BUDAPESTI CORVINUS EGYETEM

Adatelemzés és Informatika Intézet

Statisztika Tanszék

Ökonometria nagy dolgozat

Hajdu Andor – VDHTI3

Hajdu Bálint – VG8M5M

Gyuricza Ádám – HE5GV0

Tósoki Samu – H5NUH7

Gazdálkodási és menedzsment alapképzési szak

Vállalat- és gazdaságelemzés szakirány

Szemináriumvezető: Keresztély Tibor

2023.12.30.

Absztrakt

Ebben a dolgozatban a hazai járások keresztmetszeti adatelemzésének leírása szerepel, melynek fő kérdése, hogy a Gazdaság- és Vállalkozáskutató Intézet által számított járási fejlettségi mutatókat mennyire tudjuk pontosan becsülni a félév során tanult ökonometriai módszerek logikus alkalmazásával. A dolgozat során kitérünk az alkalmazott módszerekre: adattisztítás, adatvizualizálás, modellépítés és modellszelekció, heteroszkedaszticitás és főkomponenselemzés. Az információs kritériumok, Wald-teszt és Ramsey-féle reset teszt módszerek alkalmazásával a végső modell, egy interakcióval, logaritmikus transzformációkkal és egy négyzetes taggal rendelkező, modell_int4 lett, ami 93%-os pontossággal becsüli meg a járások fejlettségét.

Tartalomjegyzék

Kutatási kérdés és motiváció

A fejlettségi mutatókat (akár országokra, akár kisebb régiókra) jellemzően olyan klasszikus indikátorok segítségével készítettek, mint a jövedelem, vagyon mérete, fogyasztás. Mivel ezek sokszor nem állnak rendelkezésre, illetve figyelmen kívül hagynak rengeteg, főként szociológiai helyzetre utaló információt, ezért az újabb kutatások között gyakran szerepelnek olyan módszertannal készítettek, amelyek inkább az említett szociológiai információkat ragadják meg és helyezik előtérbe. Ilyen kutatásra példa a Gazdaság- és Vállalkozáskutató Intézet által számított járási fejlettségi mutatókat bemutató kutatás, amely például *120 fő/km2 népsűrűség feletti településeken lakók aránya*, *Tartós álláskeresők aránya* vagy *Vándorlási különbözet 100 lakosra jutó aránya* mutatók segítségével határozta meg a kompozit mutatót. Véleményünk szerint az ilyen jellegű kutatásokban rengeteg potenciál van, ezért szerettünk volna betekinteni a módszertan alkalmazásába.

A kutatásunk középpontjában az áll, hogy a félév során elsajátított ökonometriai módszerek alkalmazásával mennyire kapunk hasonló eredményeket az említett kutatóintézet kompozit mutatójához, a járási fejlettségi indexhez képest. Itt fontos megjegyezni, hogy a kutatóintézet számítási módszerével[1] ellentétben mi modelleztük a JFM mutatót.

Az adatbázis

Az adatbázis a Magyar Kereskedelmi- és Iparkamara Gazdaság- és Vállalkozáskutató Intézetének honlapjáról származik, és KSH-ról gyűjtött adatokat tartalmaz, azaz megbízhatónak számít az értékek valósságát tekintve. Ennek jelentősége az, hogy így megbizonyosodhattunk arról, hogy képesek leszünk egy valós gazdasági problémával foglalkozni, modellezni. Az adatbázis 24 változót tartalmaz, ezek közül nem magyarázóváltozó a járás neve, illetve a számított fejlettségi mutató, a JFM, hiszen az a magyarázott változó. Az alábbi felsorolás tartalmazza a változók nevét, mértékegységét, és az adatfelvétel évét.

- *Regisztrált vállalkozások 100 állandó lakosra jutó száma, 2017*
- *Regisztrált vállalkozások 100 állandó lakosra jutó számbeli változása 2014-2017 (index)*
- *1000 állandó lakosra jutó vendéglátóhelyek száma, 2017*
- *1000 állandó lakosra jutó kiskereskedelmi boltok száma, 2017*
- *Épített lakások száma (db) a lakások arányában, 2017*
- *A legközelebbi megyeszékhely elérhetősége, 2012*

[1] Az intézet által használt módszertan: „az egyes mutatók alapján kvintilisek létrehozása és a járások osztályozása **1-től 5-ig egész számmal** (az olyan esetekben, ahol az alacsonyabb érték tekinthető kedvezőbbnek pl.: elérhetőség, halálozás, ott az alacsonyabb értékkel rendelkező járások kapták a jobb „osztályzatot"). Ezt követően az egyes járásokhoz tartozó 22 osztályzat számtani átlagának kiszámolása adja a komplex mutató értékét."

• Kábeltelevíziós hálózatba bekapcsolt lakások aránya, 2017
• 1000 lakosra jutó internet előfizetések száma, 2017
• A közüzemi szennyvízgyűjtő-hálózatba (közcsatornahálózatba) bekapcsolt lakások aránya, 2017
• A lakosságtól elkülönített gyűjtéssel elszállított települési hulladék (tonna) 1000 állandó lakosra vetítve
• 1000 állandó lakosra jutó háztartások részére szolgáltatott villamosenergia mennyisége, 2017
• 1000 állandó lakosra jutó személygépkocsik száma, 2017
• Vándorlási különbözet 100 lakosra jutó aránya, 2017
• 1000 lakosra jutó halálozások száma, 2017
• Értékesített használt lakások átlagos ára (millió Ft), 2017
• 120 fő/km2 népsűrűség feletti településeken lakók aránya, 2017
• Fiatalodási index (0-18/60-X éves népesség aránya, 2017
• Települési támogatásban részesítettek száma 1000 állandó lakosra vetítve, 2017
• Rendszeres gyermekvédelmi segélyekben részesítettek száma 100 0-17 éves állandó lakosra vetítve, 2017
• Nyilvántartott álláskeresők aránya, 2017
• Tartós álláskeresők aránya, 2017
• Aktív korúak aránya, 2017

Fontos kiemelni, hogy kizárólag numerikus változók szerepelnek az adatbázisban. Ez önmagában nem jelent problémát, illetve feltehetőleg a GVI módszertanából adódóan nem szerepelnek minőségi változók. Azonban mi kíváncsiak voltunk arra, hogy tudjuk-e interakcióval javítani a modellünket, így bevontunk egy minőségi változót, a régiót, azaz minden járáshoz hozzárendeltük a régiót, amelyben elhelyezkedik, de erre az adott feladatrésznél még jobban kitérünk.

