6. előadás - Regressziószámítás II.

2016. október 10.

Specifikációs hibák

- A magyarázó- és eredményváltozók kiválasztásának alapja: szakirányú elmélet, mögöttes viselkedés ismerete, múltbeli tapasztalatok.
- Mivel a valós összefüggéseket nem ismerhetjük pontosan, így nyilván az ökonometriai modellek hibákat tartalmaznak.
- Specifikációs hiba: helytelen modellfelírás, akár a változók megválasztását, akár a függvényformák vagy az eltérésváltozók (u_t) struktúráját tekintetve.
- A harmadik esettel foglalkoztunk már (GLS becslés), most az első kettőn van a hangsúly: lényeges változó kihagyása, felesleges változó szerepeltetése, valamint a megfelelő függvényforma kiválasztása.
- De, míg eddig csak ezek matematikai hatásaival foglalkoztunk, most a struktúrális hatás is érdekel bennünket, azaz hogy ezek hogyan hatnak a modell belső struktúrájára.

Lényeges változó kihagyása - Példa

Tegyük fel, hogy a valódi modell

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

alakú, de mi az

$$Y_t = \beta_1 + \beta_2 X_{2t} + v_t$$

modellt becsültük. (Azaz a $\beta_3 = 0$ feltételezéssel éltünk, de ez hibás.) Ekkor könnyen látható, hogy $v_t = \beta_3 X_{3t} + u_t$, melyre

$$\mathrm{E}v_t = \beta_3 X_{3t} \neq 0,$$

tehát v_t megsérti a modell feltételeit. Továbbá az is megmutatható, hogy ha a $\beta_3 = 0$ feltételezés hibás, akkor az összes többi együttható becslése torzított lesz, azaz az előrejelzések is torzítottak lesznek és a tesztek is hatástalanok.

Változó bevonásának hatása a modellre - Példa

Tekintsük a következő feladatot:

- Eredményváltozó: (Y) a háztartások kiadása
- Magyarázóváltozó: (X_J) a jövedelem
- Magyarázóváltozó: (X_L) a család létszáma

Tegyük fel, hogy modellünket kétféle módon becsültük, és az alábbi eredményeket kaptuk:

$$\hat{Y}_t = 339,746 + 0,637 \cdot X_J, \qquad \bar{R} = 0,5369$$

$$\hat{Y}_t = 283,172 + 0,617 \cdot X_J + 34,17 \cdot X_L, \qquad \bar{R} = 0.5386$$

Miért változtak meg az együtthatók?

Változó bevonásának hatása a modellre - Példa

- Tegyük fel, hogy a bővebb, második modell írja le a valóságos helyzetet (ez persze csak feltételezés, hiszen a valóságot nem ismerhetjük).
- Lehet-e ez alapján következtetni az első modellbeli jövedelem együtthatóra?
- Nézzük meg, hogy van-e a két változó között kapcsolat, azaz illesszünk ezekre is egy regressziót!

Az eredmény:

$$\hat{X}_L = 1,655 + 0,000598 \cdot X_J, \quad \bar{R} = 0,2359,$$

azaz a jövedelem nem csak a kiadásra hat sztochasztikusan, hanem a család létszámára is!

Változó bevonásának hatása a modellre - Példa

Tehát azt kaptuk, hogy a szűkebb modell együtthatói előállíthatók a bővebb modell és a harmadik regresszió segítségével, mégpedig

$$339,746 = 283,172 + 34,17 \cdot 1,655$$

és

$$0,637 = 0,617 + 34,17 \cdot 0,000598$$

formában. Mi ennek az oka?

- A bővebb modellben a jövedelem direkt hatása szerepel csak, mert itt a család létszámát állandó értéken tartjuk, ezért nincs jelentősége a köztük lévő kapcsolatnak.
- A szűkebb modellben viszont a jövedelem egységnyi növekedése a létszámot is növeli tendenciájában, a növekvő létszám viszont önmagában is növeli a kiadást. Ez az ún. indirekt hatás.

