KÖZGAZDASÁGI SZEMLE, LXII. ÉVF., 2015. DECEMBER (1343–1358. o.)

GÁL RÓBERT IVÁN–TÖRZSÖK ÁRPÁD

# Háztartás-formálódás a MIDAS modellben

**A MIDAS dinamikus mikroszimulációs nyugdíjmodell magyar változata tartalmaz egy háztartás-formálódási modult, amely a szimuláció során életpályájukon végigkísért egyéneket összekapcsolja. Ez lehetővé teszi a felosztó-kirovó nyugdíjrendszer azon elemeinek modellezését, amelyekben a járulék vagy a járadék függ az egyén családi kapcsolataitól. Ilyen elemek az özvegyi és egyéb hozzátartozói nyugdíjak, a családi állapottól függő járulék- és járadékmegállapítás, a változatos gyermekkedvezmények vagy a gyermeknevelés-függő nyugdíjmegállapítás. A magyar adatkörnyezetben a háztartás-formálódási modul programozása speciális nehézségekbe ütközik, mivel nem áll rendelkezésre olyan adminisztratív vagy kérdőíves kikérdezésre épülő adatállomány, amely azon felül, hogy az egyénre nézve kapcsolattörténeti és háztartás-szerkezeti adatokat tartalmaz, kiterjed a párokra, vagy ha kiterjed is, ezt a tényt nem rögzíti. A szóba jöhető survey jellegű adatbázisok egyéni mintákat tartalmaznak, az adminisztratív adatállományok pedig, melyek a nyugdíjrendszerrel kapcsolatos releváns információkat gyűjtik, nem kapcsolják össze az egyéneket. A MIDAS modell magyar változata ezért egy retrospektív háztartás-rekonstrukciós szimulációt is elvégez.***

**Journal of Economic Literature (JEL) kód: C53, H55.**

A felosztó-kirovó nyugdíjrendszer kettős dilemmája, hogy egyénre szabott biztosítást nyújt annak ellenére, hogy a járulékfizetési teljesítmény nem egyéni, és hogy járuléknak kizárólag az aktívak által a náluk idősebbeknek fizetett transzfereket tekinti, noha a későbbi járadék forrása a fiatalabbaktól beszedett összeg.

Ami a járulékfizetési teljesítményt illeti, a családon[1] belül munkamegosztás érvényesül, tipikusan férfi és nő között. A férfiak jellemzően aktívabbak a munkaerőpiacon, amit a nyugdíjrendszer magasabb ellátással honorál, míg a nők jellemzően

* Köszönettel tartozunk *Rézmovits Ádámnak, Tóth Krisztiánnak* és egy névtelen lektornak, valamint az ONYF négy MIDAS_HU témájú műhelykonferenciáján hozzászólóknak. A tanulmányban maradt hibákért ők természetesen nem felelősek.

[1] E tanulmányban a családot és a háztartást szinonimaként használjuk.

*Gál Róbert Iván,* KSH Népességtudományi Kutatóintézet, Tárki, Budapesti Corvinus Egyetem.
*Törzsök Árpád,* Nokia Solutions and Networks.
A kézirat első változata 2015. május 26-án érkezett szerkesztőségünkbe.
DOI: http://dx.doi.org/10.18414/KSZ.2015.12.1343

többet dolgoznak a háztartásban, ami viszont nem teremt közvetlen jogosultságot. Egy tisztán járulékmeghatározott nyugdíjformula így vagy magasabb női korhatárt írna elő, vagy „garantálná" az időskori női szegénységet. A női magasabb várható élettartam ezt csak felerősíti.

A létező nyugdíjrendszerek azonban, bár ragaszkodnak az egyéni járulékfizetési teljesítmény fikciójához, kerülő utakon mégiscsak elismerik a szóban forgó aszimmetriát. Az özvegyi nyugdíj szerzett jogosultság, nem rászorultsági alapon állapítják meg. Egy ilyen nyugdíjrendszer hallgatólagosan elismeri, hogy a jogosultságszerzés alapegysége nem az egyén, hanem a házaspár. A pár tagjai közötti munkamegosztást a nyugdíjrendszer, még ha csak találomra és közelítő megoldásokkal is, de tudomásul veszi és honorálja. A jobban kereső társ munkaerőpiaci teljesítményét a másodkereső vagy kereset nélküli társ segíti elő. Aki többet keres, azért tudja megtenni, mert valaki segít neki a háztartási munkában. Aki többet dolgozik a háztartásban, azért teheti, mert valaki, legalábbis részben, eltartja. Mindkét fajta munka értéket termel, de csak az egyik teremt nyugdíjjogosultságot. Az özvegyi nyugdíj puszta létét az magyarázza, hogy a nyugdíjrendszer megalkotói valamiképp tudomásul veszik, hogy a férfiak hosszabb szolgálati karrierje és magasabb járulékai nem csupán saját teljesítményük eredménye. Ez a magyarázata annak, hogy a biztosítás kiterjed a házastársra és élettársra, de nem terjed ki, mondjuk, a barátokra. Megfordítva, a háztartásban végzett munka mennyiségének csökkenése és egyenlő elosztása lehetővé teszi a női munkavállalást, ami idővel jelentéktelenné teszi az özvegyi nyugdíjat.[2]

A másik dilemma ugyancsak a biztosítottak családi kötődéseivel kapcsolatos. A felosztó-kirovó nyugdíjrendszer egy olyan életpálya-finanszírozási intézményrendszerbe illeszkedik, amely összekapcsolja az egymást követő generációkat: cserébe saját felnevelésük költségeiért az aktív korúak eltartják az időseket, és felnevelik gyermekeiket, akik, felnövekedvén, majd őket fogják eltartani. A gyermeknevelés ráfordításainak jelentős része azonban – jelesül a háztartásban és a háztartások között zajló közvetlen újraelosztás, valamint a nem fizetett háztartási munka – nem látszik a nyugdíjrendszer nyilvántartásaiban. A létező nyugdíjformulák, ha egyáltalán foglalkoznak ilyesmivel, a járulékfizetési történetre vagy annak valamilyen közelítő változójára épülnek, tehát arra, hogy az aktív korúak mennyit adnak a náluk idősebbeknek.

Ennek ellenére a gyermekbeszámítás valamiképp általánosan elfogadott. Az Európai Unió tagállamainak nagyobbik fele alkalmaz valamilyen formulát, amely, ha csak kismértékben is, akár korhatárkedvezményként, akár járulékkedvezményként, akár kiegészítő járadékként beszámítja a felnevelt gyermekek számát. Ez megint csak egy újabb háztartási információ, amelyet bekapcsolnak a nyugdíjszámításba, még akkor is, ha az egyéni nyugdíjbiztosítás fikciója továbbra is fennmarad. Ismét kerülő úton, de a nagyobb biztosítási teljesítményt, még ha csak kismértékben is, elismeri a rendszer. Ezen túlmenően a szakirodalom és a nyugdíjakról folyó társadalmi vita több javaslatot ismer, amelyek a nyugdíj-megállapítás

[2] A hátramaradotti nyugdíjakkal kapcsolatos összefoglalást közöl *Monticone–Ruzik–Skiba* [2008].

szabályait átírva a mainál sokkal nagyobb jelentőséget adnának a felnevelt gyermekek számának és iskolai végzettségének, azaz a következő járulékfizetői nemzedék emberi tőkéjébe tett beruházásoknak.[3]

## Egyén és háztartás a MIDAS modellben

A modell képes kezelni a háztartást mint egységet, ugyanis önálló háztartás-formálódási modult tartalmaz. Ez lehetővé teszi egy sor olyan meglévő, illetve lehetséges nyugdíjszabályozás modellezését, amire a korábbi szimulációs modellek nem adtak módot.

