

# Idősorok regressziója

Ferenci Tamás  
`tamas.ferenci@medstat.hu`

Utoljára frissítve: 2023. május 9.

# Tartalom

- 1 Exogén változós idősormodellek
  - Alapgondolatok, statikus regresszió
  - Dinamikus regressziók
  - Idősoros regressziók általános modellje
  
- 2 Idősoros regresszió becslése OLS-sel
  - Standard modellfeltevések
  - Az OLS véges mintás tulajdonságai idősorokra
  - Az OLS nagymintás tulajdonságai idősorokra

# Idősorok regressziójának alapgondolata

- Az idősorunkat *más* idősor(ok)kal akarjuk magyarázni
- Lényegében tehát *ki akarjuk regresszálni* az idősorunkat (mint eredményváltozót), más idősorokkal (mint magyarázó változókkal)
- Bizonyos értelemben az eddigi AR-modellek is ilyenek voltak, csak a „más idősor” ugyanannak a késleltetettjeit jelentette
- Most viszont megengedjük, hogy tényleg eltérő idősorok (vagy azok késleltetettjei!) is belépjenek magyarázó változóként
- A fő kérdésünk az lesz, hogy e modelleknek milyen feltételeket kell teljesíteniük, hogy jó tulajdonsággal becsülhetőek legyenek a paramétereik
- És persze szokásosan az ökonometriai modellek két felhasználása: elemzés és előrejelzés

# Statikus regresszió

- A legegyszerűbb idősoros regressziós modell:

$$y_t = \beta_0 + \beta_1 z_t + u_t,$$

ahol  $y_t$  és  $z_t$  tehát két idősor *ugyanazon* időpontbeli megfigyelései

- Például: statikus Phillips-görbe (infláció vs. munkanélküliség)
- Az  $u_t$  hibatag és tulajdonságai lesznek majd vizsgálódásunk fókuszában, ami a modellfeltevéseket illeti
- Legegyszerűbb eset, ha  $z_t$  valami „teljesen más”,  $u_t$ -től külső információ (ezt majd pontosítjuk), úgy fogjuk mondani, hogy **exogén változó**
- Természetesen lehet több exogén változó is:

$$y_t = \alpha + \beta_1 z_{t1} + \beta_2 z_{t2} + \dots + \beta_k z_{tk} + u_t$$

# Statikus regresszió

- $\beta_i$  jelentése: ha az  $i$ -edik idősor értéke egy adott időszakban egy egységgel megnő (minden mást változatlanul tartva), akkor modellünk szerint várhatóan hány egységgel lesz nagyobb az eredményváltozó *ugyanazon* időszakban
- Azért hívjuk statikusnak, mert ugyanazon időszaki változásokat kapcsol össze, tehát nincs időszakok közötti hatás
- (A dinamika szó általában is időben kiterjedten lezajló dolgokra utal)

# A dinamika szükségessége

- Rengeteg helyzetben a statikusság irreális
- Nem várható, hogy egy beruházás rögtön *ugyanabban* az időpontban befolyásolja a kibocsátást, hogy egy felvilágosítókampány *azonnal* lecsökkenti a megbetegedések számát, hogy egy szociálpolitikai intézkedés *rögtön* megváltoztatja a jövedelmi viszonyokat stb.
- A legtöbb társadalmi-gazdasági jelenség csak időben elnyújtva, késleltetéssel hat, több időszakon keresztül fejti ki a hatását
- Ezért szükséges a dinamika beépítése is a regressziós modelljeinkbe

# Osztott késleltetésű modellek

- A legegyszerűbb dinamikus modell: legyen most csak egyetlen magyarázó változónk, csak épp

$$y_t = \beta_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \dots + \delta_p z_{t-p} + u_t$$

- Az idősor tárgyidőszaki értékére a korábbi  $z$ -k is hatást gyakorolnak...
- ... avagy – fordítva elmondva ugyanazt – a  $z$  mostani változásai a jövőben fognak kihatni  $y$ -ra
- $\beta_i$ : ha  $z$  most megváltozik, akkor  $i$  időszakkal később ez hogyan hat  $y$ -ra
- Amennyiben véges sok korábbi  $z$  hat  $y$ -ra (később fogjuk látni hogyan hathat végtelen sok korábbi), akkor **véges osztott késleltetésű** modellről (FDL, finite distributed lag) beszélünk

