A standard modellfeltevések, modelldiagnosztika

Ferenci Tamás
tamas.ferenci@medstat.hu

2018. február 7.

Tartalom

Emlékeztetőül

- Az erős exogenitási feltevés sérülésének három gyakorlatban tipikus esete van: kihagyott változó okozta torzítás (confounding!), mérési hiba, szimultaneitás
- Az utóbbi kettő meghaladja a mostani kereteket, így a továbbiakban az elsővel fogunk foglalkozni

Változó bevonásának hatása a modellre

Vessük össze ezt a két (demonstráció kedvéért igen kicsi) modellt az esettanulmány feladatára:

$$\widehat{\mathrm{KiadEFt}} = \underset{(13,783)}{339,746} + \underset{(0,0064924)}{0,637354}\,\mathrm{JovEFt}$$

$$T = 8314 \quad \bar{R}^2 = 0,5369 \quad F(1,8312) = 9637,2 \quad \hat{\sigma} = 662,02$$

(standard errors in parentheses)

$$\widehat{\mathrm{KiadEFt}} = \underset{(16,988)}{283,172} + \underset{(0,0074136)}{0,616911}\,\mathrm{JovEFt} + \underset{(6,0199)}{34,1727}\,\mathrm{TLetszam}$$

$$T = 8314 \quad \bar{R}^2 = 0,5386 \quad F(2,8311) = 4852,8 \quad \hat{\sigma} = 660,78$$

(standard errors in parentheses)

Miért változott meg a jövedelem becsült koefficiense?

Változó bevonásának hatása a modellre

- Mondjuk, hogy a bővebb modell írja le a valóságos helyzetet (a gyakorlatban ezt persze soha nem tudhatjuk, filozófiai kérdés)
- Azaz a valós helyzet a második regresszió
- Az érdekes, hogy ez alapján *előre* meg tudjuk mondani, hogy az első regresszióban mi lesz a jövedelem együtthatója! (. . . és ebből persze a változás okát is rögtön le tudjuk olvasni)
- A jövedelem ugyanis nem csak a kiadásra hat sztochasztikusan, hanem összefügg a taglétszámmal is:

$$\widehat{\mathrm{TLetszam}} = \underset{(0,025067)}{1,65553} + \underset{(1,1807\mathrm{e}{-005})}{0,000598206}\,\mathrm{JovEFt}$$

$$T = 8314 \quad \bar{R}^2 = 0,2359 \quad F(1, 8312) = 2566,9 \quad \hat{\sigma} = 1,2040$$

(standard errors in parentheses)

Változó bevonásának hatása a modellre

- Ebből összerakhatjuk a szűkebb regresszióban a jövedelem együtthatóját:

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

- A bővebb modellben az együttható 0,617: ennyi a jövedelem direkt hatása (ha egy egységgel nő stb.), és itt véget is ér a sztori, mert a bővebb modellben a taglétszámot állandó értéken tartjuk (v.ö.: c.p.) ezért nincs jelentősége a taglétszám és a jövedelem közti sztochasztikus kapcsolatnak
- A szűkebb modellben viszont a jövedelem egységnyi növekedése a taglétszámot is növeli tendenciájában, a növekvő taglétszám viszont (*önmagában* is!) növeli a kiadást, ez lesz az indirekt hatás
- Teljes hatás = direkt hatás + indirekt hatás(ok)