A háztartás-formálódási modul (az angol eredetiben házasságpiac, *marriage market*) a valóságos háztartás-formálódási folyamatok közül a házasságok és élettársi kapcsolatok keletkezését, fennmaradását az idők során, illetve felbomlását modellezi. A modul része a gyermekek születése (háztartásbővülés), majd kiköltözése (háztartás-összehúzódás) és önálló háztartásuk megalapítása. A modul egyelőre nem teszi lehetővé az idősek összeköltözését felnőtté vált gyermekükkel vagy elvált felnőttek összeköltözését szüleikkel, és nem kezel többgenerációs vagy többcsaládos háztartásokat, illetve azonos neműek közötti élettársi kapcsolatokat.

A modellben így kétféle entitás létezik, a személy és a háztartás. Személyek kapcsolódhatnak más személyekhez (házas- és élettársi kapcsolat, szülő–gyerek kapcsolat), valamint kapcsolódnak háztartáshoz (egyszerre csak egyhez, ahhoz, amelynek tagjai). A háztartások modellezéséhez arra van szükség, hogy ezeket az egyszerű kapcsolódásokat, illetve e kapcsolódások változásait következetesen kövessük azokban az élethelyzetekben, amelyeket modellezni kívánunk. Példaként: ha egy házaspárnak gyermeke születik, az a mikroszimulációban azt jelenti, hogy létrejön egy új személy, két új szülő–gyerek kapcsolat, és egy új személy–háztartás kapcsolat, a gyermek ugyanis a szülők háztartásának tagja lesz. Ily módon a modellben követni tudjuk a házasságkötést, a gyerekek kirepülését, a válást, a szétköltözést és a halálozást. E folyamatok mind befolyásolják a háztartások összetételét, és a személy–háztartás illetve a személy–személy kapcsolatok létrehozásával és törlésével pontosan lekövethetők.

A magyar nyugdíjnyilvántartás információtartalma tükrözi a közkeletű nyugdíjmodellek kizárólagos egyén-központúságát. Mint azt *Puskás Péter* jelen számunkban megjelent írásában részletesen bemutatja, a magyar nyugdíjnyilvántartás alapján rekonstruálható a teljes foglalkoztatási és bértörténet. Ugyanakkor a biztosított családi állapotáról, gyermekeinek számáról és azok iskolai végzettségéről semmit sem

[3] A gyermeknevelés-függő nyugdíj-megállapítás javaslata az 1980-as évek végére megy vissza (*Demény* [1986], *Bental* [1989]). Részletes érvkészletet és kidolgozott javaslatot közöl *Sinn* [2004] és *Meier–Werding* [2010]. A szakirodalom részletes áttekintése megtalálható *Cigno–Werding* [2007] könyvében. Magyarországon a kérdést elsőként *Botos–Botos* [2012] vetette fel; a témával kapcsolatos további egyetértő és vitató írásokat tartalmaz *Kovács* [2013]. A szakirodalom kapcsolódó fejezeteit (a nyugdíjrendszer termékenységi hatásai, endogén termékenység, összekapcsolódó transzferáramlatok, három életszakaszos modellek) ehelyütt nem szemlézzük.

tudunk meg belőle, annak ellenére, hogy ezek megfigyelése egyszerű, és más helyeken már keletkezik róluk adminisztratív adat. Tehát már a kiinduló helyzet rekonstruálása is szimulációs feladatot jelent.

Ráadásul e technikai jellegű retrospektív műveletre egy további okból is szükség van. A modell népességének előállítási módja miatt ugyanis még az sem lenne elegendő, ha rendelkezésünkre állnának a háztartás méretére, korösszetételére, a háztartástagok családi állapotára, rokoni viszonyaira és kapcsolataik addigi élettartamára vonatkozó információk (bár a szimuláció készítőjének dolgát kétségtelenül megkönnyítenék). A magyar MIDAS modell nem sokasági adatokon fut, hanem mintán vagy a sokasági adatok összesűrítéséből származó modellpontokon.[4]

Az első eljárás egyéneket választ ki, a második szintetikus egyéneket állít elő. Mindkét eljárás, ha megfelelően alkalmazzák, megbízhatóan reprezentálja a sokaság egyéneit, de egyik sem generál háztartásokat. A háztartás-formálódási modul viszont háztartásokkal, azaz összekapcsolt egyénekkel dolgozik. Erre az összekapcsolásra sem az egyéni mintavétel, sem a modellpontok előállítása nem alkalmas, és még akkor sem lenne az, ha a sokasági adatállomány tartalmazna adatokat az egyének háztartási viszonyaira. Ez utóbbi esetben is csak azt tudnánk meg, hogy a kiinduló állomány tagjai egyedül vagy kapcsolatban élnek-e, ha az utóbbi, akkor milyen típusú a kapcsolat (házasság vagy élettársi kapcsolat); ismernénk az érintettek párjainak életkorát is, a kapcsolat élettartamát és még egy sor további szociológiai jellemzőt. Még ebben az esetben sem lennének maguk a párok a kiinduló állomány tagjai. Ezért mielőtt hozzálátunk a mikroszimulációhoz, kiinduló adatállományunkban rekonstruálnunk kell a társadalmat mint háztartások – és nem csak mint egyének – összességét. Más szóval, a rendelkezésünkre álló ismérvek alapján össze kell párosítanunk a kiinduló állomány tagjait egymással, úgy, hogy tükrözzék a társadalom háztartási jellemzőit. Ehhez ugyancsak a háztartás-formálódási modul algoritmusait alkalmazzuk.

## A háztartás-formálódás modul működése

### *Algoritmusok*

A háztartás-formálódási modul több almodult tartalmaz. Ezek a kialakulás, a bővülés (születés), az összehúzódás (kiköltözés), a felbomlás (válás és halálozás), valamint a fentiekben említett retrospektív háztartás-formálódás almodulok.

Minden almodul két lépésből áll. Az első algoritmus kiválasztja a modell népességének megfelelő tagjait. A születés és halálozás esetében más nem is történik – mondhatni, a kiválasztás után csak végre kell hajtani a kiválasztott modellszemélyeken az ítéletet. A második algoritmus az összepárosítást hajtja végre. Ehhez természetesen az előző algoritmussal ki kell választanunk, hogy kiket rendezünk párba az adott évben.

