lehet elfogadható közelítés (bár ott sem mindig teljesül, erre egy jó példa az ún. térbeli autokorreláció – de ez túlmutat a mostani kereteken), idősoros adatoknál azonban sohasem. Ott nagyon részletesen fogunk ezzel foglalkozni.

Lineáris regresszió becslése OLS-elven

- *Hajszálpontosan ugyanaz* történik, mint az előbb, csak a sokaságra feltételezett modellünk kicsit bonyolultabb, jelesül:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- A becsült értékek adott $b_0, b_1, \ldots, b_k$ sokasági paraméterek mellett:

$$\widehat{y_i} = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$$

- A feladat tehát ugyanaz:

$$\left(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_k}\right) = \arg\min_{b_0, b_1, b_2, \ldots, b_k} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 =$$

$$= \arg\min_{b_0, b_1, b_2, \ldots, b_k} \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik})\right]^2$$

- Annyi bonyolódottság van, hogy itt most *több* paramétert kell becsülni, de ez csak a kivitelezést nehezíti, elvileg teljesen ugyanaz a feladat

Az OLS-becslési feladat vektoros-mátrixos jelölésekkel

- A jelölések egyszerűsítése érdekében fogjuk össze mindent vektorokba és mátrixokba; egyedül a magyarázó változók nem triviálisak, mert kiegészítjük őket egy csupa 1 oszloppal (ún. design mátrix):

$$\mathbf{X}_{n\times(k+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

- Így ugyanis a feladat:

$$\arg\min_{\mathbf{b}} (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

- Az $(\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$ hibanégyzetösszeget ESS-sel (error sum of squares) is fogjuk jelölni

Sajnos néhány irodalom az általunk használt ESS-re inkább az RSS-t (residual sum of squares) rövidítést használja, ami a jelölési zavarok legszerencsétlenebb típusa, ugyanis az RSS-t majd később mi is fogjuk használni, csak épp másra. Éppen ezért, ha ilyenekről olvasunk, mindig tisztázni kell, hogy a könyv vagy program írói mit értenek alatta.

2. A lineáris regresszió becslése tisztán deskriptíve

Az OLS-becslési feladat megoldása

A megoldás:

$$\arg\min_{\mathbf{b}} (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) = \arg\min_{\mathbf{b}} \left[\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right]$$

A szélsőérték-keresést oldjuk meg többváltozós deriválással (kvadratikus felület konvex, a stacionárius pont egyértelmű globális szélsőértékhely):

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{b}} \left[\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right] = \\ &= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b} = 0 \Rightarrow \widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}, \end{aligned}$$

ha $\mathbf{X}^T\mathbf{X}$ nem szinguláris

Az első lépésnél lényegében egyszerű algebrai átalakításokat végzünk (és a definíciókat használjuk), hiszen a zárójeleket felbontani, műveleteket elvégezni, mátrixokkal-vektorokkal is hasonlóan kell mint valós számokkal. (A transzponálás tagonként elvégezhető, azaz $(\mathbf{a} - \mathbf{b})^T = \mathbf{a}^T - \mathbf{b}^T$.) Egyedül annyit kell észrevenni, hogy a $\mathbf{y}^T\mathbf{X}\mathbf{b}$ egy

egyszerű valós szám, ezért megegyezik a saját transzponáltjával, $\mathbf{b}^T\mathbf{X}^T\mathbf{y}$-nal. Ezért írhattunk $-(\mathbf{Xb})^T\,\mathbf{y}-\mathbf{y}^T\mathbf{Xb}$ helyett egyszerűen – például – $-2\mathbf{b}^T\mathbf{X}^T\mathbf{y}$-t. (Itt mindenhol felhasználtuk, hogy a transzponálás megfordítja a szorzás sorrendjét: $(\mathbf{AB})^T=\mathbf{A}^T\mathbf{B}^T$.)

Itt jelentkezik igazán a mátrixos jelölésrendszer előnye. A $\mathbf{y}^T\mathbf{y}-2\mathbf{y}^T\mathbf{Xb}+\mathbf{b}^T\mathbf{X}^T\mathbf{Xb}$ lényegében egy „másodfokú kifejezés" többváltozós értelemben (az ax^2+bx+c többváltozós megfelelője), és ami igazán szép: pont ahogy az ax^2+bx+c lederiválható a változója (x) szerint (eredmény $2ax+b$), ugyanúgy ez is lederiválható a változója (azaz $\mathbf{b}$) szerint… és az eredmény az egyváltozóssal teljesen analóg lesz, ahogy fent is látható! (Ez persze bizonyítást igényel! – lásd többváltozós analízisből.) Bár ezzel átléptünk egyváltozóról többváltozóra, a többváltozós analízisbeli eredmények biztosítanak róla, hogy formálisan ugyanúgy végezhető el a deriválás. (Ezt írja le röviden a „vektor szerinti deriválás" jelölése. Egy $\mathbf{b}$ vektor szerinti derivált alatt azt a vektort értjük, melyet úgy kapunk, hogy a deriválandó kifejezést lederiváljuk $\mathbf{b}$ egyes b_i komponensei szerint (ez ugye egyszerű skalár szerinti deriválás, ami már definiált!), majd ez eredményeket összefogjuk egy vektorba. Látható tehát, hogy a vektor szerinti derivált egy ugyanolyan dimenziós vektor, mint ami szerint deriváltunk.) Ami igazán erőteljes ebben az eredményben, az nem is egyszerűen az, hogy „több" változónk van, hanem, hogy nem is kell tudnunk, hogy mennyi – mégis, általában is működik!