Megfigyelések számát tekintve 177 szerepel, ami az eredeti 198 járáshoz képest úgy jön ki, hogy a budapesti kerületeket egy, Budapesti járásnak tekinti az adatbázis.

Adattisztítás és változószelekció

Mivel az adatok közvetetten a KSH-tól származnak, ezért adattisztítás során viszonylag kevés manipulációt kellett végrehajtani. A változók adattípusa megfelelő volt, azonban az első sorban az oszlopok száma szerepelt. Mivel ezek az értékek nem számítanak megfigyelésnek, így töröltük őket. Ezen kívül a Polgárdi járás sorában kivétel nélkül hiányoztak az értékek. Ennek oka, hogy időközben megszűnt a járás (meglepő, hogy a GVI nem vette ki). Mivel imputálni értelmetlen lenne, és egyetlen megfigyelésről van szó, ezt a sort is töröltük. Outlier szűrést nem végeztünk, hiszen az a cél, hogy minden járáshoz rendeljünk egy fejlettségi értéket. így végül 175 megfigyeléssel dolgoztunk.

Ugyan szerencsésnek tartjuk, ha sok magyarázóváltozó áll rendelkezésre, a 22-t túl soknak tartjuk értelmezési és átláthatósági szempontok miatt, így egy egyszerű lineáris regressziót futtatva a JFM-ra a regresszióban szereplő p-értékek és logikus gondolkodás segítségével leszűkítettük a magyarázóváltozóink halmazát az alábbi 10 elemű halmazra.

• 1000 állandó lakosra jutó kiskereskedelmi boltok száma, 2017
• Épített lakások száma (db) a lakások arányában, 2017
• A legközelebbi megyeszékhely elérhetősége, 2012
• A közüzemi szennyvízgyűjtő-hálózatba (közcsatornahálózatba) bekapcsolt lakások aránya, 2017
• A lakosságtól elkülönített gyűjtéssel elszállított települési hulladék (tonna) 1000 állandó lakosra vetítve
• 1000 állandó lakosra jutó háztartások részére szolgáltatott villamosenergia mennyisége, 2017
• 1000 lakosra jutó halálozások száma, 2017
• 120 fő/km2 népsűrűség feletti településeken lakók aránya, 2017
• Fiatalodási index (0-18/60-X éves népesség aránya, 2017
• Nyilvántartott álláskeresők aránya, 2017

Fontos megjegyezni, hogy a kiválasztott 10 változó részben szubjektív szempontok szerint lett kiválasztva, törekedve arra, hogy minél több aspektusát megragadjuk egy járás fejlettségének. Nem biztos, hogy a legjobb 10 elemű részhalmazt választottuk, lehet, hogy fejlettebb változószelekciós eljárásokkal más eredmény jönne ki, azonban a rendelkezésünkre álló eszközökkel mi ezeket tartjuk meg.

Így végül a modellezésre használt data frame-ünk 175 megfigyelésből, 10 magyarázóváltozóból, a járások nevét viselő és a magyarázott változóból áll.

Leíró statisztika

A modellezés előtt a megértést segítő és a problémában elmélyedést elősegítő leíró statisztikát készítettünk az eredményváltozóra.

n	mean	sd	median	trimmed	min	max	range	skew	kurtosis	se
175	3	0,73	3	2,98	1,73	4,68	2,95	0,15	-1,1	0,06

1. táblázat: A JFM_19 eredményváltozó leíró statisztikái

Egy magyarországi járás átlagosan 3-as fejlettségű, azaz pont közepes (ez következik a GVI módszertanából is). Ettől átlagosan 0,73-mal tér el egy járás fejlettségi értéke. A medián érték 3, azaz, ha sorba rendeznénk az értékeket, akkor a középső érték 3-as lenne. A nyesett átlag kb. megegyezik az átlaggal. A legkevésbé fejlett járás a Cigándi, 1,73-as értékkel, a legfejlettebb pedig a Győri, 4,68-as fejlettségi mutatóval. Az eredményváltozó ferdeségi mutatója 0,15, azaz kb. szimmetrikusnak tekinthető, minimálisan jobbra elnyúló. Ez abból ered, hogy van egy megfigyelés 4,5-ös fejlettség felett, de nincsen olyan járás, amely ne ért volna el legalább 1,5-ös értéket. A csúcsossági mutató alapján inkább laposnak tekinthető a változó eloszlása. Hisztogramon ábrázolva látszik, hogy a [2;4] intervallumon kb. ugyanannyi járás figyelhető meg felenként bontva az X-tengelyt. A standard hiba pedig 0,06, ami alacsonynak számít.

Összességében az eredményváltozó leíró statisztikája alapján egy kb. normális eloszlású, alacsony szórású és standard hibájú célváltozónk van, ami szerencsésnek tekinthető.