[4] A modellpontok olyan szintetikus személyek, amelyek a modellezett sokaság tagjainak a tulajdonságaival rendelkeznek, de nem kapcsolhatók egyenként a sokaság tagjaihoz. Tulajdonságaikat a modellezett sokaságból klaszterezéssel állítjuk elő. Erről bővebben lásd *Kovács Ezsébet, Rétallér Orsolya* és *Vékás Péter* jelen számunkban közölt cikkét.

Nézzük először a *kiválasztási* algoritmust! A kiválasztás módja főként attól függ, hogy milyen információk állnak rendelkezésre. A jogszabályok ismeretében vagy szakértői információk alapján előszűrhetjük a sokaságot, például nem engedünk olyanokat házasodni, akik már házasok. Ha más információnk nincs, akkor elég annyit tudnunk, hogy a modellnép hány százalékát kell kiválasztani, majd kisorsoljuk, hogy kik legyenek azok. Talán leegyszerűsítően tudománytalanul hangzik, hogy személyként hivatkozunk a modell népességének tagjaira, de ezzel jobban érzékelhető, hogy minden egyes triviális folyamat modellszemélyeken hajtódik végre. Ha a modellnép 60 éves tagjainak 3 százaléka meghal az adott évben, akkor ki kell választani, hogy mely 60 évesek azonosítóját töröljük.

Amennyiben a kiválasztáshoz életkor-specifikus valószínűségeket is ismerünk, akkor beszélünk igazításról *(alignment)*. A MIDAS modellt futtató LIAM2 program (lásd *De Menten és szerzőtársai* [2014]) lehetővé teszi, hogy az igazításhoz az életkor mellett még egy másik jellemzőt is felhasználjunk. Például ha tudjuk, hogy az egyes kohorszokban iskolai végzettség szerint hány embert kell kiválasztani, akkor egy kétdimenziós táblához is igazíthatjuk a véletlen kiválasztást.

A lehetséges két igazítási dimenzió révén már elértük, hogy a különböző kockázatú csoportokból eltérő valószínűséggel válasszunk. Bizonyos esetekben ennél is több információ áll rendelkezésünkre. A LIAM2 modellkörnyezetben ezek felhasználására is van lehetőség, mivel a program a kiegészítő adatokból logisztikus regresszióval előállított logit egyenleteket is képes felhasználni az igazításban.

Nézzük ezt végig a szülés példáján! A szimulációban minden évre ki kell választani, hogy kinek fog gyereke születni. Ha semmilyen statisztikai információval nem rendelkeznénk, akkor is tudjuk, hogy csak nők szülnek, bizonyos életkori határok között. Ha tudjuk a születések számát, akkor azt elosztjuk az összes szülőképes korú nőre, és máris van egy egyszerű igazításunk, amelyben a férfiak és a szülőképes koron kívüli nők nulla valószínűséggel szerepelnek, a szülőképes korú nők minden egyes kohorsza egységes nem nulla valószínűséggel. Mivel vannak általános demográfiai ismereteink a szülő nők életkoráról, a szülési valószínűségeket ennek megfelelően tudjuk kiosztani. Minden egyes női kohorszról meg tudjuk mondani, hogy tagjai mekkora valószínűséggel szülnek. Ha lenne olyan felmérés, aminek révén a szülés valószínűségét hozzá tudnánk kapcsolni további, a szimulációban és a felmérésben is rendelkezésre álló változókhoz, akkor logisztikus regresszióval előállíthatnánk egy logit egyenletet, amit a szimulációban közvetlenül felhasználhatunk. A LIAM2 ezt úgy fogja értelmezni, hogy minden személy esetében behelyettesíti a logit egyenletbe a megfelelő változókat, kiszámolja a logit eredményét *(logit score)*, az e szerint sorba rendezett személyekből pedig kiválaszt annyit, amennyi az igazítás szerint az adott kohorszból kiválasztandó.

A másik felhasznált eljárás a *párosító algoritmus*.[5] A MIDAS modell magyar változata alapjául szolgáló, a belga *Federal Planning Bureau* által kifejlesztett MIDAS_BE párosító eljárás az életkorra épül. Az intuíció szerint egy kapcsolatkötésre kiválasztott nő annál nehezebben talál magának párt, minél távolabb esik életkora a

[5] A párosításra többféle eljárás is létezik, ezek összefoglalása és osztályozása *Zinn* [2012] cikkében található.

kapcsolatkötésre kiválasztott férfiak átlagéletkorától. A procedúra két lépésből áll. Első lépésként sorba rendezzük a nőket. Erre azért van szükség, mert a párosító eljárás kimenetelét – azt, hogy pontosan melyik nőhöz melyik férfit társítjuk – befolyásolja, hogy milyen sorrendben keresünk párt a nőknek. Az algoritmus másik része minden, az adott évben kapcsolatkötésre kiválasztott nőhöz kiszámolja az adott évben kapcsolatkötésre kiválasztott férfiak mindegyikének az adott nőhöz való kapcsolódási valószínűségét *(matching score),* és ennek alapján kiválasztja a legnagyobb valószínűséggel rendelkezőt. A kapcsolódási valószínűség szintén logisztikus regresszió révén előállított logit egyenletből származik. A kiválasztás során használthoz képest annyi a különbség, hogy itt nem személyek, hanem párok szerepelnek az egyenlet egyik oldalán. Ennek a regressziónak az előállításáról részletesebben is írunk a megfelelő szakaszban. Mivel előzőleg a nőket sorba rendeztük, könnyen belátható, hogy nem minden nő a legnagyobb kapcsolódási valószínűséggel rendelkező férfipárral kerül kapcsolatba. Minél közelebb van egy nő életkora a férfiak átlagéletkorához, annál valószínűbb, hogy a számára legmagasabb kapcsolódási valószínűséggel rendelkező férfi egy másik nővel már kapcsolatba került.

*Almodulok*

EGYÜTTÉLÉS • A házas- vagy élettársi kapcsolatban élőket egyszerre párosítjuk össze, és az így létrejött párokból utólag választjuk ki azokat, amelyeket házassággá alakítunk. Ennek az az oka, hogy a párosítás módja megegyezik, és egyszerre úgyis csak az egyikféle kapcsolatba léphet be valaki. A kiválasztáshoz az életkoron kívül más jellemzőt nem használtunk, a párosításhoz viszont szükség van igazításra, mert a kapcsolatkötési valószínűségek korcsoportonként jelentősen eltérnek.

HÁZASSÁG • A házasságokat a párosító algoritmus által létrehozott együtt élő párokból választjuk ki logit igazítással. Itt tehát a párosító algoritmust már nem használjuk, a házasság az élettársi kapcsolatok egy részéből jön létre. Mivel a párokhoz – adatforrás híján – nem tudunk korosztályi adatot rendelni, itt egyetlen számhoz, az adott évben megkötött házasságok számához igazítunk.

VÁLÁS • A válás szimulációja egyszerű logit igazítás. A meglévő házasságok egy meghatározott részét megszüntetjük. A logit egyenletben a párok mindkét tagjának tulajdonságai szerepelnek, az igazítás pedig a házassághoz hasonlóan nem korosztályi arányokhoz, hanem a válások éves számához történik.