Azt, hogy a megtalált stacionaritási pont tényleg minimumhely, úgy ellenőrizhetjük, hogy megvizsgáljuk a Hesse-mátrixot a pontban. A mátrixos jelölésrendszerben ennek az előállítása is egyszerű, még egyszer deriválni kell a függvényt a változó(vektor) szerint:

$$\frac{\partial^2}{\partial\mathbf{b}^2}\left[\mathbf{y}^T\mathbf{y}-2\mathbf{b}^T\mathbf{X}^T\mathbf{y}+\mathbf{b}^T\mathbf{X}^T\mathbf{Xb}\right]=\frac{\partial}{\partial\mathbf{b}}\left[-2\mathbf{X}^T\mathbf{y}+2\mathbf{X}^T\mathbf{Xb}\right]=2\mathbf{X}^T\mathbf{X}.$$

Az ismert tétel szerint a függvénynek akkor van egy pontban ténylegesen is (lokális, de a konvexitás miatt egyben globális) minimuma, ha ott a Hesse-mátrix pozitív definit. Esetünkben ez minden pontban teljesül. A $\mathbf{X}^T\mathbf{X}$ ugyanis pozitív szemidefinit (ez egy skalárszorzat-mátrix, más néven Gram-mátrix, amelyek mindig pozitív szemidefinitek), a kérdés tehát csak a határozott definitség. Belátható azonban, hogy ennek feltétele, hogy $\mathbf{X}^T\mathbf{X}$ ne legyen szinguláris – azaz itt is ugyanahhoz a feltételhez értünk! Megjegyezzük, hogy ez pontosan akkor valósul meg, ha az $\mathbf{X}$ teljes oszloprangú. (Erre a kérdésre a modellfeltevések tárgyalásakor még visszatérünk.)

Végül egy számítástechnikai megjegyzés: az együtthatók számításánál a fenti formula direkt követése általában nem a legjobb út, különösen ha sok megfigyelési egység és/vagy változó van. Ekkor nagyméretű mátrixot kéne invertálni, amit numerikus okokból (kerekítési hibák, numerikus instabilitás stb.) általában nem szeretünk. Ehelyett, a különféle programok igyekeznek a direkt mátrixinverziót elkerülni, tipikusan az $\mathbf{X}$ valamilyen célszerű mátrix dekompozíciójával (QR-dekompozíció, Cholesky-dekompozíció). Extrém esetekben még az is elképzelhető, hogy az egzakt, zárt alakú megoldás előállítása helyett valamilyen iteratív optimalizálási algoritmus (gradiens módszer, Newton–Raphson-módszer) alkalmazása a gyakorlatban járható út, annak ellenére is, hogy elvileg van zárt alakban megoldása.

A kapott eredmény nem más, mintha $\mathbf{X}$ Moore-Penrose pszeudoinverzével szoroznánk $\mathbf{y}$-t.

Pár további gondolat

- Az ún. reziduumok:
$$\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}$$
- Az előrejelzések a mintánkra:
$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$
- Ez alapján vezessük be a
$$\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$
mátrixot, ezzel $\widehat{\mathbf{y}} = \mathbf{P}\mathbf{y}$
- Emiatt szokták „hat” mátrixnak is nevezni

Az OLS geometriai interpretációja

$\mathbf{P}$ projektormátrix lesz ($\mathbf{P}^2 = \mathbf{P}$, azaz idempotens) $\rightarrow$ út az OLS geometriai interpretációjához

