A többváltozós lineáris regresszió III. Főkomponens-analízis

6. előadás

Nominális változók a lineáris modellben
Logisztikus regresszió
Főkomponens-analízis

2018. október 8.

Alapok

Kérdés: hogyan szerepeltethetünk egy minőségi (nominális) tulajdonságot (pl. férfi/nő, egészséges/beteg, szezonális hatások, korcsoportok, tapasztalat) egy lineáris regressziós modellben?

Megoldás: kódolni kell, hiszen csak számszerű értékekkel tudunk dolgozni, ezért a lehetséges (véges sok!) kimenetelt fogjuk kódolni valamilyen egészértékű változóval.

Legegyszerűbb eset: bináris kódolás, azaz csak $0 - 1$ értékű változókkal kódolunk. Ezek az ún. Dummy-változók.

Kódolás

2 kimenetel: 1 dummy változónk van 0 és 1 értékkel

k kimenetel: kell-e mindegyikre külön k darab dummy? Nem, ez ugyanis nem lenne jó megoldás minden esetben! Ugyanis

- k kimenetelre elég $(k-1)$ darab dummy változó az ún. referencia-kódolás logikája alapján. Itt egy kimenetel kap $(k-1)$ darab 0 értéket, ez lesz az ún. kontroll-csoport, a többi $(k-1)$ kimenetelen pedig mindig pontosan egy darab dummy lesz 1 értékű.
- Példa 3 kimenetel esetén: A, B, C a lehetséges kimenetelek, R_A, R_B a két dummy változó

	R_A	R_B
A	1	0
B	0	1
C	0	0

Dummy-változó csapda

Miért célszerű ezt használni a kézenfekvő "k változó - k dummy" kódolás helyett?

- Ha a modell nem tartalmaz konstans tagot, akkor használható mindkét kódolás, nem lesz belőle probléma.
- Viszont, ha a modellben van konstans tag, akkor TILOS k csoporthoz k darab dummyt használni, különben egzakt multikollenaritás lép fel! Ez az ún. dummy-változó csapda. Az ok egyszerű: ekkor ugyanis a konstans és a k darab dummy miatt X oszlopai lineárisan összefüggők lennének, ami ellentmond a modell alapfeltételeinek.

Dummy a modellben

Dummy-változó magyarázó változóként minden gond nélkül bevehető a lineáris regressziós modellbe a folytonos változók mellé. Ez nem fogja bántani a regressziót, és az OLS is gond nélkül működik. A lehetséges eseteket az alábbi három modell írja le:

- $Y = \beta_1 + \beta_D D + \beta_X X + u$: csak a tengelymetszetet téríti el, azaz pl. $+1$ egység GDP hatása ugyanaz minden csoportban, de nem ugyanannyi a 0 GDP-hez tartozó munkanélküliség
- $Y = \beta_1 + (\beta_D D + \beta_X) X + u$: a meredekséget téríti el, azaz pl. ugyanannyi a 0 GDP-hez tartozó munkanélküliség minden csoportban, de nem egyfoma a $+1$ egység GDP hatása a csoportokban. (Interakció!)
- $Y = \beta_1 + \beta_{D1} D1 + (\beta_{D2} D2 + \beta_X) X + u$: az előző kettő keverékét kapjuk. (Interakció!)

Kérdés

Mi történik akkor, ha most a 0-1 értékű változónkat nem magyarázóként, hanem magyarázott változóként kívánjuk a modellben szerepeltetni?

Példa

Minta: mérlegadataikkal adott cégek.

Feladat: csőd-előrejelzés, azaz azt kell megmondanunk, hogy az adatok alapján várhatóan melyik cég fog csődbe menni és melyik nem, a vizsgált időhorizonton belül.

- magyarázó változók: mérlegadatok - folytonos (esetleg dummy) változók
- eredményváltozó: csőd vagy sem - bináris, azaz dummy-változó

Bináris változó az eredményváltozó szerepében

Alapfeladat: osztályozást, avagy csoportba-sorolást akarunk végezni valamilyen adott szempontok szerint. Ezt nevezik klasszifikációs feladatnak.

