Kategoriális magyarázó változók

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 9.

Tartalom

Minőségi változók a regresszióban

- A kérdés, ami mostani kutatásainkat motiválja: hogyan szerepeltethetünk egy *minőségi* (nominális vagy ordinális, szokás kategoriális változónak is nevezni) tulajdonságot, pl. férfi–nő, egészséges–beteg, alapfokú–középfokú–felsőfokú végzettségű stb. egy regressziós modellben
- A regresszió csak számszerű adatokat tud felhasználni $\rightarrow$ valahogy *kódolni* kell a kategoriális tulajdonság lehetséges értékeit (kimeneteit, csoportjait)
- Eddig csak mennyiségi tulajdonságokkal foglalkoztunk, aminek kódolása triviális volt: a naturáliában kifejezett értékével (m^2, eFt stb.)
- Pl. férfi $= 0$, nő $= 1$ elég kézenfekvő, de mi van az iskolai végzettséggel?
- Az alap $= 0$, közép $= 1$, felső $= 2$ belekódolja az adatokba, hogy a felső és a közép közti különbség *kénytelen* ugyanakkora lenni, mint a közép és alap közötti (ha felső $= 3$, akkor kétszer akkora stb.)
- De mi semmi ilyet nem akarunk, hiszen azt szeretnénk, hogy ezt az adatok mondják meg!

Dummy változó fogalma

- A kódolást megvalósíthatjuk olyan változóval vagy változókkal, melyek *csak* 0 vagy 1 értéket vehetnek fel
- Az ilyen változókat nevezzük dummy (bináris vagy indikátor) változónak
- Ha két kimenet van, akkor a kódolás teljesen kézenfekvő: egy dummy változóra van szükségünk, mely (például) 0 értéket vesz fel férfira, 1-et nőre
- Bonyolultabb a helyzet, ha több kimenet van
- Triviális kódolás:

	D_A	D_B	D_C
A	1	0	0
B	0	1	0
C	0	0	1

Kódolás

Ezen „kódolási tábla” alapján a kódolás (pl. X_1: jövedelem, X_2: iskolai végzettség, X_3: életkor):

$$\begin{array}{c} \\ 1 \\ 1 \\ \vdots \\ 1 \end{array}\!\!\begin{array}{c} X_1 \\ 213 \\ 311 \\ \vdots \\ 128 \end{array}\!\!\begin{array}{c} X_2 \\ B \\ C \\ \vdots \\ B \end{array}\!\!\begin{array}{c} X_3 \\ 32 \\ 41 \\ \vdots \\ 18 \end{array} \rightsquigarrow \begin{array}{c} \\ 1 \\ 1 \\ \vdots \\ 1 \end{array}\!\!\begin{array}{c} X_1 \\ 213 \\ 311 \\ \vdots \\ 128 \end{array}\!\!\begin{array}{c} D_A \\ 0 \\ 0 \\ \vdots \\ 0 \end{array}\!\!\begin{array}{c} D_B \\ 1 \\ 0 \\ \vdots \\ 1 \end{array}\!\!\begin{array}{c} D_C \\ 0 \\ 1 \\ \vdots \\ 0 \end{array}\!\!\begin{array}{c} X_3 \\ 32 \\ 41 \\ \vdots \\ 18 \end{array}$$

Itt már minden tisztán numerikus, működhet a regresszió

Referencia-kódolás

- ...ám vegyük észre, hogy 3 csoporthoz *nem* kell 3 dummy változó, kódolható 2-vel is!
- Általában k kimenet kódolása megoldható $k-1$ dummy változóval az ún. referencia-kódolás logikájával
- Itt kiválasztunk egy kimenetet, aminél mind a $k-1$ darab dummy változó 0 értéket vesz fel (kontrollcsoport vagy referenciacsoport), és a többi $k-1$ csoportot az jelzi, hogy a $k-1$ dummy változó közül *melyik* vesz fel 1 értéket (mindig csak 1!)
- Például (3 kimenetre):

