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

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 9.

Tartalom

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:

$$\widehat{\mu} = \arg\min_{m} \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 = \arg\min_{m} \sum_{i=1}^{n} \left(y_i - m\right)^2$$

- (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)

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:
$$\left(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_k}\right) = \operatorname*{arg\,min}_{b_0, b_1, b_2, \ldots, b_k} \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2 =$$
$$= \operatorname*{arg\,min}_{b_0, b_1, b_2, \ldots, b_k} \sum_{i=1}^{n} \left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}\right)\right]^2$$
- 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:

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

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

Az OLS-becslési feladat megoldása

A megoldás:

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

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

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

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

Pár további gondolat

- Az ún. reziduumok:
$$\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}$$
- Az előrejelzések a mintánkra:
$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$
- 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

Modell jóságának viszonyítási pontjai

- A modell minősítése az ESS alapján? $\rightarrow$ kézenfekvő, de nem önmagában: viszonyítani kell! Két kézenfekvő alap:
  - Tökéletes (v. szaturált, perfekt modell): minden mintaelemre a pontos értéket becsüli $\rightarrow$ $\widehat{e_i} = 0 \Rightarrow ESS = 0$
  - Nullmodell: semmilyen külső (magyarázó)információt nem használ fel $\rightarrow$ 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:

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

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:

$$RSS = TSS - ESS.$$

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!