Mindenekelőtt emlékeztetünk rá, hogy az $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ vektorok által kifeszített alteret azok a pontok alkotják, melyek előállnak e vektorok lineáris kombinációjaként. (E pontok mindig az eredeti vektortér – ami felett a vektorokat értelmeztük – alterét alkotják, ezért jogos az elnevezés.) Ha most vektortérnek az $\mathbb{R}^n$-et tekintjük, vektoroknak pedig az $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k$ magyarázóváltozókat (és a konstanst), azaz $\mathbf{X}$ oszlopvektorait, akkor az ezek által kifeszített altér – ezt szokás egyébként az $\mathbf{X}$ mátrix oszlopterének nevezni – épp azon pontokból áll, melyek előállhatnak becsült eredményváltozó(vektor)ként valamilyen regressziós koefficiensekkel! (Hiszen a becsült eredményváltozót is e vektorok lineáris kombinációjaként állítjuk elő.) Általánosságban persze nem várható, hogy a tényleges eredményváltozó(vektor) benne legyen ebben az altérben (azaz egzaktan – értsd: minden egyes megfigyelési egységre megvalósulóan – elő lehessen állítani lineáris kombinációként), ezt fejezi ki a reziduum. Mint a tényleges és a becsült eredményváltozó(vektor) különbségvektora, a reziduum hossza megmutatja, hogy mennyire messze van a becsült és a tényleges eredményváltozó egymástól (az $\mathbb{R}^n$-ben). Mi azt szeretnénk, ha ez minimális lenne. Választva a szokásos euklideszi metrikát, visszakapjuk a legkisebb négyzetes értelmezést. A kérdés már csak az, hogy adott ponthoz (tényleges eredményváltozó) hogyan határozható meg az altér (azaz: amit lineáris regresszióval elő tudunk állítani) legközelebbi pontja… de hát ez épp a geometriai vetítés művelete! A megoldás tehát az, hogy a tényleges eredményváltozót merőlegesen rávetítjük (ortogonális projekció) a magyarázóváltozók (és a konstans) által kifeszített altérre! A vetítés eredményeként kapott pont lesz a ténylegeshez legközelebbi előállítható becsült eredményváltozó, az előállításában szereplő együtthatók pedig az optimális becsült regressziós koefficiensek. Így aztán azt is megállapíthatjuk, hogy a fenti $\mathbf{P}$ mátrix nem más, mint ami a tényleges eredményváltozót levetíti a magyarázóváltozók (és a konstans) által kifeszített altérre.

3. Modellminősítés tisztán deskriptíve

Modell jóságának viszonyítási pontjai

- A modell minősítése az ESS alapján? → kézenfekvő, de nem önmagában: viszonyítani kell! Két kézenfekvő alap:
  - Tökéletes (v. szaturált, perfekt modell): minden mintaelemre a pontos értéket becsüli → $\widehat{e_i} = 0 \Rightarrow ESS = 0$
  - Nullmodell: semmilyen külső (magyarázó)információt nem használ fel → minden mintaelemet az átlaggal becsül
- Egy adott regressziós modell teljes négyzetösszegének nevezzük, és TSS-sel jelöljük a hozzá tartozó (tehát ugyanazon eredményváltozóra vonatkozó) nullmodell hibanégyzetösszegét:

$$TSS = ESS_{\text{null}} = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

Hogyan jellemezzük modellünk jóságát?

- A minősítést képezzük a „hol járunk az úton?" elven: a tökéletesen rossz modelltől a tökéletesen jó modellig vezető út mekkora részét tettük meg
- Az út „hossza" TSS $(= TSS - 0)$, amennyit „megtettünk": $TSS - ESS$
- Egy adott regressziós modell regressziós négyzetösszegének nevezzük, és RSS-sel jelöljük a teljes négyzetösszegének és a hibanégyzetösszegének különbségét:

$$RSS = TSS - ESS.$$

Ahogy már említettük is, sajnos néhány könyv az RSS-t más néven, hogy még rosszabb legyen a helyzet, néha ESS-ként, emlegeti. (Az itteni ESS pedig épp RSS az ottani terminológiában…)

Az új mutató bevezetése

Ezzel az alkalmas modelljellemző mutató: a többszörös determinációs együttható (jele R^2):

$$R^2 = \frac{TSS - ESS}{TSS} = \frac{RSS}{TSS}.$$

Az R^2-ről bővebben

- Ha van konstans a modellben, akkor nyilván $ESS < TSS$, így minden regressziós modellre, amiben van konstans: $0 \leq R^2 \leq 1$.

- Az R^2 egy modell jóságának legszéleskörűbben használt mutatója
- Értelmezhető %-ként: a magyarázó változók ismerete mennyiben csökkentette az eredményváltozó tippelésekor a bizonytalanságunkat (ahhoz képest, mintha nem ismertünk volna egyetlen magyarázó változót sem)
- De vigyázat: nagyságának megítélése, változók száma stb.
- A belőle vont négyzetgyököt többszörös korrelációs együtthatónak szokás nevezni
- Mondani sem kell, ez az R^2 a korábban bevezetett (sokasági) R^2 mintabeli analógja

Az R^2-ről bővebben

- Ha van konstans a modellben, akkor érvényes a következő felbontás:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$$

- (Négyzetek nélkül nyilvánvaló, de négyzetekkel is!)
- Röviden tehát:

$$TSS=ESS+RSS$$

- Összevetve az előző definícióval, kapjuk, hogy

$$RSS=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$$

Egy megjegyzés a konstans szerepéről

- Az előzőek is motiválják, hogy megállapítsuk: konstanst *mindenképp* szerepeltetünk a regresszióban, ha inszignifikáns, ha nem látszik különösebb értelme stb. *akkor is*! – csak és kizárólag akkor hagyhatjuk el, ha az a modell tartalmából adódóan elméleti követelmény (erre látni fogunk nemsokára egy példát is, a standardizált regressziót)
- Ellenkező esetben (ún. konstans nélküli regresszió), a fenti felbontás nem teljesül, így a „hol járunk az úton” elven konstruált R^2 akár negatív is lehet!

Néhány könyv, az R^2 alternatív definiálása révén, a negatív esetet kizárja.