6 Gaia

Kolinealitate anizkoitza

Aurkibidea

6.1 Sarrera

Orokorrean, erregresio eredu bat estimatzeko eskuragarri diren datuak eta ereduaren oinarrizko teoria ekonomikoaren arteko baterakuntza ez da beti posible izaten, laginean datu batzuk falta direlako, datuak taldekatuta daudelako, neurketa erroreak dituztelako edota aztergai ditugun aldagai azaltzaileen arteko koerlazioa handiegia delako.

Atal honetan, azken arazo hau aztertuko dugu, hau da, kolinealitate anizkoitza edo aldagai azaltzaileen arteko **lagin** koerlazioa. Erregresio ereduko koefizienteen interpretazioa egiterakoan, konkretuki malden interpretazioa, $X_{ji}\,(j > 2)$ unitate batean handitzean aldagai azalduaren (Y_i) batezbesteko gehikuntza β_j unitatekoa dela aipatzen genuen, **beste aldagai azaltzaile guztiak konstante mantenduz**. Kolinealitate anizkoitz zehatza dugunean ordea, ezinezkoa da gainontzeko aldagai azaltzaile guztiak konstante mantentzea, beraien arteko erlazio lineala baitago. Kasu hauetan, ondoren ikusiko dugun bezala, ezinezkoa izaten da, aldagai bakoitzaren eragin isolatua bereiztea.

Bereziki bi kasu analizatuko ditugu: kolinealitate anizkoitz zehatza eta altua.

6.2 Kolinealitate anizkoitz zehatza

Ereduaren zehazpen zuzena emanik, aldagai azaltzaileetariko bat besteen konbinazio lineal zehatz bezala adierazi badaiteke, orduan kolinealitate anizkoitz zehatza dugula esango dugu. Ikusiko dugunez, ereduaren zehazpena zuzena izan arren, bere estimazioan eragina izango du.

Izan bedi hurrengo erregresio eredu orokorra:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_i X_{Ki} + u_i \quad i = 1, \ldots, N$$

non bi aldagai azaltzaileen artean konbinazio lineal zehatza ematen den, adibidez $X_{3i} = cX_{2i}$, c konstante ezaguna izanik. Informazio honekin X datu matrizea osatuko bagenu, hirugarren zutabeko balioak bigarren zutabekoak bider c eginez lortuko genituzke eta ondorioz, X matrizea ez litzateke zutabeetan hein osokoa izango (h(X)=K-1), hau da ez lituzke K zutabe independente izango. Hondar Karratuen Batura (HKB) minimo eginez lortzen diren ekuazio normalen sistema ($X'Y = X'X\beta$) ezin izango litzateke era bakar batean askatu, $X'X$ matrizearen determinantea zero izango baita eta hortaz, bere alderantzizkoa ezin izango litzateke lortu. Ondorioz, ereduko koefizienteak banaka estimatzea $\hat{\beta} = (X'X)^{-1}X'Y$ erabiliz ez litzateke posible izango.

Koefizienteen interpretazioaren ikuspuntutik ere arazoa aztertu dezakegu. Alde batetik, lehen bezala, termino independentea, aldagai azalduaren batezbestekoa izango da aldagai azaltzaile guztiak zero direnean. Bestetik ordea, β_2-ren kasuan, $E(Y)$-ren aldakuntza izango da X_2 unitate batean handitzean eta X_3 konstante mantenduz, baina kolinealitate zehatza dugunean eta gure adibidearekin jarraituz, ezin daiteke X_3 konstante mantendu X_2 unitate batean handitzean, zeren eta $X_{3i} = cX_{2i}$ baita.