Az alapmodell

Legelőször egy szelekcios_modell nevű alapmodellt hoztunk létre, amely azt a 11 változót tartalmazza, amellyel dolgozni szeretnénk a munkánk folyamán. Ebben a modellben tehát még semmilyen transzformációt nem végeztünk el a változókon. Az eredményváltozó a járások fejlettségi mutatója (JFM), a magyarázóváltozók pedig: az 1000 főre jutó kiskereskedelmi boltok száma, az épített lakások száma (a lakások arányában), a legközelebbi megyeszékhely elérhetősége, a szennyvízgyűjtő hálózatba bekapcsolt lakások aránya, a lakosságtól elszállított települési hulladék 1000 főre, a háztartások részére szolgáltatott villamosenergia 1000 főre, az 1000 lakosra jutó halálozások száma, a 120 fő/km^2 népsűrűségű településeken lakók aránya, a fiatalodási index és az álláskeresők aránya.
A magyarázóváltozók közül egy kivételével (fiatalodási index) mindegyik szignifikáns. A VIF mutatók közül mindegyik 3 alatt van, így a multikollinearitás problémája nem áll fenn. Az alapmodell hibatagjára lefuttatunk egy Kolmogorov-Smirnov tesztet, aminek p-értéke 10^(-16) nagyságrendű lett, így elvetettük, hogy a hibatag normális eloszlású. Ennek tudatában lefuttatuk a modellre a Breusch-Pagan próba Koenker korrekcióval vett verzióját. A p-érték ~0.025 lett. Ez azt jelenti, hogy nem minden szokásos alfán mondható heteroszkedasztikusnak az alapmodell. 2%-os alfán még homoszkedasztikus, de 5%-os alfán már heteroszekdasztikus.

Transzformációk detektálása

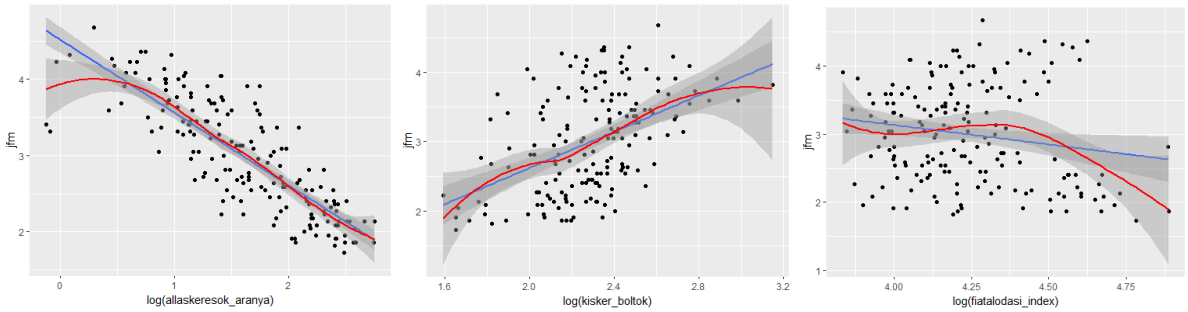
A következő feladatunk a transzformációs tagok detektálása volt. Ebben a részben azt vizsgáltuk meg, hogy szükséges-e az eredményváltozó vagy a magyarázóváltozók nem-lineáris transzformáltjával dolgoznunk és emellett megvizsgáltuk az interakció szükségességét is. Ahogy már említettük eredetileg az adatbázisban csak numerikus változóink voltak, így nem lett volna lehetőségünk interakciós kapcsolatok vizsgálatára. Mi ennek elkerülése érdekében létrehoztunk egy faktor változót, amely a járásokat régió szerint csoportosítja és ezt *regio* változónak neveztük el az adatbázis bővítésekor. Ez a faktor típusú változó természetesen 7 kategóriát tartalmaz, pont annyit amennyi régióval Magyarország is rendelkezik. A referenciakategóriának Közép-Magyarországot állítottuk be, mivel ez az ország központja és a viszonyítások általában ehhez a térséghez szoktak kapcsolódni.

Változók eloszlásának vizsgálata hisztogramon

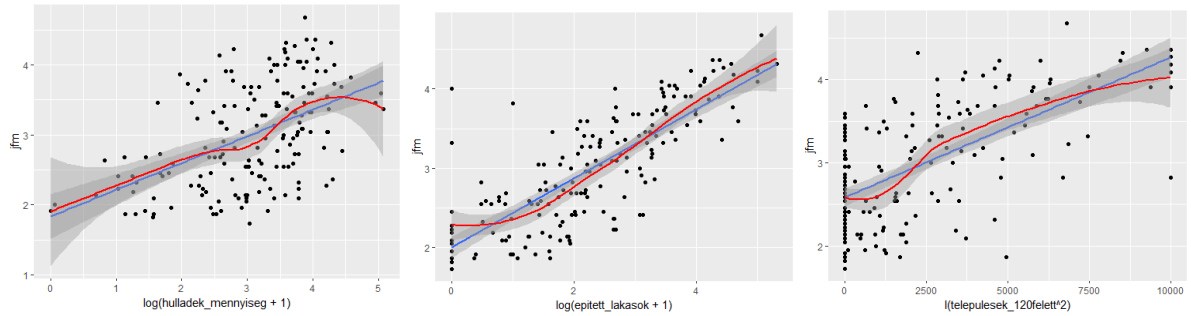
Első lépésben a numerikus tagok eloszlását vizsgáltuk meg hisztogramon. Az eredményváltozó normális eloszlást követ, így a nagy minta mellett ez a modellfeltevés is érvényesül. Ez azt jelenti, hogy nem kell logaritmizálnunk ezt a változót, az eredményeink nem lesznek torzítottak. A magyarázóváltozók között 5 változó szintén normális eloszlást követ, szóval itt sem kell logaritmizációhoz folyamodnunk. Ezek a változók a következőek: *legkozelebbi_megyeszkh, szennyviz_csatlakozas, villamosenergia, halalozas, telepulesek_120felett.* A maradék 5 magyarázóváltozó esetében azt vettük észre, hogy lehet azokat logaritmizálni. Ez a következő változókat jelenti: *kisker_boltok, epitett_lakasok, hulladek_mennyiseg, fiatalodasi_index, allaskeresok_aranya.* Természetesen a magyarázóváltozók eloszlása nem olyan kulcsfontosságú kérdés, mint az eredményváltozóé, viszont mi azt preferáljuk jobban, hogyha minden változó normálishoz közeli eloszlást követ és ennek érdekében, ha szükséges, akkor logaritmizálni fogjuk az adott változót.