	R_A	R_B
A	1	0
B	0	1
C	0	0

- Itt C a referenciacsoport, R_A és R_B a két szükséges (ugye $k=3$!) magyarázó változó
- Vegyük észre, hogy $R_A \equiv D_A$ és $R_B \equiv D_B$ (tehát a két kódoláshoz pontosan ugyanazon dummykra van szükség, csak a referencia-kódolásnál eldobjuk az egyiket – ez lesz a kontrollcsoport)

Dummy változó csapda

- Ha van konstans a modellben, akkor *tilos* is k csoporthoz k dummyt használni a kódoláshoz
- Ellenkező esetben egzakt multikollinearitás jön létre (gondoljuk végig, hogy a dummy változókhoz mi tartozik a design mátrixban, ld. előbb!); ez az ún. *dummy változó csapda*
- Ha k csoportot mégis k dummyval kódolunk („triviális kódolás"), akkor viszont nem szerepeltethetünk konstanst

Triviális kódolás konstans nélkül

- A két kódolási mód (k darab dummy, nincs konstans és $k-1$ darab dummy, van konstans) jól szemléltethető egy csak a nominális tulajdonsággal magyarázó regresszióval
- k darab dummy, nincs konstans:

	D_A	D_B	D_C
A	1	0	0
B	0	1	0
C	0	0	1

$$Y = \beta_A D_A + \beta_B D_B + \beta_C D_C + \varepsilon$$

- Együtthatók értelmezése: ha az A csoportban vagyunk, akkor a fenti egyenlet $Y = \beta_A + \varepsilon$ lesz $\Rightarrow$ β_A az A csoport csoportátlaga (legkisebb négyzetes elv!); hasonlóan a többi

Referencia-kódolás konstanssal

- $k-1$ darab dummy, van konstans:

	D_A	D_B
A	1	0
B	0	1
C	0	0

$$Y = \beta^* + \beta_A^* D_A + \beta_B^* D_B + \varepsilon$$

Együtthatók értelmezése referencia-kódolásnál

- Értelmezésnél egy dolgot tartsunk mindig szem előtt: ugyanarra a csoportra ugyanannak az értéknek kell kijönnie, akárhogy kódolunk!
- Így $\beta_C = \beta^*$
- Továbbá (a B csoport példáján):

$$\beta_B = \beta^* + \beta_B^* = \beta_C + \beta_B^* \Rightarrow \beta_B^* = \beta_B - \beta_C$$

- Tehát az együtthatók az *eltéréseket* jelentik a referenciacsoporttól (ami pedig a konstansba kerül)
- Vegyük észre, hogy a változónkénti szignifikanciák eltérhetnek (mert másra fognak vonatkozni), de az előrejelzése – és így a modellminősítő mutatók – nem

Fontos hipotézisvizsgálatok

- Egyrészt: szignifikáns-e egy adott csoport átlagának eltérése a referenciacsoport átlagától
- Ez itt nem más, mint β_A^* vagy β_B^* relevanciája
- Egyszerűen t-próbával ellenőrizhető!
- Másrészt: van-e egyáltalán bármilyen csoportok közötti eltérés:

$$H_0 : \beta_A^* = \beta_B^* = \ldots = 0$$
$$H_1 : \exists j : \beta_j^* \neq 0$$

- Több csoport átlaga eltér-e? De hát az az ANOVA!
- Az egyezés nem pusztán formai, teljes tartalmazi egyezés van (ez nem csak hasonló, hanem ugyanaz: az ANOVA elmondása regressziós „keretben")