Változó bevonásának hatása a modellre

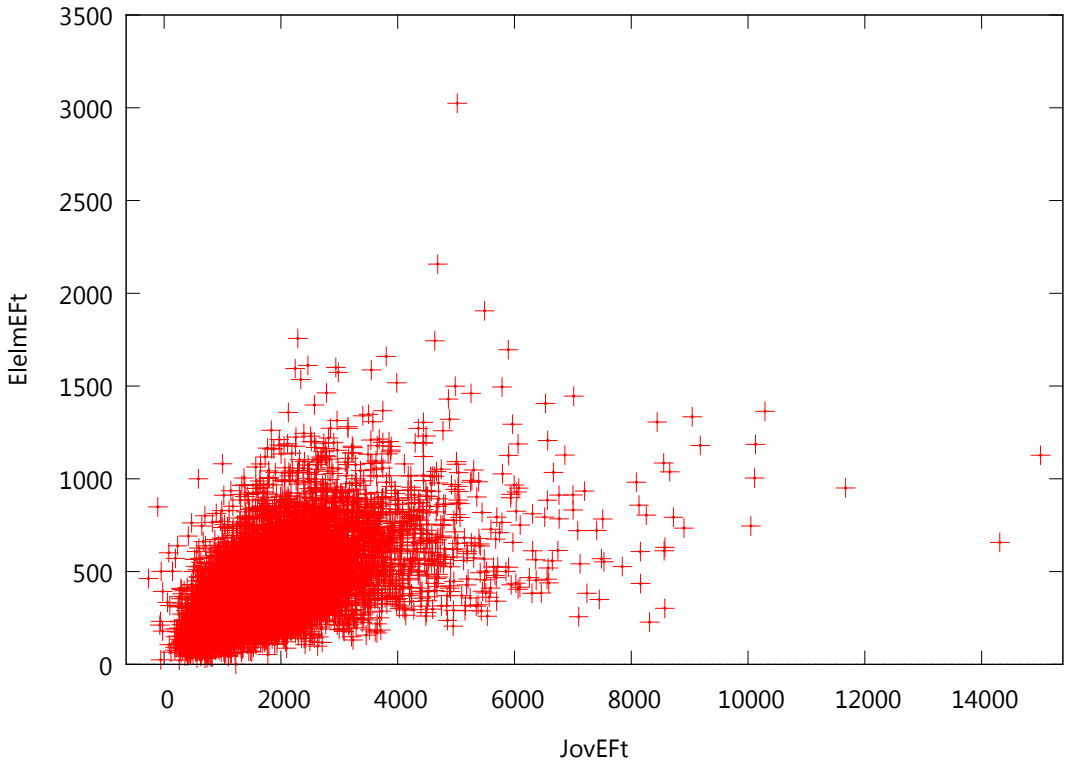
- A szűkebb regresszióban nem tudjuk *izolálni* a taglétszám hatását: ha a jövedelem nő, az a bővebb modellben nem társul a taglétszám növekedésével (v.ö. a paraméter c.p. értelmezésével), a szűkebb modellben viszont igen (hiszen ott nem endogén változó a taglétszám) $\rightarrow$ a szűkebb modellben a kihagyott változón keresztül terjedő hatások is *beépülnek* az együtthatóba
- Azaz: a bővebb regresszióval, az új változó bevonásával védekeztünk a confounding ellen (kiszűrtük a hatását: kontrolláltunk az újonnan bevont változóra)
- A gyakorlatban persze nem tudhatjuk, hogy mi a „kihagyott változó"

A specifikációs torzítás és iránya

- Akkor van tehát kihagyott változó okozta torzítás, ha *egyszerre* fennáll két feltétel: a kihagyott változónak van – önmagában – hatása az eredményváltozóra (tehát a β-ja nem nulla), *és* korrelált a bennmaradt magyarázó változóval
- Ebből adódóan a torzítás iránya negatív és pozitív is lehet attól függően, hogy ez a két tényező milyen előjelű
- Belátható, hogy többváltozós esetben, ha csak egy változónak is van endogenitási baja, akkor is torzított lesz az *összes* változó becsült koefficiense

Példa a heteroszkedaszticitásra

Először próbáljunk szemléletes képet kapni a heteroszkedaszticitásról:



Emlékeztetőül

- A feltétel: $\sigma_i^2 := \mathbb{D}^2\left(\varepsilon_i \mid \underline{\underline{X}}\right) = \sigma^2$ i-től függetlenül minden $i = 1, 2, \ldots, n$
- Vagy, ezzel egyenértékűen: $\mathbb{E}\left(\varepsilon_i^2 \mid \underline{X}_i\right) = \sigma^2$

A heteroszkedaszticitás okai

A heteroszkedaszticitás oka lehet:

1. A jelenség természetes velejárója (ld. az élelmiszerfogyasztás, vagy általában a kiadások példáját: „bővülő lehetőségek az ízlés kiélésére")
2. Csoportosított adatok használatakor: például háztartásonként átlagoljuk a jövedelmet és az élelmiszerekre fordított kiadást $\rightarrow$ még ha egy háztartástag szintjén állandó σ^2 is a szórásnégyzet, a csoportosított adatokban ez σ^2/n_i lesz, ahol n_i az i-edik háztartás létszáma, ami nagyon is eltérő lesz háztartásról-háztartásra (azaz megfigyelési egységenként)