Magyarázóváltozók és az eredményváltozó páronkénti kapcsolata pontdiagramon

A hisztogramok megvizsgálása után a következő lépés az volt, hogy ábrázoltuk a magyarázóváltozók és az eredményváltozó kapcsolatát páronként egy pontdiagrammon. Ezzel igazából az volt a célunk, hogy ne csak szimplán egy eloszlás alapján döntsünk egy változó transzformálásáról, hanem azután, miután már más vizualizációval is megbizonyosodtunk arról, hogy valóban helyes döntés lesz a transzformáció. A pontdiagramok ábrázolása után arra jutottunk, hogy a hisztogramnál talált 5 lehetséges változót valóban logaritmizálni lehet és emellett a *telepulesek_120felett* változónak pedig a kvadratikus transzformáltjával lenne érdemes dolgozni a jobb becslés érdekében. Ezek a kapcsolatok a pontdiagramokon itt láthatóak:



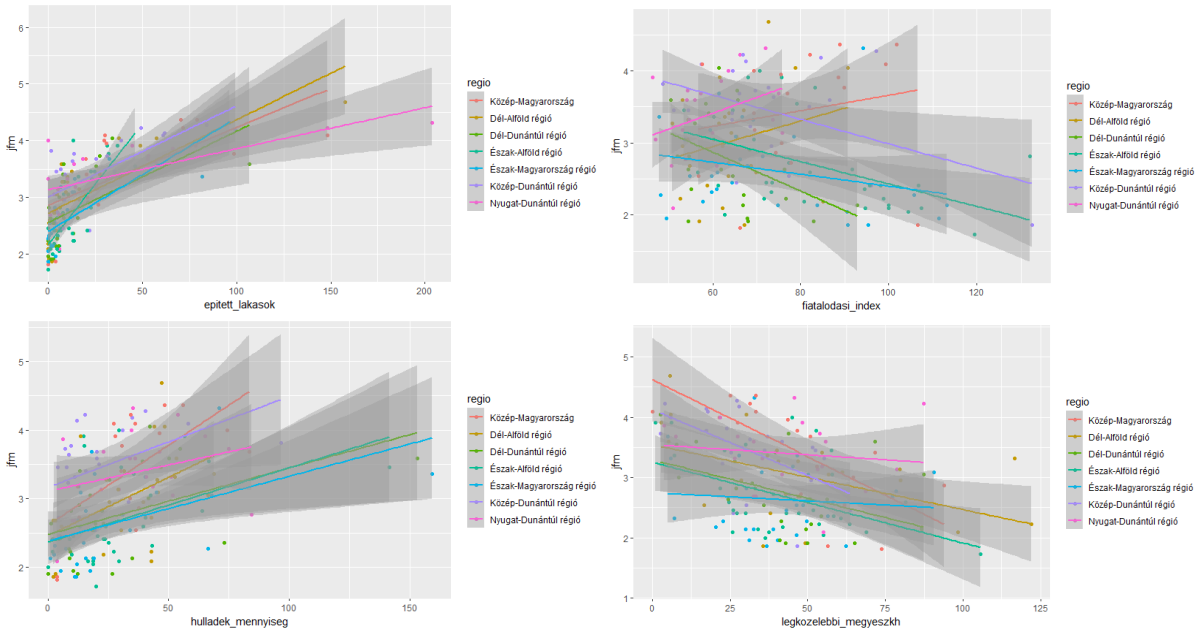
1., 2. és 3. ábra: Transzformált tagok és eredményváltozó páronkénti kapcsolata



4., 5. és 6. ábra: Transzformált tagok és eredményváltozó páronkénti kapcsolata

Interakciók detektálása

Az utolsó lépés ebben a feladatrészben az interakció detektálása volt. Itt az említett *regio* faktor típusú változót és a többi numerikus magyarázóváltozó kapcsolatát vizsgáltuk. Az interakciót akkor tartottuk szükségesnek, ha a különböző régiókban drasztikusak a meredekségkülönbségek és valóban nagyok az eltérések a térségek között. A kisebb eltérésekkel, meredekségkülönbségekkel nem foglalkoztunk. Így ezek után arra jutottunk, hogy 4 interakciót kell megvizsgálni majd a modellépítés folyamán. Ezek a lehetséges interakciók a következőek: *fiatalosodasi_index*regio, hulladek_mennyiseg*regio, epitett_lakasok*regio, legkozelebbi_megyeszkh*regio.* A hozzájuk tartozó ábrák pedig itt láthatóak:



7., 8., 9. és 10. ábra: Interakciók detektálása

Modellépítés

Ezt követően megalkottuk a modell_1-et. Itt még nem dobtunk ki semmit, csak annyit változtattunk, hogy egy új magyarázóváltozót emeltünk be, ez pedig a már említett *régió* változó. A modellre futtatott regresszió adatai itt láthatóak:

	Együttható	p-érték
(Intercept)	2,7959	6,78E-12
kisker_boltok	0,0272	0,000932
epitett_lakasok	0,0026	0,000125
legkozelebbi_megyeszkh	-0,0032	2,25E-05
szennyviz_csatlakozas	0,0067	7,28E-06
hulladek_mennyiseg	0,0023	0,00107
villamosenergia	0,0005	3,18E-05
halalozas	-0,0609	1,36E-06
telepulesek_120felett	0,0040	7,13E-09
fiatalodasi_index	0,0022	0,196638
allaskeresok_aranya	-0,0973	1,28E-24
regioDél-Alföld régió	-0,0217	0,745435
regioDél-Dunántúl régió	-0,0199	0,774499
regioÉszak-Alföld régió	-0,0816	0,2289
regioÉszak-Magyarország régió	-0,0016	0,981295
regioKözép-Dunántúl régió	-0,0192	0,765214
regioNyugat-Dunántúl régió	0,0817	0,241606
R-négyzet	93,59%	Magas magyarázóerő
F-próba	< 2,2E-16	A modell szignifikáns
RESET	9,29E-08	A modell rosszul specifikált

2. Táblázat: A modell_1 regresszió eredményei

Bár a régió változó nem lett szignifikáns, még nem biztos, hogy érdemes teljesen elvetni, ugyanis ez az egyetlen minőségi változónk, így interakciós szempontból még hasznos lehet. A globális F-próba alapján a modellünk továbbra is szignifikáns a mintán kívüli világban. Illetve, a modellre lefuttatott RESET teszt szerint a modell_1 rosszul specifikált.

Ezt követően elvégeztünk a változókon pár transzformációt. A kiskereskedelmi boltok, az épített lakások, a hulladék mennyisége, a fiatalodási index és az álláskeresők aránya változókat logaritmizáltuk (apró módszertani megjegyzés: bizonyos együtthatók értékeihez hozzáadtunk 0.001-et, hogy az esetleges 0 értékek miatt ne ütközzünk hibába), a települések 120 felett változót pedig négyzetesen tettük bele a modellbe. A fiatalodási index és a régió továbbra sem lett szignifikáns, minden más változó pedig szignifikáns maradt. A modell_2 RESET tesztjének p-értéke 5.595*10^(-8) lett. Így a modell_2 minden transzformálás ellenére kevésbé specifikált, mint a modell_1.

Következő körben úgy döntöttünk, hogy kiszedjük a fiatalodási indexet, így létrehozva a modell_3-at. Sok változás nem állt be, de a Nyugat-Dunántúl régió most már szignifikáns lett 10%-on (a többi p-értéke ugyanúgy magas maradt).

	modell_2	modell_3
RESET	5,60E-08	5,17E-08
BIC	25,11	19,96
Wald	0,898	Megérte elhagyni

3. Táblázat: Modell_2 és modell_3 összehasonlítása

A RESET teszt szerint valamivel kevésbé specifikált a modellünk, mint a modell_2. Viszont a BIC mutató a modell_3-nál alacsonyabb, mint a modell_2-nél, így mégis a modell_3-at részesítjük előnyben. Ezt a döntést a Wald próba 0.898-as p-értéke is alátámasztja.

Ezt követően úgy döntöttünk, hogy interakciókkal még tovább bővítjük a modellünket. Bár a fiatalodási indexet korábban kihagytuk, de itt ismét indokolt lehet visszahozni, hátha szignifikáns lesz valamivel interaktálva. Ahogyan már említettük, az egyetlen minőségi változónk a régió, (minden más numerikus), így a régió minden interakciónak része lesz. Az ábrák alapján négy lehetséges interakció jöhet szóba és ezeket fogjuk most megvizsgálni a különböző modellekben. Ezek a régió és a fiatalodási index, a régió és a hulladék mennyiség, a régió és az épített lakások, illetve a régió és a legközelebbi megyeszékhely interakciója.

Az új modellünk a modell_int1, amiben mind a négy interakció szerepel.

	Együttható	p-érték
(Intercept)	2,011751	2,39E-04
log(kisker_boltok)	0,266083	0,001939
log(epitett_lakasok + 0.001)	0,014201	0,029302
legkozelebbi_megyeszkh	0,000636	0,786329
szennyviz_csatlakozas	0,00504	5,70E-04
log(hulladek_mennyiseg + 0.001)	0,04497	0,011332
villamosenergia	0,000697	1,93E-07
halalozas	-0,05918	3,24E-06
I(telepulesek_120felett^2)	0,0000517	1,37E-09
log(allaskeresok_aranya)	-0,43092	9,26E-17
regioDél-Alföld régió	0,013673	0,974207
regioDél-Dunántúl régió	0,100987	0,812601
regioÉszak-Alföld régió	0,564021	0,131253
regioÉszak-Magyarország régió	0,557327	0,124521
regioKözép-Dunántúl régió	0,988738	0,010254
regioNyugat-Dunántúl régió	1,192991	0,011335
fiatalodasi_index	0,004697	0,250457
hulladek_mennyiseg	0,003483	0,122454
epitett_lakasok	-0,00061	0,694805
regioDél-Alföld régió:fiatalodasi_index	0,00039	0,943246
regioDél-Dunántúl régió:fiatalodasi_index	0,005318	0,347785
regioÉszak-Alföld régió:fiatalodasi_index	-0,00887	0,035675
regioÉszak-Magyarország régió:fiatalodasi_index	-0,00564	0,183125
regioKözép-Dunántúl régió:fiatalodasi_index	-0,00774	0,081913
regioNyugat-Dunántúl régió:fiatalodasi_index	-0,01196	0,068029
regioDél-Alföld régió:hulladek_mennyiseg	0,002521	0,380457
regioDél-Dunántúl régió:hulladek_mennyiseg	-0,00503	0,119877
regioÉszak-Alföld régió:hulladek_mennyiseg	-0,00304	0,244609
regioÉszak-Magyarország régió:hulladek_mennyiseg	-0,00506	0,05953
regioKözép-Dunántúl régió:hulladek_mennyiseg	-0,00607	0,037987
regioNyugat-Dunántúl régió:hulladek_mennyiseg	-0,00093	0,75566
regioDél-Alföld régió:epitett_lakasok	0,00311	0,110896
regioDél-Dunántúl régió:epitett_lakasok	0,007715	0,035343
regioÉszak-Alföld régió:epitett_lakasok	0,016151	1,17E-05
regioÉszak-Magyarország régió:epitett_lakasok	0,007686	0,002996
regioKözép-Dunántúl régió:epitett_lakasok	0,001216	0,566711
regioNyugat-Dunántúl régió:epitett_lakasok	0,001467	0,398389
legkozelebbi_megyeszkh:regioDél-Alföld régió	-0,00259	0,325189
legkozelebbi_megyeszkh:regioDél-Dunántúl régió	-0,00806	0,005367
legkozelebbi_megyeszkh:regioÉszak-Alföld régió	-0,00093	0,737647
legkozelebbi_megyeszkh:regioÉszak-Magyarország régió	-0,00187	0,518212
legkozelebbi_megyeszkh:regioKözép-Dunántúl régió	-0,00579	0,087458
legkozelebbi_megyeszkh:regioNyugat-Dunántúl régió	-0,00663	0,036359
R-négyzet	95,61%	Magas magyarázóerő
F-próba	< 2,2E-16	A modell szignifikáns
RESET	1,51E-05	A modell rosszul specifikált