- Gyakorlati jelentősége órási: halálozási adatok, demográfiai adatok, felvételi adatok, stb. elemzése során.
- A probléma legelemibb megoldási módszere: logisztikus regresszió.
- Más megoldás is létezik az ilyen feladatok megoldására: gépi tanulás, klaszterezés, diszkriminancia analízis, stb., de ezekkel most nem foglalkozunk.

Módszer

Az elméleti modell:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u, \quad t = 1, \ldots, T$$

ahol az Y eredményváltozó bináris (dummy) változó.

Kérdés: működik-e most is az OLS becslés? Nem valószínű, hiszen az OLS eredménye skálás ($\pm\infty$ közti érték), ebből nyilván nehéz lesz bináris változót becsülni.

Trükk: a lineáris struktúrát megőrizve a célváltozónk "ügyes" transzformáltjára alkalmazzuk a lineáris regressziós modellt.

Módszer - 1. trükk

A célváltozó tehát bináris (dummy) változó, azaz

$$Y = \begin{cases} 1 & \text{"siker" esetén} \\ 0 & \text{"kudarc" esetén} \end{cases}$$

1. trükk: nem a siker tényét, hanem annak

$$P_{\underline{X}} = P(Y = 1|\underline{X})$$

(a mintára vonatkozó) feltételes valószínűségét modellezzük. Ezzel $\{0, 1\}$ helyett már $[0, 1]$-beli változót kell modelleznünk.

Ez persze még mindig kevés a lineáris regresszióhoz!

Módszer - 2. és 3. trükk

2.trükk: újabb transzformáció - az esélyhányados bevezetése, amely a siker és a kudarc valószínűségének aránya, azaz

$$\text{odds}_{\underline{X}} = \frac{P_{\underline{X}}}{1 - P_{\underline{X}}} \in [0, \infty)$$

Visszafejtve a transzformációt

$$P_{\underline{X}} = \frac{P_{\underline{X}}}{P_{\underline{X}} + 1 - P_{\underline{X}}} = \frac{\frac{P_{\underline{X}}}{1-P_{\underline{X}}}}{\frac{P_{\underline{X}}}{1-P_{\underline{X}}} + 1} = \frac{\text{odds}_{\underline{X}}}{1 + \text{odds}_{\underline{X}}}$$

Ez már majdnem jó lesz, csak a negatív értékek maradnak ki!

3. trükk: logaritmálás, azaz a logit bevezetése

$$\text{logit}_{\underline{X}} = \log(\text{odds}_{\underline{X}}) \in (-\infty, \infty).$$

Ezzel a siker és kudarc eloszlását is szimmetrizáltuk!

Modell

Erre illesztünk tehát lineáris modellt

$$\text{logit}_{\underline{X}} = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k = \underline{X}^T \beta$$

alakban. Ezt nevezik logit avagy logisztikus regressziónak.

Innen, a becslések elvégzése után,

$$\text{odds}_{\underline{X}} = e^{\hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k}$$

és

$$P_{\underline{X}} = \frac{e^{\hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k}}{1 + e^{\hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k}} = \frac{e^{\underline{X}^T \beta}}{1 + e^{\underline{X}^T \beta}}$$

Becslés és az eredmények értelmezése

- A paraméterek becslése a ML-módszerrel történik, mert ekkor elegendő a feltételes valószínűségek előállítása. A likelihood függvény:

$$L(\beta_1, \ldots, \beta_k) = \prod_{Y_i=1} P_{\underline{X},i} \prod_{Y_i=0} (1 - P_{\underline{X},i})$$

- Átlagos növekedési ráta:

$$\frac{\text{odds}_{X_1,\ldots,X_{l-1},X_l+1,X_{l+1},\ldots,X_k}}{\text{odds}_{X_1,\ldots,X_{l-1},X_l,X_{l+1},\ldots,X_k}} = e^{\beta_l}$$

- Marginális hatás:

$$\frac{\partial P_{\underline{X}}}{\partial X_i} = \beta P_{\underline{X}}(1 - P_{\underline{X}})$$

Eredmények értelmezése

Kérdés: a kapott becslések valószínűségek. Hogyan kell ezek alapján elvégezni a klasszifikációt?