Ikusi dugunez, ezin izango litzateke koefiziente guztiak estimatu banan-banan, baina bai ordea beraien konbinazio linealen bat. Gure adibidearekin jarraituz, $X_{3i} = cX_{2i}$ (c=konstantea)

ereduan barneratzen dugunean eta ordenatzerakoan ondorengo eredua lortuko genuke:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 (cX_{2i}) + \ldots + \beta_K X_{Ki} + u_i$$

$$Y_i = \beta_1 + (\beta_2 + c\beta_3) X_{2i} + \ldots + \beta_K X_{Ki} + u_i$$

$$Y_i = \beta_1 + \gamma X_{2i} + \ldots + \beta_K X_{Ki} + u_i$$

Lortutako eredu honetan ez dugu kolinealitate arazorik eta bertako koefiziente guztiak interpretagarriak dira eta estimagarriak KTAn bitartez. Eredu honetako koefiziente guztiak estimatu daitezke eta horien artean γ, hau da, hasierako ereduko koefizienteen konbinazio lineal hau, baina ez ordea β_2 eta β_3 bi koefizienteak banaka. Koefiziente guztiak estimatzeko informazio independente nahiko ez dago, zeren eta **lagina izanik**, X_3 aldagaiak ez baitu X_2 aldagaiak eskeintzen duen informazioaren gehigarri esanguratsurik ematen.

Nola estimatzen du Gretlek kolinealitate anizkoitz zehatza duen eredu bat?

Demagun Ramanathaneko (2002) data4-1 fitxategian SQFT aldagaiaren konbinazio lineala den aldagai berri bat sortzen dugula, SQFT1 = 5 × SQFT eta hurrengo eredua zehazten dugula:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 SQFT1 + u_i$$

Kolinealitate anizkoitz zehatza dagoenean, Gretlek kolinealitatea sortarazten duen aldagaietariko bat omititzen du, hurrengo emaitza erakutsiz:

```
Eredua 1: KTA estimazioak 14 behaketak erabiliz 1-14

Aldagai azaldua: price

Omitituta kolinealitate zehatzagatik sqft1

     ALDAGAIA      KOEFIZIENTEA      DESB.TIP        T ESTAT    P-BALIOA

  const               121,179          80,1778         1,511   0,15888
  sqft                  0,148314        0,0212080     6,993   0,00002 ***
  bedrms              -23,9106          24,6419       -0,970   0,35274

  Aldagai azalduaren batezbestekoa = 317,493
  Aldagai azalduaren desbiderazio tipikoa = 88,4982
  Hondar Karratuen Batura = 16832,8
  Hondarren desbiderazio tipikoa = 39,1185
  R-karratua = 0,834673
  Zuzendutako R-karratua = 0,804613
  F-estatistikoa (2, 11) = 27,7674 (p-balioa = 5,02e-005)
  Log-egiantza = -69,5093
  Akaike Informazio Irizpidea (AIC) = 145,019
  Schwarz Bayesian Irizpidea (BIC) = 146,936
  Hannan-Quinn Irizpidea (HQC) = 144,841
```

Hau da, Gretlek ohartarazten du erregresioan aldagai bat omititu duela, gure adibidean SQFT1, eta aldagaia ez duen ereduaren estimazioaren emaitzak ematen ditu. Teorikoki, SQFT1 = 5 × SQFT ereduan barneratu ondoren lortzen den eredua honakoa da:

$$PRICE_i = \beta_1 + (\beta_2 + 5\beta_4)SQFT_i + \beta_3 BEDRMS_i + u_i$$

non β_1 eta β_3 banaka estimatu daitezkeen baina $\widehat{\beta_2}$ eta β_4 ez. Beraz, estimazio emaitzetan, SQFT-ren koefiziente estimatua (0,148314) $(\beta_2 + 5\beta_4)$ konbinazioari dagokio.