4. Táblázat: Interakciós modell: modell_int1 regressziós eredményei

Az interakciók között vegyesen vannak szignifikáns és nem szignifikáns kapcsolatok. A RESET teszt p-értéke viszont jelentősen nőtt. Bár a modell még mindig rosszul specifikált, de eddig ez a legjobb. A BIC mutató viszont a modell_3-at preferálja (a modell_3 értéke 19,96, a modell_int1-é pedig 64,72).

Ezután úgy döntöttünk, ideje végleg elbúcsúznunk a fiatalodási indextől. Ezt a modellt modell_int2-nek neveztük el és kiszedtük a fiatalodási indexet és annak a régióval vett interakcióját is.

	modell_3	modell_int2
RESET	5,60E-08	1,64E-05
BIC	19,96	50,71

5. Táblázat: Modell_3 és modell_int2 összehasonlítása

A RESET teszt p-értéke minimálisan megnőtt. De a BIC mutató jobban érdekelt minket. A modell_int2 BIC értéke még mindig jóval nagyobb, mint a modell_3-é. Összességében tehát nem érte meg betenni még mindig az interakciókat.

Következő lépésben létrehoztuk a modell_int3-at, amelyből kiszedtük a régió és a legközelebbi megyeszékhely távolságának interakcióját is.

	modell_3	modell_int3
RESET	5,60E-08	1,34E-06
BIC	19,96	35,47

6. Táblázat: Modell_3 és modell_int3 összehasonlítása

A modellünk kicsit rosszabbul specifikált a modell_int2-höz képest, de jobban specifikált a modell_3-hoz képest. Viszont még mindig olybá tűnik, hogy feleslegesen kerültek be az interakciók.

Legvégső lépésben létrehoztuk a modell_int4-et, amelyből kiszedtük a régió és a hulladékszállítás interakciót.

	modell_3	modell_int4
RESET	5,60E-08	3,962E-08
BIC	19,96	17,54
Wald	1,69E-05	Nem érné meg elhagyni az interakciót

7. Táblázat: Modell_3 és modell_int4 összehasonlítása

Az eredmények alapján látható, hogy az interakciók nélküli modell_3-hoz képest rosszabbul specifikált a modell_int4, viszont a BIC információs kritérium alapján ez a preferált modell és a Wald-teszt is azt mutatja nekünk, hogy szükség van az interakcióra. Most már végre kijelenthetjük, hogy volt értelme interakciót tenni a modellbe. A BIC és a Wald-teszt fényében

mi a modell_int4-et választjuk, annak ellenére is, hogy a modell_3-nál rosszabbul specifikált. Számunkra az előbbi 2 szempont fontosabb, mivel mindkét esetben a modell rosszul specifikált és nagyon minimális a különbség csak a kettő RESET-teszt értéke között. Az alapmodellhez képest viszont a modell_int4 az jobban specifikált, szóval ahhoz viszonyítva is sikerült javítani a helyzetet.

A végső modellünk tehát a modell_int4. Ebben a kiskereskedelmi boltok logaritmusa, az épített lakások logaritmusa, a legközelebbi megyeszékhely, a szennyvíz csatlakozás, a hulladékmennyiség logaritmusa, a villamosenergia, a halálozás, a települések 120 felett négyzete, az álláskeresők arányának logaritmusa, a régió, illetve a régió és az épített lakások interakciója szerepelnek.

