Grafikus modellek társadalomtudományi alkalmazása mobilitási adatokon

PhD értekezés

Németh Renáta

Témavezető:
Rudas Tamás, DSc
Statisztika Tanszék

Szociológia Doktori Program
Eötvös Loránd Tudományegyetem
Társadalomtudományi Kar

Budapest, 2009

Köszönetnyilvánítás

A disszertáció elkészítésében nyújtott segítségért témavezetőmön, Rudas Tamáson kívül köszönet illeti Wicher Bergsmát; az elmúlt években velük végzett közös munka képezi az alapját az itt bemutatott új eredményeknek. Róbert Péter a mobilitás-kutatási alkalmazások megvalósulását kísérte tanácsaival szintén éveken át. Szeretnék továbbá köszönetet mondani egy-egy probléma megoldásában való részvételéért Thomas Richardsonnak, aki a láncgráf modellek kurrens megközelítéseinek összevetésében segített, David Firthnek, aki a mobilitás-kutatás egyik alternatív eszközével, a unidiff modellel való összehasonlításban működött közre, végül Jonathan Kelleynek, aki rendelkezésemre bocsátotta Magyarországon beszerezhetetlen publikációját és annak recenzióit.

TARTALOMJEGYZÉK

1 Bevezetés ... 6
1.1 Motiváció ... 6
1.2 A kutatás fókusza és az értekezés felépítése ... 8
2 Marginális loglineáris modellek ... 12
2.1 Hagyományos loglineáris modellek ... 12
2.2 Marginális modellek, marginális loglineáris modellek ... 18
2.3 A paraméterezés jó tulajdonságai ... 24
2.3.1 Variációs függetlenség ... 24
2.3.2 A modellek létezése ... 28
2.3.3 Simaság ... 29
2.3.4 A becslések standard aszimptotikus tulajdonságai ... 33
3 Grafikus modellek ... 35
3.1 Háttér ... 35
3.2 Gráfelméleti definíciók és jelölések ... 36
3.3 Feltételes függetlenség ... 42
3.4 Irányítatlan gráfok ... 44
3.5 Irányított körmentes gráfok ... 47
3.6 Láncgráfok ... 48
3.6.1 Bevezető ... 48
3.6.2 Páronkénti Markov-tulajdonságok ... 50
3.6.3 Lokális Markov-tulajdonságok ... 54
3.6.4 Blokk-rekurzív Markov-tulajdonságok ... 56
3.6.5 Globális Markov-tulajdonságok ... 59
3.7 A gráfok és a feltételes függetlenségek megfeleltetésének problémái: ekvivalens gráfok, ekvivalens Markov-tulajdonságok ... 61
3.8 Szociológiai példák grafikus modellekre ... 65
4 Grafikus modell marginális modellként kategoriális adatokon ... 72
4.1 Egy előzmény: Goodman módosított útmodellje ... 72
4.2 Az általános megoldás: grafikus modellek marginális loglineáris modellekként ... 78
4.3 A feltételes függetlenségek és a paraméterek megfeleltetésének problémái ... 79
4.4 Irányítatlan gráfok ... 81
4.5 Irányított körmentes gráfok ... 82
4.6 Láncgráfok ... 85
4.7 Útmodellek ... 88
4.8 MLH becslés, illeszkedésvizsgálat ... 89
5 Nyitva maradt módszertani kérdések ... 92
6 Alkalmazás ... 94
6.1 A grafikus modellek általános alkalmazási problémái ... 94
6.1.1 A gráf felállítása ... 94
6.1.2 Konfigurációk, melyek klasszikus asszociációs problémákat fednek ... 95
6.1.3 Direkt és indirekt hatás ... 96
6.1.4 A grafikus modell és a strukturális egyenletek modell (SEM) összevetése ... 98
6.2 Mobilitás-kutatási alkalmazás: használati feltételek, alternatívák ... 99
6.2.1 Történeti áttekintés ... 99

6.2.2 A történeti áttekintés tanulságai ... 104
7 Saját alkalmazások ... 108
7.1 Klasszikus modell, friss adatok: Treiman hipotézise a magyar rendszerváltásban ... 108
7.1.1 Kutatási motivációk, kutatási kérdések ... 108
7.1.2 Elméleti háttér, korábbi vizsgálatok ... 109
7.1.3 Módszertan ... 112
7.1.4 Eredmények ... 114
7.1.5 Összegzés ... 116
7.2 A Treiman-modellnek és a Boguszak és társai (1990) által javasolt módosításának összevetése nemzetközi adatokon ... 118
7.3 Klasszikus adatok, új modell: Sewell és Shah (1968) vizsgálata a felsőfokú továbbtanulásról ... 122
7.4 Szerhasználat amerikai fiatalok körében ... 126
7.5 A demokráciával ill. az életszínvonallal való elégedettség kapcsolata, az MHP három hulláma alapján ... 130
8 Függelék: számítástechnikai megvalósítás ... 137
8.1 Modellbecslés ... 137
8.2 Illeszkedésteszt ... 143
8.3 Paraméterbecslés ... 143
9 Irodalom ... 145
10 Felhasznált adatbázisok ... 152
11 Ábrák jegyzéke ... 153
12 Táblázatok jegyzéke ... 154
13 A szerző publikációi ... 155
13.1 Folyóirat cikkek és könyvfejezetek ... 155
13.2 Konferencia kötetek ... 156
13.3 Konferencia kivonatok ... 156
13.4 Munkaanyagok, kutatási jelentések ... 157

Mottó:

„Forradalom ment végbe a társadalmi mobilitás kutatásában: az egykoron uralkodó Blau-Duncan paradigmát a loglineáris modellezés döntötte le a trónról. [...] A loglineáris forradalom figyelemre méltó kísérlet volt, és kezdetben úgy tűnt, fényes új jövőt kínál Blau-n és Duncan-en túllépve. [...] Végül, én úgy gondolom, Goldthorpe kísérlete elbukott. Azóta több mint egy évtized telt el. Számtalan modellt illesztettek, design-mátrixok seregeit vették számba, khi-négyzetek ezreit sorakoztatták fel. Mégis, kevés lényegi újdonság derült ki. Teljes részletességgel írhatók már le a szűk kereszttáblába szorított apai és fiúi foglalkozás cellagyakoriságai, ám az ehhez kapcsolódó értelmezés ad hoc, az elmélet szegényes. [...] Mindez visszalépést jelent a statisztikailag egyszerűbb, de koncepcionálisan összetettebb Blau-Duncan paradigmához képest."[1]

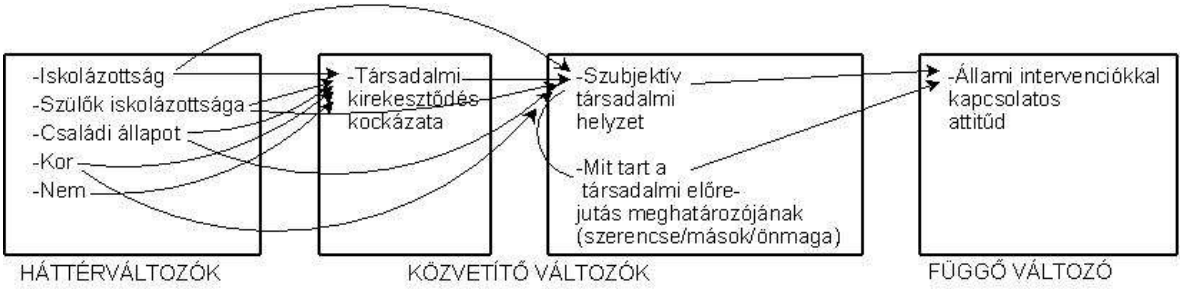
(Idézet Jonathan Kelley 1990-es *The failure of a paradigm* című, több kritikát provokált tanulmányából)

[1] A szerző fordítása.

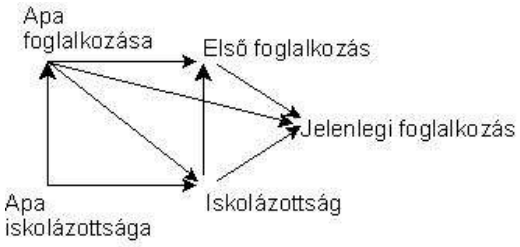
1 BEVEZETÉS

1.1 Motiváció

A mottóbeli idézet jól rávilágít a dolgozat tárgyát képező új statisztikai módszerek megjelenése előtti helyzetre, amikor Goldthorpe illetve Blau és Duncan megközelítése még nem volt egyszerre megvalósítható. Ezeknek az új módszereknek a birtokában azonban már nem zárja ki egymást a két megközelítés. Ugyanis a **grafikus modellek** kategoriális változókon értelmezettek, mint Goldthorpe loglineáris paradigmája, ugyanakkor a Kelley által preferált Blau-Duncan-paradigmához hasonlóan oksági láncok kezelésére, direkt és indirekt hatások megkülönböztetésére is alkalmasak. A grafikus modellek könnyen vizualizálható és jól interpretálható módon jelenítenek meg komplex rendszereket, és modulárisak, vagyis e komplex rendszereket egyszerűbb részek összekapcsolásával építik fel. Lehetőséget adnak például az alábbihoz hasonló bonyolult asszociációs struktúrák leírására (itt két pont között az él / nyíl hiánya mindig feltételes függetlenségi állításoknak feleltethető meg)[2]:



Az elméleti modell változtatása egyszerűen követhető a gráf módosításával. Például az alábbi klasszikus modellben[3]



megfontolandó az iskolázottság és a jelenlegi foglalkozás közötti egyirányú nyíl irányítatlan élre cserélése, hiszen ma a műszaki-technikai haladás révén már egy adott szakmában is állandósulhat a változás, amellyel az érintetteknek lépést kell tartaniuk, amit viszont csak az ismeretek folyamatos gyarapításával lehetséges. Vagy más példával: ha a modellt a rendszerváltás előtti kelet-európai országokra alkalmazzuk, akkor az apa iskolázottsága és a jelenlegi foglalkozás közé nyíl húzható, mivel a magas iskolázottságú apáktól származó, negatívan diszkriminált utódok

[2] A gráf forrása: Cox és Wermuth (2001).
[3] Blau-Duncan modell (Blau, Duncan, 1967).

munkába állásukat követően kulturális tőkéjük és érvényesülési törekvéseik révén gyakran mégis kedvező pozíciókat szereztek a munkaerőpiacon.[4]

A grafikus modellek statisztikai leírása még nem tekinthető minden elemében lezártnak, ez a terület jelenleg is igen aktív (Lauritzen, Wermuth, 1989, Whittaker, 1990, Frydenberg, 1990a, Andersson, Madigan, Perlman, 2001, Lauritzen, Richardson, 2002, Cox, Wermuth, 1996, Wermuth, Cox, 2004, Wermuth, 2005, Drton, 2008). Wicher Bergsma[5] és Rudas Tamás a kategoriális változókon értelmezett grafikus modellekkel kapcsolatban az utóbbi években publikált fontos eredményeket (2002, 2004). Én 2003-ban kapcsolódtam be a munkájukba, Rudas Tamást választva témavezetőmül. Eleinte önálló mobilitás-kutatási alkalmazásokat készítettem (Németh 2004, 2006a, 2006b, Németh, Rudas, Bergsma 2006), munkámat ebben az időben több éven át Róbert Péter is tanácsaival kísérte. Később a módszertani alapkutatásba is bekapcsolódtam (Rudas, Bergsma, Németh, 2006). Mint az értekezésből látható lesz, a grafikus modellekkel kapcsolatos fontos kérdések egy része már megoldódott ezeknek az utóbbi években született eredményeknek a révén. A legfontosabb kérdések egyike a modellek paraméterezésének megválasztása. Kihasználva, hogy jól használható általános feltételek születtek a modell marginális loglineáris paraméterekkel történő megfelelő paraméterezésére (Bergsma, Rudas, 2002), a grafikus modellek paraméterezése kézenfekvő módon interpretálható és fontos reguláris tulajdonságokkal bíró paraméterekkel oldható meg. Szerzőtársaimmal jelenleg befejezéséhez közeledő kutatásunk (Rudas, Bergsma, Németh, 2009) pedig reményeink szerint választ ad majd az értekezés végén továbbra is nyitva maradt kérdések közé sorolt problémák jelentős részére.

Kutatási témám megválasztásának motivációja tehát elsősorban az volt, hogy általa egy jelentős nemzetközi kutatási áramlatba kapcsolódhattam be. Vonzott a kutatási terület matematikai szépsége is: a valószínűségszámításon és statisztikán kívül olyan egymástól távol eső területek eredményei kerülnek itt együttes felhasználásra, mint a gráfelmélet, logika, algebra, geometria vagy topológia. Saját hozzájárulásként magam a módszer szociológiai, elsősorban mobilitás-kutatási alkalmazhatóságát szerettem volna megmutatni. A mobilitás-kutatáson belül, a státuszelérési folyamat különféle modelljeivel kapcsolatban valóban felmerülhet a grafikus modellek felhasználása. Ezek a megközelítések ugyanis tipikusan oksági folyamatmodellekkel követik a társadalmi dinamikát, kategoriális változókat használva – mint a származási osztálypozícióból az elért osztálypozícióba vezető út társadalmi mobilitási modelljei, vagy a származási osztálypozícióból a megszerzett iskolai végzettséghez vezető út egyenlőtlenség-kutatási modelljei.

A mobilitás-kutatás választásával, azt hiszem, nem szűkítem le túlságosan tárgyamat, hiszen a mobilitás-kutatás módszereinek bővülése általában szorosan követi az általános statisztikai módszerek fejlődését, sőt néha generálja is azt. Ez a kérdés ugyanakkor nem redukálható a statisztikai előnyök vizsgálatára, a grafikus modelleknek, más statisztikai módszerekhez hasonlóan, a koncepcionális előfeltételei is fontosak. E koncepcionális előfeltételeket a kutatás elméleti paradigmái képviselik - a mottó megválasztásával, Kelly idézésével is célom volt, hogy a módszernek a mobilitás-kutatás paradigmái között való elhelyezkedését felvillantsam. Goldthorpe elméleti megközelítése a társadalmi osztályokat kategoriális változóval operacionalizálja, a strukturális változásoktól független mobilitási esélyeket helyezve a

[4] Ez utóbbi Boguszak et al. (1990) a Blau-Duncan modell módosítására vonatkozó javaslata, az értekezésben részletesen is foglalkozom majd vele.
[5] Wicher Bergsma, holland szerzőtársam jelenleg a London School of Economics and Political Science oktatója.

középpontba, ezért volt szüksége a loglineáris módszerre. Blau és Duncan mobilitási csatornákat kívánt követni, ezért használt útmodelleket. A grafikus modellekkel mindkét cél megvalósítható. Az alábbi értekezésben a módszert koncepcionálisan pontosabban is pozícionálom majd, a bemutatott (többségében már publikált) alkalmazások pedig reményeim szerint azt bizonyítják, hogy a módszer beilleszthető a létező keretek közé, eredményei összevethetők más kutatásokéval.

1.2 A kutatás fókusza és az értekezés felépítése

A grafikus Markov-modellek (röviden: grafikus modellek) előnye, hogy segítségükkel gráffal reprezentálható (1) többlépcsős, összetett folyamatokat írhatunk le, (2) ahol egy él hiánya közvetlenül megfeleltethető egy feltételes függetlenségnek. A feltételes függetlenség tulajdonságaival a 3.3. fejezet foglalkozik. E fejezet véleményem szerint különösen érdekes, hiszen a feltételes függetlenség, mint logikai rendszer – fontos következményekkel járó - karakterizációját, illetve különféle, az információelméleti alkalmazások kiindulópontját megmutató általánosításait is tartalmazza.

A grafikus modellek közé az irányítatlan gráfokkal (3.4. fejezet), az irányított körmentes gráfokkal (3.5. fejezet), illetve a legáltalánosabb láncgráfokkal (3.6. fejezet) leírható modellek sorolhatók. Az első típus a változók között csak szimmetrikus, a második csak aszimmetrikus relációkat enged meg, a harmadik típus szimmetrikus és aszimmetrikus relációk megjelenítését is lehetővé teszi. Az irányítatlan gráfokra épülő grafikus modelleknek (másnéven a Markov-féle véletlen mezőknek) például térstatisztikai és képfeldolgozási felhasználásai ismertek. Az irányított körmentes gráfokat napjainkban elsősorban a mesterséges intelligenciakutatás szakértői rendszerekkel (*expert systems*) foglalkozó ága alkalmazza (pl. Lauritzen, Spiegelharter, 1988, Pearl, 1988), illetve ok-okozati viszonyok feltárásával kapcsolatos kutatásokban kerülnek elő gyakran (pl. Spirtes et al., 1993). Az első, illetve részben a második modell alapvető statisztikai tulajdonságai is tisztázódtak az utóbbi években, ezeket az eredményeket a disszertációban ismertetem (4.4. és 4.5. fejezet). A láncgráfok az első két típus matematikai általánosításai, de létrehozásukat nem csak elméleti, hanem gyakorlati igények is motiválták. Gyakori ugyanis az olyan alkalmazási kontextus, amikor bizonyos változócsoportok *között* aszimmetrikus kapcsolat van (longitudinális vizsgálat időpont szerint rendezett változói, vagy keresztmetszeti vizsgálat ok-okozati sorba rendezhető magyarázó / közbülső / kimeneti változóhalmazai), míg a változócsoportokon *belül* szimmetrikusak a relációk. A láncgráfok leírása még nem kanonizálódott, több kutató foglalkozik velük párhuzamosan, gyakran más-más megközelítésekben (Lauritzen, Wermuth, 1989, Whittaker, 1990, Frydenberg, 1990a, Andersson, Madigan, Perlman, 2001, Lauritzen, Richardson, 2002, Cox, Wermuth, 1996, Wermuth, Cox, 2004, Wermuth, 2005, Drton, 2008). A disszertációban az ő munkájukat is megkísérlem rendszerezve bemutatni (4.6. fejezet), kiemelve azt a szerzők által gyakran figyelembe nem vett tényt, hogy az eltérő megközelítések következményeként sem a leírható modellek köre, sem a leírt modellek statisztikai tulajdonságai nem kompatibilisek.

A kategoriális változókon értelmezett gráfmodellek bizonyos szempontból gazdagabb struktúrát adnak, mint a többváltozós normális eloszláson értelmezettek. Ennek a különbségnek az oka, hogy a páronkénti $A \perp B$ és $A \perp C$ feltételes függetlenségek csak az utóbbi esetben ekvivalensek az együttes függetlenséggel, $A \perp BC$-vel. Ezért van jelentősége kategoriális esetben több különböző szabály definiálásának a feltételes függetlenségek gráfból történő kiolvasására. Ha pl. a gráfban A nincs összekötve sem B-vel, sem C-vel akkor az egyik szabály

(un. páronkénti Markov-tulajdonság) szerint csak $A \perp B$ és $A \perp C$ teljesül, de egy másik szabály (a lokális Markov-tulajdonság szerint) $A \perp BC$ is igaz. Ez a két szabály normális eloszlás mellett nem különbözik egymástól.