6.3 Kolinealitate anizkoitz altua

Errealitatean, aldagai ekonomikoen artean nolabaiteko erlazioa izaten da, nahiz eta erlazio hori zehatza izatea ohikoa ez izan. Oso garrantzitsua da erlazio hori zein gradukoa den aztertzea, oso sakona baldin bada, hau da, kolinealitate anizkoitz altua baldin badago, ondorioak izango baititu adibidez ereduko koefizienteen zehaztasunean eta kontrasteetan. Hala ere, kontuan izan behar da askotan datuetan ematen den arazoa izaten dela, hau da, aukeratutako laginekoa eta ez eredukoa.

Aldagai azaltzaile batzuren arteko erlazioa sakona baldin bada baina ez zehatza, X datu matrizea hein osokoa izango da zutabeetan. Hau da, ereduan K koefiziente baldin baditugu, $h(X) = K$ izango da eta hortaz, $X'X$ matrizearen determinantea zeroren desberdina izango denez, koefiziente guztiak estimagarriak dira.

Aurreko adibidean oinarrituz, kolinealitate anizkoitz altua hurrengo moduan ikusi daiteke:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$X_{3i} \approx cX_{2i}$$

non X_2 eta X_3 aldagaien arteko erlazioa zehatza ez den. Aldagai azaltzaileen artean kolinealitate altua denean, ohikoa izaten da ondorengo **arazoak** izatea:

- Datuetan aldaketa txiki batzuk, koefizienteen estimazioetan aldaketa handiak sortaraztea.
- Koefizienteen estimazioek espero den zeinua ez izatea edota magnitute sinesgaitza.
- Koefizienteen estimatzaileek bariantza handia izaten dute, hau da, zehaztasun txikiarekin estimatuta daude. Ondorioz, aldagaien banakako esanguratasun kontrastea egitean, estatistikoaren balioa oso txikia izango da eta ondorioz, hipotesi hutsa ez baztertzearen probabilitatea izugarri handitzen denez, aldagaia banaka nabaria ez dela ondorioztatuko dugu. Hau da, *gezurrezkoak* diren hipotesi hutsak ez baztertzearen probabilitatea handiagoa izatea eragingo du. Ondorioz, koefiziente baten konfidantza tartea lortuz gero, tartea oso zabala izango da.
- Eredua estimagarria da eta erregresioaren mugatze koefizientea altua izaten da. Horrela, aldagai azaltzaileen baterako esanguratasun kontrastea burutuz gero, hipotesi hutsa ez da baztertzen eta aldagaiak batera nabariak direla ondorioztatzen da, baina lehen ikusi dugunez, banaka ez esanguratsuak direla irtengo da.

Kolinealitatearen arazo nagusienak hauek izanik, emaitza horiek izatea lagungarri izan daiteke kolinealitatea dagoen edo ez susmatzeko. Ondorio edo arazo horien arrazoitariko bat $X'X$ matrizearen determinantean datza. Aldagaien arteko erlazioa sakona izan arren, ez da zehatza eta beraz, X matrizearen zutabeen artean ez dagoenez konbinazio lineal zehatzik, determinantea ez da zero izaten. Hala ere, kolinealitate altuagatik, determinantea zerotik oso hurbileko balio bat izango da.

Dena den, kontu handia izan behar da, gertagarria delako aldagaien neurri unitateagatik determinantea ia zero izatea eta ez kolinealitate arazoagatik. Bestalde, posible da aldagaien banakako esanguratasun kontraste batean hipotesi hutsa ez baztertzea, benetan aldagai ez nabaria delako aldagai azaldua azaltzeko. Garbi izan behar dugu, egoera hauek ematen diren guztietan ez dagoela beti kolinealitate anizkoitzaren arazoa eta beraz, kontu handia izan behar da emaitzak aztertu eta interpretatzerakoan.