Változó bevonásának hatása a modellre - Példa

- A szűkebb modellben nem tudjuk izolálni a létszám hatását, a jövedelem változásával a létszám is változik, míg a bővebb modellben a jövedelem növekedése vagy csökkenése nem jár a létszám megváltozásával.
- Tehát a szűkebb modellben a kihagyott változón keresztül terjedő hatások is beépülnek az együtthatókba.
- A becsült paraméterekre gyakorolt hatás nehezen következtethető ki ebből, mert az indirekt hatástól függ a torzítás iránya és nagysága.

A marginális hatás

- A marginális hatás: a magyarázó változó egységnyi növelésének hatására mennyit változik az eredményváltozó
- Mértékegységekkel nem kell törődni.
- Matematikai definíció:

$$\frac{\partial Y}{\partial X_j}$$

- Ha a modellünk az

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u,$$

akkor

$$\frac{\partial Y}{\partial X_j} = \beta_j,$$

azaz a marginális hatás és a becsült regressziós együttható kvázi szinonimák! (Persze csak akkor, ha a modell minden korábbi feltételnek eleget tesz!)

Interakció

- Térjünk vissza a korábbi példához: igaz-e az, hogy adott egységnyi pluszjövedelem a család létszámától függetlenül azonos többletkiadást jelent? Nyilván nem...
- Ilyenkor azt modjuk, hogy a két változó között interakció van, azaz az egyik marginális hatásának nagyságát befolyásolja a másik szintje.
- Tehát a kapcsolat a marginális hatás és a szint között van, nem pedig a marginális hatások közt avagy a szintek közt külön-külön!

Interakció

Tegyük fel, hogy a kapcsolat lineáris, azaz az egyik változó szintje lineárisan hat a másik változó marginális hatására. Például

$$(\beta_J + \beta_{JL} X_2) X_1,$$

ahol β_{JL} az interakció hatását kifejező együttható. Ekkor

$$\begin{aligned} Y &= \beta_1 + (\beta_J + \beta_{JL} X_2) X_1 + \beta_L X_2 + u = \\ &= \beta_1 + \beta_J X_1 + (\beta_L + \beta_{JL} X_1) X_2 + u \end{aligned}$$

azaz az interakció szükségképpen szimmetrikus, tehát a hatás kölcsönös! Ezért írható helyette egyszerűbben a

$$\beta_J X_1 + \beta_L X_2 + \beta_{JL} X_2 X_1$$

összefüggés is, a marginális hatás ebből már adódik.

Marginális hatás interakciók esetén

Ha interakció van a modellben, például az l-dik és m-dik tag közt, akkor

$$\begin{aligned}\frac{\partial Y}{\partial X_l} =& \frac{\partial}{\partial X_l}(\beta_1 + \beta_2 X_2 + \ldots + \\ &+\beta_l X_l + \ldots + \beta_m X_m + \ldots + \beta_k X_k + \beta_{lm} X_l X_m + u) = \\ &=\beta_l + \beta_{lm} X_m,\end{aligned}$$

azaz az egyik változó szerinti marginális hatás tényleg függ a másik változó szintjétől!

Nemlinearitás - alapgondolatok

- Az élet általában nem lineáris.
- Mégis, a lineáris modellek sokszor nem térnek el (nagyon) a valóságtól, ráadásul matematikailag sokkal könnyebben kezelhetők, mint a nemlineáris modellek.
- Az esetek többségében persze ez csak egy közelítés lesz, de ha "ügyesek" vagyunk, akkor ez nem nagy baj.
- Érdemes vizsgálni a modellek érvényességi határait.

Nemlinearitás fajtái

Változójában nemlineáris modell:

- Továbbra is fennáll a modell-struktúra, azaz csak alapműveleteket végzünk a paraméterek és a változók közt.
- Az OLS szempontjából mindegy, hogy egy változó esetlegesen egy másik valamilyen transzformáltjaként áll elő.
- Ide tartozik a kvadratikus és polinomiális hatás, a logaritmikus és exponenciális hatás, stb...
- Az interakció szintén változóbeli nemlinearitás.
- Minden korábbi becslés és technika ugyanúgy működik, mint korábban.
- De, vigyázni kell, paramétereiben lineáris marad a modell!