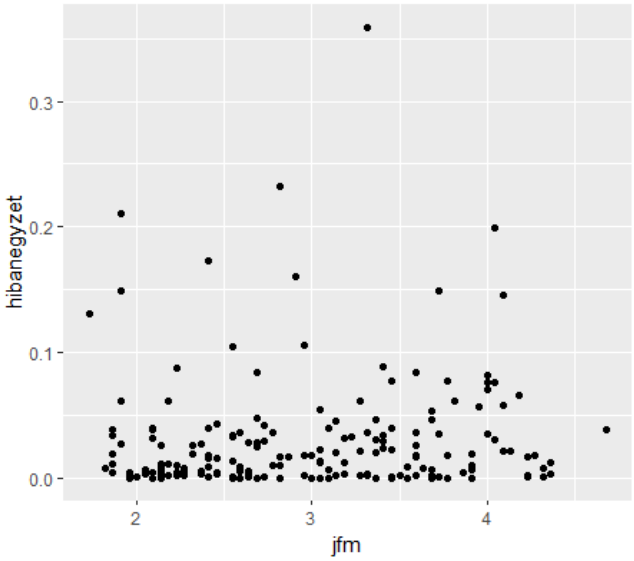
	Együttható	p-érték
(Intercept)	2,645333	1,63E-16
log(kisker_boltok)	0,291178	0,000305
log(epitett_lakasok + 0.001)	0,014631	0,028031
legkozelebbi_megyeszkh	-0,00248	0,001414
szennyviz_csatlakozas	0,005765	0,00011
log(hulladek_mennyiseg + 0.001)	0,041347	0,003088
villamosenergia	0,000635	1,03E-06
halalozas	-0,05651	1,82E-07
I(telepulesek_120felett^2)	4,66E-05	1,06E-07
log(allaskeresok_aranya)	-0,46837	1,53E-20
regioDél-Alföld régió	-0,09825	0,241313
regioDél-Dunántúl régió	-0,07987	0,345442
regioÉszak-Alföld régió	-0,29739	0,001273
regioÉszak-Magyarország régió	-0,10118	0,225745
regioKözép-Dunántúl régió	-0,02405	0,79222
regioNyugat-Dunántúl régió	0,088531	0,320459
epitett_lakasok	-0,0002	0,882851
regioDél-Alföld régió:epitett_lakasok	0,003698	0,028998
regioDél-Dunántúl régió:epitett_lakasok	0,003429	0,140004
regioÉszak-Alföld régió:epitett_lakasok	0,015409	4,46E-06
regioÉszak-Magyarország régió:epitett_lakasok	0,005567	0,006026
regioKözép-Dunántúl régió:epitett_lakasok	0,00107	0,572995
regioNyugat-Dunántúl régió:epitett_lakasok	8,25E-05	0,956254
R-négyzet	93,95%	Magas magyarázóerő
F-próba	< 2,2E-16	A modell szignifikáns
RESET	3,96E-08	A modell rosszul specifikált

8. Táblázat: Végső modell: modell_int4

RESET teszt alapján a végső modellünk továbbra is rosszul specifikált. Ezen sajnos nem tudtunk segíteni. Ugyanakkor azt is meg kell említeni, hogy a RESET teszt eléggé szigorú, így nem egyszerű jól specifikált modellt összerakni.

Heteroszkedaszticitás

Megvizsgáltuk a végső modellünket homoszkedaszcitás szempontjából is. A Kolmogorov-Smirnov teszt p-értéke 10^(-16) nagyságrendű lett, szóval a hibatag nem normális eloszlású, így érdemes a Koenker korrekciózott Breusch-Pagan tesztet alkalmazni. A teszt p-értéke 0.1469 lett. Ebből az következik, hogy a modellünk homoszkedasztikus, szóval a hibatagnégyzetek véletlenszerűek.



11. ábra: Hibanégyzetek a JFM függvényében

Multikollinearitás

Ezután a multikollinearitást is leellenőriztük. A korrelációs mátrixon látható, hogy a változók között nincsen nagy korreláció, ez arra utal, mint az alapmodellnél, hogy nem lesz multikollinearitás a modellünkben. A VIF mutatók értékei után azt kaptuk, hogy 3 változó kivételével mindegyik változó a zavaró 5-ös érték alatt helyezkedik el. Az ebből kilógó értékek, ahogyan sejteni lehetett az interakció és annak két tagja volt. A régió közel 53-as, az épített lakások változó pedig közel 8-as VIF értéket vett fel. A két tag interakciója 90-es VIF mutató értékkel rendelkezett. Ezek az értékek mesterségesen lettek generálva általunk, mivel interakciót vettünk be a modellbe, ilyenkor elkerülhetetlen a magas VIF érték ezeknél a mutatóknál, szóval mi úgy gondoljuk, hogy ezt nem kell kezelni és ez nem okoz problémát a becslésnél. Ezen döntésünket az is igazolja, hogy megnéztük a modell_3 VIF mutató értékeit, amely ugyanezeket a változókat tartalmazza az interakció kivételével és ott minden változó a 4-es érték alatt szerepelt, szóval szó sincs multikollinearitásról. Továbbá a Gyökös VIF-mutató értékei is 1-hez közeliek a végső modellünkben, amely azt jelenti, hogy nincsen multikollinearitás és nem kell kezelni semmit sem.

Ennek ellenére elvégeztük a főkomponenselemzést a teljesség érdekében és megnéztük, hogy a sajátérték-szabály szerint mennyi főkomponenst kellene bevenni a modellbe és ezeknek mi lenne a jelentéstartalmuk. Itt csak az utolsó lépést hagytuk ki, amely az, hogy valóban kezeljük a multikollinearitást a modellben, mivel ahogyan már említettük erre nincsen szükség. A sajátérték szabály szerint 3 főkomponenst kell bevenni. PC1 erősen negatívan korrelál az épített lakásokkal, a szennyvíz csatlakozással, a kiskereskedelmi boltokkal és a 120 fő/km2 feletti településekkel. Ez a főkomponens szerintünk egy népsűrűségi mutató. Ha nő PC1, tehát nő a