Kolinealitatea hautematzeko modu eragingarri bat, aldagai azaltzaile bakoitza besteekiko eginiko erregresio laguntzaileen mugatze koefizienteak erabiliz izango da. Erregresio laguntzaile hauetako R_j^2-tariko bat edo batzuk handiak badira, erregresioan kontuan hartutako aldagaien arteko koerlazioa handia dela esan nahiko du eta beraz, aldagai hauen arteko kolinealitate anizkoitz posiblea dugula. Gure adibidean, oso lagungarria izango litzateke X_2 eta X_3ren arteko erregresio laguntzaile bat burutzea:

$$X_{2i} = \alpha_1 + \alpha_2 X_{3i} + v_i$$

Kolinealitate altua badute, erregresio honen mugatze koefizientea oso altua izango da, erregresio lineal eredu bakuna izanik, mugatze koefizientea X_2 eta X_3 aldagaien arteko lagineko koerlazio koefiziente linealaren karratua baita ($R^2 = (r_{X_2X_3})^2$).

Beste aukera bat aldagaien arteko lagineko koerlazio koefizienteak kalkulatzean datza. Aldagaien arteko koerlazio koefizienteek -1 eta 1 bitarteko balioak har ditzakete, positiboa baldin bada erlazio zuzena dela adierazten duelarik eta negatiboa bada ordea, alderantzizkoa. Bestalde, balio absolutuan bat baliotik hurbileko balioa bada, erlazioa sakona da eta zero baldin bada, koerlazio linealik ez dutela esango dugu. Aurreko gaietan ikusi bezala, etxebizitzen prezioaren adibideko kasuan (**A eredua** hartuz), ondorengo koerlazioak lortuko genituzke:

6.1 Taula: **A ereduko** aldagaien koerlazio matrizea

```
Koerlazio Koefizienteak, 1 - 14 behaketak erabiliz
        %5eko esanguratasuna (alde bikoa) = 0,5324  n = 14 -rentzat

         price          sqft        bedrms         baths
        1,0000        0,9058        0,3156        0,6696  price
                      1,0000        0,4647        0,7873  sqft
                                    1,0000        0,5323  bedrms
                                                  1,0000  baths
```

Ikus dezakegunez, aldagai azaltzaileen arteko erlazio linealak zuzenak dira, koefiziente guztien zeinuak positiboak baitira. Erlazio sakonena, SQFT eta BATH-en artekoa litzateke 0.7873 balioarekin, zuzena eta nahiko sakona. Hala ere, BATH eta BEDRMS aldagaien koerlazio koefizientea ikusirik (0.5323) eta aurreko ikasgaian egin ditugun banakako eta baterako

esanguratasun kontrasteak gogoratuz, azterketa sakonago bat egingo dugu. Ikusi dugunez, eredu honetan aldagai azaltzaileak batera nabariak direla ondorioztatu dugu eta SQFT aldagaia ezik, beste biak banaka ez nabariak direla. Kolinealitate arazoak izan ditzakegu eta horren ondorioz eman al dira emaitza hauek? Formalki aztertuko dugu Gretleko aukerak aprobetxatuz.

Badira kolinealitatea **hautemateko prozedura formalak** ere. Belsley, Kuh eta Welsch (1980) egileen prozedura adibidez, matrize baten Baldintza Zenbakian (γ) oinarritzen da, hau da, matrize horren balio propio handienaren eta txikienaren ($\lambda_{max}, \lambda_{min}$) zatiduraren erro karratuan: $\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$. Gure kasuan, $X'X$ matrizearen balio propioak kalkulatuko genituzke eta aipaturiko Baldintza Zenbakia kalkulatuko. Aldagaien arteko koerlazioa handituz gero, zenbaki hau handiagoa izango litzateke. Egile hauek diotenez, zenbaki hau 30 baino handiago denean, gradu altuko kolinealitate arazoak egon daitezke.