SZÉTKÖLTÖZÉS • Az élettársi kapcsolatok megszűnését nevezzük így. A váláshoz hasonlóan kezeljük. A fő különbség, hogy a válások számáról hatósági adatokkal rendelkezünk, a szétköltözéshez csak kérdőíves felmérésből származó adataink vannak.

MEGÖZVEGYÜLÉS • A megözvegyülés nem jelent külön számítást a szimulációban. Aki meghal, annak a partnere özvegy lesz, és a partnerkapcsolatot töröljük.

KIREPÜLÉS *(get-a-life)* • A fiatal felnőttek önálló háztartást alapítanak. Ezt az almodult a MIDAS modell belga változata nyomán egyszerű feltételvizsgálattal szimuláljuk. Ha valaki eléri a 24 éves korhatárt, és a szüleivel közös háztartásban él, akkor leválasztjuk onnan, és új, egyszemélyes háztartást generálunk számára.

A kiinduló állapotot szimuláló retrospektív modulra az alábbiakban külön kitérünk.

## A szimulációs eljárás

Az egyszerűség kedvéért először a szimulációban használt párosító modult tárgyaljuk részletesen. A retrospektív párosítást ugyan előbb kell lefuttatni, de a házasságkötések modellezésénél szerzett tapasztalatokat a retrospektív modulnál is használtuk, így érthetőbb lesz a magyarázat. Mindenekelőtt azonban bemutatjuk az adatokat, amelyeket a számítások során használtunk.

### *A párosítandók kiválasztása*

A párosítandók kiválasztása úgy történik, hogy az alapsokaságra (a 18 és 75 év közötti nők és férfiak közül azok, akik még egyedülállóak, vagyis elérhetők) a fentebb leírt módon végigterítünk egy valószínűségi korprofilt, vagyis ismert valószínűségekhez igazítunk. Megjegyezzük, hogy itt az adminisztratív statisztika nem használható. Egyszerre kívánjuk ugyanis létrehozni a házas- és élettársi kapcsolatokat, és egyrészt az utóbbiakról nem rendelkezünk teljes körű adminisztratív adattal, másrészt nem a teljes népességhez, hanem az elérhető, tehát házas- vagy élettársi kapcsolatban nem lévő személyekhez arányosítunk.

A házassági piac esetében a limitek meghatározásának elsődleges adatforrása a népszámlálás. Jelen munkához közvetlenül a népszámlálási adatbázisból kérdeztük le a megfelelő kereszttáblákat. Fontos, hogy nem a párkapcsolatban élők, hanem az egyik évről a másikra párt találók arányára van szükségünk.

Mivel a párosító regresszióban használt adatforrás (lásd lejjebb) csak 18 és 75 év közöttieket vizsgál, erre az alapsokaságra számoljuk ki a gyakoriságokat, és a szimulációban is csak ezekkel a korosztályokkal számolunk.

Az így előállított koréves relatív gyakoriságok szerepelnek a szimulációban az igazítás alapjaként. A szimulációban ez azt fogja jelenteni, hogy egy korosztályból csak az adott valószínűséggel kerül be valaki a párosítandók közé.

### *A párosító regresszió*

ADATOK • Az optimális adatállomány egy hosszú időszakot lefedő panelállomány, amelyben kellő esetszámunk van a kapcsolatformálódásra és a kapcsolatok felbomlására, továbbá amely tartalmazza a kérdezett személyek korábbi családi állapotát is, és vizsgálható, hogy kik voltak az előző évben a családi állapotuk szerint

egyedülállók vagy különélők. Ilyen adatállomány híján keresztmetszeti adatokat tartalmazó állományból is tudunk dolgozni, amely kapcsolattörténeti információt tartalmaz.

A modellezéshez két adatállományt is használtunk. A módszer kifejlesztésében és tesztelésében a *Generation and Gender Survey* (GGS) első és második hullámának magyar adataira támaszkodtunk. A GGS (http://www.ggp-i.org) egy 19 ország kutatócsoportjait koordináló program panel-adatfelvétele. Az adatállomány, sok egyéb mellett, részletes kapcsolattörténeti információt tartalmaz a válaszolókról. A megkérdezettek között nincsenek párok, az együtt élők nem mintatagokkal élnek párban, de a házas- vagy élettársi kapcsolatban élők esetében a felvétel kitért a megkérdezett partnerének egyes adataira is. A magyar adatállomány első hulláma 16 363 egyén, azon belül 10 363 házas- vagy élettársi kapcsolatban élő személy és 6000 egyedülálló adatait tartalmazza. Bár a GGS panelállomány, a rendelkezésünkre álló idősor egyelőre túl rövid, és nem szolgáltat kellő esetszámot, ezért jelenleg egy év keresztmetszeti adataiból dolgozunk. Ez a minta teljes egészében a rendelkezésünkre állt, a modell tesztelése, a változók kipróbálása ennél fogva sokkal egyszerűbb volt. Később részletezett okokból ez a minta kicsinek bizonyult, ezért a magyar MIDAS modellbe bekerülő egyenleteket népszámlálási adatállományra illesztettük.

A REGRESSZIÓS MODELL • A párosító algoritmus futtatásához egy olyan modell szükséges, amely a megfigyelhető szociológiai jellemzők alapján kiszámolja a lehetséges párokhoz tartozó kapcsolatkötési valószínűségeket. Az első lépés olyan kiinduló adatbázis létrehozása, amelyben a potenciális és a ténylegesen létrejött párok is szerepelnek. A potenciális párok az elvileg elérhető személyekből (egyedülállók, elváltak, özvegyek stb.) állnak. A GGS-ből ismerjük a párkapcsolatban élők jelenlegi kapcsolatának kezdő évét (a házasságkötés, illetve összeköltözés évét), így tudjuk, hogy mely párok azok, amelyek kevesebb mint egy éve formálódtak. Ehhez a regresszióhoz tehát nem szűrjük elő a potenciális házasodókat, mint ahogy a szimulációs modellben teszszük majd, hanem fordítva következtetünk: ha a kapcsolat egyéves, akkor feltételezzük, hogy a benne részt vevők előtte elérhetők voltak.

Ezután az ezekben a párokban szereplő személyekből létrehozzuk az összes lehetséges férfi–nő rendezett párt a megfelelő demográfiai jellemzőkkel oly módon, hogy a listában megjelöljük a ténylegesen létrejött párokat. Ez lesz az az adatsor, amelyen a logisztikus regressziót futtatni fogjuk, a célváltozó pedig az a változó, amely 1 értéket vesz fel a ténylegesen létrejött párkapcsolatok esetében, és nullát az összes többi potenciális párnál.

Itt ki kell térni arra, hogy a GGS adataiból ezt az adatbázist nem egyértelmű előállítani. A kérdőívben az aktuális partnerre vonatkozó adatokat nehéz kinyerni, ha az illetőnek több házassága is volt. Ha az aktuális partnerre vonatkozó adatok már rendelkezésre állnak, a férfiakat és a nőket külön kezelve kell szétválasztani a párokat, és újrarendezett (nő–férfi) párokba rakni őket minden lehetséges kombinációban, valamint létre kell hozni azt a változót, ami a tényleges és a potenciális kapcsolatokat megkülönbözteti egymástól.