A disszertáció módszertani alapkérdése a kontingenciatábla olyan paraméterekkel történő felparaméterezése, amelyekkel könnyű a grafikus modellek megadása. A felparaméterezés a táblát egy-egyértelműen meghatározó mutatók megadását jelenti. Pl. 2X2-es tábla esetén lehetséges paraméterezés a négy cellagyakoriság megadása, egy másik paraméterezés a mintanagyság, az első sormarginális, az első oszlopmarginális és az első cella a megadása; egy harmadik lehetőség az esélyhányados és három cella megadása. Mindhárom esetben leírható az a modell, amely a táblát definiáló két változó függetlenségét állítja, de a specifikálás az utolsó esetben a legkönnyebb, az esélyhányados 1-re történő állításával. Megkerülhetetlen kérdés tehát, hogy a rendelkezésre álló paraméterezések közül melyiket válasszuk a kontingenciatábla gyakoriságainak leírására. Ugyanis, (1) a grafikus modell ezeknek a paramétereknek bizonyos megszorításaival (általában néhány paraméter nullára állításával) adható meg. Továbbá (2) azokban a gyakori esetekben, amikor a cél egy jól illeszkedő gráf (azaz az ezáltal prezentált modell) élei (a modellben szereplő hatások) elhagyhatóságának a tesztelése, akkor ezt a paramétermegszorítások fokozatos feloldásával (*stepwise* - lépegetve) érhetjük el. Ugyanennek a az esetnek a megfelelője, amikor rosszul illeszkedő gráf javítása a cél, és bizonyos élek behúzásának hatását teszteljük a paramétermegszorítások fokozatos feloldásával. Végül (3) a modellhez tartozó eloszlás értelmezése a becsült paraméterek interpretációjával történik. Mindhárom esetben azon múlik az eredményünk, hogy a tábla paraméterezése milyen paramétereket tartalmaz. Nem megfelelő paraméterezés esetén előfordulhat, hogy (1) nem világos, a modellt hogyan, a paraméterek milyen megszorításával definiálhatjuk, (2) a lépegetős módszer egyes lépéseinél az él elhagyását/behúzását nem tudjuk, hogyan feleltethetjük meg a paraméterek megszorításának/feloldásának, (3) a modellhez tartozó eloszlás paraméterei nem a modellezett jelenség minket érdeklő jegyeihez tartoznak.

A tábla felparaméterezésének egyik lehetősége a **marginális loglineáris paraméterek** használata (a paraméterezést a 2. fejezetben részletezem). Ezek a hagyományos loglineáris paraméterektől abban térnek el, hogy nem csak az összes változó együttes eloszlása, hanem bármely marginális eloszlás mellett definiálhatunk paramétereket, így finomabban tudunk hatásokat modellezni velük.. Ezzel az általánosítással ugyanakkor a hagyományos loglineáris paraméterezésnél sokkal összetettebb keretet kapunk. Például az irodalomban többször előkerül az a probléma (mint Bergsma és Rudas (2002) felsorolja: Liang et al. (1992), Glonek és McCullagh (1995) és Kaumermann (1997)), hogy bizonyos paraméterek értékének előírása a többi paraméterre is hatással van, korlátozza lehetséges értékeik körét. Emiatt előfordulhat, hogy egy modell megadásánál olyan értékeket adunk paramétereinknek, amelyek nem egyeztethetők össze, így tudtunkon kívül üres, nem létező modellt definiálunk. A disszertációban ajánlott marginális loglineáris paraméterezések bizonyos kritériumok teljesülése esetén ebből a szempontból is előnyös tulajdonságokkal bírnak: nem fordulhat elő, hogy a paraméterértékek inkompatiblitása miatt a modell nem létezik, továbbá a paraméterek értékei egymástól függetlenül interpretálhatók. Ezen kívül a modellek illeszkedése valószínűség-hányados próbával tesztelhető, a modelleknek megfelelő maximum likelihood-becslés egyértelműen létezik és előállítható, és konfidencia-intervallum konstruálható hozzá. Ez utóbbi tulajdonságok is csak a disszertációban tárgyalt feltételek esetén teljesülnek, de általánosságban nem; pl. már olyan egyszerű modell esetén sem, amely két változó marginális és feltételes függetlenségét egyszerre

állítja („az egyetemi felvételi sikeressége független a felvételiző nemétől, összességében és szakonként is”).

Mint láttuk, a **grafikus modellek** paraméterezése általános esetben több problémát vet fel. E modellek **marginális loglineáris paraméterekkel** történő megadása egészen új megközelítés. Megfelelő paraméterezés esetén nem csak az előbbiekben felsorolt előnyös tulajdonságok biztosíthatók, hanem a grafikus modellek által támasztott fenti elvárások is teljesülnek, azaz az élek elhagyása egyszerűen megfeleltethető bizonyos paraméterek nullára állításának, és a modell paraméterei kézenfekvő módon interpretálhatók. A dolgozat fő módszertani eredményei a paraméterek megfelelő megválasztásával és interpretálásával kapcsolatosak.

Az irányított körmentes gráfok (4.5. fejezet), a belőlük levezetett útmodellek (4.7. fejezet), illetve a láncgráfok (4.6. fejezet) itt bemutatott paraméterezése jórészt szerzőtársaimmal elért eredményekre támaszkodik. Hogy ne csak a letisztult végeredményeket mutassam be, hanem mélyebb megértésük érdekében a hozzájuk vezető utat is, két önálló fejezetet szenteltem az elmúlt években szerzőtársaimmal végzett kutatásom jellegzetes, részben választ nyert, részben zsákutcában maradt problémáinak (3.7. és 4.3. fejezet). Ezt a két fejezetet különösen fontosnak tartom a még nyitva maradt, az 5. fejezetben felsorolt módszertani kérdések nehézségének és lehetséges megoldási útjainak a megismertetése szempontjából.

A strukturális egyenletek modell, amely az eredeti Blau-Duncan-féle útelemzés általánosítása, a grafikus modellhez sok szempontból hasonlít: mindketten gráffal szemléltetik az asszociációs struktúrát, a változók közötti oksági kapcsolatok útvonalai ugyanúgy követhetők, és a direkt és indirekt hatások is ugyanúgy megfoghatók mindkét esetben. Így bizonyos értelemben vetélytársaknak tekinthetők, ezért szántam önálló fejezetet összevetésüknek (6.1.4. fejezet). Hasonlóan, külön fejezetben foglalkozom (4.1. fejezet) a grafikus modellek előzményének tekinthető, ám sok kérdést nyitva hagyó Goodman-féle módosított útmodelleknek.

Az általános alkalmazást bemutató 6.1. fejezetben a hétköznapi gyakorlati praxis olyan konkrét problémáit veszem sorra, mint a gráf kialakítása, a gráf módosításának következményei, a paraméterek interpretációja. A modellillesztés és paraméterbecslés módját a 4.8. fejezetben, ezek (Mathematica és SPSS programjaimmal történő) számítástechnikai megvalósítását a 8. fejezetben mutatom be.

A grafikus modellek kvantitatív empirikus szociológiai alkalmazhatóságát az 6.2. fejezetben, a mobilitás-kutatáson belül vizsgálom. A módszer kontextusba helyezését a mobilitás-kutatási módszerek és fogalmak több évtizedes történetének szempontomból releváns pontjainak kiemelésével, majd a grafikus modellek alkalmazási feltételeinek és alternatíváinak ezek tükrében végzett vizsgálatával valósítom meg. A két legismertebb mobilitási gráfmodellt a 3.8. fejezetben elemzem, külön kitérve a gráf különböző szerzők által javasolt módosításainak következményeire is.

A 7. fejezetben néhány saját alkalmazást részletesebben is bemutatok, ezek egy része már publikált mobilitás-kutatási alkalmazás. E példákon keresztül szeretném megmutatni, hogy mobilitás-vizsgálatok kategoriális adataira azért lehet érdemes használni ezt a módszertant, mert ezen a területen gyakran alkalmaznak oksági modelleket (státuszelérési modellek), és a szakterület tipikus megközelítései statisztikailag megfogalmazhatóvá válnak marginális loglineáris paraméterek használatával. Már publikált kutatásaim összefoglalásánál arra a célra koncentráltam, hogy megmutassam, a grafikus modellek módszere beilleszthető az adott kutatási

hagyomány keretei közé, eredményei összevethetők más kutatásokéval. A nem publikált, módszertanilag bonyolultabb alkalmazásokban inkább a statisztikai munka hangsúlyos. Itt mutatok majd példát klasszikus szociológiai gráfmodellek friss adatokon történő illesztésére, a modellek módosításának tesztelésére, látunk majd példát klasszikus adatok új modellel történő magyarázatára, illetve paneladatokon történő alkalmazásra is.

Az eredményeket a társadalomtudományi alkalmazhatóságra koncentrálva ismertetem, a matematikai apparátust ezért megpróbálom a lehető legszűkebbre vonni, a megfogalmazást csak a szükséges mértékig formalizálni. A mélyebb matematikai háttértudást igénylő magyarázatokat csak lábjegyzetben közlöm, a tételek bizonyítását pedig kihagyom, vagy csak alapötletüket vázolom. Ha egy szakkifejezés magyar fordítása nem kanonizálódott még, illetve ha egy jelölés rövidítését az angol elnevezés ismerete segít értelmezni, akkor az angol eredetit is megadom.

2 MARGINÁLIS LOGLINEÁRIS MODELLEK

2.1 Hagyományos loglineáris modellek

A disszertáció tárgyát képező, kategoriális változókra alkalmazott grafikus modellek[6] (Rudas, Bergsma, 2004; Rudas, Bergsma, Németh, 2006ab) az utóbbi évtizedekben megjelent két statisztikai terület, a marginális modellek, illetve a grafikus modellek metszéspontján helyezkednek el. E két területet tárgyalom részletesebben a következőkben. Az alábbi fejezet tételei, definíciói és jelölései, hacsak nem jelzem másként, Lauritzen (1996), Rudas (1998), Bergsma és Rudas (2002) ill. Rudas, Bergsma, Németh (2009) munkájára támaszkodnak.

Legyen $V = \{V_1,...,V_p\}$ kategoriális változók egy halmaza, V_j értékeit az I_j, $1 \le j \le p$ nemüres véges halmazon veszi fel. A $T = \times_{j=1}^{p} I_j$ Descartes-szorzatot nevezzük p-dimenziós *kontingencia-tábla*nak, melynek egy *cellá*ja az $\mathbf{i} = (i_1,...,i_p), i_j \in I_j$. A nem-negatív valós $\mu(\mathbf{i})$ értéket nevezzük az **i** cellához tartozó *cellagyakoriság*nak.

1. példa A formalizmus könnyebb megértéséhez vegyünk egy példát, a későbbiekben (pl. a 7.1. fejezetben) szereplő társadalmi státusz-változókkal. Legyen a V változóhalmazunk háromelemű: V_1 iskolázottság, V_2 foglalkozási státusz, V_3 jövedelem. Értékkészlet-halmazuk: I_1 ={diplomás, diploma nélküli}, I_2 ={alsó-, közép-, felső osztály}, I_3={átlag alatt kereső, átlag felett kereső}. A kontingencia-tábla ennek a három halmaznak a kereszt-kategóriáiból, vagyis Descartes-szorzatából áll, cellái közé tartozik pl. az **i**=(diplomás, középosztálybeli, átlag alatt kereső) hármas. Az ehhez a cellához tartozó $\mu(\mathbf{i})$ az ezzel a három tulajdonsággal rendelkező egyének számát adja meg.

Legyen F a T-n értelmezett pozitív[7] μ gyakoriságí eloszlások osztálya. Az F osztály és a k hosszú valós vektorok közötti $\theta : F \to \mathbf{R}^k (k \ge 1)$ leképezést nevezzük F egy paraméterének, egy konkrét eloszlás mellett felvett értékét *paraméterérték*nek. Ilyen paraméter például a cellagyakoriságok összege, ilyenkor k=1, hiszen a valós számokra képez a leképezés. Néha kényelmesebb lesz vektorértékű paraméter helyett paraméterek halmazáról beszélni, ilyenkor az eredetileg vektorba rendezett paramétereket külön vesszük, és a paraméterérték helyett is paraméterértékekről beszélünk. *Paraméterezés*nek nevezzük mindazon θ_i paraméterek megadását, amik a tábla egy-egyértelmű megadását teszik lehetővé: $\theta=(\theta_1, ..., \theta_t)$. A θ leképezés ilyenkor invertálható. Pl. 2X2-es tábla esetén lehetséges paraméterezés a négy cellagyakoriság megadása, egy másik paraméterezés az összes megfigyelés számának, az első sormarginálisnak, az első oszlopmarginálisnak és az első cellának a megadása; egy harmadik lehetőség az esélyhányados és három cellagyakoriság megadása.

---

[6] A „grafikus modell" az irányítatlan gráfokkal reprezentálható hierarchikus loglineáris modellek egy osztályára is használatos elnevezés, a disszertációban (és sok más helyen, pl. Lauritzen, 1996-ban) használt másik értelmében a kifejezés általános, nem csak irányítatlan gráffal megadható modellekre vonatkozik, a pontos definíciót lásd később.

[7] Értsd: a gyakoriság T minden cellájában nagyobb 0-nál, ezt szigorúan pozitív eloszlásnak is nevezik. E pozitivitási feltevés fontossága a későbbiekben fog kiderülni.

Ha $M \subseteq V$ a változók egy részhalmaza, akkor $\mathbf{i}_M$ jelöli az $\mathbf{i}$-ből azon indexek vektorát, amelyek M-hez tartoznak, tehát $(.)_M$ projekciót definiál. A marginális táblát T_M jelöli. Azt mondjuk, hogy ilyenkor *marginalizál*juk a táblát az M-hez nem tartozó változók szerint.

A T_M celláihoz tartozó $\mu_M(\mathbf{i}_M)$ *marginális gyakoriságok* a $\mu(\mathbf{i})$ cellagyakoriságok megfelelő összegzésével áll elő:

$$\mu_M(\mathbf{i}_M) = \sum_{\mathbf{j} \in T : \mathbf{j}_M = \mathbf{i}_M} \mu(\mathbf{j}).$$

Az 1. példa visszatérve: ha a háromdimenziós kereszttáblát marginalizáljuk a jövedelem, V_3 szerint, akkor a $(.)_{V_1 V_2}$ projekciót alkalmazzuk. A marginális $T_{V_1 V_2}$ tábla kétdimenziós, egy cellagyakorisága így áll elő:

$$\mu_{V_1 V_2}(\text{diplomás, középosztálybeli}) = \sum_{\mathbf{j} \in T : \mathbf{j}_{V_1, V_2} = (\text{diplomás, középosztálybeli})} \mu(\mathbf{j}) =$$

$$= \mu(\text{diplomás, középosztálybeli, átlag alatt kereső}) + \mu(\text{diplomás, középosztálybeli, átlag felett kereső}).$$

Most már definiálhatjuk a loglineáris paramétereket. A V változók alkotta kontingencia-tábla celláihoz tartozó loglineáris paraméterek[8,9] rekurzív módon állíthatók elő:

$$\lambda_{\varnothing}^{V}(\mathbf{i}_{\varnothing}) = \frac{1}{|T|} \sum_{\mathbf{j} \in T} \log \mu(\mathbf{j}), \tag{1}$$

$$\lambda_{L}^{V}(\mathbf{i}_{L}) = \frac{1}{|T_{V \setminus L}|} \sum_{\mathbf{j} \in T : \mathbf{j}_L = \mathbf{i}_L} \log \mu(\mathbf{j}) - \sum_{L' \subset L} \lambda_{L'}^{V}(\mathbf{i}_{L'}). \tag{2}$$

Az alsó indexben szereplő $L \subseteq V$ halmazt nevezzük a paraméterhez tartozó *hatás*nak. Látható, hogy az üres halmazhoz tartozó hatás esetén a cellagyakoriságok logaritmusának átlaga adja a paramétert. A többi esetben a kifejtés első tagja szintén átlagot ad: az adott $\mathbf{i}$ cellával L-indexükben megegyező cellák loggyakoriságának átlagát. Példánkra visszatérve, a $\lambda_{V_1 V_3}^{V_1 V_2 V_3}(\text{diplomás,átlag felett kereső})$ paraméter értéke:

$$\lambda_{V_1 V_3}^{V_1 V_2 V_3}(\text{diplomás, magas jövedelem}) = \frac{1}{3} \sum_{k \in \{\text{alsó, közép, felső}\}} \log \mu(\text{diplomás}, k, \text{magas jövedelem}) -$$

$$-\lambda_{V_1}^{V_1 V_2 V_3}(\text{diplomás}) - \lambda_{V_3}^{V_1 V_2 V_3}(\text{magas jövedelem}) - \lambda_{\varnothing}^{V_1 V_2 V_3}(\mathbf{i}_{\varnothing}).$$

---

[8] Ez az un. λ-típusú paraméterezés alapelve szerint mindig az egyenletes eloszlásnak megfeleltetett „nincs hatás” szituációhoz viszonyítja egy-egy hatás erősségét. Több más paraméterezés létezik, a szintén gyakran használt μ-típusú paraméterezés mellett pl. a „nincs hatás” szituáció az, amikor csak a referenciacella gyakorisága nem nulla; a paraméterek azt mérik, hogy a többi cella gyakorisága mennyire tér el a referenciacelláétól.

[9] Az egyszerűség kedvéért itt és a disszertáció egészében a gyakorisági- és nem a valószínűség-eloszlás paraméterezését tárgyalom. A valószínűség-eloszlás paraméterezése egyébként könnyen megkapható ebből az $\emptyset$ hatáshoz tartozó un. főhatás elhagyásával.