## Az FDL-modell

- Ha  $z$  konstans, majd egy időszakra felugrik eggyel nagyobbra, majd után visszaáll a konstans szintre, akkor a tárgyidőszaki  $y$   $\beta_0$ -al lesz nagyobb mint az állandósult szintje, az eggyel később  $y$   $\beta_1$ -gyel, ..., a  $p$ -vel későbbi időszakban  $\beta_p$ -val, és a  $p$  utáni időszakokra már nem hat ez a módosulás
- Ezeket hívjuk **rövid távú hatásmultiplikátornak**, értéke tehát az  $i$ -edik időszakra épp  $\beta_i$
- Éppen ezért szokás kiplottolni  $\beta_i$ -t a  $i$ -vel szemben
- A másik tipikus értelmezési keret, hogy  $z$  egy adott időszakban felugrik eggyel és *úgy is marad*, kérdés, hogy hosszú távon mi történik  $y$ -nal
- Ugyanabban az időszakban  $\beta_0$ -al nő meg, a következőben  $\beta_0 + \beta_1$ -gyel és így tovább
- Ezt hívjuk **hosszú távú hatásmultiplikátornak**, értéke tehát  $\sum_{i=0}^p \beta_i$



# FDL-modell strukturálatlan becslése

- Ha a fenti módon egyszerűen megbecsüljük  $\beta$ -kat, akkor **strukturálatlan becslésről** beszélünk (mert semmit nem tettünk fel a  $\beta$ -k értékeiről)
- A probléma ezzel, hogy nagyon sok esetben egy idősor egymást követő értékei nagyon korreláltak  $\rightarrow$  a fenti modellben rendkívül erős multikollinearitás lesz, a  $\beta$ -kat csak nagyon bizonytalanul (hatalmas CI-vel) tudjuk csak becsülni
- (*Együttesen* vizsgálhatóak, például  $F$ -teszttel, vagy a hosszú távú hatásmultiplikátort is jól meg tudjuk becsülni, csak külön-külön nem)

## FDL-modell struktrált becslése, Almon-lag

- Éppen ezért gyakori, hogy nem teljesen szabadon becsüljük  $\beta_i$ -ket, hanem feltételezünk valamilyen struktúrát
- Lényegében: átcseréljük az eredeti paramétereket kisebb számú, kevésbé multikollineáris paraméterekre
- (Persze ennek az az ára, hogy a struktúrát el kell találnunk, az ugyanis nem az adatokból jön, hanem mi mondjuk meg kívülről)
- Az egyik népszerű választás az **Almon késleltetési struktúra**, amikor azt tételezzük fel, hogy a  $\beta_i$ -k az  $i$ -ben polinomiálisak:

$$\beta_i = \sum_{j=0}^n w_j i^j,$$

ahol  $n$  tipikusan kicsi (pl. 2-3)

- Akármennyi is  $p$ , nekünk csak  $n$  darab – általában már nem túl multikollineáris – paramétert kell becsülnünk
- De még egyszer: fontos, hogy  $\beta_i$  tényleg kvadratikuss ( /köbös/stb.) legyen  $i$ -ben

## FDL-modell strukturált becslése, Koyck-lag (GDL)

- Egy másik népszerű választás, hogy  $\beta_i$  geometriailag lecsengő  $i$ -ben:

$$\beta_i = \beta_0 \rho^i,$$

ahol természetesen  $|\rho| < 1$

- (Azt is mondhattuk volna, hogy  $\beta_i = \rho \beta_{i-1}$ )
- Ezt hívják **Koyck késleltetési struktúrának**
- Ami nagyon érdekes, hogy ehhez igazából az sem kell, hogy csak véges sok késleltetés lépjen be a modellbe!
- Nyugodtan lehet az a modellünk, hogy

$$y_t = \beta_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \dots + u_t,$$

*nem* lesz végtelen sok becsülendő paraméterünk, hiszen a DL részhez tartozó paraméterek száma *mindenképp* 2 ( $\beta_0$  és  $\rho$ )

- Tehát értelmesen megbecsülhető a fenti specifikáció is, mintegy végtelen osztott késleltetésű modellként, a neve **geometriai osztott késleltetű** modell (GDL, geometric distributed lag)

# A GDL-modell értelmezése

- A rövid távú hatásmultiplikátor tehát  $\beta_i = \beta_0 \rho^i$
- A hosszú távú hatásmultiplikátor izgalmasabb:

$$\sum_{i=0}^{\infty} \beta_i = \sum_{i=0}^{\infty} \beta_0 \rho^i = \beta_0 \sum_{i=0}^{\infty} \rho^i = \frac{\beta_0}{1 - \rho}$$