Gai honetan proposatu eta erabiliko duguna ordea, Neter, Wasserman eta Kutner (1990) proposaturiko Jasankortasuna (JAS) eta Bariantzaren Puztea (VIF) izango dira, Gretlek prozedura hau jarraitzeko aukera ematen baitu.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_i X_{Ki} + u_i \qquad i = 1, \ldots, N.$$

Aurreko erregresio lineal eredu orokorreko koefiziente baten ($\hat{\beta}_j$) bariantza ondorengo espresioaren bitartez kalkulatu daiteke:

$$bar(\hat{\beta}_j) = \frac{\sigma^2}{\sum (X_{ji} - \bar{X}_j)^2} \frac{1}{(1 - R_j^2)} = \frac{\sigma^2}{\sum (X_{ji} - \bar{X}_j)^2} VIF_j$$

non β_j, X_j aldagaiaren koefizientea den, R_j^2 koefizientea X_j aldagaia beste aldagai azaltzaile guztien funtzioan adierazitako erregresio laguntzailearen mugatze koefizientea den eta azkenik, $VIF_j = \frac{1}{(1-R_j^2)}$ bariantzaren puztea den. Era laburrean azalduz, R_j^2 batetik gero eta hurbilago badago, hau da, X_j aldagaiak beste aldagai azaltzaileekin duen kolinealitatea gero eta handiagoa bada, VIF_j ere handitu egiten da ($1 \leq VIF_j \leq \infty$), ikusi dugunez, kolinealitateak bariantza "puztu" egiten baitu. Egile hauen arabera, $VIF_j > 10$ bada, X_j aldagaiaren kolinealitatea besteekiko handia dela esango dugu. (OHARRA: jasankortasuna VIF-aren alderantzizkoa da: $JAS_j = \frac{1}{VIF_j}$).

Aipatu beharra dago, kolinealitatearen azterketa VIF-ak edota jasankortasuna erabiliz, zenbait kritika jasan dituela, askotan hauen arabera lortutako emaitzak ez baitira erabatekoak. Ikusi bezala, koefizienteen estimatzaileen bariantzak, VIF_j, σ^2 eta $\sum (X_{ji} - \bar{X}_j)^2$-en arabera daudenez, VIF_j altu bat ez da baldintza nahikoa eta beharrezkoa estimatzaileen bariantza handiak lortzeko. Nahiz eta VIF_j altu bat izan, baliteke σ^2 txikia edota $\sum (X_{ji} - \bar{X}_j)^2$ oso handia izatea eta kolinealitatearen ondorioen arraztorik ez izatea kontraste eta tartezko estimazioetan adibidez.

Etxebizitzen prezioen adibidearekin jarraituz, aldagaien artean kolinealitatea dugun edo ez aztertzeko bariantzaren puzte koefizientea kalkulatuko dugu. Horretarako, eredua KTA bitartez estimatutako lehiatilan *Kontrasteak* → *Kolinealitatea* klikatuko dugu, bertan azaltzen baitira kolinealitate arazoa emateko aztertu eta konparatu beharreko balioak. Ondorengoa lortzen da:

Ikus dezakegunez, VIF_j-aren arabera, kolinealitate arazorik ez dugula ondorioztatuko genuke. Hortaz, ereduko BATHS eta BEDRMS aldagaiak ez nabariak dira etxebizitzaren prezioa

6.2 Taula: Kolinealitate kontrastearen emaitzak

```
Bariantzaren Inflazio Faktoreak

 Balio minimo posiblea = 1.0
 Balioak > 10.0 badira, agian kolinealitate arazoren bat dago

   2)            sqft    2,651
   3)          bedrms    1,406
   4)           baths    2,900

VIF(j) = 1/(1 - R(j)^2), non R(j) j aldagaiaren eta beste aldagai
azaltzaileen arteko koerlazio koefiziente anizkoitza den

X'X matrizearen propietateak:

 1-norma = 55654161
 Determinantea = 1,366856e+008
 Elkarrekiko baldintza zenbakia = 2,808412e-009
```

azaltzeko eta SQFT ordea, nabaria da eta aurreko gaian ikusi bezala, proposatzen genuen **E eredua** erabiliko genuke etxebizitzen prezioa azaltzeko.