A T tábla hagyományos loglineáris paraméterekkel történő felparaméterezésének nevezzük a λ_P megadását, ahol $P=\{(L,V), L\subseteq V\}$ rendezett párok halmaza. Formálisan λ_P:

$$\lambda_P=\left\{\lambda_L^V : L\subseteq V\right\}.$$

Példánk esetén

$$P=\left\{\begin{matrix}(\varnothing,V_1V_2V_3),(V_1,V_1V_2V_3),(V_2,V_1V_2V_3),(V_3,V_1V_2V_3),(V_1V_2,V_1V_2V_3),(V_1V_3,V_1V_2V_3),(V_2V_3,V_1V_2V_3),\\(V_1V_2V_3,V_1V_2V_3)\end{matrix}\right\},$$

$$\lambda_P=\left\{\lambda_{\varnothing}^{V_1V_2V_3},\ \lambda_{V_1}^{V_1V_2V_3},\ \lambda_{V_2}^{V_1V_2V_3},\ \lambda_{V_3}^{V_1V_2V_3},\ \lambda_{V_1V_2}^{V_1V_2V_3},\ \lambda_{V_1V_3}^{V_1V_2V_3},\ \lambda_{V_2V_3}^{V_1V_2V_3},\ \lambda_{V_1V_2V_3}^{V_1V_2V_3}\right\}.$$

A paraméterezés definíciójából közvetlenül következik, hogy adott **i** cellához tartozó paraméterértékek összege éppen a log-cellagyakoriságot adja:

$$\log\mu(\mathbf{i})=\sum_{L'\subseteq V}\lambda_{L'}^V(\mathbf{i}_{L'}), \tag{3}$$

és hogy 0-t kapunk, ha egy adott paramétert úgy összegezünk, hogy tetszőleges értéken rögzítjük a hozzá tartozó L hatás valamely $M\subseteq L$ részhalmazának koordinátáit:

$$\sum_{\mathbf{j}_L\in T_L:(\mathbf{j}_L)_M=(\mathbf{i}_L)_M}\lambda_L^V(\mathbf{j}_L)=0. \tag{4}$$

Az utóbbi tulajdonság egyszerű esetben egy adott sorhoz/oszlophoz tartozó paraméterértékek összegének 0 voltát jelenti, pl.:

$$\sum_{k\in\{alsó,\,közép,\,felső\}}\lambda_{V_1V_2}^{V_1V_2V_3}(\text{diplomás},k)=0.$$

Vagyis elég az alsó- és a középosztályhoz tartozó paraméter megadása, a felsőosztályé redundáns, mert ezekből megkapható. Általánosan is, a $T=\times_{j=1}^p I_j$ tábla felparaméterezéséhez a paraméterértékek közül elég néhánynak a megadása, a többiek redundánsak. Konkrétan: a λ_P paraméterezés redundanciája kiküszöbölhető, ha minden változó értékkészletéből elhagyunk egy értéket, és az így kapott kisebb kontingencia-táblához tartozó cellák λ_L^V paramétereit vesszük csak figyelembe:

1. definíció Legyen $v_i\in I_i$, $\tilde{I}_i=I_i\setminus\{v_i\}$, $\tilde{T}=\times_{j=1}^p\tilde{I}_j$. Ekkor az $\mathbf{i}\in\tilde{T}$ cellákon értelmezett, de az eredeti T táblából számolt paraméterek összessége $\tilde{\lambda}_L^V$, a paraméterezést $\tilde{\lambda}_P$ jelöli: $\tilde{\lambda}_P=\left\{\tilde{\lambda}_L^V : L\subseteq V\right\}$.

A $\tilde{\lambda}_L^V$ paraméternek már nincs redundáns értéke. A $\tilde{\lambda}^V_{L=\bigcup_{i=1}^k V_i}$ paraméterhez éppen $\prod_{j=1}^k \left(\left|I_j\right|-1\right)$ nem-redundáns paraméterérték tartozik. Könnyen kiszámolható, hogy a $\tilde{\lambda}_P$ paraméterezés összesen $\prod_{j=1}^p \left|I_j\right|$ paraméterértékből áll, ami éppen a T tábla celláinak számával egyezik meg. Az SPSS loglineáris modellezési parancsai (a Hierarchical Loglinear Model és a General Loglinear Model) éppen így járnak el a paraméterbecslések takarékos megadásánál: csak a nem redundánsakat tüntetik fel közülük, mindig a változók utolsó értékét elhagyva.

Példánk esetén a T (diplomás, diploma nélküli)×(alsó, közép-, felső osztály)×(magas jövedelem, alacsony jövedelem) 12 cellás táblázat helyett változónként egy-egy érték elhagyásával a $\tilde{T}$ (diplomás)×(alsó, középosztály)×(magas jövedelem) 2 cellás táblázatot tekintjük. A

$$\tilde{\lambda}_P(\mu)=\left\{\begin{array}{l}\tilde{\lambda}_{\varnothing}^{V_1V_2V_3}\left(\text{mindegyik cella}\right),\ \tilde{\lambda}_{V_1}^{V_1V_2V_3}\left(\text{diplomás}\right),\ \tilde{\lambda}_{V_2}^{V_1V_2V_3}\left(\text{alsó osztály}\right),\ \tilde{\lambda}_{V_2}^{V_1V_2V_3}\left(\text{középosztály}\right),\\ \tilde{\lambda}_{V_3}^{V_1V_2V_3}\left(\text{magas jöv.}\right),\ \tilde{\lambda}_{V_1V_2}^{V_1V_2V_3}\left(\text{diplomás,alsó osztály}\right),\ \tilde{\lambda}_{V_1V_2}^{V_1V_2V_3}\left(\text{diplomás,középosztály}\right),\\ \tilde{\lambda}_{V_1V_3}^{V_1V_2V_3}\left(\text{diplomás,magas jöv.}\right),\ \tilde{\lambda}_{V_2V_3}^{V_1V_2V_3}\left(\text{alsó osztály, magas jöv.}\right),\\ \tilde{\lambda}_{V_2V_3}^{V_1V_2V_3}\left(\text{középosztály, magas jöv.}\right),\ \tilde{\lambda}_{V_1V_2V_3}^{V_1V_2V_3}\left(\text{diplomás,alsó osztály, magas jöv.}\right),\\ \tilde{\lambda}_{V_1V_2V_3}^{V_1V_2V_3}\left(\text{diplomás, középosztály, magas jöv.}\right)\end{array}\right\}$$

paraméterezéshez 12 paraméterérték tartozik, melyekkel az eredeti táblázat 12 cellája felparaméterezhető.

A $\tilde{\lambda}_P$ paraméterezés esetén a paraméterek nem csupán nem redundánsak, de értéktartományukon belül egymástól függetlenül bármilyen értéket felvehetnek. Ezt a tulajdonságot nevezik variációs függetlenségnek, mely formálisan az alábbi módon definiálható:

2. definíció Legyen $\theta=(\theta_1, ..., \theta_t)$ t paraméterből álló paraméterezés. Azt mondjuk, hogy a θ_i-k *variációsan függetlenek*, ha $R(\theta)=R(\theta_1)\times...\times R(\theta_t)$, ahol $R(.)$ az értékkészletet jelöli.

λ-paraméterek esetén az értékkészlet a (-∞,+∞) intervallum. Példánkon illusztrálva: mivel tudjuk, hogy a $\tilde{\lambda}_P$ paraméterezés variációsan független, a paraméterek bármilyen negatív vagy pozitív értéket felvehetnek. Pl. a

$$\tilde{\lambda}_{V_3}^{V_1V_2V_3}(\text{magas jövedelem})=0,3,$$
$$\tilde{\lambda}_{V_1V_2}^{V_1V_2V_3}(\text{diplomás, középosztály})=-8,$$
$$\tilde{\lambda}_{V_1V_2}^{V_1V_2V_3}(\text{diplomás, felső osztály})=6$$

paraméterekhez is létezik nekik megfelelő gyakorisági eloszlás. A variációs függetlenség fontosságáról a marginális loglineáris paraméterezés kapcsán írok bővebben, a 2.3.1 fejezetben. Ott láthatunk majd példát variációsan nem független paraméterekre is.

Fontos tulajdonsága továbbá a loglineáris paraméterezésnek, hogy a paraméterértékek jól értelmezhetők: a paraméterek a hatáshoz tartozó változók közötti feltételes kapcsolat átlagos erősségének mértékeként interpretálhatók, az összes többi változóra, mint feltételre nézve. Például ha a hatáshoz tartozó változók binárisak, a paraméter könnyen megfeleltethető a vonatkozó feltételes esélyhányadosok bizonyos átlagának. Szemléltetésül: az előbbi $\lambda_{V_1V_3}^{V_1V_2V_3}\left(\text{diplomás,átlag felett kereső}\right)$ paraméterérték az alábbi formára hozható:

$$\lambda_{V_1V_3}^{V_1V_2V_3}\left(\text{diplomás, magas jöv.}\right)=$$

$$=\frac{1}{3}\sum_k \log\left(\left(\frac{\mu\left(\text{diplomás, }k\text{, magas jöv.}\right)\mu\left(\text{nem diplomás, }k\text{, alacsony jöv.}\right)}{\mu\left(\text{diplomás, }k\text{, alacsony jöv.}\right)\mu\left(\text{nem diplomás, }k\text{, magas jöv.}\right)}\right)^{1/4}\right)=$$

$$=\frac{1}{3}\sum_k \log\left(EH\left(V_1=\{1,2\},V_3=\{1,2\}\mid V_2=k\right)^{1/4}\right),$$

ami a foglalkozási státusz egyes kategóriáinak rögzítése mellett számolt feltételes esélyhányadosok negyedik gyökének átlaga. Az esélyhányados ismert asszociációs mérőszám, így valóban azt mondhatjuk, hogy a paraméter az iskolázottság és jövedelem foglalkozásra feltételes kapcsolatának átlagos erősségét mutatja. Ha a $\lambda_{V_1V_3}^{V_1V_2V_3}$ értéke 0, akkor az „átlagos hatás" 0, tehát V_1 és V_3 között V_2 bizonyos értékei melletti pozitív feltételes kapcsolat van, bizonyos értékei mellett negatív feltételes kapcsolat van, amelyek átlagoláskor kioltják egymást. Szokásos szóhasználattal ezt úgy is mondhatjuk, hogy a V_2-re történő *korrigálás* után, vagy a V_2 *hatását kiszűrve* a két változó között nincs asszociáció.

Értelemszerűen, ha a $\lambda_{V_1V_3}^{V_1V_2V_3}$ paraméter értéke pozitív (negatív), akkor a feltételes kapcsolatok inkább pozitívak (negatívak).

Fontos itt kitérnem arra a mobilitás-kutatási alkalmazásoknál, az 6.2.2. és 7.1. fejezetekben még többször előkerülő tényre, hogy az esélyhányados és a marginálisok variációsan függetlenek (Rudas, 1998). Ez nem csak kétdimenziós táblára igaz, hanem általánosan, magasabbrendű esélyhányadossal és nála alacsonyabbrendű marginálisokkal megfogalmazva is. Ez a tulajdonság asszociációs mérőszámok esetén elsődleges fontosságú, hiszen az asszociáció tisztán az együttes eloszlásban meglevő információhoz kapcsolódik, a marginálisokban meglevő információhoz nem. Tehát a loglineáris paramétereknek az esélyhányadosokra történő visszavezethetősége azért is előnyös tulajdonság, mert így öröklik azoknak az alacsonyabbrendű marginálisoktól való függetlenségét is.

Változók feltételes függetlensége a későbbiekben központi jelentőségű fogalom lesz, ezért ennek a fogalomnak a paraméterekkel való kapcsolatát külön is érdemes megvizsgálni.

3. definíció *X*, *Y* és *Z* diszkrét valószínűségi változók esetén *X* és *Y feltételesen függetlenek Z*-re nézve, ha:

$$P\left(X=x,Y=y\mid Z=z\right)=P\left(X=x\mid Z=z\right)P\left(Y=y\mid Z=z\right),$$

Z minden *z* értéke mellett, amire $P\left(Z=z\right)>0$. Jelöléssel: $X\perp Y\mid Z$.

Vegyük észre, hogy a definíció változtatás nélkül értelmezhető változóhalmazok függetlenségére is.

A feltételes függetlenségnek a paraméterekkel való kapcsolatának vizsgálatához vegyünk először azt a speciális esetet, amikor *X, Y* és *Z* bináris változók. Láttuk, hogy a λ_{XY}^{XYZ} paraméter a *Z*-re, mint feltételre vett, szigorúan monoton függvénnyel (negyedik gyökök logaritmusa) transzformált feltételes esélyhányadosok átlagát adja. Belátható, hogy a λ_{XYZ}^{XYZ} paraméter ugyanakkor a feltételes esélyhányadosok transzformáltjának különbségével egyenlő:

$$\lambda_{XYZ}^{XYZ}(i,j,k) = (-1)^{1+ijk}\left(\log\left(EH(X,Y \mid Z=1)^{1/4}\right) - \log\left(EH(X,Y \mid Z=2)^{1/4}\right)\right).$$

Ha a λ_{XYZ}^{XYZ} paraméter értéke 0, akkor a két feltételes esélyhányados megegyezik, vagyis *Z*-nek nincs hatása az *XY* kapcsolatra. Itt a három változó szerepe szimmetrikus, vagyis ugyanez igaz az *X* feltétel melletti *YZ* kapcsolatra is stb.

Ha $X \perp Y \mid Z$, akkor nyilván a feltételes esélyhányadosok értéke 0, ezért az ezek függvényeként definiálható λ_{XYZ}^{XYZ} és λ_{XY}^{XYZ} paraméterek értéke is 0 minden cellára. Ugyanez fordítva is igaz: a két paraméter 0 volta, mint láttuk, a feltételes esélyhányadosok azonosságát és átlagos 0 értékét jelenti, ez a két feltétel viszont 0 voltukat implikálja, ami éppen feltételes függetlenséget jelent.

Általában is igaz, hogy loglineáris paraméterezés mellett változók feltételes függetlensége könnyen megfeleltethető bizonyos paraméterek 0 voltának:

1. tétel Legyen X és Y a V két diszjunkt részhalmaza, legyen $Z = V \setminus (X \cup Y)$. Ekkor $X \perp Y \mid Z$ akkor és csak akkor, ha

$$\lambda_L^V = 0,\ \forall L \in \mathbf{P}(X \cup Y \cup Z) \setminus \left(\mathbf{P}(X \cup Z) \cup \mathbf{P}(Y \cup Z)\right),$$

ahol **P**(.) az összes részhalmaz halmazát, azaz a hatványhalmazt jelöli. Itt a $\lambda_L^V = 0$ kifejezés úgy értelmezendő, hogy a λ_L^V leképezés null-leképezés, azaz a hozzá tartozó marginális táblázat minden cellájára 0 az értéke.

Az eddigiekben a kontingencia-tábla felparaméterezését tárgyaltam, de nem említettem még a modell fogalmát. *Modell*nek a *T*-n értelmezett *F* gyakorisági eloszlások *G* részhalmazait nevezzük. Fenti példánknál maradva pl. a $V_1 \perp V_3 \mid V_2$ modell (a jövedelem és az iskolázottság foglalkozásra vett függetlensége) csak azokat a pozitív gyakorisági eloszlásokat tartalmazza, ahol a cellagyakoriságok megfelelnek a 3. definíció feltételének. Egy G_1 modell tartalmaz egy G_2 modellt, ha a megfelelő gyakorisági eloszlás-halmazok tartalmazzák egymást. Ilyenkor azt mondjuk hogy a G_2 *szűkebb modell*, mint a G_1, vagy azt, hogy G_2 a G_1 *almodellje (submodel),* vagy hogy G_1 a G_2 *tartalmazó modellje* (*supermodel*). Az almodellt és a tartalmazó modellt egymásbaágyazottnak (*nested*) is nevezik. Ilyen konstelláció áll elő pl., amikor G_1-et definiáló feltételes függetlenségi állítások között a G_2-t definiáló feltételes függetlenségi állítások is szerepelnek. Pl. a $V=\{V_1, V_2, V_3, V_4\}$ változók esetén, ahol V_4 az apa foglalkozása a V_2-vel megegyező kategóriákkal, ha G_1-et a $V_1 \perp V_3 \mid V_2$ definiálja, G_2-*t* a ($V_1 \perp V_3 \mid V_2$ és $V_4 \perp V_2 \mid V_1$), akkor G_2 szűkebb, mint G_1.

A disszertációban szereplő modellek szempontjából fontos loglineáris és log-affin modellekhez tartozó eloszlások azonosíthatók a loglineáris paraméterekre előírt bizonyos összefüggésekkel, ezeket definiálom az alábbiakban.

4. definíció Azt mondjuk, hogy $G \subseteq F$ *loglineáris modell,* ha elemei olyan eloszlások, amiknek a $\tilde{\lambda}_P$ paraméterezéssel kapott paraméterértékeinek bizonyos lineáris kombinációi nullával egyenlők. Formálisan:

$$\mu \in G \Leftrightarrow \sum_{L \subseteq V, \mathbf{i}_L \in \tilde{T}_L, j=1\ldots J} \alpha_{L\mathbf{i}_L j} \tilde{\lambda}_L^V(\mathbf{i}_L) = 0,$$

ahol J számú lineáris kombinációt írtunk fel, a G által meghatározott $\alpha_{L\mathbf{i}_L j}$ konstansokkal (itt az alsóindex az L hatáshalmaztól, az $\mathbf{i}_L$ cellától és a j. egyenlettől való függést jelöli). Láttuk, hogy azok a $X \perp Y \mid Z$ feltételes függetlenségi modellek, ahol a Z feltételben az X-en és Y-on kívül minden, a V-hez tartozó változó szerepel, könnyen megadhatók bizonyos loglineáris paraméterek 0-ként való előírásával. Ezek tehát loglineáris modellek: az α-k ilyenkor egy kivétellel mind 0-k.

Ha nem 0-ban, hanem valamely más értékben rögzítjük a paraméterek lineáris kombinációját (pl. adott erősségű feltételes kapcsolatot írunk elő egy változópárra), ennél általánosabb modellosztályt definiálhatunk:

5. definíció Azt mondjuk, hogy $G \subseteq F$ *log-affin* modell, ha olyan eloszlásokat tartalmaz, amiknek a $\tilde{\lambda}_P$ paraméterezéssel kapott paraméterértékeink bizonyos lineáris kombinációi egy-egy előírt β_j konstanssal egyenlők. Formálisan:

$$\mu \in G \Leftrightarrow \sum_{L \subseteq V, \mathbf{i} \in \tilde{T}, j=1\ldots J} \alpha_{L\mathbf{i}j} \tilde{\lambda}_L^V(\mathbf{i}_L) = \beta_j,$$

ahol $\alpha_{L\mathbf{i}j}$ és β_j a G által meghatározott konstansok.

Az elnevezés onnan jön, hogy a modelleket definiáló paramétermegkötések az első esetben lineáris kombinációk, a második esetben affin kombinációk. Geometriailag a loglineáris modellek a log-cellagyakoriságok $|T|$ dimenziós terének lineáris altereivel, a log-affin modellek ennek affin altereivel (a lineáris alterek eltoltjaival) azonosíthatók.

2.2 Marginális modellek, marginális loglineáris modellek

Az alábbi tételek, definíciók és jelölések, hacsak nem jelzem másként, Bergsma és Rudas (2002) ill. Rudas és Bergsma (2004) munkájára támaszkodnak.