# Az eddigiek kombinációja

- Az eddigiek természetesen kombinálhatóak is: lehet benne *több*  $z$  is, akár késleltetve
- Természetesen lehet vegyesen is (bizonyosak késleltetés nélkül, mások késleltetéssel, a rend sem kell, hogy azonos legyen)
- A jobb oldalra berakhatjuk az eredményváltozó késleltettjeit is (itt nyilván egyidejű tagot nem rakhatunk be...), ezzel AR-hatást is létrehozhatunk

# Egy általános modell felé

- Láttuk tehát, hogy a magyarázó változó lehet:
  - Exogén  $z$  (mint a statikus regresszióban)
  - Exogén  $z$  késleltetettje (mint a DL-ben)
  - Az eredményváltozó késleltetettje (mint az AR-ben)
- Külön-külön mindegyiket néztük már – lásd a zárójeles megjegyzéseket – de semmi akadálya, hogy többet (vagy akár az összeset egyszerre) berakjuk egy modellbe!

# Az általános modell

- Mindent összetéve:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + u_t = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_t + u_t,$$

ahol  $x_{t,i}$  egyaránt *lehet* exogén változó, késleltett exogén változó vagy késleltetett eredményváltozó

- Formailag teljesen olyan, mint a regresszió keresztmetszetben, van eredményváltozó és vannak magyarázóváltozók
- Egyszerűen behúzzuk őket egy modellbe – történetesen nem keresztmetszeti adatok, hanem idősorok, de hát az OLS-nek mindegy, számok vannak így is, úgy is – és simán megbecsüljük OLS-sel... jó ötlet ez?
- A következőkben ezzel fogunk foglalkozni: mi történik akkor, ha a  $\beta$ -kat egyszerűen megbecsüljük OLS-sel, milyen feltételek mellett milyen tulajdonságúak lesznek az így kapott becslések?

# A modellfeltevések és szerepük

- A helyzet, és a kérdés teljesen analóg a keresztmetszetről látottakkal: milyen modellfeltevések mellett garantálhatóak, hogy az OLS szolgáltatott becsléseknek jó tulajdonságaik legyenek?
- Úgy fogjuk végignézni, hogy mindenhol a keresztmetszettel rakjuk párhuzamba
- Ugyanúgy 5 (+1) modellfeltevés lesz



# Linearitás

- Keresztmetszetnél ez volt:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$$

és ez igaz mindegyik megfigyelési egységre, és így az egész mintára is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + u_i$$

- Idősornál *pontosan ugyanez* a feltétel (legfeljebb  $i$  helyett  $t$ -t szokás írni)

# Nincs egzakt multikollinearitás

- Keresztmetszetnél ez volt:

$$\mathbb{P}(\text{rank } \underline{\underline{X}} = k) = 1$$

- Idősornál *pontosan ugyanez* a feltétel

# Szigorú exogenitás

- Keresztmetszetnél, ha trehányak voltunk, ez volt:

$$\mathbb{E}(u_i | \underline{X}_i) = 0,$$

ha precízek, akkor ez:

$$\mathbb{E}(u_i | \underline{\underline{X}}) = 0$$

- Idősornál, ha precízek voltunk, *pontosan ugyanez* a feltétel
- Tehát: *minden* időszaki hiba várható érték független *minden* (akár más időszaki!) magyarázó változótól

## Szigorú és egyidejű exogenitás

- A trehányság azért volt megengedhető, mert ha fae a mintavétel – ami keresztmetszetnél egy elfogadható feltevés lehet – akkor a precíz tényleg a trehányra egyszerűsödik
- Ez teljesen logikus: ha a különböző mintaelemek függetlenek, akkor egy adott időszaki hiba az összes többi időszaki magyarázó változótól nyilván várható érték független lesz, tehát csak az ugyanazon időszakiktól való függetlenséget kell megkövetelni
- Idősornál, mivel a fae mintavétel itt már nem elfogadható általánosságban, ez a trehányság nem lesz megengedhető
- A továbbiakban a  $\mathbb{E}(u_i | X_i) = 0$  feltételt **egyidejű exogenitásnak**, az – erősebb –  $\mathbb{E}(u_i | \underline{X}) = 0$  feltételt **szigorú (vagy erős) exogenitásnak** nevezzük
- (Most válik érthetővé, hogy a szigorú exogenitás elnevezésben mit jelent a szigorú!)
- A standard modellfeltevésben tehát a szigorú exogenitás szerepel