Nemlinearitás fajtái

Paramétereiben nemlineáris modell:

- Megsérti a lineáris kombináció struktúráját, hiszen a paraméter nem csak együtthatóként szerepel a modellben.
- Ez már nem becsülhető az OLS módszerrel, mert az eredményváltozó nem állítható elő mátrixműveletekkel.
- Nemlineáris legkisebb négyzetek (NLS) módszerével becsülhetőek ezek a modellek, de ezek általában nehezen, vagy csak lokálisan megoldható feladatok. (Iteratív algoritmusokat használunk, de ezeknek gond lehet a konvergenciájával, stabilitásával és az egyértelműségével is.)
- Kezelése: algebrai linerizációval és az OLS használatával történik. (Ez persze nem mindig működik, de a gyakorlatban fontos esetekben azért általában használható.)

Log-Log modell

- Kedvelt a termelési és keresleti függvények becslésénél, pl. a jól ismert Cobb-Douglas termelési függvény is ilyen:

$$Y = \beta_1 L^{\beta_L} K^{\beta_K} u$$

ahol Y a kibocsátás, L a munka, K a tőke felhasználása.

- Linearizálás: vegyük mindkét oldal logaritmusát. Ekkor

$$\log Y = \log \beta_1 + \beta_L \log L + \beta_K \log K + u'$$

Log-Lin modell

- Tipikusan jövedelmek, bérek alakulásának modellezése a tanulással vagy tapasztalatszerzéssel töltött évek függvényében.
- A modell:

$$Y = e^{\beta_1 + \beta_2 X + u}$$

azaz

$$\log Y = \beta_1 + \beta_2 X + u$$

- Tehát az eredményváltozót logaritmizálva a magyarázóváltozók maradnak szintben.

Lin-Log és reciprok modellek

- Lin-Log modell: tipikusan a termés és a megművelt terület nagysága, vagy a terület és a kínálati ár összefüggésének jellemzésére szolgál.

$$Y = \beta_1 + \beta_2 \log X + u$$

- Reciprok modell: tipikusan a keresleti modellek ilyenek. Itt

$$Y = \beta_1 + \frac{\beta_2}{X} + u$$

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

$$P_X = P(Y = 1|X)$$

(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

$$\text{odds}_X = \frac{P_X}{1-P_X} \in [0, \infty)$$

Visszafejtve a transzformációt

$$P_X = \frac{P_X}{P_X + 1 - P_X} = \frac{\frac{P_X}{1-P_X}}{\frac{P_X}{1-P_X} + 1} = \frac{\text{odds}_X}{1 + \text{odds}_X}$$

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

3. trükk: logaritmálás, azaz a logit bevezetése

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

Ezzel a siker és kudarc eloszlását is szimmetrizáltuk!

Modell

Erre illesztünk tehát lineáris modellt

$$\text{logit}_X = \alpha + \beta X + u$$

alakban. Ezt nevezik logit avagy logisztikus regressziónak.

Innen, a becslések elvégzése után,

$$\text{odds}_X = e^{\hat{\alpha}+\hat{\beta}X}$$

és

$$P_X = \frac{e^{\hat{\alpha}+\hat{\beta}X}}{1 + e^{\hat{\alpha}+\hat{\beta}X}}$$

Becslés és az eredmények értelmezése

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

$$L(\alpha, \beta) = \prod_{Y_i=1} P_{X,i} \prod_{Y_i=0} (1 - P_{x,i})$$

- Átlagos növekedési ráta:

$$\frac{\text{odds}_{X+1}}{\text{odds}_X} = \frac{e^{\alpha+\beta(X+1)}}{e^{\alpha+\beta X}} = e^{\beta}$$

- Marginális hatás:

$$\frac{\partial P_X}{\partial X_i} = \beta P_X (1 - P_X)$$

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: $\frac{A+D}{A+B+C+D}$