A marginális modellek a kontingencia-tábla marginálisaira tett megszorításokkal definiálhatók. Marginális modellt definiál például az a feltevés, hogy egy populációban adott vélemény megoszlása nem változik valamely beavatkozás ellenére sem[10]. Ilyen kísérleti elrendezésre példa a *deliberative poll* (kb.: gondolkodtató közvéleménykutatás, nincs bevett

[10] Az ennek megfelelő elrendezés a biostatisztikában gyakoribb, ahol a tipikus klinikai kísérleteknél kezelés előtti és utáni állapotokat különböztetnek meg, és a betegek állapotának változása a megfigyelt esemény.

magyar elnevezése), amit Magyarországon elsőként Örkény Antal és társai próbáltak ki 2005 őszén. A mintába vont személyek cigánykérdéssel kapcsolatos véleményét mérték a kérdések közös megvitatása előtt és után. Két változóval kódolva a vélemény beavatkozás előtti (A) és beavatkozás utáni (B) megoszlását, egy lehetséges marginális modell a két változó által meghatározott kontingencia-tábla megfelelő sor- és oszlop-marginálisainak páronkénti egyenlőségét, formálisan a T_A ill. T_B marginális táblák cellagyakoriságainak azonosságát tételezi fel. Marginális modell megadásához tehát a marginális eloszlások felparaméterezésére van szükségünk, ami pl. marginális loglineáris paraméterekkel oldható meg. Ezek a hagyományos loglineáris paraméterektől abban térnek el, hogy nem csak az összes változó együttes eloszlása, hanem bármely marginális eloszlás mellett definiálhatunk paramétereket. Ezzel az általánosítással ugyanakkor a hagyományos loglineáris paraméterezésnél sokkal összetettebb keretet kapunk. Látni fogjuk, hogy a hagyományos esetre jellemző előnyös tulajdonságok itt nem mindig teljesülnek.

Tekintsük a korábban már definiált, egy $M \subseteq V$ marginálishoz tartozó $(.)_M$ projekciót, a T_M marginális táblát és a $\mu_M(\mathbf{i}_M)$ marginális gyakoriságokat. A marginális loglineáris paraméterek a hagyományos loglineáris paraméterekhez (lásd (1) és (2) formulákat) hasonlóan, rekurzív módon definiálhatók, azzal a különbséggel, hogy a változók V halmaza helyett csupán annak M marginálisát tekintjük, s az szerepel az indexekben is:

$$\lambda_{\varnothing}^{M}(\mathbf{i}_{\varnothing}) = \frac{1}{|T_M|} \sum_{\mathbf{j} \in T_M} \log \mu_M(\mathbf{j}), \tag{5}$$

$$\lambda_{L}^{M}(\mathbf{i}_{L}) = \frac{1}{|T_{M \setminus L}|} \sum_{\mathbf{j} \in T_M : \mathbf{j}_L = \mathbf{i}_L} \log \mu(\mathbf{j}) - \sum_{L' \subset L} \lambda_{L'}^{M}(\mathbf{i}_{L'}). \tag{6}$$

A loglineáris paraméterekhez hasonlóan értelmezhetők a marginális paraméterek is: az (alsó indexben szereplő L) hatáshoz tartozó változók közötti feltételes kapcsolat átlagos erősségének mértékeként interpretálhatók, a (felső indexben szereplő M) marginálishoz tartozó összes többi változóra, mint feltételre nézve.

A paraméter értelmezése és értéke nyilván függ attól, hogy melyik marginálison belül definiáltuk. Itt elég csak a Simpson paradoxonra utalni. A paradoxon arra a tényre hívja fel a figyelmet, hogy egy hatás erőssége, sőt előjele is függ attól, hogy marginális vagy feltételes hatásként vettük azt figyelembe[11]. Korábbi, a társadalmi státusz három indikátorával (V_1 bináris iskolázottság, V_2 három értékű foglalkozási státusz, V_3 bináris jövedelem) kapcsolatos példánkhoz visszatérve, azt az apa foglalkozásával (V_4) bővítve pl. az alábbi három, ugyanazon hatás mellett definiált paraméter mást és mást jelent, más és más értéket vehet fel:

[11] A biostatisztikai és szociológiai módszertani irodalomban standard példának számító paradoxon lényege, hogy (első látásra intuíciónknak ellentmondó módon) két változó kapcsolatának előjele egy harmadik változó figyelembe vételét követően megváltozhat. Pl. az A vállalat nagyobb arányban alkalmaz roma dolgozókat, mint a B vállalat, de a dolgozók iskolai végzettsége szerint külön vizsgálva mindegyik feltételes táblában a B vállalat roma dolgozói aránya a nagyobb. A példa megmagyarázhatóvá válik, ha feltesszük, hogy a B vállalat magasabb végzettséghez kötődő munkaköröket kínál, és hogy a romák átlagosan alacsonyabb iskolázottságúak. A paradoxon nevét Edward H. Simpsonról, egyik első leírójáról kapta.

$$\lambda_{V_1V_3}^{V_1V_3}(1,1) = \log EH\left(V_1 = \{1,2\}, V_3 = \{1,2\}\right)^{1/4},$$

$$\lambda_{V_1V_3}^{V_1V_2V_3}(1,1) = \frac{1}{3}\sum_{j=1..3} \log EH\left(V_1 = \{1,2\}, V_3 = \{1,2\} \mid V_2 = j\right)^{1/4},$$

$$\lambda_{V_1V_3}^{V_1V_2V_3V_4}(1,1) = \frac{1}{9}\sum_{\substack{j=1..3,\\ l=1..3}} \log EH(V_1 = \{1,2\}, V_3 = \{1,2\} \mid V_4 = j, V_2 = l)^{1/4}.$$

Előfordulhat pl., hogy az első paraméter nagy pozitív értéket vesz fel, tehát a kereset és az iskolázottság között erős pozitív kapcsolatot mutat, míg a második paraméter értéke kisebb pozitív szám, jelezve, hogy a diplomások magasabb keresete részben, de nem egészében az általuk elért jobb állásokkal magyarázható. Ha a harmadik paraméter értéke 0, akkor azt mondjuk, hogy az iskolázottság és a kereset között a többi változóra történt korrigálás után nincs (feltételes) kapcsolat[12]. Ilyen szituáció az, amikor a diplomások általában jobb állásokhoz jutnak, továbbá a magasabb foglalkozási státuszú apák utódai olyan kulturális előnyöket örökölnek (beszédstílus, megjelenés), amik jobb keresethez juttatják őket, köreikben magasabb a diplomások aránya is, de a jövedelmet az iskolázottság közvetlenül, ezeken a csatornákon kívül már nem befolyásolja.

A hagyományos loglineáris modellekre érvényes (3) és (4) összefüggések itt is állnak, tetszőleges M marginális rögzítésével

$$\log \mu_M\left(\mathbf{i}_M\right) = \sum_{L' \subseteq M} \lambda_{L'}^{M}\left(\mathbf{i}_{L'}\right), \tag{7}$$

azaz az M-hez tartozó összes paraméter összege a megfelelő marginális cellagyakoriságokat adja. Fennáll továbbá, hogy

$$\sum_{\mathbf{j}_L \in T_L : (\mathbf{j}_L)_{M'} = (\mathbf{i}_L)_{M'}} \lambda_L^M\left(\mathbf{j}_L\right) = 0, \tag{8}$$

vagyis 0-t kapunk, ha egy M-lel és L-lel definiált paraméter értékeit a T_L tábla azon celláin összegezzük, amiknek az L halmaz valamely $M' \subseteq L$ részhalmazához tartozó koordinátái a rögzített **i** cella megfelelő koordinátáival egyenlők.

Ugyanannak a kontingencia-táblának a felparaméterezése marginális loglineáris paraméterekkel többféleképpen történhet. Pl. belátható, hogy az alábbi két paraméterezés egyaránt egy-egyértelmű megadása a $T_{V_1V_2V_3}$ táblához tartozó eloszlásnak:

$$\lambda_{\varnothing}^{V_1}, \lambda_{V_1}^{V_1}, \lambda_{V_2}^{V_1V_2}, \lambda_{V_1V_2}^{V_1V_2}, \lambda_{V_3}^{V_1V_2V_3}, \lambda_{V_1V_3}^{V_1V_2V_3}, \lambda_{V_2V_3}^{V_1V_2V_3}, \lambda_{V_1V_2V_3}^{V_1V_2V_3}, \tag{9}$$

$$\lambda_{\varnothing}^{\varnothing}, \lambda_{V_1}^{V_1}, \lambda_{V_3}^{V_3}, \lambda_{V_1V_3}^{V_1V_3}, \lambda_{V_2}^{V_2}, \lambda_{V_2V_3}^{V_2V_3}, \lambda_{V_1V_2}^{V_1V_2}, \lambda_{V_1V_2V_3}^{V_1V_2V_3}. \tag{10}$$

[12] Mint a 2.1 fejezetben láttuk, ilyenkor a feltételes függetlenség még nem feltétlenül áll: a 0 átlagos értékként adódik, akár negatív és pozitív értékek átlagaként.

A (10)-es paraméterezés, amikor minden hatás a vele megyegyező marginálison belül adott, éppen Glonek és McCullagh többváltozós logisztikus transzformációjához tartozó paraméterekből áll (Glonek, McCullagh, 1995).

Marginális loglineáris paraméterek egy halmaza a paraméterekhez tartozó *M* marginálisból és *L* hatásból álló rendezett párok *P* halmazával definiálható:

$$\lambda_P = \left\{ \lambda_L^M : (L, M) \in P \right\}.$$

λ_P-re eddig még semmilyen megkötést nem tettünk, akárhány paramétert tartalmazhatott, a kontingenciatábla egyértelmű felparaméterezéséhez képest akár túl keveset vagy feleslegesen sokat is. A következőkben ennek megoldására először bizonyos szempontból „elegendő” számú paramétert biztosítunk, majd ezek közül elhagyjuk a feleslegeseket.

6. definíció Az (*L,M*) rendezett párok *P* halmazát *teljes*nek (*complete*) nevezzük, ha minden $L \subseteq V$ hatáshoz van olyan (és csak egy olyan) $M \subseteq V$ marginális, amire (*L,M*) a *P*-nek eleme.

A hagyományos loglineáris paraméterezéshez tartozó *P* pl. teljes, hiszen minden hatáshoz pontosan egy marginális, a *V* tartozik, és minden $M \subseteq V$ szerepel a hatások között. A $V = \{V_1, V_2, V_3\}$ változókra teljes a (9) és (10) paraméterezések *P* halmaza is is.

A teljes *P* halmazokhoz tartozó λ_P paraméterezések a (8)-as egyenletből következően feleslegesen sok paramétert tartalmaznak. A redundáns paraméterek eltávolítása a hagyományos loglineáris paramétereknél látott módon itt is elvégezhető.

7. definíció Legyen $v_i \in I_i$, $\tilde{I}_i = I_i \setminus \{v_i\}$, $\tilde{T} = \times_{j=1}^{p} \tilde{I}_j$. Ekkor az $\mathbf{i} \in \tilde{T}$ cellákon értelmezett, de az eredeti *T* táblából számolt

$$\tilde{\lambda}_L^M(\mathbf{i}_L) : F \to \mathbb{R}^{|\tilde{T}|}$$

paraméterek összességét $\tilde{\lambda}_P$ jelöli: $\tilde{\lambda}_P = \left\{ \tilde{\lambda}_L^M : L \subseteq M \subseteq V \right\}$.

A konstrukcióból következően $\tilde{\lambda}_P$ már nem tartalmaz redundáns paramétert.

Pl. ha a λ_P a (9)-es paraméterezéssel egyezik meg, $\tilde{\lambda}_P$ a következőképpen áll elő:

$$\tilde{\lambda}_P(\mu) = \left\{ \begin{array}{l} \tilde{\lambda}_{\varnothing}^{V_1}(\text{mindegyik cella}),\ \tilde{\lambda}_{V_1}^{V_1}(\text{diplomás}),\ \tilde{\lambda}_{V_2}^{V_1V_2}(\text{alsó osztály}),\ \tilde{\lambda}_{V_2}^{V_1V_2}(\text{középosztály}), \\ \tilde{\lambda}_{V_1V_2}^{V_1V_2}(\text{diplomás,alsó osztály}),\ \tilde{\lambda}_{V_1V_2}^{V_1V_2}(\text{diplomás,középosztály}), \tilde{\lambda}_{V_3}^{V_1V_2V_3}(\text{magas jöv.}), \\ \tilde{\lambda}_{V_1V_3}^{V_1V_2V_3}(\text{diplomás,magas jöv.}),\ \tilde{\lambda}_{V_2V_3}^{V_1V_2V_3}(\text{alsó osztály, magas jöv.}), \\ \tilde{\lambda}_{V_2V_3}^{V_1V_2V_3}(\text{középosztály, magas jöv.}),\ \tilde{\lambda}_{V_1V_2V_3}^{V_1V_2V_3}(\text{diplomás, alsó osztály, magas jöv.}), \\ \tilde{\lambda}_{V_1V_2V_3}^{V_1V_2V_3}(\text{diplomás, középosztály, magas jöv.}) \end{array} \right\}$$

A loglineáris ill. log-affin marginális modellek meghatározása is a hagyományos esetben látott definíciót követi.

8. definíció Legyen adott egy P halmaz. Azt mondjuk, hogy $G \subseteq F$ *marginális loglineáris modell,* ha elemei olyan eloszlások, amiknek a $\tilde{\lambda}_P$ paraméterezéssel kapott paraméterértékeinek bizonyos lineáris kombinációi nullával egyenlők. Formálisan:

$$\mu \in G \Leftrightarrow \sum_{\tilde{\lambda}_L^M \in \tilde{\lambda}_P, \mathbf{i}_L \in \tilde{T}_L, j=1\ldots J} \alpha_{\tilde{\lambda}_L^M \mathbf{i}_L j} \tilde{\lambda}_L^M \left(\mathbf{i}_L\right) = 0,$$

ahol J számú lineáris kombinációt írtunk fel, a G által meghatározott $\alpha_{\tilde{\lambda}_L^M \mathbf{i}_L j}$ konstansokkal. Az α három alsóindexe azt mutatja, hogy minden $\tilde{\lambda}_L^M$ paraméterhez és minden i_L cellához más konstansszorzó tartozhat az 1...J egyenletek mindegyikében.

9. definíció Legyen adott egy P halmaz. Azt mondjuk, hogy $G \subseteq F$ *marginális log-affin modell,* ha elemei olyan eloszlások, amiknek a $\tilde{\lambda}_P$ paraméterezéssel kapott paraméterértékeinek bizonyos lineáris kombinációi adott konstansokkal egyenlők. Formálisan:

$$\mu \in G \Leftrightarrow \sum_{\tilde{\lambda}_L^M \in \tilde{\lambda}_P, \mathbf{i}_L \in \tilde{T}_L, j=1\ldots J} \alpha_{\tilde{\lambda}_L^M \mathbf{i}_L j} \tilde{\lambda}_L^M \left(\mathbf{i}_L\right) = \beta_j,$$

ahol J számú lineáris kombinációt írtunk fel, a G által meghatározott $\alpha_{\tilde{\lambda}_L^M \mathbf{i}_L j}$ és β_j konstansokkal.

Megfelelő módon felparaméterezett kontingenciatábla esetén a modell a paraméterekre tett egyszerű megkötésekkel határozható meg. A társadalmi státusszal kapcsolatos 1. példa véve ismét: az 1. tétel szerint az a modell, ami az Iskolázottság (V_1) és Jövedelem (V_3) Foglalkozásra (V_2), mint feltételre vett függetlenségét állítja, a $\lambda_{V_1 V_3}^{V_1 V_2 V_3}$ és $\lambda_{V_1 V_2 V_3}^{V_1 V_2 V_3}$ paraméterek 0-ra állításával definiálható[13]. Ez a (9)-es paraméterezés mellett közvetlenül megadható, hiszen a paraméterek szerepelnek a paraméterezésben. Viszont a (10)-es paraméterezés esetén a modell ennél nehézkesebb módon (paraméterek összetett, nem feltétlenül lineáris függvényével) definiálható. Ugyanakkor a V_1 és V_3 marginális függetlensége éppen a (10)-es paraméterezéssel megy könnyen, hiszen ott szerepel a (0-ra állítandó) $\lambda_{V_1 V_3}^{V_1 V_3}$ paraméter.

Ha az apa foglalkozásáról (V_4) és iskolázottságáról (V_5) is van információnk, akkor az iskolázottság-foglalkozás kapcsolat intergenerációs változatlanságát feltételező modell az alábbi marginális asszociációk azonosságával definiálható:

$$\lambda_{V_4 V_5}^{V_4 V_5} = \lambda_{V_1 V_2}^{V_1 V_2}.$$

Ez marginális loglineáris modell, hiszen a két paraméter különbségének 0 voltával azonosítható, egyetlen lineáris kombinációval állítható elő, ahol $\alpha_1 = 1$ és $\alpha_2 = -1$.

2. példa A mobilitáskutatásban gyakran használt *log-multiplicative layer effect* vagy másnéven *unidiff* modellt (Xie, 1992, Erikson, Goldthorpe, 1992) is érdemes itt megemlíteni (a módszerről részletesebben lásd a 6.2.1. fejezetet). A modell nevét onnét, kapta, hogy (a) időbeli

[13] Az egyértelműség kedvéért: az 1. tétel a sok lehetséges közül az egyek paraméterezését adja meg a feltételes függetlenségnek. Így a tétel használatakor szintén csak az egyik szóba jöhető (de a legegyszerűbb, legkézenfekvőbb) paraméterezést határozunk meg.

összehasonlításoknál két időpont között egyenletesen (*uniform*) változó mobilitási esélyeket tételez fel, vagy - technikailag azonos módon - (b) országok összevetésénél az országonként változó intenzitású mobilitási esélyek azonos mintázatát teszi fel. Ez a feltevés az apa foglalkozása (*F'*) – utód foglalkozása (*F*) feltételes asszociációnak egy φ_k konstanssal való szorzástól eltekintve állandó voltával azonosítható, ahol a feltétel a *T* (időpont az első esetben vagy ország a másik esetben), a *k* pedig a feltételváltozó egy értékéhez tartozik. Paraméterekkel:

$$\lambda_{F'F}^{F'FT}(i,j)+\lambda_{F'FT}^{F'FT}(i,j,k)=\psi_{ij}\varphi_k.$$

Itt a ψ_{ij} mutatja az apa-utód kereszttábla (i,j) cellájában a kapcsolat minden időpontra (országra) közös értékét, φ_k pedig ennek a kapcsolatnak az intenzitásának időbeli (országok közötti) változását. Rögzített φ_k–k mellett ezek az egyenletek marginális log-affin modellt definiálnak, az alábbi lineáris egyenletrendszerekkel:

$$\frac{1}{\psi_{ij}}\lambda_{F'F}^{F'FT}(i,j)+\frac{1}{\psi_{ij}}\lambda_{F'FT}^{F'FT}(i,j,k)=\varphi_k.$$

Csökkenő (növekvő) φ_k–kkal az asszociáció gyengülése (erősödése), vagyis társadalmi nyitódás (záródás) tesztelhető. A tényleges *unidiff* modell ennél még kevésbé meghatározott: a φ_k–kat is ismeretlennek tételezi fel, és ezek becslése a cél, feltéve, hogy a modell illeszkedik.