Egynél több kategoriális magyarázó változó

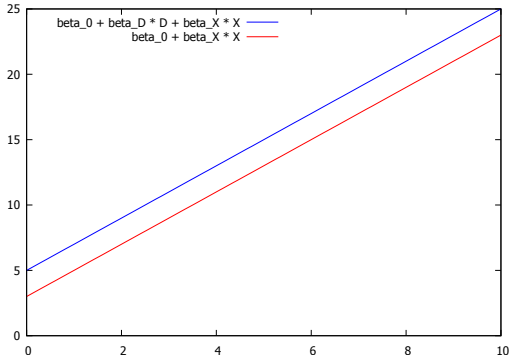
- Ha egynél több kategoriális magyarázó változó van, akkor nem kódolható mindegyik triviálisan, ilyenkor már a konstans eltávolítása sem segít
- (Nem az lesz a baj, hogy valamelyik összege a konstans, hanem, hogy a kettő összege ugyanaz – ez elvileg is megoldhatatlan)
- Referencia-kódolás minden további nélkül használható
- A kétszempontos ANOVA megfelelője regressziós keretben!
- Természetesen feltételezhető interakció is, ez esetben a dummy-kat az összes lehetséges kombinációban szorozni kell

Dummyzás folytonos magyarázó változó jelenléte mellett

- Amit eddig csináltunk az lényegében az volt, amit *konstans dummyzásának* nevezhetünk: csoportonként eltérő (de konstans) értékkel becsültük az eredményváltozót
- Mi van, ha bevonunk egy magyarázó változót?
- Azaz ekkor már nem egy konstanst becsülünk az egyes csoportokra, hanem egy egyenest (a folytonos magyarázó változó függvényében)
- Dummyzással (tehát a csoporttagság szerint) eltéríthetjük az egyenesek tengelymetszetét és meredekségét is!
- Lehet csoportonként különböző
  1. $+1$ egység magyarázó változó hatása
  2. a 0 magyarázó változóhoz tartozó eredményváltozó
- E feladat neve: ANCOVA

Eltérő tengelymetszet

Ha csak a tengelymetszetet térítjük el (+1 egység magyarázó változó hatása ugyanaz minden csoportban, de nem ugyanannyi a 0 magyarázó változóhoz tartozó eredményváltozó):

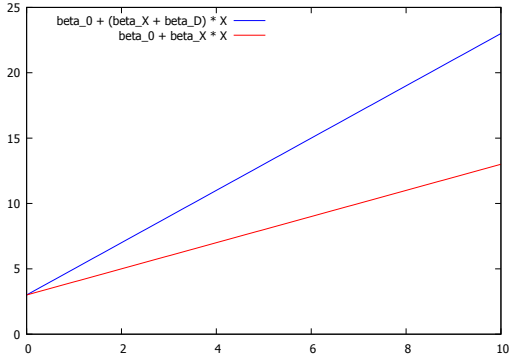


Algebrailag:

$$Y = \beta_0 + \beta_D D + \beta_X X + \varepsilon$$

Eltérő meredekség

Ha csak a meredekséget térítjük el (0 magyarázó változóhoz ugyanaz az eredményváltozó tartozik, de $+1$ egység magyarázó változó hatása csoportonként eltérő):



Algebrailag:

$$Y = \beta_0 + (\beta_X + \beta_D D)\, X + \varepsilon$$

Eltérő tengelymetszet és meredekség

- Akár a tengelymetszet *és* a meredekség is lehet különböző
- Ahogy előbb láttuk, csak a módszereket kell kombinálni: a konstanst és a meredekséget is megdummyzzuk:

$$Y = \beta_1 + \beta_2 X + \varepsilon,$$

de úgy, hogy $\beta_1 = \alpha + \alpha_A D_A + \alpha_B D_B$ és $\beta_2 = \gamma + \gamma_A D_A + \gamma_B D_B$

- Nagyon fontos észrevenni, hogy a meredekség dummyzása a dummy és a mennyiségi változó közti interakcióra vezet:

$$Y = \alpha + \alpha_A D_A + \alpha_B D_B + \gamma X + \gamma_A (D_A X) + \gamma_B (D_B X) + \varepsilon$$