Az adatfelvétel a nemekre nézve szimmetrikus volt, tehát ugyanazokat a kérdéseket tették fel minden válaszadónak.[6] Magyarázó változóként csak azok jönnek szóba, amelyek a megkérdezett személyről és partneréről is rendelkezésre állnak. Jelenleg ezek az adatok az életkor, az iskolai végzettség és a munkaerő-piaci státus. Mivel a megkérdezettek 37 százaléka nem tudta, vagy nem akarta megmondani a házas- vagy élettársa iskolai végzettségét, még ez az információ is hiányosan áll rendelkezésre.

Ennek az adatsornak tehát minden rekordja tartalmazza egy bizonyos férfi és egy bizonyos nő életkorát, iskolai végzettségét és munkaerő-piaci státusát, valamint azt a logikai változót, ami megmondja, hogy ez a bizonyos két személy tényleges pár, vagy sem. Ezekből az adatokból még létre kell hozni azokat a változókat, amelyek a logisztikus regresszió magyarázó változói lesznek. Fontos, hogy a változóknak nem az egyénekre kell vonatkozniuk, hanem a párokra. Az egyénekről viszonylag pontosan tudható, hogy hány éves korukban házasodnak a legnagyobb valószínűséggel, de ez az információ nem segít abban, hogy a teljes házassági piacot egy modellel modellezzük. Ezeknek a változóknak nem a párkapcsolatokról meglévő tudásunk leképezésére kell alkalmasnak lenniük, egészen más elvárásaink vannak velük kapcsolatban. Az a fontos, hogy a szimulációban könnyen előállíthatók legyenek, és használhatók legyenek a logisztikus regresszióban magyarázó változóként. Az előállt modell persze könnyebben ellenőrizhető, ha a változók egyszerűen értelmezhetők, ezért erre is törekedtünk. A magyarázó változók nagy száma a logisztikus regresszió esetében nem jelent gondot: ha nincs túl erős lineáris összefüggés, akkor az algoritmus kiszűri azokat a változókat, amelyek nem játszanak szerepet.

Az elkészített változók némi magyarázatot igényelnek. Az egyetlen numerikus, arányskálán mérhető adat az életkor. Mivel a rendezett párok tulajdonságairól van szó, ez valójában két adat: a férfi és a nő életkora. Ezek közvetlenül, illetve transzformációk után a logisztikus regresszió számára felhasználhatók. Mivel az életkor hatása nem lineáris, és nem is feltétlenül monoton, az életkorok második és harmadik hatványa egyaránt lehetséges változó. Hasonlóan egyszerű numerikus jelölt az életkorok különbsége, ami szintén szerepel második és harmadik hatványon.

Az iskolai végzettség és a munkaerő-piaci státus különbségét többféleképpen is próbáltuk megragadni. Ezek nem mérhetők arányskálán. Az iskolai végzettségből képzett ordinális változó esetében még talán van értelme arról beszélni, hogy a házas felek között két „lépcsőnyi" különbség van (tehát például egyikük iskolázatlan, a másik pedig középiskolai végzettséggel rendelkezik), a munkaerő-piaci státusokat kevésbé intuitív így sorba rakni. Mivel a logisztikus regresszió számára a kategoriális változókat vakváltozókká alakítjuk, azt a megoldást választottuk, hogy az összes

[6] Abban az esetben, ha a felmérés például csak nőket tartalmaz, és a partnerekről szóló információ korlátozott, a nőkről más adatok mennek be a regresszióba, mint a férfiakról. Így, hogy szimmetrikus, vagyis férfiakat és nőket egyaránt kérdeztek, választhatunk: vagy csak a nőktől lekérdezett adatokat használjuk, és akkor róluk többféle adatot is használhatunk, vagy használjuk a férfiaktól felvett kérdőíveket is, de akkor csak olyan változókat használhatunk, amelyek a megkérdezettről és a partnerről egyaránt rendelkezésre állnak.

lehetséges párosításból képezünk vakváltozókat.[7] E szerint tehát az FMEDU43 *(Female-Male-education)* változó értéke 1, ha az adott rendezett pár nőtagja érettségizett (4) a férfi pedig szakmunkás végzettséggel rendelkezik (3), minden más esetben 0. A változók jelentését a *Függelékben* részletezzük.

A bemenő adatsor az összes potenciális férfi–nő párt tartalmazza, a célváltozó pedig azokban az esetekben kap 1 értéket, amikor a pár ténylegesen létezik. Mivel a potenciális párok száma lényegesen nagyobb, mint a ténylegesen létrejövőké, a célváltozó egyik értéke teljesen eluralja az adatsort. Erre a problémára az egyik lehetséges megoldás az lenne, hogy lehetőleg minél kisebb mintán dolgozunk. Mivel a GGS mintavételen alapult, nem akartuk tovább növelni a mintavételi hibát, a mintát nem állt módunkban csökkenteni.

Ilyen esetekben a kiválasztott regressziós egyenlet validálása nem magától értetődő. A pozitív esetek aránya annyira kicsi, hogy mindenféle modell nélkül is nagyon jó találati arányt érhetünk el, ha senkit nem párosítunk senkivel. Az optimális modell klaszszifikációs táblája is ennek megfelelő: egyetlen potenciális párt sem házasít össze, tehát az illeszkedés jóságát vizsgáló Hosmer–Lemeshow-próba sem használható.

A szimulációhoz használt egyenletekkel kapcsolatban azonban nincsenek olyan elvárásaink, mint egy magyarázó modellel kapcsolatban. A mikroszimuláció céljai számára nem a valóság egy szeletét kell megmagyaráznunk, amihez a modell megfelelő szintű illeszkedése elengedhetetlen volna. Feladatunk egyszerűbb, mint egy magyarázat készítése és empirikus igazolás: olyan egyenletet kell felírnunk, amelynek értéke szerint a potenciális párokat sorba rendezve a tényleges párok előrébb szerepelnek, mint a nem párok.