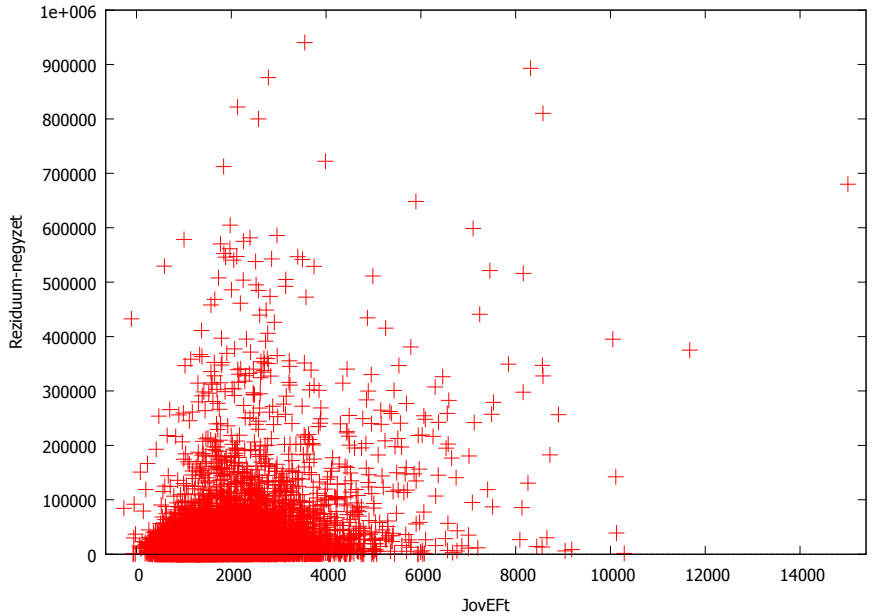
Heteroszkedaszticitás következményei

Mi történik, ha a heteroszkedaszticitással nem törődve, továbbra is a szokásos OLS-t alkalmazzuk a becslésre?

- Ahogy láttuk, az OLS szolgáltatta paraméter-becslések továbbra is torzítatlanok és konzisztensek lesznek...
- ...de már nem lesz hatásos (elveszíti a BLUE-ságot) $\rightarrow$ lesz olyan lineáris torzítatlan becslés, aminek kisebb a varianciája
- Ráadásul a becsült standard hibák (illetve általában a paraméterek becsült kovariancia-mátrixa) még torzított és inkonzisztens is lesz!
- A t- és F-statisztikáknak még aszimptotikusan sem lesz t-, illetve F-eloszlásuk
- Azaz a tesztek és a paraméterekre adott konfidencia-intervallumok érvényüket vesztik
- Az előrejelzések torzítatlanok lesznek ugyan, de nem hatásosak

Grafikus módszerek

- Nem analitikus, de benyomás-szerzésre jó lehet
- Reziduum-négyzetek kiplottolása különféle magyarázóváltozókkal (vagy a becsült eredményváltozóval) szemben:



Goldfeld–Quandt (GQ) próba

- Alapötlet: rendezzük a mintát azon magyarázó változó szerint, ami mentén nem állandó a feltételes szórás (az előbbi példán mondjuk a jövedelem), vágjuk szét három részre a mintát eszerint (kis, közepes és nagy értékű e változó), és hasonlítsuk össze (F-próbával) az alsó és a felső régióban a szórást
- Előnye: egyszerű, intuitív, könnyen átlátható
- Hátrányai:
  - Tudni kell, hogy mely változó mentén nem állandó a szórás (és muszáj, hogy egyetlen ilyet mutassunk)
  - Csak akkor jó, ha e változó mentén monoton módon változik a szórás
  - Gazdaságtalan, nem használ fel minden információt a mintából (a középső részt egyszerűen kidobja a kukába!)
- A `gretl` nem is tartalmaz rá előre megírt parancsot

LM-próbák

- Mi volna, ha a modell paraméterének tekintenénk, hogy az i-edik megfigyelési egységnél mekkora a σ_i^2 feltételes variancia?
- Önmagában nyilván rossz ötlet: lehetetlen lesz megbecsülni, hiszen minden paraméterre csak egyetlen megfigyelésünk (az $\widehat{u}_i^2$ reziduum-négyzet) lesz
- De: egyszerűsítsük a struktúráját! Azaz: feltételezzünk egy kevesebb paraméterre redukált formát, mely meghatározza a feltételes szórást

LM-próbák

- Tehát: van elképzelés, hogy mely változók „felelősek" potenciálisan a heteroszkedaszticitásért, melyek mozgatják a hibatag szórását $\rightarrow$ rakjunk erre egy (lineáris regressziós) modellt; pár lehetséges példa erre:

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP} + e_i$$
$$\sigma_i = \alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP} + e_i$$
$$\ln\left(\sigma_i^2\right) = \alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP} + e_i$$

- Itt Z_i-k ismert változók, melyek körét mi határozzuk meg, mint amik „felelhetnek" a nem-állandó szórásért (ezek természetesen részben vagy egészben magyarázó változók is lehetnek az eredeti regresszióban)
- A σ_i feltételes szórás helyébe annak a becslőjét, az $|\widehat{u}_i|$ reziduumot írjuk a segédregresszióban