3. példa A *deliberative poll*-lal kapcsolatos példára térve: legyen *A* a beszélgetés előtti, *B* a közös beszélgetés utáni vélemény (az egyszerűség kedvéért bináris változóval kódolva). A vélemény megoszlásának változatlanságát tételező modell az *AB* kereszttábla *A* és *B* marginálisának egyenlőségével azonosítható. Az *A* marginálisa a $\tilde{\lambda}_L^A$, $L \subseteq A$ paraméterekkel (7)-ből következően könnyen megadható:

$$\log\mu_A(1)=\tilde{\lambda}_\varnothing^A(\mathbf{i}_\varnothing)+\tilde{\lambda}_A^A(1),$$

hasonlóan

$$\log\mu_B(1)=\tilde{\lambda}_\varnothing^B(\mathbf{i}_\varnothing)+\tilde{\lambda}_B^B(1),$$

ebből

$$\log\mu_A(1)=\log\mu_B(1)\Leftrightarrow\tilde{\lambda}_A^A(1)=\tilde{\lambda}_B^B(1),$$

vagyis

$$\mu\in G\Leftrightarrow\tilde{\lambda}_A^A(1)-\tilde{\lambda}_B^B(1)=0.$$

Tehát a modell marginális loglineáris modell, és könnyen megadható, ha az *(A,A)* és *(B,B)* szerepel a *P*-ben.

4. példa Ugyanebben a keretben maradva tegyük fel, hogy az a feltevésünk, hogy a romákkal szembeni elutasító/elfogadó attitűd a romákról meglévő ismeretekkel szorosan összefügg, és a közös beszélgetés nem változtat ennek az összefüggésnek az erősségén. Ekkor, ha C jelöli a romákkal kapcsolatos információk mennyiségét a beszélgetés előtt, D pedig a beszélgetés után, akkor a

$$\lambda_{AC}^{AC}=\lambda_{BD}^{BD}$$

egyenlőséggel fejezhető ki a marginális asszociációk állandósága. Ez tehát szintén marginális loglineáris modell. Ha azt gondoljuk továbbá, hogy elsősorban ez az informáltsági szint számít az

attitűd kialakulásánál, és az elfogadó attitűd pozitív kapcsolata az iskolázottsággal csak azért tűnik erősnek, mert az iskolázottság befolyásolja az informáltsági szintet is, akkor az attitűd és az iskolázottság (*I*) ismeretszintre vett feltételes függetlenségét tesszük fel, ami szintén marginális loglineáris modell, és az 1. tétel segítségével így fejezhető ki:

$$\lambda_{AI}^{ACI} = \lambda_{ACI}^{ACI} = \lambda_{BI}^{BDI} = \lambda_{BDI}^{BDI} = 0,$$

tehát könnyen megadható, ha ezek a paraméterek szerepelnek a paraméterezésben.

2.3 A paraméterezés jó tulajdonságai

Lényeges kérdés tehát az, hogy olyan paraméterezést válasszunk, amellyel a vizsgált modell egyszerűen megadható. A paraméterekre tett megkötések értelmezésénél felmerülő másik fontos kérdés a paraméterek interpretálhatósága: láttuk, hogy marginális loglineáris paraméterek esetén a paraméterek kézenfekvő módon, feltételes asszociációkra történő visszavezetéssel értelmezhetők. Az alábbiakban a marginális loglineáris paraméterezés további, alapvető fontosságú tulajdonságait tárgyalom. Hacsak nem jelzem másként, a tételeket Bergsma, Rudas (2002)-ből és Rudas, Bergsma (2004)-ből emelem ki, a definíciók, jelölések és magyarázatok esetében e forrásokon kívül Lauritzen (1996)-ra is támaszkodom majd.

2.3.1 Variációs függetlenség

Mint láttuk, a hagyományos loglineáris paraméterek variációsan függetlenek, vagyis az egyes paraméterek a többiektől függetlenül bármilyen értéket felvehetnek (a tulajdonság pontos meghatározását a 2. definíciónál láttuk). A marginális loglineáris paraméterezésekre ez nem áll általánosan. Pedig alapvető fontosságú lenne a tulajdonság teljesülése, hiszen ha a paraméterek nem függetlenek variációsan, az nemlétező modellek megadásához vezethet, problémát okoz interpretációjukban, és gyakran okoz számítási problémákat is.

5. példa Rudas és Bergsma (2004) példájára támaszkodva, annak egy társadalomtudományi megfelelőjén illusztrálom a variációsan nem független paraméterek interpretációs problémáit. Vegyük a már látott *deliberative poll* példáját. Tekintsük a közös vitában részt vett, és az abban részt nem vett (ezekbe a csoportokba a vitát megelőzően véletlen módon besorolt) személyek csoportjait, külön a diplomásokat és a legalacsonyabb iskolai végzettséggel rendelkezőket, és nézzük meg, milyen a romákkal szembeni attitűdjük (két értékben: elfogadó vagy elutasító) a közös vitát követően. A háromdimenziós táblát az alacsonyabb dimenziós marginálisokkal és egy, a vita hatását iskolai végzettségenként külön mérő mérőszámmal paraméterezzük fel. Az utóbbi mérőszámot az elfogadó attitűddel rendelkezők arányának a vitán részt vett és a vitán részt nem vett csoportokban mért különbségeként definiáljuk. Az alacsony iskolai végzettségen belül az alábbi táblában a zárójelek előtt feltüntetett gyakoriságokkal számolva $\frac{10}{10+40} - \frac{5}{5+45} = 0,1$ ez az érték, a diplomásokra $\frac{30}{30+20} - \frac{20}{20+30} = 0,2$. A mérőszám nem független variációsan a marginálisoktól, azok meghatározzák maximum értékét. A táblában zárójelben tüntettük fel azokat a gyakoriságokat, amelyek esetén a marginálisok változatlanok lennének, és a mérőszámok maximálisak. A mérőszám maximuma eszerint a legfeljebb általános iskolát végzetteknél 0,3, a diplomásoknál 1. Ennek következtében a két végzettségre kapott mérőszám nem vethető össze érvényes módon, pl. a most kapott 0,1-es és 0,2-es értékek esetén nem

mondhatjuk, hogy az alacsony iskolázottságúakon belül fele olyan hatásos a közös vita, mint a diplomásokon belül, hiszen a 0,2 a lehetséges 1-es maximumtól jóval távolabb van, mint a 0,1 a saját 0,3-as maximumától.

1. táblázat. Hipotetikus példa egy *deliberative poll* végeredményéről. Zárójelben a vita maximális hatása esetén fennálló gyakoriságok.

	Közös vitába bevont	*Közös vitába be nem vont*
	Legfeljebb 8 általános	
Elfogadó	10 (15)	5 (0)
Elutasító	40 (35)	45 (50)
	Diplomás	
Elfogadó	30 (50)	20 (0)
Elutasító	20 (0)	30 (50)

A példák után a $\tilde{\lambda}_P$ típusú marginális loglineáris paraméterekre visszatérve: belátható, hogy ha V legalább három változót tartalmaz, akkor a Glonek és McCollugh-féle marginális loglineáris paraméterekből (mint amilyenek a (10)-ben látott paraméterek) előállított $\tilde{\lambda}_P$ paraméterei sem függetlenek variációsan. Kérdés, milyen kritérium alapján állapítható meg egy $\tilde{\lambda}_P$ paraméterezésről, hogy variációsan független-e. Bergsma és Rudas (2002) fontos tétele szerint a paraméterezés bizonyos kombinatorikus tulajdonságainak teljesülésekor fennáll a paraméterezés komponenseinek variációs függetlensége. Ehhez előbb definiálnunk kell a P rendezett pár-halmaz két tulajdonságát, a hierarchikusságot (*hierarchy*) és a rendezett módon való dekomponálhatóságot (*ordered decomposability*).

Legyen P az (L, M), $L \subseteq M \subseteq V$ rendezett párok halmaza. Legyen

$$\mathbf{M} = \{M \mid \exists L \subseteq V : (L, M) \in P\}$$

a paraméterezéshez tartozó marginálisok halmaza. A marginálisok egy adott $M_1, \ldots, M_s$, $M_i \in \boldsymbol{M}$ rendezéséhez legyen

$$\mathbf{L}_i = \{L \mid (L, M_i) \in P\}$$

az i. marginálishoz tartozó hatások halmaza, és legyen

$$\mathbf{K}_1 = \mathbf{P}(M_1),$$
$$\mathbf{K}_i = \mathbf{P}(M_i) \setminus \left(\mathbf{P}(M_1) \bigcup \ldots \bigcup \mathbf{P}(M_{i-1})\right),$$

ahol **P**(.) szokott módon a hatványhalmazt jelöli.

10. definíció A P halmazt *hierarchikus*nak nevezzük, ha az **M** elemeinek van olyan $M_1, \ldots, M_s$ rendezése, ahol

$$M_i \not\subseteq M_j, \text{ ha } i>j \tag{11}$$

és

$$\mathbf{L}_i \subseteq \mathbf{K}_i. \tag{12}$$

Azaz a sorrend nem lehet növekvő (11), és minden hatást a lehető legkorábbi marginális mellett kell szerepeltetni (12). Ilyenkor az **M** rendezését is hierarchikus rendezésnek, és magát a paraméterezést is hierarchikus marginális loglineáris paraméterezésnek nevezzük.

A (9) és (10) paraméterezések hierarchikusak. Ezzel szemben az alábbi teljes paraméterezés nem hierarchikus:

$$\lambda_P = \left\{ \lambda_{\varnothing}^{A},\ \lambda_{A}^{AB},\ \lambda_{B}^{AB},\ \lambda_{AB}^{AB} \right\},$$

mert a (11) eléréséhez **M** ({A}, {AB}) sorrendű lehet csak, ekkor $\mathbf{L}_1$={$\emptyset$}, $\mathbf{L}_2$={{A}, {B}, {AB}}, $\mathbf{K}_1$={$\emptyset$,{A}}, $\mathbf{K}_2$={{B}, {AB}}. Viszont így $\mathbf{L}_2 \not\subseteq \mathbf{K}_2$. A problémát az okozza, hogy az A hatás nem a lehető legkorábbi marginális (A) mellett szerepel.

A P halmaz másik fontos tulajdonsága a rendezett módon való dekomponálhatóság:

11. definíció Marginálisok **M** halmazának egy $M_1, ..., M_s$ hierarchikus rendezését *rendezett módon dekomponálhatónak* nevezzük, ha **M**-nek legfeljebb két eleme van, vagy ha minden k=3, ..., s-re az { $M_1, ..., M_k$ } r számú maximális elemére[14] vagy r≤2, vagy az eredeti rendezést tartó $M_1, ..., M_r$ olyan, hogy minden l=3, ..., r-re van olyan $j_l<l$, hogy:

$$\left(\bigcup_{i=1}^{l-1} M_i\right) \bigcap M_l = M_{j_l} \bigcap M_l,$$

vagyis a rendezés olyan, hogy minden tagjának az összes korábbival való metszete benne van valamelyik korábbi tagban. A P halmazt akkor nevezzük rendezett módon dekomponálhatónak, ha a hozzá tartozó **M** elemeinek van olyan rendezése, ami hierarchikus és rendezett módon dekomponálható. Ilyenkor magát a paraméterezést is rendezett módon dekomponálhatónak nevezzük.

Fontos, hogy a tulajdonság az **M** egy rendezéséhez köthető: előfordul, hogy van olyan rendezés, ami megfelel a definíciónak, és van olyan is, ami nem, de a tulajdonság meglétéhez elegendő egyetlen megfelelő rendezést találni.

Triviális, hogy P rendezett módon dekomponálható, ha az **M** egyelemű. Ilyen a hagyományos loglineáris rendezés, ahol **M={** V **}.** A (9)-es paraméterezéshez tartozó P is rendezett módon dekomponálható. Marginálisai V_1, V_1V_2, $V_1V_2V_3$, a keresett rendezésnek ezt a sorrendet kell tartania ahhoz, hogy (11) teljesüljön. A rendezés, mint láttuk, hierarchikus. Az első 3 tagjának maximális elemei közé egyetlen halmaz, $V_1V_2V_3$ tartozik, így a rendezett dekomponálhatóság triviális módon teljesül.

Ezzel szemben a (10)-es paraméterezés nem rendelkezik a tulajdonsággal. Akárhogyan rendezzük is az **M** = { $\emptyset$, V_1, V_3, V_1V_3, V_2, V_2V_3, V_1V_2, $V_1V_2V_3$ } halmaz elemeit, sosem teljesülnek a feltételek. Pl. a $\emptyset$, V_1, V_3, V_1V_3, V_2, V_2V_3, V_1V_2, $V_1V_2V_3$ rendezés mellett az első hét tag maximális elemeit, V_1V_3, V_2V_3, V_1V_2-t véve, $V_1V_3 \cup V_2V_3$ és V_1V_2 metszete, V_1V_2 nem részhalmaza sem V_1V_3-nak sem V_2V_3-nek. Általában is igaz, hogy a Glonek-McCollugh-féle paraméterezés, amilyen a (10)-es is, nem dekomponálható rendezett módon. Még általánosabban:

[14] Egy halmaz maximális elemeinek nevezzük azokat az elemeket, amikhez nincsen őket tartalmazó halmazbeli elem.

könnyen belátható, hogy ha **M**=**P**(V) és legalább három eleme van, akkor P nem dekomponálható rendezett módon.

Most már megadható a paraméterezésre olyan szükséges és elégséges feltétel, ami komponenseinek variációs függetlenségét biztosítja:

2. tétel Egy hierarchikus és teljes P halmazhoz tartozó $\tilde{\lambda}_P$ paraméterezés komponensei akkor és csak akkor függetlenek variációsan, ha a P rendezett módon dekomponálható.

A tétel segítségével tehát könnyen eldönthető bizonyos paraméterezésekről, hogy komponensei variációsan függetlenek-e. Ha igen, akkor a paraméterek értékei bárhogyan előírhatók, interpretációjuk is egyszerű. Felhasználható továbbá a tétel megfelelő paraméterezés előállítására is. Tekintsük pl. ismét a társadalmi státusz három indikátorával (V_1 iskolázottság, V_2 foglalkozási státusz, V_3 jövedelem) kapcsolatos példánkat. Tegyük fel, hogy csak a jövedelmet meghatározó marginális kapcsolatok, azaz a V_1V_3 ill. V_2V_3 marginálisok érdekelnek minket, ezekkel a paraméterekkel:

$$\lambda_{V_1}^{V_1V_3},\ \lambda_{V_3}^{V_1V_3},\ \lambda_{V_1V_3}^{V_1V_3},\ \lambda_{V_2}^{V_2V_3},\ \lambda_{V_2V_3}^{V_2V_3}.$$

Ez egy hierarchikus paraméterhalmaz, de nem teljes. Teljessé tehető a $V_1V_2V_3$ marginális megjelentetésével, és az összes eddig még nem szerepelt hatáshalmaznak a lehető legkorábbi helyen történő bevonásával:

$$\lambda_{\varnothing}^{V_1V_3},\ \lambda_{V_1}^{V_1V_3},\ \lambda_{V_3}^{V_1V_3},\ \lambda_{V_1V_3}^{V_1V_3},\ \lambda_{V_2}^{V_2V_3},\ \lambda_{V_2V_3}^{V_2V_3},\ \lambda_{V_1V_2}^{V_1V_2V_3},\ \lambda_{V_1V_2V_3}^{V_1V_2V_3}.$$

Mivel az így előállított P rendezett módon dekomponálható is, a tétel biztosítja, hogy a megfelelő $\tilde{\lambda}_P$ paraméterezés komponensei variációsan függetlenek. Ez a konstrukció teljessé tétellel általában is működik: bármely hierarchikus, de nem teljes paraméterezés teljessé tehető a V halmaznak az **M** marginálisok közé vonásával, és a hiányzó hatásoknak a sorban lehető legkorábbi szerepeltetésével.

Használható továbbá a tétel variációsan nem független paraméterezés „kijavítására” is. A (10) -es paraméterezés, mint láttuk, rendezett módon nem dekomponálható. Azonban ha a problémát okozó $\lambda_{V_1V_2}^{V_1V_2}$ paramétert a $\lambda_{V_1V_2}^{V_1V_2V_3}$ -re cseréljük, a kapott

$$\lambda_{\varnothing}^{\varnothing}, \lambda_{V_1}^{V_1}, \lambda_{V_3}^{V_3}, \lambda_{V_1V_3}^{V_1V_3}, \lambda_{V_2}^{V_2}, \lambda_{V_2V_3}^{V_2V_3}, \lambda_{V_1V_2}^{V_1V_2V_3}, \lambda_{V_1V_2V_3}^{V_1V_2V_3}$$

paraméterezés teljes, hierarchikus, és rendezett módon dekomponálható, tehát a hozzá tartozó $\tilde{\lambda}_P$ variációsan független.

A problémát okozó paraméterek kicserélgetése bármikor elvégezhető. Ez persze nem jelenti azt, hogy ha ezek a kutató számára koncepcionálisan fontos paraméterek, akkor nem hagyhatók változatlanul, de a tétel ilyen esetben is segítségünkre van azzal, hogy a variációs függőség bizonyításával figyelmeztet minket a paraméterek interpretációs korlátaira.

2.3.2 A modellek létezése

Mint már említettem, a variációs függetlenséggel kapcsolatba hozható másik lényeges kérdés a modellnek megfelelő eloszlás létezése. Általános esetben ugyanis a paraméterek előírt értékei ellentmondhatnak egymásnak. Fontos probléma tehát, hogy milyen általános feltételek mellett létezik a paraméterekre tett megszorításoknak megfelelő tábla.

6. példa Az eloszlás létezésének és a variációs függetlenségnek a kapcsolatára Rudas és Bergsma (2004) mutat egy egyszerű példát: tekintsük az *A*, *B* és *C* bináris változókat, és az alábbi paraméterértékeket:

$$\lambda_{\varnothing}^{A} = \log 8,\ \lambda_{A}^{A}(1) = 0,\ \lambda_{B}^{B}(1) = 0,\ \lambda_{C}^{C}(1) = 0,\ \lambda_{AB}^{AB}(1,1) = \frac{1}{4}\log\frac{1}{9},$$

$$\lambda_{AC}^{AC}(1,1) = \frac{1}{4}\log\frac{1}{9},\ \lambda_{BC}^{BC}(1,1) = \frac{1}{4}\log 9.$$

Mind a hét érték a marginális loglineáris paraméterek értékkészletébe esik, hiszen értékkészletük a (-∞,+∞) intervallum. Tehát a paraméterekhez külön-külön létezik megfelelő eloszlás. Ha a hét paraméter variációsan független lenne, létezne ezeknek a paraméterértékeknek együttesen megfelelő eloszlás is. Belátható ugyanakkor, hogy ilyen eloszlás nincs. A második, harmadik és negyedik paraméterérték szerint ugyanis a változók marginális eloszlása egyenletes. Az első paraméterérték szerint az *A* két kategóriájának log-gyakoriságának átlaga log8, és mivel az *A* eloszlása egyenletes, ez azt jelenti, hogy 16 megfigyelésünk van. Vagyis az *A*, *B* és *C* marginális eloszlása egyaránt (8, 8). Az utolsó három paraméter alapján az *AB* és *AC* marginális táblákban 1/9 az esélyhányados értéke, míg a *BC* táblában 9. Mivel a kétváltozós marginális kereszttáblák marginálisai és esélyhányadosai is ismertek, a táblák egyértelműen meghatározottak. Azonban nem kompatibilisek: egyszerű számolással adódik, hogy a három kétdimenziós táblából nem rakható össze háromdimenziós tábla. Tehát a megadott paraméterek nemlétező eloszlást (vagy más szóval üres modellt) definiálnak. Elmondhatjuk, hogy a paraméterek nem függetlenek variációsan: a megadott értékek együttese nem esik a paraméterezés értékkészletébe, tehát az nem áll elő egyszerűen az egyes paraméterek értékkészletének Descartes-szorzataként.