népsűrűség, akkor az említett mutatók értékei csökkennek. Ez teljesen logikus. Ahol többen laknak ott több az épített lakás, a szennyvíz csatornával egybekötött lakás, a kiskereskedelmi boltok száma és természetesen maga a népsűrűség is. A PC2 erősen negatívan korrelál a halálozással, viszont erősen pozitívan a fiatalosodási index változókkal. Ez a főkomponens egyfajta demográfiai mutató, azon belül is pontosabban egy korcsoport/korszerkezeti mutató. A megfogalmazása így elég nehéz ennek a főkomponensnek, ha PC2 nő, akkor kisebb a halálozások száma és nagyobb a fiatalok aránya egy járáson belül. Ez logikusnak mondható, mivel Magyarország egy modern és fejlett egészségüggyel rendelkezik (az afrikai országokhoz képest mindenképp) és normális esetben elmondható az, hogy ha nagyobb a fiatalok aránya egy járáson belül, az kisebb halálozással jár, mint annál a járásnál, ahol ez az arány fordított. Végezetül pedig a PC3 erősen negatívan korrelál a legközelebbi megyeszékhellyel és a hulladék mennyisége változókkal. Ez a főkomponens érdekesen talán úgy írható le, mint egy vidékiségi mutató. Ez azt jelenti, hogyha PC3 nő, akkor messzebb kerülünk az adott megyeszékhelytől és ezáltal csökken a hulladék mennyisége is. Ez teljesen logikus, mivel a megyeszékhelyek körül általában a nagyobb, jobban lakott járások vannak, és ebből adódóan több ember nagyobb mennyiségű hulladékot termel. Ha innen kicsit messzebb megyünk, egy vidékiesebb, kevésbé lakott helyre, messzebb a megyeszékhelytől ott természetesen kevesebb a hulladék.

Regressziós egyenlet és előrejelzés

A végső modellre felírtuk a regressziós egyenletet is, azonban a sok változóra való tekintettel azt táblázatos formában közöljük.

Változók	Becsült béta koefficiensek
(Intercept)	2,65
log(kisker_boltok)	0,291
log(epitett_lakasok + 0.001)	0,0146
legkozelebbi_megyeszkh	-0,00248
szennyviz_csatlakozas	0,00577
log(hulladek_mennyiseg + 0.001)	0,0413
villamosenergia	0,000635
halalozas	-0,0565
I(telepulesek_120felett^2)	0,0000466
log(allaskeresok_aranya)	-0,468
regioDél-Alföld régió	-0,0983
regioDél-Dunántúl régió	-0,0799
regioÉszak-Alföld régió	-0,297
regioÉszak-Magyarország régió	-0,101
regioKözép-Dunántúl régió	-0,0241
regioNyugat-Dunántúl régió	0,0885
epitett_lakasok	-0,000196
regioDél-Alföld régió:epitett_lakasok	0,0037
regioDél-Dunántúl régió:epitett_lakasok	0,00343
regioÉszak-Alföld régió:epitett_lakasok	0,0154
regioÉszak-Magyarország régió:epitett_lakasok	0,00557
regioKözép-Dunántúl régió:epitett_lakasok	0,00107
regioNyugat-Dunántúl régió:epitett_lakasok	0,0000825

9. Táblázat: A becsült béta koefficiensek

A táblázatban szereplő tengelymetszetet nem értelmezzük, hiszen az azt jelentené, hogy minden numerikus változóból 0 értékű járás eredményét értelmezzük, azonban könnyű belátni, hogy egy ilyen járásban nem lenne élet, hiszen semmilyen jellegű fogyasztás nem merül fel (az

általunk használt változókat tekintve). Azonban azt érdemes megemlíteni, hogy a tengelymetszet 2,65-os értékében található a Közép-Magyarország régió dummy értéke. Általánosságban elmondható, hogy igen alacsony béta értékeket becsült a modell az erős szignifikancia ellenére, de érdemes szerintünk kitérni pl. a legközelebbi megyeszékhely változó -0,00248-es bétájára, ami azt jelenti, hogy átlagosan ha 1 kilométerrel távolabb helyezkedik el a megfigyelt járás a legközelebbi megyeszékhelytől, akkor várhatóan 248 százezreddel fog csökkeni a fejlettségi mutató, ami a várakozásainknak megfelel és logikus, hiszen a megyeszékhelyeken jellemzően több a munkalehetőség, több a beruházás, és ezek közelsége pozitívan hat.

A lineáris regressziós modell építésének egyik fő célja, hogy a múlt magyarázásán kívül a jövőről is kapjunk egy lehetőleg reális képet. A modellépítés során az egyik fő célunk a kimaradt budapesti kerületek fejlettségének prediktálása volt, hiszen kifejezetten izgalmasnak gondoljuk a vízfejű országunk általában legfejlettebbnek gondolt fővárosának részletesebb elemzését. A modellünkhöz felhasznált változókhoz szükséges adatok azonban sajnos nem álltak rendelkezésre (feltehetően ezért maradtak ki az eredeti adatbázisból is), így az egyes kerületekhez nem tudtunk értéket rendelni. Mivel a járásokról szóló adatbázis nem csupán egy minta, hanem ezen körülmények mellett maga a sokaság is, így véleményünk szerint értelmes előrejelzést nem tudtunk készíteni. Az érdekesség kedvéért azért megnéztük a minden szempontból medián értékekkel rendelkező, elképzelt „Medián járás” -t (a régiót tekintve Közép-Magyarország értéket rendeltünk hozzá). Várakozásainknak megfelelően a JFM egy tizedesjegyre 3 lett, ami a JFM változó leíró statisztikájával összhangban van.

Összefoglalás

Érdekes és egyben pozitív is, hogy nem csak gazdasági mutatók alapján határozható meg egy térség fejlettsége. Ez a mutató egyéb, szociális jellegű adatokat is tömörít, ami kifejezetten üdítő. A mi célunk ennek a mutatónak egyfajta megközelítése volt. A magas magyarázóerő alapján azt gondoljuk, hogy jó munkát végeztünk. Nyitott kérdés, hogy vajon ezen magyarországi minta alapján mennyire tudnánk megbecsülni más országok régióinak a fejlettségét. Sajnos más országok régióira nem találtunk ennyire részletes adatokat és terjedelmi okok miatt sem lenne szerencsés most ezt a kérdést bővebben feszegetni. De a jövőben egy érdekes kutatás alapja lehet.