LM-próbák

- Akkor nincs heteroszkedaszticitás, ha a segédregresszióban teljesül a $H_0 : \alpha_2 = \alpha_3 = \ldots = \alpha_P = 0$ nullhipotézis (hiszen ekkor a σ_i speciálisan állandó lesz, nekünk épp ez kellett)
- Ezt ún. LM-elven vizsgáljuk (részletesen lásd később), a lényeg, hogy az erre irányuló próba:

$$LM_{\text{emp}} = nR^2 \overset{H_0}{\sim} \chi^2_{p-1},$$

ahol R^2 természetesen a *segéd*regresszió többszörös determinációs együtthatója

- A fenti modellekhez tartozó próbák nevei rendre: Breusch–Pagan-próba, Glejser-próba, Harvey–Godfrey-próba (ún. multiplikatív heteroszkedaszticitásra)
- (Valójában a próbák eredetileg kicsit más alakúak, de nagy mintán egységesen a fenti formára hozhatóak)

LM-próbák

- Előnyeik:
  - Ezek már minden információt felhasználnak
  - Nem muszáj, hogy a heteroszkedaszticitásért egyetlen változó legyen felelős
- Hátrányaik:
  - Továbbra is nekünk kell tudnunk, hogy mely változó(k) felelős(ek) a nem-állandó szórásért
  - Hibanormalitást igényelnek és erre érzékenyek

Breusch–Pagan (BP) próba

- Ezt tartalmazza a heteroszkedaszticitásra irányuló LM-próbák közül beépítetten a `gretl`
- Még egyszer, a segédregressziója:

$$\widehat{u}_i^2 = \alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP} + e_i$$

- (Valójában a jobb oldal helyett egy $f\left(\alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP}\right)$ transzformáltat is tekinthetnénk valamilyen f függvénnyel, a próba végeredménye ugyanis beláthatóan ugyanaz lesz, ez tehát erre általánosodik)
- Ami fontos: a `gretl` beépített próbája *mindenképp pontosan* a magyarázó változók körét adja meg heteroszkedaszticitásért potenciálisan felelős változókként (tehát nem lehet sem egy részhalmazát venni, sem további, külső változót bevonni)
- A hibanormalitásra robusztusabb változata: Koenker-próba

White–teszt

- Az összes eddigi tesztnek még mindig hátránya, hogy tudni kell, hogy mi mozgatja a heteroszkedaszticitást
- A White-próba ötlete: ha nincs ötletünk, használjunk „mindent", ami a változóinkból kinyerhető (a valódi ok inkább az, hogy a homoszkedaszticitási feltétel gyengíthető arra, hogy az interakciókkal és a kvadratikus hatásokkal nincs összefüggése ε^2-nek)
- Minden: az összes magyarázó változó, az összes interakció, az összes kvadratikus hatás (persze csak ahol van értelme)
- Innentől olyan, mint a BP-próba
- Nagy mintás
- További előnye, hogy a hibanormalitásra sem annyira érzékeny
- Hátránya: itt is érvényesül a „minél kevesebb előfeltevésre épít egy próba, annál gyengébb" elv (itt szemléletesen: nagyon megnő a segédregresszióban a magyarázó változók száma) $\rightarrow$ ha van *a priori* információnk, használjuk! (itt: ha ismerjük, mi felel a heteroszkedaszticitásért, erősebb a BP-próba)

Modellspecifikáció változtatása

- Ötlet: úgy módosítjuk a modellspecifikációt, hogy az új specifikációban szereplő hiba már ne legyen heteroszkedasztikus
- Nem-statisztikai jellegű korrekció: szakmai ismeretet (is) igényel arról, hogy vajon mi a jó módosított specifikáció
- Nem is univerzális (nem feltétlenül alkalmazható minden esetben)
- Például:
  - Logaritmálás (ld. a nemlinearitásokról szóló részt a 6. fejezetben)
  - „Deflálás” (áttérés valamilyen méret-jellegű mutatóra leosztott változóra, pl. népességről népsűrűségre)

Heteroscedasticity Consistent Covariance Matrix, HCCM

- Ötlet: a becsült értékek torzítatlanok, azokat hagyjuk békén: maradjanak ugyanazok, mint a heteroszkedaszticitás figyelmen kívül hagyásával becsült modellben
- A standard hibákkal kéne valamit kezdeni
- HCCM nevű eljárás képes ezeket korrigálni: robusztus (vagy Huber–White–Eicker) standard hibák
- Matematikai részletekkel nem törődünk