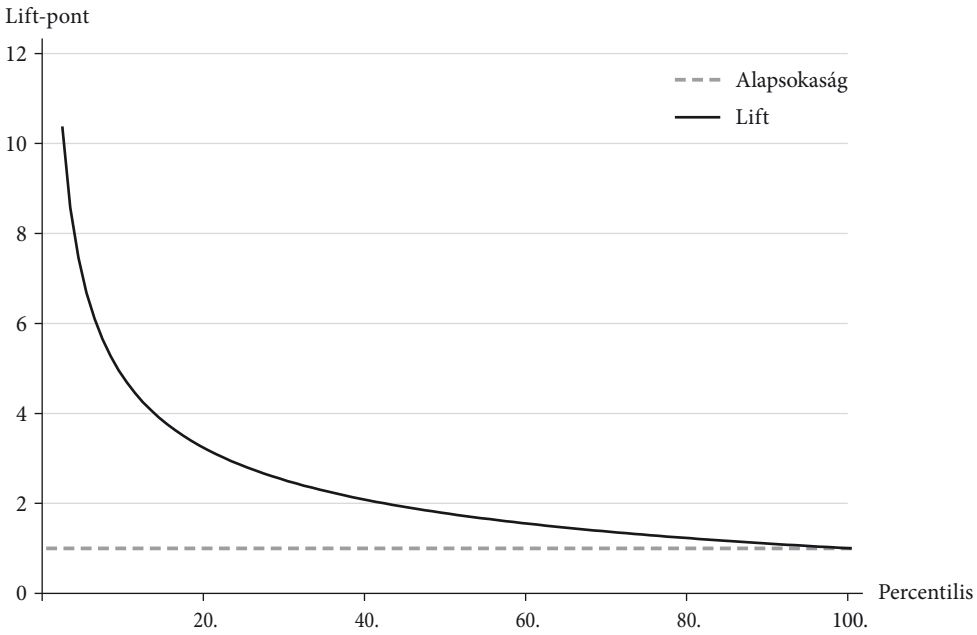
A modell fejlesztéséhez, a különböző opciók kipróbálásához inkább az adatbányászatban használatos keresztellenőrzés alapú mérőszámok használhatók (Lift, AUC stb.).[8] Ismét hangsúlyozzuk, hogy nem feltétlenül kívánjuk a függő és a független változók közötti összefüggést feltárni. A modell által megmagyarázott varianciát modellezzük, a többiről pedig a véletlen gondoskodik, a párosító algoritmus ezt magától elvégzi, egyszerűen a modell által előresorolt potenciális párok között nagyobb a tényleges párok előfordulásának a valószínűsége, mint a hátrasoroltaknál. A valóságot teljes egészében nem tudjuk reprodukálni, és mindegy, hogy a modell tökéletlensége vagy a hiányzó adatok miatt nem tudjuk. Ilyen értelemben a modell elfogadását vagy elvetését sem alapozhatjuk a szokásos mérőszámokra, csak egyfajta praktikus mérlegelésre, hogy a modell által megmagyarázott variancia indokolja-e, hogy beletegyük a szimulációba, vagy hasonlóan jó eredményt adna egy megfelelően paraméterezett véletlenszám-generátor.

[7] Ezzel a megoldással egyelőre lemondtunk a közvetett hatások vizsgálatáról (tehát nem vizsgáljuk külön, hogy – mondjuk – a 40 feletti diplomás férfiak mennyivel valószínűbben házasodnak diplomás nőkkel, mint a 30 év alattiak). A közvetett hatások vizsgálatához a változókat más algoritmusokkal kell előszűrni (például döntési fa vagy Bayes-háló).

[8] A keresztellenőrzés-alapú mérőszámok kiszámításához az adathalmazt véletlenszerűen két csoportra osztjuk, az egyik részére illesztjük a modellt, a másik részén pedig teszteljük: ellenőrizzük, hogy a modell előrejelzése mennyire volt pontos. Ezt akár sokszor egymás után is elvégezhetjük, és a mérőszámokat átlagolhatjuk, a véletlen csoportosítás miatt mindig más lesz a tanító és a tesztadatbázis.

Szemléltetésül bemutatunk egy Lift-görbét. A lift-görbe azt mutatja meg, hogy a *score* szerint sorba rendezett sokaság kvantiliseiben mennyivel nagyobb valószínűséggel fordul elő a célváltozó keresett értéke (esetünkben az, hogy a potenciális pár tényleges párt alkot), mint az alapsokaságban. Az *1. ábrán* az 5 és 10 év közötti házas- és élettársi kapcsolatok regressziójának Lift-görbéje látható. A vízszintes tengelyen a percentiliseket, a függőlegesen a relatív valószínűségeket tüntettük fel. Azt láthatjuk, hogy a *score* szerint sorba rendezett minta első 20 százalékában összesen körülbelül háromszor több házas található, mint a teljes sokaságban. Az adatbányász projekteknél gyakran a 30 százalékos pontot szokták vizsgálni, ami esetünkben 2,5. Ne felejtsük el azonban, hogy a potenciális párok és a tényleges párok közötti arányszám itt is nagyon magas, mindössze a lista legfelső 1 százalékát használjuk. A jelen példa esetében az 1 százalékos pont 11-nél van, tehát a felső 1 százalékban 11-szer nagyobb a párkapcsolat létrejöttének valószínűsége, mint a teljes sokaságban.

*1. ábra*
A GGS-minta (1. hullám) 5–10 éves együttélési kapcsolatainak Lift-görbéje




*Forrás:* saját számítás.

*Szétköltözés és válás, megözvegyülés és kirepülés*

Ezeknek az eseményeknek a modellezését nem tárgyaljuk részletesen, mert algoritmikus szempontból roppant egyszerűek, és nagyon hasonlítanak a korábban leírt események modellezéséhez. A válás és a szétköltözés egyszerű, életkor alapú igazítás, a rendelkezésre álló adatokból nem tudtunk olyan logisztikus regressziót előállítani,

ami segített volna a válásra ítélt kapcsolatok kiválasztásában. Az özvegyülés nem jelent mást, mint a halálozás modellezésének végigvezetését a háztartáson. A kirepülés egyszerű feltételvizsgálat: aki eléri a megfelelő életkort (24 év), annak önálló, egyszemélyes háztartást hozunk létre.

## A kiinduló adatállomány háztartásokká alakítása (retrospektív háztartás-formálódás)

Miként a bevezetésben már jeleztük, a retrospektív háztartás-formálódási almodulra azért van szükségünk, mert a szimuláció során felhasznált, adminisztratív adatokat tartalmazó adatállományunk egyéneket tartalmaz, háztartásokat nem. Egyrészt az adatgyűjtés nem terjed ki egy sor háztartási információra, így az egyének családi állapotára, kapcsolataik időtartamára, gyermekeik számára, és még egy sor jellemzőre. Másrészt még ha ezek az adatok ismertek is lennének egyéni szinten, nem tudnánk, hogy az adatállományban kire vonatkoznak. Ehhez ugyanis nemcsak a családi hátteret kell ismernünk, de össze is kell tudnunk kapcsolni a háztartástagokat. E nélkül nem lehetséges a hozzátartozói és a gyermekszámtól függő ellátások szimulációja.

A retrospektív háztartás-formálódási almodul annyiban különbözik a szimulációhoz felhasznált, a fentiekben bemutatott háztartás-formálódástól, hogy nem az egy évben létrejött kapcsolatokat kell reprodukálnia, hanem a szimuláció kiinduló pillanatában együtt élőket kell kiválasztania.

A retrospektív háztartás-formálódás rekonstrukciója során nagyon heterogén sokaságot kell modelleznünk. Az utóbbi 50 évben megváltoztak a házasságok és élettársi kapcsolatok létrejöttének okai és körülményei. Ezeket a kapcsolatokat vagy egy modellben kell tudnunk megragadni, vagy olyan csoportokra kell felbontanunk az adatállományt, amelyek között nincs átjárás, tehát amelyeken belül az emberek elérhetők a párosító algoritmus számára, a csoporton kívüliek viszont elérhetetlenek.