Kolinealitate arazoa zuzentzeko proposaturiko **soluzioak** asko dira, baina orokorrean ez dira asebetegarriak.

- Lagin horretan ematen den arazoa bada, datuak aldatuz posible da (ez da beti ematen) kolinealitate arazoa konpontzea. Ideia, aurrekoak baino gutxiago erlazionaturiko datuak barneratzea izango da, datu bakar batzuk barneratuz edota lagina aldatuz. Edonola ere, oso gutxitan izan ohi da hobeagoak diren datuak lortzeko aukera.

- Askotan, koefizienteeen informazioa barneratuz arazoa desagertu egiten da. Hala ere, informazio gehigarri hori kolinealitate arazoa hauteman baino lehen kontuan hartu beharko genuke eta ez ondoren, informazio gehiagorekin eredua era efizienteagoan estimatuko baitugu.

- Kolinealitate anizkoitza sortarazten duten aldagairen bat eredutik kendu. Hala ere, soluzio honekin kontu handia izan behar da, ereduaren zehazpen oker bat izan dezakegu eta, hau da aldagai nabari baten omisioa. Ondorioz, lortuko genituzkeen koefizienteen KTA estimatzaileak eta perturbazioaren bariantzaren estimatzailea alboratuak izango lirateke eta proposatutako kontrasteak (inferentzia) ez lirateke baliogarriak izango.

- Aldagaien arteko kolinealitatea altua bada, intuitiboki baliogarria eta ulergarria den soluzio posible bat **Osagai Nagusiko Metodoa** erabiltzean datza. Hau da, metodo honekin X matrizetik aldagai kopuru txikiago bat kanporatuko da, konkretuki aldagai azaltzailearen aldakuntz edo bariantza guztia edo gehiena jasoko duten aldagaiak. Horrela lortzen diren estimatzailearen bariantza, KTA estimatzailearena baino txikia-

goa izan arren, alboratua izango da eta gainera, aldagaien neurri unitateen menpekoa. Bestalde, osagai nagusiak aukeratzerakoan ez dugu aldagai azaltzaile eta azalduaren arteko erlazioa kontuan hartzen eta azkenik, horrela lortzen diren estimatzaileak, hasierako ereduko koefizienteen konbinazioak dira eta ondorioz, oso zaila izaten da emaitzak interpretatzea.

- **Cresta Erregresio Estimatzailea**: $(\hat{\beta}_\lambda) = [X'X+\lambda D]^{-1}X'Y$ non D matrize diagonal bat den $X'X$ matrizeko diagonal nagusiko elementuak izanik, λ parametro ezezaguna aukeratu egin behar da. Normalean parametro hau aukeratzeko balio txiki batekin hasten gara $\lambda = 0,01$ eta bere balioa handitzen joaten gara estimatzailea egonkortu arte. Aurkeztutako estimatzailea alboratua izan arren, bere bariantza eta kobariantza matrizea, KTA estimatzailearena baino "txikiagoa"da eta Batezbesteko Errore Koadratiko (BEK) txikiagoa du. KTAko estimatzailea kasu berezi bat bezala lor daiteke $\lambda = 0$ denean. Estimatzailearen batezbestekoa eta bariantza honakoak dira:

$$E(\hat{\beta}_\lambda) = \sigma^2[X'X + \lambda D]^{-1}X'X\beta$$
$$Bar(\hat{\beta}_\lambda) = \sigma^2[X'X + \lambda D]^{-1}X'X[X'X + \lambda D]^{-1}$$

Bibliografia

Belsley, D., E. Kuh eta R. Welsch (1980). *Regression Diagnostics.* Wiley, New York.

Neter, J., W. Wasserman eta M.H. Kutner (1990). *Applied Linear Statistical Models*, 3. ed., Boston, M.A.: Irwin, Boston.

Ramanathan, R. (2002). *Introductory Econometrics with Applications*, 5. ed., South Western, Ohio.