A példában látott összefüggés a paraméterek variációs függetlensége és az általuk definiált marginális loglineáris modellnek megfelelő eloszlás létezése között általánosan is teljesül:

3. tétel A variációsan független paraméterekkel definiált marginális log-affin modellek nem üresek.

Láttuk, hogy a variációs függetlenség hierarchikus és teljes *P* esetén a rendezett módon való dekomponálhatósággal érhető el, sőt a teljesség nem is feltétlenül kell, hiszen hierarchikus és rendezett módon dekomponálható, nem teljes paraméterezés a korábban már említett módon könnyen teljessé tehető. Tehát a tételt kimondhatnánk rendezett módon dekomponálható hierarchikus *P*-hez tartozó $\tilde{\lambda}_P$-re is.

A marginális loglineáris modellekre nézve egyszerűbb feltétel is előírható a modellnek megfelelő eloszlás létezésére:

4. tétel Ha az üres halmaz legfeljebb egy marginális mellett szerepel hatásként a *P*-ben, akkor a *P*-nek megfelelő paraméterekkel definiált marginális loglineáris modell nemüres.

A tétel bizonyításakor az használandó ki, hogy egyenletes eloszlás esetén minden, az üres halmaznál nagyobb hatást tartalmazó paraméter értéke 0, tehát ezeknek a paramétereknek bármely lineáris kombinációja 0.

A disszertációban marginális loglineáris modellekkel azonosítható grafikus modellekkel foglalkozom, a fenti tétel tehát ezeknek a modelleknek általánosságban is biztosítja a nemürességét.

2.3.3 Simaság

A paraméterek simasága (*smoothness*) esetén a cellagyakoriságok és a paraméterértékek közötti leképezés „szép leképezés". Az elnevezés onnan ered, hogy bizonyos értelemben a függvény képe ilyenkor valóban sima, egyenletes, kiszögellések és mélyedések nélküli. A paraméterek interpretációja szempontjából a simaság lényeges tulajdonság, a modell simasága pedig az aszimptotikus eloszlások teljesüléséhez szükséges (pl. hogy a maximum likelihood becslés elég nagy minta esetén normális eloszlással, a valószínűségi hányados tesztstatisztika khi-négyzet eloszlással közelíthető). Nem sima modellre ezek nem mindig teljesülnek (Drton, 2008). A disszertáció tárgyát képező feltételes függetlenségi modellek nem feltétlenül simák (erre a következőkben mutatok példát), ezért lényeges kérdés a simaságot garantáló feltételek létezése.

Precízen a simaság definíciója a következő:

12. definíció Legyen $\mathbf{B} \subseteq \mathbb{R}^k$ nyílt halmaz. Ekkor a θ: $F \to \mathbf{B}$ paramétert F t-dimenziós $(1 \le t \le k)$ *sima paraméterezés*ének nevezzük, ha az alábbi három tulajdonsággal rendelkezik:

R1: θ **B**-re képező homeomorfizmus[15]

R2: θ kétszeresen folytonosan differenciálható

R3: θ elsőrendű parciális deriváltjaihoz tartozó mátrix teljes rangú, rangja t.

A θ paramétert *sima paraméter*nek nevezzük, ha R2 és R3 teljesül. A $G \subseteq F$ modellt *sima modell*nek nevezzük, ha van a G és egy $\mathbf{B} \subseteq \mathbb{R}^k$ között értelmezett sima paraméterezése. Az R3 a paramétertér lokális dimenziójának konstans voltát állítja, ez a konstans az R3-ban szereplő t, amit a modell dimenziójának is szokás nevezni. A t értéke a modell illeszkedéstesztje szempontjából fontos kérdés, ugyanis a modell dimenziójának ismerete szükséges a valószínűséghányados-próba szabadságfokának megállapításához.

Marginális loglineáris paraméterek esetén a paraméterezés simaságára létezik elégséges feltétel:

5. tétel Hierarchikus és teljes P halmazhoz tartozó $\tilde{\lambda}_P$ paraméterezés sima paraméterezése F-nek.

A tétellel kapható $\tilde{\lambda}_P$ sima paraméterezés komponenseinek affin kombinációja sima paraméterezése a megfelelő marginális log-affin modellnek, vagyis a modell ilyenkor sima, és dimenziója is közvetlenül adódik:

[15] Homeomorfizmus: olyan kölcsönösen egyértelmű leképezés, ami folytonos, és az inverze is folytonos.

6. tétel A hierarchikus és teljes P halmazhoz tartozó $\tilde{\lambda}_P$ paraméterezésre épülő nemüres marginális log-affin modell sima modell, dimenziója az őt definiáló lineárisan független megkötések száma.

Az 3. tételből tudjuk, hogy variációsan független paraméterekkel nemüres modellt kapunk, hierarchikus és teljes P mellett pedig a variációsan függetlenség eléréséhez a 2. tétel szerint rendezett módon való dekomponálhatóság kell. Vagyis ha a tétel hierarchikus és teljes P-je rendezett módon dekomponálható is, akkor a marginális log-affin modell létezése is és simasága is garantált.

A teljesség a fenti két tételben megint elhagyható, ahogyan a 2. tételnél is láttuk, vagyis bármely hierarchikus P-hez létezik a P kibővítésével kapott sima paraméterezés. A hierarchikusság viszont nem hagyható el, ennek illusztrálására használjuk fel Dawid (1980) ismert példáját.

7. példa Definiáljuk a G modellt a következő két állítással:

$$A \perp B,\ A \perp B \mid C.$$

Ez a modell az 1. tétel szerint pl. a következő paramétermegkötésekkel adható meg:

$$\lambda_{AB}^{AB}(1,1) = \lambda_{AB}^{ABC}(1,1) = \lambda_{ABC}^{ABC}(1,1,1) = 0. \tag{13}$$

Ezek a paraméterek nem hierarchikus paraméterezést adnak, hiszen az AB hatás két különböző marginális mellett is szerepel. Azt viszont nem tudjuk, hogy sima-e a paraméterezés, mert tételünkben a hierarchikusság elégséges, de nem szükséges feltétele a simaságnak. Dawid lehet segítségünkre, aki bebizonyította, hogy a modell a következőkkel ekvivalens:

$$A \perp BC \text{ vagy } B \perp AC \text{ vagy mindkettő.} \tag{14}$$

Eszerint a modell két különböző, egymást metsző modell uniójaként áll elő. Vannak pozitív eloszlások, amelyek az első modellhez hozzátartoznak, de a másodikhoz nem, és fordítva. Ezért a modell nem lehet nyílt halmaznak sima leképezésnél vett őse (a precízebb indoklásra alább, az algebrai statisztikai megközelítés tárgyalásánál visszatérek). Vagyis a modellnek nem létezik sima paraméterezése. Tehát a nem hierarchikus, (13)-beli paraméterek nem szerepelhetnek sima paraméterezésben.

Zárójelben: a példa annak illusztrálására is kiváló, hogy ugyanannak a modellnek a marginális loglineáris paraméterekkel történő felparaméterezésének egyszerűsége/bonyolultsága erősen függ a marginálisok megválasztásától. Itt ha az AB, ABC marginálisok helyett csak az ABC marginálist használnák, a modell (14) alapján, az 1. tétel szerint a következő paraméter-megszorításokkal lenne megadható:

$$\lambda_{ABC}^{ABC}(1,1) = \lambda_{AB}^{ABC}(1,1) = 0 \text{ és } \lambda_{AC}^{ABC}(1,1) * \lambda_{BC}^{ABC}(1,1,1) = 0,$$

ami nem lineáris megkötés.

A példa általánosítható is:

7. tétel Ha ugyanaz a hatás két különböző marginális mellett is szerepel a paraméterezésben, akkor $\tilde{\lambda}_P$ nem sima.

A tétel segítségével példát mutathatunk arra (a definíciókból sejthető) állításra, hogy sima modellnek létezhet sima és nem sima paraméterezése is. Az

$$A \perp B,\ A \perp BC$$

modell ugyanis az 1. tétel szerint az alábbi megkötésekkel definiálható:

$$\lambda_{AB}^{AB} = \lambda_{AB}^{ABC} = \lambda_{ABC}^{ABC} = 0.$$

A 7. tétel szerint ez nem sima paraméterezés. Ugyanakkor mivel a második feltételes függetlenségből következik az első, a paramétermegkötések közül az első redundáns, éppen az, ami a nem-simaságot okozta. Ezt elhagyva a maradék két paraméter egyszerűen egészíthető ki hierarchikus és teljes, tehát sima paraméterezéssé.

Máskor nem ilyen könnyű a sima paraméterezésre való áttérés. Vegyük a következő két feltételes függetlenséget:

$$EF \perp AC \mid B,\ FG \perp AB \mid C. \tag{15}$$

Az *FA* hatásnak az 1. tétel szerinti paraméterezésben két különböző marginális mellett kellene szerepelnie, azaz az így adódó paraméterezés nem lenne sima. Ugyanakkor a két feltételes függetlenség az 3. fejezet Feltételes függetlenség c. alfejezetében található (21)-es feltételes függetlenségi ekvivalencia segítségével „kisebb”, problémamentes állításokra bontható. Az első állítás felbontása:

$$EF \perp AC \mid B \leftrightarrow \left(F \perp AC \mid B\right) \text{ és } \left(E \perp AC \mid BF\right),$$

míg a másodiké:

$$FG \perp AB \mid C \leftrightarrow \left(G \perp AB \mid CF\right) \text{ és } \left(F \perp AB \mid C\right).$$

A problematikus *FA* hatás most már ugyanazon marginális mellett szerepel a két esetben, és valóban, a négy kisebb feltételes függetlenséghez az 1. tétellel generált paraméterek könnyen egészíthetők ki hierarchikus és teljes paraméterezéssé, hiszen a szükséges marginálisok (*ABCF, ABCFE, ABCFG, ABCFGE*) sorrendje hierarchikus.

A (C1)-(C4) levezetési szabályok segítségével az is belátható, hogy a modellből levezethető az

$$F \perp A \mid CB,\ F \perp A \mid C$$

modell, ami a 30. oldalon szereplő Dawid-féle 7. példa egy alesete, tehát nem sima modell. Mint láttuk ugyanakkor, e modell által tartalmazott szűkebb (15)-ös modellnek van sima paraméterezése, vagyis a szűkebb modell sima. Tehát egy modell lehet sima akkor is, ha nem sima modell almodellje. És megfordítva is: nem sima modell tartalmazó modellje is lehet sima, erre triviális példa az, hogy a Dawid-féle nem sima modellt eredetileg definiáló, (15)-ben

található két feltételes függetlenség által külön-külön meghatározott modellek simák. Tehát nincsen a tartalmazásra visszavezethető olyan reláció, ami segítene eldönteni a simaságot.

A 7. tétel tehát segítségünkre lehet paraméterezések nem-simaságának bizonyítására. Az 5. tétel elégséges feltételt adva alkalmas egy elég általános osztályhoz tartozó paraméterezés simaságának bizonyítására, de szükséges és elégséges feltétel a paraméterezés simaságára nem ismert. Ezért a simaság cáfolata a 7. tétel alesetein kívül jelenlegi ismereteink szerint csak konkrét esetekben (adott modell mellett, a változók kategóriaszámának rögzítésével) végezhető el, pl. a Dawid példájában látott módon, expliciten megadva az unióként előállítható modell komponenseit. Ennek kivitelezése algebrai statisztikai eszközökkel történik. Az algebrai statisztika új és érdekes megközelítés (pl. Drton, Sullivant, 2007), a statisztikai modelleket algebrai szemszögből vizsgálja, a modell statisztikai tulajdonságait algebrai, topológia, geometriai tulajdonságokra visszavezetve. Jelen problémánk esetében az algebrai megközelítés azért merül fel, mert kategoriális változók feltételes függetlensége az együttes eloszláshoz tartozó valószínűségekre tett polinomiális megkötésekhez vezet, ezeknek a 0-val való egyenlőségét előírva (Pachter, Sturmfels, 2005, Stillman, 2005). Pl. belátható, hogy $X \perp Y \mid Z$ esetén a cellavalószínűségekre a 3. definícióban felírt összefüggés ekvivalens az alábbi egyenlőséggel:

$$P\left(X = x, Y = y, Z = z\right) = P\left(X = x, Z = z\right) P\left(Y = y, Z = z\right) / P(Z = z),$$

vagyis az alábbi polinom 0-val való egyenlőségével:

$$P\left(X = x, Y = y, Z = z\right) P(Z = z) - P\left(X = x, Z = z\right) P\left(Y = y, Z = z\right) = 0.$$

Egy feltételes függetlenségi modell ennek megfelelően polinomok egy ideáljával[16] azonosítható. A polinomideálok véges sok primér ideál[17] (*primary ideal*) metszeteként állíthatók elő, egy ilyen primér felbontás (*primary decomposition*) létezése esetén a függetlenségi modell bizonyos polinomiális megkötések uniójaként áll elő. A 30. oldalon szereplő Dawid-féle 7. példa modelljéről is belátható ezen a módon, hogy a megfelelő ideál két primér ideálra bontható, azaz két polinomhalmaz 0-val való egyenlőségének vagylagosan való teljesülésével definiálható, tehát előállítható két modell uniójaként. Az algebrai megközelítést geometriaira váltva elmondhatjuk, hogy a modellbe tartozó eloszlások halmaza tehát két komponens uniójaként áll elő, és a két komponens metszetében levő pontok nem reguláris tulajdonságai miatt az unió nem lehet sima.

A primér felbontás megtalálására egyszerűbb modellek esetén léteznek szabadon letölthető algebrai szoftverek, pl. a SINGULAR[18] szoftver Markov.lib nevű könyvtára[19] használható erre a célra.

[16] Az ideál algebrai fogalom. A feltételes függetlenségi modellbe tartozó, az F egy alterét alkotó eloszlásokon 0 értéket felvevő polinomok ideált alkotnak, ami annyit tesz, hogy (1) bármely két, az ideálhoz tartozó polinom különbsége is az ideálban van, illetve (2) az ideálba tartozó polinomok tetszőleges polinommal való szorzata is az ideálban van.

[17] Egy polinomideál primér ideál, ha minden *a, b* polinomra, amikor az *ab* szorzat az ideálban van, akkor vagy *a*, vagy *b*-nek önmagával véges számszor vett szorzata (véges hatványa) is az ideálban van.

[18] http://www.singular.uni-kl.de/

[19] Luis Garcia írta: http://www.shsu.edu/~ldg005/software.html

2.3.4 A becslések standard aszimptotikus tulajdonságai

A paraméterezés itt utolsóként felsorolt előnyös tulajdonsága a standard nagymintás elmélet alkalmazhatósága. A valószínűségek maximum likelihood becslésének és a valószínűség-hányados tesztstatisztikának aszimptotikus (elég nagy minta esetén várható) viselkedése tartozik ide.

Ahogy a többi tulajdonságok, úgy ez a tulajdonság sem teljesül általános esetben a marginális loglineáris modellekre. A 30. oldalon szereplő, sokat idézett 7. példa modellje, mint láttuk, nem volt sima, két modell uniójaként állt elő. A két modell metszetében levő elméleti eloszlások esetén akármilyen nagy mintát veszünk is, a likelihoodnak pozitív valószínűséggel a modell mindkét komponensében lokális maximuma van. A valószínűséghányados statisztika pedig aszimptotikusan nem khi-négyzet eloszlású, hanem két khi-négyzet eloszlás minimumaként áll elő. Figyelmeztető, hogy ezek szerint olyan egyszerű modellek sem kezelhetők feltétlenül jól, mint amilyen ez is ($A \perp B,\ A \perp B \mid C$ - pl. „az egyetemi felvételi sikeressége független a felvételiző nemétől, összességében és szakonként is”).

Megfelelően megválasztott marginális loglineáris paraméterekre ugyanakkor teljesülnek a standard aszimptotikus tulajdonságok.

8. tétel Tegyük fel, hogy egy nemüres log-affin marginális modell sima paraméterezéssel van megadva. Ekkor akár Poisson, akár multinomiális eloszlás[20] esetén

a) A mintanagyság végtelenhez tartásával[21] 1-hez tart annak a valószínűsége is, hogy a π cellavalószínűség-vektor maximum likelihood becslése, $\hat{\pi}$ létezik, és hogy előáll a likelihood egyenlet stacionárius pontjaként[22].

b) Ha N a mintanagyság, akkor a $N^{1/2}\left(\hat{\pi}-\pi\right)$ aszimptotikusan normális eloszlású, 0 várható értékkel.

c) A valószínűség-hányados statisztika aszimptotikusan khi-négyzet eloszlású, szabadságfoka pedig a modell dimenziójával, tehát a modellt definiáló lineárisan független megkötések számával egyezik meg.

---

[20] Poisson eloszlás feltételezése esetén egy adott cella gyakorisága független a többi cellák gyakoriságától, a cellagyakoriságok és azok összege is Poisson eloszlást követ. Ebből következően ilyenkor az összgyakoriság nem rögzített. Multinomiális eloszlás esetén az összgyakoriság rögzített, így a cellagyakoriságok nem függetlenek. Az előbbi főleg biostatisztikai és biztosítási matematikai alkalmazásokban fordul elő (pl. adott idő-intervallumban megjelenő új betegek vizsgálata egy populációban, ahol a rizikófaktorok adják a változókat, vagy adott időintervallum baleseti statisztikája, ahol a változók a baleset típusai és a megelőző intézkedések fajtái lehetnek). Multinomiális eloszlást tételezhetünk fel a rögzített mintanagyságot célzó társadalomtudományi felmérések többsége esetében.

[21] A mintanagyság növelése úgy értendő, hogy a kontingenciatábla egy adott (Poisson vagy multinomiális) eloszlásból származik, ennek egy megvalósulását (tehát magát a táblát) tekintjük egy mintaelemnek, és ebből az eloszlásból veszünk többet, *N*-et (ekkor N darab táblánk van), majd ezt az N-t növeljük.

[22] Ez a következő feltételes szélsőérték-keresési probléma megoldásában játszik szerepet. A likelihood-egyenlet, pontosabban a Lagrange-féle likelihood-egyenlet valósértékű többváltozós függvény, a cellavalószínűségek vektorán és az un. Lagrange multiplikátoron, mint rendezett páron van értelmezve. A Lagrange-multiplikátor szereplésének oka, hogy segítségével az egyenlet maximumhelye a modellt definiáló megszorítások mellett kereshető meg. Az egyenlet stacionárius pontjai pedig azok az értelmezési tartománybeli pontok, ahol a gradiens (a parciális deriváltak vektora) 0-t vesz fel, vagyis ezek az egyenlet potenciális lokális szélsőérték-helyei.