A vizsgált sokaság ilyenfajta szétdarabolása viszont nem magától értetődő. Olyan jellemző mentén kell elvágni a mintát, amely a párokat jellemzi, nem az egyéneket, mert a külön csoportban lévő egyéneket a párosító algoritmus nem tudja összepárosítani, tehát egymás számára elérhetetlenek lesznek. Mivel eleve az volt a sejtésünk, hogy a különböző történelmi korokban létrejött párkapcsolatok szociológiai jellemzőiket tekintve eltérnek egymástól, úgy döntöttünk, hogy a kapcsolat élettartama szerint csoportosítjuk a párokat. Ez megfelel céljainknak, hiszen a különböző élettartamú kapcsolatban élő emberek ténylegesen nem elérhetők egymás számára: aki egy 50 éve fennálló házasságban él, az nem élhet egy 10 éve fennálló házasságban.

Először is leválogattuk az 50 évnél hosszabb kapcsolatban élő párokat. A felhasználható változók szórása kisebb, mint a teljes sokaság esetében (például az életkor, korkülönbség, de akár az iskolai végzettség tekintetében is), nem volt meglepő, hogy a regressziós egyenlet magyarázó ereje csekély. A továbbiakban tízéves kapcsolat-életkori csoportokat hoztunk létre, vagyis a 40 és 50 év közötti hosszúságú kapcsolatokat válogattuk egybe stb. A 10 évnél fiatalabb kapcsolatokat további két részre bontottuk, az 5 évnél fiatalabb és 5 évnél régebbi kapcsolatokra.

A retrospektív szimuláció során az így előállított regressziós egyenleteket külön párosító eljárásokban használjuk fel. A sorrend logikailag tulajdonképpen mindegy lenne, de itt a gyorsabb futás érdekében kihasználjuk, hogy a régóta fennálló kapcsolatokban értelemszerűen idősebb emberek élnek. Aki egy 50 éve fennálló házasságban él, az legalább 68 éves. Ha az 50 éve fennálló házas- és élettársi kapcsolatokat szimuláljuk először, akkor az alapsokaságnak csak egy kis részén kell futtatnunk, a 68 év felettieken. Akiket a szimuláció ezen részében összepárosítottunk, azok a szimuláció következő lépése számára már nem lesznek elérhetők, mert már házasok, vagyis a következő lépésben párosítandó sokaságot sikerült csökkenteni. A következő lépésben a 40 és 50 év közötti élettartamú kapcsolatok csak azokat érintik, akik 58 év felettiek, és a szimuláció fordított időrendje szerint még nem házasok (tehát lehetnek 68 év felettiek is, de nem élhetnek 50 évnél régebbi kapcsolatban). Ezzel az egymást követő párosító eljárásoknak ugyan egyre bővebb életkori tartományból kell a párok számára az egyéneket kiválasztaniuk, de legalább a korábbi lépésekben már összepárosított emberekkel csökkenteni tudtuk a párosítandók számát.

Az egész retrospektív kapcsolatmodellezés alapja az, hogy a különböző élettartamú kapcsolatokat külön-külön kapcsolati piacokon hozzuk létre. Vagyis minden kapcsolatélettartam-csoporthoz külön igazítást, valamint kiválasztási és párosító egyenletet készítünk, és ezeket meghatározott sorrendben futtatjuk. A korábbiakban leírtak szerint 7 csoportot hoztunk létre: 0–5, 6–10, 11–20, 21–30, 31–40, 41–50 év tartamú, illetve 50 évnél hosszabb kapcsolatokból, és a legrégebbi kapcsolatokat hoztuk létre először, majd sorrendben az egyre fiatalabbakat.

*Hivatkozások*


Bental, B. [1989]: The old age security hypothesis and optimal population growth. Journal of Population Economics, Vol. 1. No. 4. 285–301. o. http://dx.doi.org/10.1007/bf00166069.

Botos Katalin–Botos József [2012]: A nyugdíjreform alapkérdései. Pénzügyi Szemle Online. http://www.penzugyiszemle.hu/vitaforum/a-nyugdijreform-alapkerdesei-5-egy-uj-magyar-nyugdijrendszer-alapjai.

Cigno, A.–Werding, M. [2007]: Children and Pensions. MIT Press, Cambridge, MA–London.

De Menten, G.–Dekkers, G.–Desmet, R.–Bryon, G.–Liègeois, Ph.–Wagener, R.–O'Donoghue, C. [2014]: LIAM2: a new open source development tool for the development of discrete-time dynamic microsimulation models. Journal of Artificial Societies and Social Simulation, Vol. 17. Vol. 3. http://jasss.soc.surrey.ac.uk/17/3/9.html.

Demény Pál [1986]: Pronatalist policies in low-fertility countries. Patterns, performance, and prospects. Megjelent: *Davis, K.–Bernstam, M. S.–Ricardo-Campbell, R.* (szerk.): Below-replacement fertility in industrial societies: Causes, consequences, policies. Cambridge University Press, Cambridge MA, 335–358. o.

Kovács Erzsébet (szerk.) [2013]: Nyugdíj és gyermekvállalás. Gondolat, Budapest.

Meier, V.–Werding, M. [2010]: Ageing and the welfare state: securing sustainability. Oxford Review of Economic Policy, Vol. 26. No. 4. 655–673. o. http://dx.doi.org/10.1093/oxrep/grq031.

Monticone Ch.–Ruzik, A.–Skiba, J. [2008]: Women's pension rights and survivors' benefits. ENEPRI Research Reports 53.
Sinn, H.-W. [2004]: The pay-as-you-go pension system as fertility insurance and an enforcement device. Journal of Public Economics, Vol. 88. No. 7. 1335–1357. o. http://dx.doi.org/10.1016/s0047-2727(03)00015-x.
Zinn, S. [2012]: A Mate-Matching Algorithm for Continuous-Time Microsimulation Models. International Journal of Microsimulation, 5 (1. No. 31–51. o.)

# Függelék

## A népszámlálási adatbázis használata

A korábban bemutatott GGS adatbázis nagy segítségünkre volt a modellezés módszerének kidolgozásában, de bizonyos részmintákon nagyon kevés esetet tartalmazott, ezért a modellbe kerülő egyenleteket népszámlálási adatokra illesztettük. Ezek frissebbek is, több esetet is tartalmaztak, valamint kevesebb volt a hiányzó adat. Mivel a retrospektív házassági piacot kapcsolat-élettartam szerint szegmentáltuk, ezért az első lépés az, hogy a kapcsolat-élettartam szerinti csoportokban tényleges házas- és élettársi kapcsolatban élő párokat kell kinyernünk a népszámlálási adatbázisból, kapcsolatélettartam-kategóriák szerinti csoportosításban.

A népszámlálási adatbázisban az egyes személyeket öt mezőből álló összetett kulcs azonosítja: területi kulcs, számlálókörzet kulcs, címsorszám (háztartás), családsorszám, valamint a személy sorszáma a családon belül. Az első három mező alapján össze tudjuk kapcsolni az azonos háztartásban[9] élő személyeket. A párok kinyerése azonban nem magától értetődő feladat, ezért részletesen leírjuk a lépéseket.