Heteroscedasticity Consistent Covariance Matrix, HCCM

- Univerzálisan működőképes, nem igényel semmilyen feltevést a heteroszkedaszticitás struktúrájáról (legalábbis nagy mintán: itt emiatt gyakran automatikusan robusztus standard hibát adnak meg, esetleg mindkét standard hibát)
- Viszont ha fennáll a homoszkedaszticitás, akkor jobban járunk a szokásos standard hibával, mert annak a kismintás viselkedése is garantált (már csak ezért is érdekes a tesztelés)
- Alapozható rá más teszt is, nem csak a t-próba

Általánosított legkisebb négyzetek módszere (GLS)

- Ez már a teljes modellt újrabecsüli: a becsült koefficiensek is mások lesznek
- Alapötlet: a hibák kovarianciamátrixa nem skalármátrix $\rightarrow$ semmi baj, feltételezzünk egy általánosabb mátrixot, és számoljuk azzal végig a legkisebb négyzetes becslést
- Matematikai részletek nélkül a végeredmény:

$$\widehat{\beta_{\mathrm{GLS}}} = \left(\mathbf{X}^T \mathbf{\Omega}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{\Omega}^{-1} \mathbf{y},$$

  ahol $\mathbf{\Omega}$ a – teljesen általános – feltételes kovarianciamátrix:
  $\mathbf{\Omega} = \mathbb{E}\left(\underline{U U^T} \mid \underline{\underline{X}}\right)$
- A baj, hogy ez önmagában csak akkor alkalmazható, ha ismerjük ezt a $\mathbf{\Omega}$ mátrixot, azaz az egyes – feltételes – szórásokat

Súlyozott legkisebb négyzetek módszere (WLS)

- Van gyakorlati példa arra, amikor – legalábbis konstans szorzó erejéig – ismerjük a feltételes szórásokat: ha tudjuk, hogy azok mely változóval arányosak
- Például: fogyasztási egységek számával arányos a feltételes szórás: $\sigma_i = \sigma \cdot F_i$ (a fogyasztási egységek F_i száma minden megfigyelési egységre ismert)
- Ennyi (tehát a konstans szorzó erejéig ismert feltételes szórás) már elég: ha

$$Y_i = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

heteroszkedasztikus is, a

$$\frac{Y_i}{\sqrt{F_i}} = \beta_1 \frac{1}{\sqrt{F_i}} + \beta_2 \frac{X_{i2}}{\sqrt{F_i}} + \ldots + \beta_k \frac{X_{ik}}{\sqrt{F_i}} + \frac{\varepsilon_i}{\sqrt{F_i}}$$

könnyen belátható, hogy nem lesz az

Súlyozott legkisebb négyzetek módszere (WLS)

- Ez az ún. súlyozott legkisebb négyzetek módszere (WLS, weighted least squares); nem keverendő össze a megfigyelési egységek súlyozásával (erre is látni fogunk példát)
- Fontos, hogy a súlyok *nem* becslésből származtak, hanem ismertek voltak (ld. a fogyasztási egységek példáját)
- Természetesen egy változónál több is felhasználható (fogyasztási egységek száma *és* település stb.), hogy leírjuk a σ_i-t és a függvényforma is lehet akármilyen bonyolult (fogyasztási egységek négyzetével arányos feltételes szórás stb.), a lényeg, hogy olyan kifejezést konstruáljunk *kizárólag* ismert változókból, mellyel egyenesen arányos lesz a feltételes szórás
- A kulcs az, hogy visszaredukáljuk egyetlen ismeretlen paraméterre a feltételes szórásokat (ugyanúgy, ahogy homoszkedaszticitásnál lenne)

Kivitelezhető általánosított legkisebb négyzetek módszere (FGLS)

- Ha nem ismert a heteroszkedaszticitás struktúrája, akkor más megoldás kell, hogy a gyakorlatban alkalmazható legyen a GLS
- Egy segédregresszióban az eredeti regresszió reziduumainak négyzeteit regresszáljuk ki a White-tesztnél látott módon, innen kapjuk a hiba becsült varianciáit
- Ezek felhasználásával egy súlyozott regressziót számítunk, amivel újra közelítjük a hibák varianciáit
- Így nyerünk – ismert struktúra nélkül is – becslést a feltételes szórásra, amit az alapregresszióban a WLS-hez hasonló módon alkalmazhatunk a heteroszkedaszticitás korrigálására
- (A segédregresszióban reziduum-négyzet helyett mást is alkalmazhattunk volna, ahogy a heteroszkedaszticitásra irányuló LM-próbáknál is volt)