Annak, hogy a maximum likelihood becslés a likelihood egyenlet stacionárius pontja, a becslés gyakorlati megvalósításában van szerepe. Ilyenkor megfelelő algoritmusok léteznek a maximum megkeresésére. A modell illeszkedése tehát (a harmadik állítás szerint) valószínűség-hányados próbával tesztelhető, a modellnek megfelelő maximum likelihood-becslés (az első állítás szerint) egyértelműen létezik és előállítható, továbbá a becslés (a második állítás alapján) torzítatlan és konfidencia-intervallum konstruálása is lehetséges hozzá.

A 2. tétel, 3. tétel, 4. tétel és 6. tétel felhasználásával az a)-c) tulajdonságok teljesüléséhez elegendő, ha a modellt definiáló marginális loglineáris paraméterek hierarchikus és teljes P-hez tartozik. Egyéb esetekben a tulajdonságok teljesülése nem garantált.

3 GRAFIKUS MODELLEK

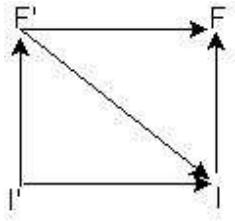
3.1 Háttér

A grafikus modellek használata ma az alkalmazott statisztika számtalan területén – statisztikus fizika, genetika, mesterséges intelligencia – elterjedt. Az elmélet történetileg is több szaktudományhoz kapcsolható: három, egymástól független szerző munkáira vezethető vissza. S. Wright a genetika, W. Gibbs a fizika, míg A. A. Markov a valószínűség-számítás területén vezette be a grafikus modell fogalmát (Wermuth 2005). Wright a mai szociológiában elerjedt szóhasználattal *útelemzés*nek nevezte azt a módszert, amikor irányított gráfot használt bizonyos genetikai jellemzők kialakulásának megjelenítésére. Gibbs fizikai rendszerek energiájának meghatározásakor a rendszert alkotó részek (pl. gázatomok) *szomszéd*sági relációit használta fel. A grafikus modellek alapfogalmát, a *feltételes függetlenség*et Markov komplex struktúrák egyszerűbb függőségi láncokkal történő leírására használta. Az 1970-es években indult meg a grafikus modellek elméletének kidolgozása, irányítatlan (csak szimmetrikus relációkat megengedő) gráfok, illetve irányított (csak aszimmetrikus relációkat megengedő) körmentes gráfok vizsgálatával. Napjainkban az irányítatlan gráfoknak például térstatisztikai és képfeldolgozási felhasználásai ismertek. Az irányított körmentes gráfok, melyek Bayes-féle hálózatokként (*Bayesian networks*) is ismertek, elsősorban ok-okozati viszonyok feltárásával kapcsolatos kutatásokban kerülnek elő gyakran (pl. Spirtes et al., 1993), és a mesterséges intelligenciakutatás szakértői rendszerekkel (*expert systems*) foglalkozó ága alkalmazza őket (pl. Lauritzen, Spiegelharter, 1988, Pearl, 1988). A tipikus kérdésfeltevések közé tartozik, hogy hogyan konstruálható hatékony algoritmus (un. tanuló algoritmus) egy adott eloszláshoz illeszkedő gráf konstruálására (pl. Spirtes et al., 1993).

Az utóbbi évtizedben az elméletet a mindkét relációfajtát megengedő láncgráf modellekre is általánosították. A láncgráfok létrehozásának gyakorlati motivációja az volt, hogy gyakori az olyan alkalmazási kontextus, amikor bizonyos változócsoportok között aszimmetrikus kapcsolat van, míg a változócsoportokon belül szimmetrikusak a relációk. A változócsoportokat pl. longitudinális vizsgálat időpont szerint rendezett változóhalmazai, vagy keresztmetszeti vizsgálat ok-okozati sorba rendezhető magyarázó, közbejövő és kimeneti változóhalmazai alkothatják. A láncgráfok leírása még nem véglegesedett, több kutató foglalkozik velük párhuzamosan, gyakran más-más megközelítésekben (Whittaker, 1990, Frydenberg, 1990a, Andersson, Madigan, Perlman, 2001, Richardson, 2003, Cox, Wermuth, 1996, Wermuth, Cox, 2004, Wermuth, 2005, Drton, 2008).

A grafikus modellek matematikai leírására térve: a modellek a változókat gráfpontokként jelenítik meg, a köztük levő irány nélküli/iránnyal rendelkező kapcsolatot élekkel/nyilakkal reprezentálva. (Az iránnyal rendelkező kapcsolat nem feltétlenül jelent ok-okozati viszonyt, lásd a 6.1.1. fejezetet.) Két pont között az él/nyíl hiánya feltételes függetlenségi állításoknak feleltethető meg. A grafikus modellek előnye, hogy könnyen vizualizálható és jól interpretálható módon jelenítenek meg komplex rendszereket, illetve, hogy modulárisak, vagyis e komplex rendszereket egyszerűbb részek összekapcsolásával építik fel.

Példaként tekintsük az alkalmazások között, a 7.1. fejezetben is szereplő irányított körmentes gráfot, ami a státuszelérés egy klasszikus modelljét ábrázolja.



1. ábra. Egy gráf-modell (I: iskolázottság, F: foglalkozás, I': apa iskolázottsága, F': apa foglalkozása)

A szereplő változók a korábbi fejezetekben már sokszor szerepelt státusz-indikátorok. A modell az Apa iskolázottságával, mint háttérváltozóval „indul", a Foglalkozással, mint függő változóval „végződik", és két közbejövő változót tartalmaz, amelyek potenciálisan a korábbiaktól függenek, ők maguk pedig a későbbiek lehetséges magyarázó változói. A rendszer modularitása jól látható: az Apa foglalkozásával, a Foglalkozással és az Iskolázottsággal, mint függő változóval meghatározott modulokat három külön regressziós modellként foghatjuk fel.

Másik társadalomtudományi alkalmazásként említhetjük a 1.1., Motiváció c. fejezetben látott gráfot, melyet Cox és Wermuth (2001) állított fel az állami intervenciókkal kapcsolatos attitűd kialakulását modellezve. A modell az attitűd meghatározódását jeleníti meg az objektív társadalmi-demográfiai tényezők csoportjától kiindulva az azoktól függő szubjektív és objektív társadalmi státusz-indikátorok csoportján át. A láncgráf lehetővé teszi, hogy az említett csoportok között iránnyal rendelkező, a csoportokon belül irány nélküli kapcsolatokat tételezzünk fel. Ez a megközelítés mélyebb betekintést enged ebbe a folyamatba annál, mint ha csak a szokott módon többszörös regressziót alkalmaznánk az attitűddel, mint függő változóval, és az összes többi változót azonos szintű magyarázó változóként kezelve.

3.2 Gráfelméleti definíciók és jelölések

Mivel a következőkben gráfokkal reprezentálható modelleket fogok vizsgálni, szükséges lesz néhány gráfelméleti fogalom és jelölés bevezetése.

Gráfnak nevezzük a $G = (V, E)$ rendezett párt, ahol V a gráf *csúcsainak* véges halmaza, az E élek pedig a $V \times V$ különböző elemekből álló rendezett párok részhalmazát adják. Ebből következően az itt tárgyalt gráfok egyszerűek, tehát nincs többszörös élük és nincsenek adott csúcsból adott csúcsba visszakanyarodó hurkaik.

Ha egy $(\alpha, \beta) \in E$ esetén (β, α) is eleme E-nek, akkor az élt *irányítatlan él*-nek vagy *vonal*-nak (*line*) nevezzük, ellenkező esetben *irányított él*ről vagy *nyíl*ról beszélünk. A gráfok megjelenítésekor a csúcsokat pontokkal, az irányítatlan éleket vonalakkal, míg az (α, β) irányított éleket α–ból β–ba mutató nyilakkal jelöljük majd.

G' a *G* *részgráf*ja, ha *G*-ből élek és pontok elhagyásával kapjuk, az egyetlen feltétel, hogy ha elhagyunk egy pontot, akkor természetesen elhagyjuk az összes belőle induló élt is. *Teljes* részgráfnak nevezzük a *G'*-t, ha bármely két csúcsa között fut él.

α–ból β–ba mutató nyíl esetén α–t a β *szülő*jének, β-t az α *gyerek*ének nevezzük. β szüleinek halmazát $pa(\beta)$ jelöli (*parents*), α gyerekeinek (*children*) halmazát $ch(\alpha)$. Ha α és β között

vonal van, akkor ők *szomszéd*ok (*neighbours*). α szomszédjainak halmazát $ne(\alpha)$–vel jelöljük. Ha sem vonal sem nyíl nincs közöttük, akkor *nem-szomszédok (non-adjacent)*.

Ha nem csúcsról, hanem csúcsok egy $A \subset V$ részhalmazáról beszélünk, akkor az A szülő-, gyerek- és szomszéd-halmazaiba maga A sosincs beleértve:

$$pa(A) = \bigcup_{\alpha \in A} pa(\alpha) \setminus A,$$

$$ch(A) = \bigcup_{\alpha \in A} ch(\alpha) \setminus A,$$

$$ne(A) = \bigcup_{\alpha \in A} ne(\alpha) \setminus A.$$

Csúcsok egy A részhalmazának *határa (boundary)* az a $bd(A)$ halmaz, amihez A szülei vagy szomszédjai tartoznak:

$$bd(A) = pa(A) \cup ne(A).$$

Az A csúcshalmaz *lezártja (closure)* a $cl(A) = bd(A) \cup A$ halmaz.

Egy α-ból β-ba vezető n hosszú *út* az $\alpha = \alpha_0, ..., \alpha_n = \beta$, páronként különböző csúcsok sorozata, ahol $(\alpha_{i-1}, \alpha_i) \in E$, minden $i = 0, ..., n$-re. Ha az út csúcsai között csak nyilak/csak vonalak vannak, akkor *irányított/irányítatlan út*ról beszélünk.

Egy $C \subset V$ részhalmazt (α, β)*-elválasztó*nak (*separator*) nevezünk, ha az α-ból β-ba vezető utak mindegyike átmegy C valamely csúcsán. A C *elválasztja* az A halmazt a B-től, ha minden olyan (α, β) párra elválasztó, ahol $\alpha \in A, \beta \in B$. Azok a csúcsok, amelyekből út vezet β-ba, de amelyekbe nem vezet út β-ból, alkotják β *ősei*nek (*ancestors*) $an(\beta)$ halmazát. Fordítva: azok a csúcsok, amelyekbe út vezet α-ból, de amelyekből nem vezet út α-ba, alkotják α *leszármazottai*nak (*descendants*) $de(\beta)$ halmazát. α *nem-leszármazottai* (*non-descendants*) pedig a többi csúcs közül természetesen azok, amelyek nem leszármazottai: $nd(\alpha) = V \setminus (de(\alpha) \cup \{\alpha\})$. Nagybetűvel jelöljük a leszármazottak halmazát, ha az az α-t is tartalmazza: $Nd(\alpha) = nd(\alpha) \cup \{\alpha\}$. A leszármazottak és a nem-leszármazottak halmazát értelemszerűen A csúcshalmazokra is definiálhatjuk: $de(A)$, $nd(A)$, $Nd(A)$.

Ha $bd(\alpha) \subseteq A, \forall \alpha \in A,$ azt mondjuk, hogy A *őshalmaz (ancestral set)*. Irányított gráfban A akkor és csak akkor őshalmaz, ha $an(\alpha) \subseteq A, \forall \alpha \in A.$ Őshalmazok metszete is őshalmaz, ezért bármely A csúcshalmazra meghatározható az őt tartalmazó legkisebb őshalmaz, ezt $An(A)$-val jelöljük.

A *kör* az út definíciójának azzal a módosításával kapható, hogy a kezdő és végző csúcs ugyanaz kell, hogy legyen. *Irányított/irányítatlan/félig irányított kör*nek nevezzük a csak nyilat/csak vonalakat/legalább egy nyilat tartalmazó köröket.

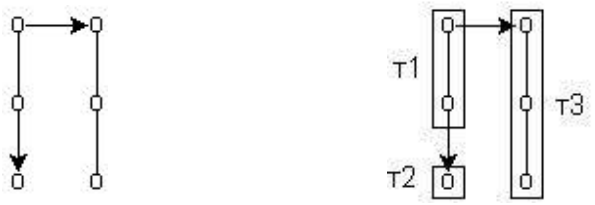
*Irányítatlan gráf*nak (IG, angol rövidítéssel UG az *undirected graph*-ból) nevezzük a csak vonalakat tartalmazó gráfot, míg *irányított körmentes gráf*nak (röv. IKG, angol rövidítéssel DAG a *directed acyclic graph*-ból) a csak nyilakat tartalmazó, irányított körrel nem rendelkező gráfot.

A disszertációban bemutatott harmadik speciális gráftípus a *láncgráf* (LG, angol rövidítéssel CG a *chain graph*-ból). Egészen friss munkák foglalkoznak ezzel a gráftípussal (Whittaker, 1990, Frydenberg, 1990a, Lauritzen, 1996, Cox, Wermuth, 1996, Wermuth, Cox, 2004, Wermuth, 2005, Andersson et al., 2001, Lauritzen, Richardson, 2002, Drton, 2008). Az egyes szerzők, bár ritkán reflektálnak erre a tényre, nem használnak egységes definíciót. A definíciók különbözőségének komoly következményei vannak a gráfra épülő statisztikai modellek szempontjából, ezt az 3.6., Láncgráfok c. fejezetben részletesen is kifejtem.

Alapvetően kétfajta definíció rajzolódik ki a fent idézett munkákból. Lauritzen, Frydenberg, Drton ill. Andersson és társai szerint a láncgráf olyan gráf, ami vonalakat és nyilakat is tartalmazhat, azzal a megkötéssel, hogy nincsen benne félig-irányított kör. A továbbiakban *1-es típusú láncgráf*nak nevezem ezt a típust. Vegyük észre, hogy az IG-k és IKG-k a láncgráfok speciális (vagy csak vonalakat vagy csak nyilakat tartalmazó) típusai. A *láncgráf komponensei*nek vagy *blokk*oknak (*component, block*) nevezzük az egymásból irányítatlan úton elérhető csúcsok halmazát. Precízen:

Jelöljük $\alpha \sim_G \beta$-val azt az esetet, amikor a G gráfban α és β elérhető egymásból irányítatlan úton. Legyen T a $\sim_G$ ekvivalencia-reláció által V-ben indukált ekvivalenciaosztályok halmaza. Minden $\alpha \in V$ csúcshoz egyértelműen tartozik egy $\tau(\alpha) \in T$ komponens.

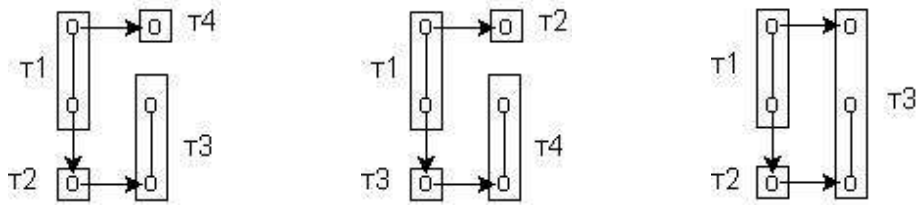
Az alábbi ábra egy 1-es típusú láncgráfot és a neki megfelelő τ_1, τ_2, τ_3, komponenseket ábrázolja. felhívnám rá a figyelmet, hogy a komponensek indexelése ennél a típusnál esetleges.



2. ábra. 1-es típusú láncgráf, és a gráf komponensei

Whittaker (1990) illetve Cox és Wermuth (pl. 1996) meghatározása szerint a láncgráf (*2-es típusú láncgráf*) csúcsai n számú komponens valamelyikéhez vannak hozzárendelve, a komponenseken pedig teljes rendezés van bevezetve: $\tau_1 < \ldots < \tau_n$. A komponensek ennek megfelelő $1 \ldots n$ indexekkel vannak ellátva, mindegyikhez egy vagy több csúcs tartozik. A komponensek rendezése részbenrendezést indukál a csúcsokon – két különböző komponenshez tartozó csúcsról tudjuk, hogy melyik a korábbi/későbbi, két azonos komponensbe tartozó csúcs nem hasonlítható össze. Két azonos komponensbe tartozó csúcs között vonal húzható. Két különböző komponensbe tartozó csúcs között nyíl lehet, de a nyíl a korábbi csúcsból kell, hogy a későbbi csúcsba mutasson. Vegyük észre, hogy a félig-irányított köröknek az 1-es típusú láncgráfok definíciójában megkövetelt hiánya itt is teljesül.

Az alábbi három gráf 2-es típusú gráf:



3. ábra. 2-es típusú láncgráfok, $\tau_1<\tau_2<\tau_3<\tau_4$

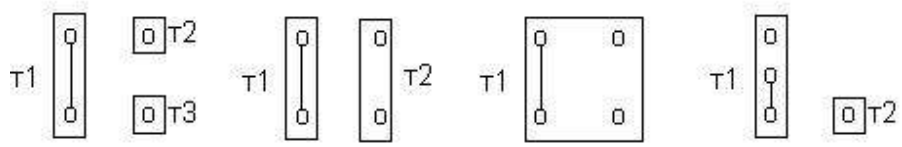
Ez a példa közvetlenül rámutat az 1-es és 2-es típusú láncgráfok közötti fontos különbségekre. Míg az 1-es típus esetén a gráfot pusztán a csúcsok és élek definiálják (a komponensek ebből adódnak), addig a 2-es típus esetén a csúcsok komponensekhez tartozása ill. a komponensek rendezése ehhez adódó pluszinformáció, nem olvasható ki a csúcsok-élek megadásából. Így fordulhat elő az, hogy a 3. ábra három gráfja ugyanazokon a csúcsokon és éleken van értelmezve, de más a komponensek megadása (1. vs. 3. gráf), illetve más a komponensek rendezése (1. vs. 2. gráf).

A 2-es típusú láncgráfokkal szemben az 1-es típusúaknál nincsen a komponenseken teljes rendezés bevezetve. Ha két komponens nincs nyíllal nincs összekötve, akkor azok nem hasonlíthatók össze. A 2. ábra mutatott gráf τ_2 és τ_3 komponense például nem hasonlítható össze. Ebből a gráfból úgy készíthetnénk 2-es típusú láncgráfot, ha expliciten megadnánk a τ_2 és τ_3 közötti rendezést.

Az 1-es típusú láncgráfoknál a gráfból közvetlenül adódik a csúcsok komponensekhez rendelése, a 2-es típusúaknál nem. Emiatt a különbség miatt fordulhat elő, hogy van olyan gráf (pl. a 3. ábra 3. gráfja), ami a 2-es definíció szerint láncgráf, de az 1-es szerint nem az, hiszen van olyan csúcsa, amelyik egyik saját komponensbelijével sincsen összekötve.