- Cut-off point definiálása: $\hat{Y} = 1$, ha $P_X > C$, ahol C értéke előre adott, de változtatható.
- Különböző C értékekhez különböző klasszifikáció tartozik.
- Jóság mérése: klasszifikációs mátrix, ami a megfigyelt és a becsült értékek kontingencia táblája.

	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$	A	B
$Y = 0$	C	D

A, D: a helyes osztályozások száma
B, C: elsőfajú hibák száma
Helyes osztályozási ráta: $\frac{A+D}{A+B+C+D}$

Példa

Magyarországi mentőautók állománya: $Y = 1$, ha a mentő balesetmentesen közlekedett egész évben, egyébként $Y = 0$; $Kor = 1$, ha a mentőautót 40 évesnél fiatalabb sofőr vezeti, 0 egyébként; $Hely = 1$, ha a mentőautó budapesten van szolgálatban, 0 egyébként; továbbá $O1 = 1$, ha betegszállító, $O2 = 1$, ha rohammentő, és $O3 = 1$, ha gyermekmentő az autó típusa, egyébként $O1 = O2 = O3 = 0$. A balesetmentesség valószínűségét modellező logisztikus regresszió paraméterbecslése a következő:

Változó	b	s.e.(b)	exp(b)
O1	0.858	0.157	2.358
O2	-0.160	0.174	0.852
O3	-0.920	0.149	0.398
Kor	-1.062	0.244	0.346
Hely	-2.420	0.140	0.089
Konstans	2.248	0.299	9.466

Határozza meg egy 38 éves sofőr vezetésével vidéken szolgálatot teljesítő betegszállító balesetmentességének valószínűségét!

Főkomponens-analízis

Célok

- Az adatok dimenziójának (azaz a változók számának) csökkentése úgy, hogy lényeges információt ne veszítsünk.
- Lényegkiemelés, avagy nehezen megfogható fogalmak (pl. gazdasági fejlettség) definiálása összetett mutatórendszerrel való jellemzés útján - **Látens változók előállítása**
- Osztályozási feladatok adott közös faktorok alapján.
- Független komponensek előállítása

Az ok az esetek legnagyobb részében a változók függetlenségi és közel azonos szórású tulajdonságainak hiányában rejlik.

Eszköz: $k < n$ darab új változó bevezetése úgy, hogy ezek már korrelálatlanok legyenek, és a minta teljes varianciáját (azaz a változók összes szórását) a lehető legjobban adják vissza.

Matematikai háttér

Legyen X egy olyan n dimenziós valószínűségi vektorváltozó, melyre $\mathrm{E}X = 0$ és $\mathrm{Var}X = D$.

Cél: az n dimenziós térben olyan új koordináta-rendszer bevezetése, melyben véletlen vektorunk koordinátái már korrelálatlanok

Megoldás: ha $u_1, \ldots, u_n$ ortonormált bázis $\mathbb{R}^n$-ben, akkor X i-dik koordinátája ebben a bázisban $Y_i = u_i^T X$ lesz, azaz

$$Y = U^T X, \quad \text{ahol} \quad U = [u_1, \ldots, u_n] \quad \text{ortonormált mátrix.}$$

Azaz $X = UY$, vagyis X-et egy korrelálatlan komponensű véletlen vektor elforgatottjaként kívánjuk előállítani. Ha tehát $\mathrm{Var}\,Y = \Lambda$ diagonális mátrix, akkor

$$D = \mathrm{Var}X = \mathrm{Var}(UY) = U\Lambda U^T,$$

ami éppen D spektrálelőállítása.