# A szigorú exogenitás sérülései

- Természetesen minden, amit keresztmetszetenél is láttunk (pl. kihagyott változó, mérési hiba)
- Itt azonban más okok is lehetnek a háttérben:
  - Rosszul megragadott dinamika: például statikus regressziót becslünk, miközben FDL lenne a helyes
  - Az  $u$ -beli változás nem befolyásolhatja a későbbi  $x$ -et (ez meg hogy lehetne? úgy, ha  $y$  értékei visszahatnak a későbbi  $x$ -kre, például bűnözés regresszálása a rendőri erők létszámával)

# Homoszkedaszticitás

- Keresztmetszetnél ez volt:

$$\sigma_i^2 := \mathbb{D}^2(u_i \mid \underline{X}) = \sigma^2$$

- Idősornál *pontosan ugyanez* a feltétel

# Autokorrelálatlanság

- Keresztmetszetnél ez volt: ha  $u_i$  a mintavétel, akkor automatikusan teljesül, különben

$$\text{cov}(u_i, u_j \mid \underline{X}) = 0$$

minden  $i, j = 1, 2, \dots, n, i \neq j$

- Idősornál *pontosan ugyanez* (az utóbbi) a feltétel

# Hibanormalitás

- Keresztmetszetnél ez volt:

$$\underline{u} \mid \underline{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

- Idősornál *pontosan ugyanez* a feltétel



# Összefoglalva

- Ha precízen fogalmaztunk, akkor igazából a keresztmetszetről látott feltételek egy-az-egyben ugyanazok, mint amire itt is szükség van
- Most már elárulható, hogy ez nem véletlen: a precíz fogalmazás *épp* azért kellett, hogy az ott látott dolgok valójában *univerzálisak* legyenek, tehát ne csak keresztmetszetre vonatkozzanak, hanem ugyanúgy idősorra is
- ...ami tulajdonképpen jól érthető is: a tiszta elmélet egységes kell legyen, hiszen a változóknak „mindegy”, hogy ők most idősorok, vagy keresztmetszeti adatok, vagy micsodák

# Az OLS véges mintás tulajdonságai idősorokra

- Az első három feltétel teljesülése esetén az OLS szolgáltatta becslések torzítatlanok
- Ha mind az öt feltétel teljesül, akkor az OLS szolgáltatta becslések ezen felül hatásosak is (azaz BLUE-k is)
- Ha mind az öt feltétel teljesül, akkor a  $\sigma^2$  és a hibák kovarianciamátrixának OLS szolgáltatta becslése torzítatlan
- Ha még a hibanormalitás is teljesül, akkor az OLS szolgáltatta becslések eloszlása normális, a  $t$  ( $F$ ) statisztikák nulleloszlásai tényleg  $t$ -k ( $F$ -ek), a szokásos tesztek és a konfidenciaintervallumok validak

## A keresztmetszeti esethez való viszony

- Mindez lényegében azt jelenti, hogy ezen feltevések teljesülése esetén az idősoros adatokkal *pontosan ugyanúgy* hajthatunk végre regressziót, mintha keresztmetszetiek lennének!
- Persze látni kell, hogy ezek rettentő erős feltevések voltak, a gyakorlatban ritkán teljesülnek

## Véges (vagy kis-) mintás tulajdonság mivolt

- A tulajdonságoknál nem mondtuk semmit a mintanagyságról: ez azt jelenti, hogy mintanagyságtól függetlenül – azaz minden mintanagyságra – igazak
- Ilyenkor azt szokták mondani, hogy ezek „kis” mintás (véges mintás) tulajdonságok voltak
- (A kismintás elég szerencsétlen elnevezés, hiszen természetesen nagy mintára is igazak, gyakorlati szempontból persze érthető a kifejezés oka)

# A nagymintás tulajdonság értelme és szükségessége

- Nagymintás: nem minden  $n$ -re igaz, hanem csak  $\lim_{n \rightarrow \infty}$  értelemben (szokás még aszimptotikus tulajdonságnak is nevezni)
- Fontos, mert a gyakorlatban a véges mintás tulajdonságokhoz tartozó feltételek sokszor nem teljesülnek, de nagy mintát néha van módunk venni, így nagyon lényeges annak vizsgálata, hogy ezzel mit tudunk „kiváltani”
- Igazából már keresztmetszetenél is láttunk egy nagymintás tulajdonságot: amikor azt mondtuk, hogy az első három tulajdonság fennállása esetén az OLS szolgáltatja a becslések konzisztensek