Le szeretnénk válogatni minden egyes kapcsolatélettartam-csoportra 100-100 családot. Valójában először nőket választunk ki, első körben velük azonosítjuk a háztartásokat. Ehhez be kell töltenünk a népszámlálás személyi állományát, és a következő szűrőfeltételeket beállítani. Egyrészt egyelőre a nőket keressük (NEME = 2), másrészt olyan nőkre van szükségünk, akiknek a családjában a családtípus megjelölése szerint párkapcsolat van. Ezek a 2011-es népszámlálási adatbázisban az LCSTIP változó következő értékeivel azonosíthatók:

– házaspár gyermek nélkül (01),[10]
– házaspár csak nőtlen/hajadon gyermekkel (02),
– házaspár nőtlen/hajadon és nem nőtlen/hajadon gyermekkel (03),
– házaspár csak nem nőtlen/hajadon gyermekkel (04),
– élettársi kapcsolat gyermek nélkül (05),
– élettársi kapcsolat csak nőtlen/hajadon gyermekkel (06),

[9] Lehetne érvelni amellett, hogy a háztartáson belül a család azonosítójával kössük össze a párokat, de egyrészt a nem család háztartások száma elenyésző, másrészt a végső modellben használt háztartásfogalom nagyon szűk, maximum kétgenerációs családokat kezelünk.

[10] A népszámlálási adatbázisban minden változó numerikus értékeket vesz fel, de némelyiket ennek ellenére szövegként tárolták, erre a lekérdezések írásakor figyelemmel kell lenni.

– élettársi kapcsolat nőtlen/hajadon és nem nőtlen/hajadon gyermekkel (07),
– élettársi kapcsolat csak nem nőtlen/hajadon gyermekkel (08).

A családban betöltött szerep szerint is szűrni kell, hogy biztosan a párkapcsolatban élő nőt válasszuk ki a családból, ehhez az LCSFO változó következő értékeit használjuk:
– családfő, nő (2),
– családfő házastársa (3),
– családfő élettársa, nő (5).

Ezután kiválasztjuk a megfelelő kapcsolatélettartam-kategóriába tartozó nőket, és ebből a csoportból veszünk egy 100 fős mintát. A területi kulcs, a számlálókörzet kulcsa, valamint a címsorszám összefűzésével létrehozunk egy háztartás-azonosítót, majd az összes többi változót töröljük, és a háztartás-azonosítókat sorba rendezzük az összekapcsoláshoz.

A következő lépésben lekérdezzük a fent kiválasztott háztartás-azonosítóhoz tartozó személyeket, nőket és férfiakat egyaránt. A szűréshez ugyanígy felhasználjuk a nők esetében már használt változókat, pontosabban a LCSTIP változót ugyanazokkal az értékekkel, az LCSFO változót pedig az összes családfőre és a családfő házas- vagy élettársára mutató értékekkel:
– családfő, férfi (1),
– családfő, nő (2),
– családfő házastársa (3),
– családfő élettársa, férfi (4),
– családfő élettársa, nő (5).

Így leválogattunk 200 embert, akik a megfelelő tulajdonságú 100 család párkapcsolatban álló tagjai.

Ezek után létrehozzuk az összes lehetséges párt ezen a halmazon. Mivel ez elég összetett feladat, lépésről lépésre bemutatjuk a http://www.ats.ucla.edu/stat/spss/faq/all_possible_pairs.htm oldalon található leírás alapján. Az összes lehetséges párban az összes férfi–férfi és nő–nő pár is benne lesz, de ez csak az egyszerűbb kódolás miatt van így, később csak a férfi–nő párokkal fogunk dolgozni. Az összes lehetséges pár létrehozásához szükség van egy egyszerű sorszámra mint azonosítóra. Majd új változóként hozzáadjuk az 1 értéket, és a teljes halmaz elemszámát ($n = 200$) minden sorhoz. Ezekre a segédváltozókra azért van szükség, mert az összes lehetséges párt 1-től $n$-ig futó ciklussal hozzuk létre. A ciklusban új változóként minden rekordhoz hozzáírjuk a ciklusváltozót (NID, 1-től $n$-ig), és kiírjuk az adatokat egy fájlba az *xsave* paranccsal. Az *xsave* nem írja felül a fájlt, hanem minden ciklusban hozzáfűzi az új sorokat. Ennek eredményeképpen a fájlban létre fog jönni egy olyan adathalmaz, amelyben az eredeti rekordok mindegyike 200-szor szerepel, és minden példány meg van számozva a NID változóval. Most már csak az dolgunk, hogy ezt a fájt sorba rendezzük NID szerint, majd újra összekapcsoljuk saját magával úgy, hogy az összekapcsoló azonosító ezúttal az előbb ciklusváltozóként beállított NID legyen, és az új változókat valami megkülönböztető elnevezéssel lássuk el. Így egy olyan fájlt kapunk,

amelyikben az eredeti mintánk 200 tagja egyenként 2 × 200-szer szerepel. A fájl szerkezetét az *F1. táblázat* szerint kell elképzelni.

*F1. táblázat*
Az összes lehetséges párt tartalmazó fájl szerkezete

| ID | NEME | LCSFO | HH-ID | ... | F-ID | F-NEME | F-LCSFO | F-HH-ID | F-... |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | | | 1 | 1 | 5 | | |
| 2 | 2 | 3 | | | 1 | 1 | 5 | | |
| 3 | 1 | 4 | | | 1 | 1 | 5 | | |
| ... | | | | | ... | | | | |
| 1 | 1 | 2 | | | 2 | 2 | 2 | | |
| 2 | 2 | 3 | | | 2 | 2 | 2 | | |
| 3 | 1 | 4 | | | 2 | 2 | 2 | | |
| ... | | | | | ... | | | | |

A 2 × 200 előfordulás tehát azt jelenti, hogy egy személy az eredeti változónevekkel is szerepel 200-szor, és az F előtagú változónevekkel is szerepel 200-szor, az összes lehetséges párosításban.

Mint említettük, ebben a 40 ezer sorban az összes férfi–férfi és nő–nő pár is szerepel, sőt mindenki saját magával is össze van kapcsolva. Ha most kiválasztjuk a 40 ezer sorból azt a tízezret, amelyben az eredeti változónevekkel nők szerepelnek (NEME = 2), a F előtagú változónevekkel pedig férfiak (F-NEME = 1), akkor az az összes lehetséges férfi–nő párt adja ki az eredeti mintahalmazon.

Azokat a párokat, amelyek ténylegesen léteznek, megjelöljük, és ez lesz a kiszámolandó logisztikus regresszió célváltozója.

Ezt az eljárást az összes kapcsolatélettartam-csoportra újra végre kell hajtani, így végül lesz 7 külön regressziós egyenletünk, amit a LIAM2 modellkörnyezetben a párosító algoritmus argumentumaként használni tudunk.