Néhány fontos megjegyzés

- $Y = U^T X$ az X ún. főkomponens-vektora, ennek i-dik koordinátája X i-dik főkomponense.
- Ők már páronként korrelálatlanok, 0 várható értékűek és szórásaik éppen D sajátértékeinek négyzetgyökei, azaz $DY_i = \sqrt{\lambda_i}$. Ezek X ún. kanonikus szórásai.
- Y forgatásinvariáns, azaz ha V ortonormált mátrix, akkor X és VX főkomponens-vektora megegyezik.
- Y viszont nem skalárinvariáns, azaz érzékeny a mértékegység megváltoztatására.
- Ha rang$D = n$, akkor az összes főkomponens nullától különböző. Ha rang$D = k < n$, akkor csak az első k főkomponens nem nulla.
- Ha D sajátértékei különbözőek, akkor a főkomponensek előjeltől eltekintve egyértelműek. Ha nem, akkor a többszörös sajátértékhez tartozó sajátvektorok tetszőlegesen forgathatók az általuk meghatározott sajátaltérben, azaz nem lesznek egyértelműek.

Szemléletes jelentés

- Az új koordinátarendszer első, u_1 irányának megfelelő tengelye éppen azt az irányt jelöli ki, melynek irányában X szórása a legnagyobb az összes lehetséges n dimenziós irány közül, és ez a maximum éppen X kanonikus szórása. Ez az első főtengely.
- A második főtengely az elsőre merőleges irányok közül az, melyre nézve X szórása a legnagyobb. Ez éppen a második kanonikus szórás lesz.
- Ezt tovább folytatva az alábbi eredmény fogalmazható meg:

Tétel 1.

$$\max\{e^T De : \|e\| = 1\} = u_1^T Du_1 = \lambda_1$$

$$\max\{e^T De : \|e\| = 1, e^T u_j = 0, 1 \leq j < i\} = u_i^T Du_i = \lambda_i$$

Dimenzió-csökkentés

Feladat: egy n dimenziós véletlen vektor közelítése alacsonyabb dimenziós változóval.

Megoldás: legyen $\mathcal{L}_k$ azon n dimenziós Z valószínűségi változók tere, melyek egy valószínűséggel egy (legfeljebb) k dimenziós hipersíkban veszik fel értékeiket (azaz rang $\mathrm{Var} Z \leq k$). Keressük azt a $Z \in \mathcal{L}_k$ valószínűségi változót, amely négyzetes hibában a legjobban közelíti X-et, azaz

$$\mathrm{E}\|X - Z\|^2 \to \min!$$

Megmutatható, hogy $\mathrm{E}Z = \mathrm{E}X = 0$, és Z éppen X merőleges vetülete lesz az $u_1, \ldots, u_k$ vektorok által generált k dimenziós altérre. Azaz

$$Z = \sum_{i=1}^{k} Y_i u_i.$$

Ekkor $\mathrm{E}\|X - Z\|^2 = \lambda_{k+1} + \ldots + \lambda_n$.

k meghatározása

- X teljes varianciája:

$$\mathrm{E}\|X\|^2 = \lambda_1 + \ldots + \lambda_n.$$

  Ez megegyezik az Y főkomponens-vektor teljes varainciájával, hiszen a forgatás normatartó, azaz $\mathrm{E}\|Y\|^2 = \lambda_1 + \ldots + \lambda_n$.

- Z teljes varianciája:

$$\mathrm{E}\|Z\|^2 = \lambda_1 + \ldots + \lambda_k,$$

  azaz az első k főkomponens ennyit magyaráz meg X teljes varianciájából.

- k megválasztása: a

$$\psi_k = \frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_n}$$

  variancia-hányados sorozat alapján.

Példa - gépkocsi jellemzők vizsgálata

Változók:

- MPG - fogyasztás (mérföld/gallon)
- ENGINE - motor űrtartalma (inch-köb)
- HORSE - lóerő
- WEIGHT - súly (font)
- ACCEL - gyorsulási idő

Főkomponens-analízist szeretnénk végezni ezen 5 folytonos változó segítségével!

Eszköz: SPSS - eredmények az output-on.