## Kitérő: idősorok ergodicitása

- Egy idősort **ergodikusnak** nevezünk, ha az időben távoli tagjai – bármely időpontból indulva – függetlenbe tartanak az időbeni távolságuk növekedtével (aszimptotikusan függetlenek)
- Egy ergodikus idősorra, ha még stacioner is (és így létezik  $\mu$ ) teljesül, hogy

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{i=1}^T Y_i \xrightarrow{\text{m.b.}} \mu$$

- Néha ezzel definiálják az ergodicitást (pontosabban szólva a várható értékben ergodicitást) – ilyenkor a stacionaritást meg kell követelni, vagy legalábbis óvatosan eljárni
- (Természetesen mindig definiálhatjuk az  $I_{\{Y_t \in A\}}$  idősort, ilyenkor a várható érték valószínűség lesz)

# Az ergodicitás tartalma

- Lényegében azt mondja ki, hogy időátlag tart a sokasági – összességi – átlaghoz:
  - Azért fontos, mert azt mondja, hogy elég sok elemet megfigyelve (az időben – ugye mi csak ezt tudjuk megtenni!) *tényleg* tudunk következtetni a várható értékekre/valószínűségekre (ami igazából érdekel minket!)
  - A nagy számok törvényének megfelelője, illetve általánosítása (nem kellett a teljes függetlenséget feltenni)
- Néha szokás ezt gyenge függőségnek is nevezni

# Ergodicitás és az autokovarianciák

- Érezhető, hogy ha egyszer az ergodicitás olyasmit követel meg, hogy az egyre távolabbi értékek egyre függetlenebbek legyenek (a teljes függetlenséghez tartva), akkor összefügg az autokovarianciákkal – hiszen azok is valami függetlenséggel kapcsolatban lévő dolgot mérnek
- Csakugyan, belátható, hogy egy idősor ergodik (a várható értékre), ha a kovarianciái nullába tartanak, mégpedig olyan gyorsan, hogy abszolút összegezhetőek is:

$$\sum_{i=1}^{\infty} |\gamma_i| < \infty$$



# Stacionaritás és ergodicitás

- Egy stacioner idősor nem feltétlenül ergodikus:  $Y_t = X$  (ahol  $X$  egy valószínűségi változó), azaz az idősor konstans
- Egy ergodikus idősor nem feltétlenül stacioner:  $Y_t = \alpha t + u_t$ , ahol  $\alpha \neq 0$  és  $u_t \sim \mathcal{N}(0, \sigma_u^2)$  függetlenül
- Nagyon sok esetben azonban a kettő ugyanaz (néhol keveredés is van emiatt a szóhasználatban)

## Az új modellfeltevések

- Pluszban megköveteljük a linearitásnál, hogy az idősorok legyenek stacionerek és ergodikusak is
- Cserében viszont
  - Szigorú (erős) exogenitás helyett elég lesz az egyidejű exogenitás:  $\mathbb{E}(u_i | \underline{X}_i) = 0$
  - Szigorú (erős) homoszkedaszticitás helyett elég lesz az egyidejű homoszkedaszticitás:  $\mathbb{D}^2(u_i | \underline{X}_i) = \sigma^2$
  - Szigorú (erős) autokorrelálatlanság helyett elég lesz az egyidejű autokorrelálatlanság:  $\text{cov}(u_i, u_j | \underline{X}_i, \underline{X}_j) = 0$
- (A hibanormalitásról nem tettünk fel semmit: nem kellett, mert úgyis aszimptotikus eredményeink lesznek, ahol a centrális határeloszlás tétel kihasználható – ugyanis a fenti feltételek mellett az is működni fog, nem csak a nagy számok törvénye)

# Az OLS nagymintás tulajdonságai

- Az előbb vázolt modellfeltevések közül az első három teljesülése esetén az OLS szolgáltatta becslések konzisztensek
- Ha mind az öt teljesül, akkor az OLS szolgáltatta becslések aszimptotikusan normálisak, a  $t$  ( $F$ ) statisztikák nulleloszlásai tényleg  $t$ -k ( $F$ -ek) aszimptotikusan, a szokásos tesztek és a konfidenciaintervallumok aszimptotikusan validak