- Logikus is: az egyik változó (folytonos) hatása eltér a szerint, hogy a másik változónak (kategoriális) mi a szintje: különböző meredekségek
- Avagy fordítva elmondva (egyenértékűen, hiszen az interakció ugye szimmetrikus): az egyik változó (kategoriális) hatása eltér a szerint, hogy a másik változónak (folytonos) mi a szintje: az egyenesek közti különbség függ attól, hogy hol nézzük

Eltérő tengelymetszet és meredekség

- De hát ez megoldható a minta szétszedésével is!
- A két módszer – természetesen – ugyanarra az eredményre vezet
- A dummyzás mégis jobb a minta szétszedésénél; vajon miért?
  - Messzemenően több lehetőségünk van a dummyzott (egybenlévő) modellel → gazdaságilag releváns hipotézisek vizsgálhatóak egyszerűen (ld. mindjárt)

Hipotézisvizsgálat a dummyzott modellben

- Pl.: van-e egyáltalán bármilyen eltérés a csoportok között? (Értsd: eltér-e a becsült egyenes (bármilyen szempontból) a csoportok között, vagy mindegyikben teljesen ugyanaz?)
- Ez az ún. *strukturális törés*, hipotézispárja: $H_0 : \alpha_A = \alpha_B = \gamma_A = \gamma_B = 0$, H_1: valamelyik ezek közül nem nulla, tehát van strukturális törés
- És most jön a szép rész: ha a fenti modellt megbecsültük (sima OLS-sel), akkor ez a hipotézis egyszerűen egy közönséges Wald- (vagy hasonló) próbát jelent!
- Hasonlóképp: nem lehet, hogy csak a tengelymetszetek eltérőek? $\rightarrow$ ez az ún. *párhuzamos ráták* hipotézise, $H_0 : \gamma_A = \gamma_B = 0$; szintén Wald-teszttel elintézhető
- Minden hasonló, gazdasági kérdés *lefordítható* ökonometriailag, például változó vagy változók relevanciájának tesztelésére

Kontraszt-kódolás

- Kontraszt-kódolás: trükkös kódolás úgy kitalálva, hogy a dummy-k együtthatója ne a referencia-csoporthoz, hanem az átlaghoz képesti eltérést jelentse
- A megoldás:

	C_A	C_B
A	1	0
B	0	1
C	-1	-1

- (A dummy változó nem 0 és 1 értéket vehet csak fel)
- Miért fog ez működni?

Kontraszt-kódolás

Mert:

$$\beta_0 + \beta_{C_A} + 0 = \overline{y}_A \quad (1)$$
$$\beta_0 + 0 + \beta_{C_B} = \overline{y}_B \quad (2)$$
$$\beta_0 - \beta_{C_A} - \beta_{C_B} = \overline{y}_C \quad (3)$$

És így:

- (1)+(2)+(3) $\Rightarrow 3\beta_0 = \overline{y}_A + \overline{y}_B + \overline{y}_C \Rightarrow \beta_0$ tényleg a főátlag (ha azonosak a csoportok elemszámai! különben ún. súlyozott kontraszt kellene, ahol a dummy változók már nem is feltétlenül egész értékeket vennének fel)
- (2)+(3) $\Rightarrow$
  $2\beta_0 - \beta_{C_A} = \overline{y}_B + \overline{y}_C \Rightarrow \beta_{C_A} = 2\beta_0 - (\overline{y}_B + \overline{y}_C) = 2\beta_0 - (3\beta_0 - \overline{y}_A) \Rightarrow \beta_{C_A} = \overline{y}_A - \beta_0$
  $\Rightarrow$ tényleg az átlagtól való eltérés (és hasonlóan a másik)

Egy terminológiai megjegyzés

- Az angol irodalomban az általunk kontrasztkódolásnak nevezett módszert nagyon gyakran „effect coding"-nak nevezik...
- ...a kontraszt pedig az, amikor a csoportok tetszőleges – általunk meghatározott – lineáris kombinációját teszteljük