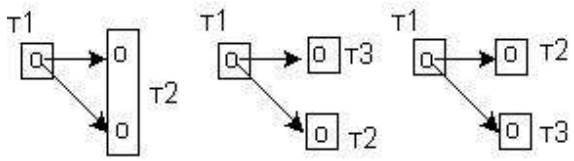
Azok a gráfok tartoznak a két típus mindegyikébe, ahol a komponensek csúcsai irányítatlan úton elérhetők egymásból, és a komponensek közötti nyilak teljes rendezést indukálnak a komponenseken.

Míg, ahogyan említettük, az 1-es típusú láncgráfnak az IG és IKG speciális esete, addig egyik gráf sem tartozik a 2-es típusú láncgráfok közé, mivel az IG-ből vagy IKG-ből nem lehet egyértelműen kiolvasni sem a komponensek rendezését, sem a komponenshez tartozást. Az alábbi ábrán olyan 2-es típusú láncgráfok találhatók, melyek ugyanahhoz az IG-hez tartoznak. A komponensekhez rendelés variálásával jött létre ez a négy különböző változat. A komponensek rendezésének megváltoztatásával további variánsokat lehetne előállítani.



4. ábra. Ugyanahhoz az IG-hez rendelhető 2-es típusú LG-k, $\tau_1<\tau_2<\tau_3$

Az alábbi ábrán pedig ugyanahhoz az IKG-hez rendelhető összes lehetséges 2-es típusú láncgráfot mutatom be:



5. ábra. Ugyanahhoz az IKG-hez rendelhető 2-es típusú LG-k, $\tau_1<\tau_2<\tau_3$

Mint láttuk, a két definíció alapvető különbsége, hogy a 2-es típus esetén (1) a csúcsok komponensekhez tartozása ill. (2) a komponensek rendezése a gráfból ki nem olvasható pluszinformáció. Láttuk, hogy ez a két eltérés a két definíciónak megfelelő gráfhalmaz különbözőségét implikálja. Az igazi kérdés ugyanakkor az, hogy a megfelelő statisztikai modellek is különböznek-e. Erre a kérdésre a 3.6. fejezetben, a gráfok és a modellek megfeleltetésének tárgyalása után térek vissza.

A szakirodalom gyakran eltekint a két típusú lángráf közötti különbségtől; az idézett szerzők közül egyedül Whittaker (1990) említi az (1)-es, Lauritzen és Richardson (2003) pedig a (2)-es különbséget, utóbbiak kritikáját is adják a komponensek 2-es típusú láncgráfoknál kötelező elő-rendezésének, mint modellezési technikának. Az (1)-es különbséggel kapcsolatban felhívnám arra a figyelmet, hogy 1-es típusú láncgráfoknál össze nem kötött csúcsok ugyanazon komponensbe való besorolása nem lehetséges. Ha a komponenseket időbeli egymásutánjuk definiálja, akkor eszerint ennél a gráftípusnál egyidőben, de egymástól függetlenül megvalósult események modellezése nem lehetséges.

Itt kell még megemlítenünk, hogy Wermuth és Cox (pl. 2004) a 2-es típusú láncgráfok további specifikálásaként megengednek nem csak teli, hanem szaggatott vonalakat és nyilakat is, azzal a megkötéssel, hogy egy adott komponensen belül csak egyfajta (vagy csak teli, vagy csak szaggatott) vonalak szerepelhetnek, ill. az egy adott komponensbe belépő nyilak is csak azonos fajták lehetnek. A szaggatott és teli vonalak/nyilak megkülönböztetésének oka, hogy ezek segítségével egyetlen szabállyal (később bevezetendő szóhasználattal: egyetlen Markov-tulajdonsággal) több különböző feltételes függetlenség-kiolvasási szabályt definiálhatnak. Mi ugyanezt a célt több különböző szabály bevezetésével fogjuk elérni (lásd az 3.6 fejezetet).

Szükséges lesz még néhány, a láncgráfokhoz kapcsolódó jelölés bevezetésére. A G láncgráf $K_1, ..., K_n$ komponensei, mint csúcsok alkotta irányított körmentes gráfon (jel.: G_{IKG}) kézenfekvő módon definiálhatók adott K_i komponens ősei, szülei, leszármazottai vagy nem-leszármazottai:

- ősök:

  $K_{an}(i) := \{K_j \mid j \in \{1...n\},$ a G_{IKG} gráfban K_j–ből irányított út vezet K_i-be$\}$,

- ősök, K_i-t is beleértve:

  $K_{An}(i) := \{K_j \mid j \in \{1...n\},$ a G_{IKG} gráfban K_j–ből irányított út vezet K_i-be vagy $j=i\}$,

- szülők:

  $K_{pa}(i) := \{K_j \mid j \in \{1...n\}, \},$ a G_{IKG} gráfban K_j–ből nyíl mutat K_i-be$\}$,

- leszármazottak:

  $K_{de}(i) := \{K_j \mid j \in \{1...n\},$ a G_{IKG} gráfban K_i–ből irányított út vezet K_j-be $\}$,

- nem-leszármazottak:

  $K_{nd}(i) := \{K_j \mid j \in \{1...n\}, K_j$ nem eleme $K_{de}(i)$-nek és $j \neq i\}$,

- és nem-leszármazottak, K_i-t is beleértve:

  $K_{Nd}(i) := \{K_j \mid j \in \{1...n\}, K_j$ nem eleme $K_{de}(i)$-nek$\}$.

Adott láncgráfból kapható irányítatlan gráf két megadási módjára lesz még szükségünk az elválasztás fogalmának láncgráfokra való majdani kiterjesztéséhez.

*Erkölcstelenség*nek (*immorality*) nevezzük a (α, β; γ) ponthármast, ha a közös γ gyerekkel rendelkező α és β csúcspár nincsen vonallal összekötve (úgymond nincs összeházasodva – innét az elnevezés). Egy G LG-hez vagy IKG-hoz tartozó *erkölcsös gráfnak* (*moral graph*) nevezzük és G^m-mel jelöljük azt az IG-t, ami a közös gyerekkel rendelkező szülőknek vonallal való összekötésével (összeházasításával) és az összes nyíl vonalra cserélésével áll elő.

Legyen α, β, és γ egy láncgráf három különböző csúcsa. *Zászló*nak (*flag*) nevezzük az [α, β; γ] hármast, ha $\alpha \rightarrow \beta - \gamma$ teljesül. *2-kettőszászló*nak (*2-biflag*) nevezzük az [α, β; γ, δ] négyest, ha $\alpha \rightarrow \gamma - \delta \leftarrow \beta$ teljesül; ahol α és β között bármi lehet (vonal is, bármilyen irányú nyíl is, vagy lehetnek összekötetlenek is), de más vonal vagy nyíl nincs a csúcsok között. *Hármas*nak (*triplex*) nevezzük az $\{(\alpha, \beta), \gamma\}$ rendezett párt, ha a következő három feltétel valamelyike teljesül: (α, β; γ) erkölcstelenség, vagy [α, β; γ] zászló, vagy [β, α; γ] zászló. Egy hármas vagy egy 2-kettőszászló kibővítésekor (*augmentation*) az alakzatot alkotó csúcsok között húzódó nyilakat vonalakra cseréljük, és minden hiányzó él helyére vonalat teszünk. Egy G LG-hez vagy IKG-hoz tartozó *kibővített gráfnak* (*augmented graph*) nevezzük és G^a-val jelöljük azt az IG-t, ami a hármasok és a 2-kettőszászlók kibővítésével és az összes nyíl vonalra cserélésével áll elő.

3.3 Feltételes függetlenség

A grafikus modellek valószínűségi változóknak a gráfból kiolvasható feltételes függetlenségeivel definiálhatók, ezért a feltételes függetlenség a következőkben központi fogalom lesz.

X, *Y* és *Z* diszkrét valószínűségi változók feltételes függetlensége a 3. definícióban látott módon határozható meg, jelöléssel: $X \perp Y \mid Z$. A feltételes függetlenségi reláció rendelkezik az alábbi négy tulajdonsággal (*X-et*, *Y-t* és *Z-t* akár változóknak, akár változóhalmazoknak tekinthetjük):

($C1$) szimmetria: ha $X \perp Y \mid Z$, akkor $Y \perp X \mid Z$, (16)

($C2$) felbontás: ha $X \perp Y \mid Z$ és $U = h(X)$, akkor $U \perp Y \mid Z$, (17)

($C3$) gyenge unió: ha $X \perp Y \mid Z$ és $U = h(X)$, akkor $X \perp Y \mid (Z,U)$, (18)

($C4$) összevonás: ha $X \perp Y \mid Z$ és $X \perp W \mid (Y,Z)$, akkor $X \perp (W,Y) \mid Z$. (19)

(C2) és (C3) alkalmazásakor speciális esetként az U lehet X részhalmaza is.

Pozitív eloszlás[23] esetén, vagyis ha

$$P(X = x, Y = y, Z = z) > 0 \ \forall x, y, z,$$

az alábbi, gyakran használt tulajdonság is teljesül:

($C5$) metszet: ha $X \perp Y \mid Z$ és $X \perp Z \mid Y$, akkor $X \perp (Y,Z)$. (20)

A pozitív eloszlás kitételét a korábbiakban is feltettük (a 2. fejezetben végig a pozitív eloszlások *F* halmazát tekintettük), és legtöbbször a továbbiakban is fel fogjuk tenni.

A (C1)-(C4) tulajdonságokból további tulajdonságok vezethetők le. Például az alábbi, a feltételes függetlenségi állítás szétdarabolhatóságát megmutató, gyakran hasznosnak bizonyuló ekvivalencia

$$C \perp BD \mid A \leftrightarrow (C \perp B \mid A) \text{ és } (C \perp D \mid AB) \tag{21}$$

könnyen előáll a segítségükkel. A jobboldalból (C4)-et felhasználva következik a baloldal. A baloldali feltételes függetlenségi állításból a (C2)-t felhasználva következik $C \perp B \mid A$. A

[23] Ez a tulajdonság a változók keresztáblájára nézve azt jelenti, hogy az az elméleti gyakoriság mellett nem tartalmaz üres cellát, vagy másik szóhasználattal: nincsen benne strukturális nulla. Ez egy eléggé természetes kitétel, és általános esetben valóban teljesül is. Tipikusan olyan esetekben nem igaz, amikor a változók között valamilyen szükségszerű, logikai kapcsolat van. Pl. ilyen kapcsolatban van a gazdasági aktivitás és a legutóbbi foglalkozás típusa, ha a tanulók mindkét változónál önálló kategóriát alkotnak, tehát sorukban/oszlopukban egy kivétellel minden cellagyakoriság strukturálisan nulla, vagy a nem és az abortuszok száma (férfiaknál értelmezhetetlen) stb. A strukturális nullák egy része a változók átdefiniálásával kiküszöbölhető (pl. keresztkategóriáikkal egy harmadik változót képzünk).

baloldali állításból (C3) segítségével jön $C \perp BD \mid AB$, ebből pedig (C2)-vel $C \perp D \mid AB$. A tulajdonság felhasználhatóságára láttunk példát a 2.3 fejezet Simaság c. alfejezetében, ahol egy látszólag nem sima modell simaságát láttuk be a segítségével, úgy, hogy a modellt definiáló nemhierarchikus paraméterezést adó feltételes függetlenségi állításokat kisebb, hierarchikus paraméterezést lehetővé tevő állításokká bontottuk a segítségével.

Belátható, hogy a (21)-beli tulajdonság nem csak szintaktikailag vezethető le a négy szabály segítségével, hanem valóban igaz is[24]. Kérdés, hogy (1) ez minden, a (C1)-(C4) segítségével levezetett állításra igaz-e? Kérdés a fordított irány is (2): levezethető-e a négy szabály segítségével minden (igaz) feltételes függetlenségi tulajdonság?

Dawid (1979) bizonyította be, hogy az első kérdésre „igen" a válasz, vagyis ha (C1)-(C4)-et a feltételes függetlenség, mint logikai rendszer levezetési szabályainak tekintjük, akkor az így definiált rendszer *helyes*: minden, a négy szabály használatával levezethető állítás igaz. Bár sokan a második kérdésre is „igen" választ vártak (pl. Pearl, 1988), végül mégis a „nem" nyert bizonyítást. A rendszer *nem teljes* (Studený, 1992), vagyis van olyan igaz állítás, ami szintaktikailag nem vezethető le belőlük. Sőt az is belátható, hogy a feltételes függetlenségnek nem létezik véges axiomatizációja, vagyis sem a (C1)-(C4) kibővítésével, sem más levezetési szabályok segítségével nem lehet az összes, egyébként igaz feltételes függetlenségi tulajdonságot előállítani. Ez utóbbi állításnak fontos következményei lesznek a disszertáció későbbi fejezeteiben.

Rendkívül érdekes, és a grafikus modellek széleskörű alkalmazhatóságát mutatja, hogy a (C1)-(C4) tulajdonságok milyen sok, látszólag független fogalom karakterizációjára alkalmasak. Tekintsük ehhez a (C1)-(C4) kifejezéseket formális kifejezésekként, a valószínűséghez nem kötődő jelentéssel. Ha az *irrelevancia* axiómáiként tekintjük őket, akkor tudásmennyiségek megszerzése (pl. nyelvkönyvek elolvasása) közötti kapcsolat leírására alkalmazhatók – valóban, ismertek a grafikus modellek információelméleti alkalmazásai is. Ebben az esetben a $X \perp Y \mid Z$ kifejezés így interpretálható: *Z ismeretének birtokában Y elolvasása (értsd: ismerete) irreleváns X elolvasása (értsd: megértése) szempontjából.* Ekkor pl. a C2 így olvasandó ki: *ha Z ismeretének birtokában Y elolvasása irreleváns X elolvasása szempontjából, akkor X bármely U fejezetének elolvasása szempontjából is az.*

Másik fontos példa az euklidészi vektorterekben definiált geometriai merőlegességé. Ekkor *X*, *Y* és *Z* lineáris alterek, és $X \perp Y \mid Z$ így olvasandó ki: *X és Z metszetének ortogonális kiegészítő altere merőleges Y és Z metszetének ortogonális kiegészítő alterére.*

Végül a véges irányítatlan gráfokon definiált elválasztás fogalma említhető (definícióját lásd a következő fejezetben). Ekkor *X*, *Y* és *Z* a gráf csúcsainak részhalmazai, és $X \perp Y \mid Z$ így olvasandó ki: *Z elválasztja X-et Y-tól*. Később látni fogjuk, hogy a globális Markov-tulajdonságokkal definiált grafikus modellek valóban a függetlenség és az elválasztás megfeleltetésére épülnek.

A (C5) analogonja a feltételes függetlenséghez hasonlóan az utóbbi három modell (irrelevancia, merőlegesség, elválasztás) esetében sem áll általánosan, csak bizonyos feltételek

[24] Ebben a megközelítésben egy formula igaz, ha minden valószínűségi mérték mellett kiértékelve igaz.

mellett. Van azonban olyan tulajdonság, ami nem viselkedik azonosan a különböző modellekben. Az elválasztásra teljesül, a feltételes függetlenségre vagy az irrelevanciára viszont nem:

$$(C6)\ \left(A \perp B \mid C\right) \text{ és } \left(A \perp D \mid C\right) \rightarrow A \perp BD \mid C \tag{22}$$

Erről a grafikus modellekkel kapcsolatos irodalomban és a disszertációban is gyakran előkerülő tulajdonságról könnyen belátható, hogy pozitív sűrűségfüggvény esetén sem feltétlenül teljesül. Többváltozós normális eloszlás esetén viszont mindig igaz. Ez egy alapvető differenciát jelent a többváltozós kategoriális és a többváltozós normális eloszlás között. Ez az eltérés az alapja annak, hogy kategoriális esetre érdemes a páronkénti Markov-tulajdonság (ami a fenti implikáció bal oldalán található páronkénti feltételes függetlenségeket enged meg) mellett egyéb Markov tulajdonságokat is bevezetni, normális eloszlás esetére viszont nem.

A (C5) egy variánsát is meg kell itt említenem, mert a továbbiakban szintén szerepelni fog:

$$\text{(C5*) ha } A \perp B \mid (C \cup D) \text{ és } A \perp C \mid (B \cup D), \text{ akkor } A \perp (B \cup C) \mid D), \tag{23}$$

A (C5*), akárcsak a (C5), például akkor teljesül, ha az eloszlás pozitív.

A továbbiakban a feltételes függetlenségi állításokkal definiálható gráfmodellek három típusát (irányítatlan gráf, irányított körmentes gráf, láncgráf) veszem sorra. Az első két típus bemutatása, a kapcsolódó tételekkel, hacsak nem jelzem másként, Lauritzen 1996-os összefoglaló munkájára épül.

3.4 Irányítatlan gráfok

A grafikus modellek legegyszerűbb típusa irányítatlan gráfokkal reprezentálható. A gráf minden éle irányítatlan, ennek megfelelően minden változót egyenrangúnak tekintünk.

A gráf csúcsainak és a hozzájuk rendelt változóknak a formális megfeleltetése a következő módon adható meg. Tekintsük a $G = (V, E)$ gráfot. A gráfhoz a csúcsokkal indexelt $(X_\alpha)_{\alpha \in V}$ diszkrét valószínűségi változók tartoznak, $(\chi_\alpha)_{\alpha \in V}$ véges értékkészletekkel. Ha $A \subset V$ a csúcsok egy halmaza, akkor a hozzá tartozó $X_A := (X_\alpha)_{\alpha \in A}$ változó értékkészlete az $\chi_A := \times_{\alpha \in V} \chi_\alpha$ Descartes szorzat, ennek egy adott értékét $x_A := (x_\alpha)_{\alpha \in A}$.–val jelöljük röviden. A V-hez tartozó változó értékkészletét $\chi := \chi_V$ jelöli. A továbbiakban az egyszerűség kedvéért az

$$A \perp B \mid C$$

jelölést alkalmazzuk az

$$X_A \perp X_B \mid X_C$$

esetre.

Most már formálisan is definiálhatjuk a grafikus Markov-modelleket: ezek egy $G = (V, E)$ gráfot felhasználva, annak V csúcshalmaza segítségével valamely szabály (un. Markov-tulajdonság) szerint feltételes függetlenségi állításokat definiálnak a csúcsoknak megfeleltetett $X_\alpha, \alpha \in V$ változók között. Irányítatlan gráf esetén háromfajta Markov-tulajdonságot szokás

megkülönböztetni. Ezek a gráf hasonló értelmezését adják: valamilyen módon az „egymással közvetlenül nem érintkező” csúcsok feltételes függetlenségét mondják ki a „velük közvetlenül érintkező” csúcsokat feltéve. Amiben különböznek, az a két feltételesen független halmaz számossága: az első tulajdonság két csúcsról, a második egy csúcsról és egy (tartalmazásra nézve lehető legnagyobb) csúcshalmazról, a harmadik két csúcshalmazról beszél. A második tulajdonság esetén adott változó független attól a változóhalmaztól, ami a vele az első tulajdonság alapján független változókat tartalmazza. Mint a 3.3. fejezetben már említettük, a második, lokális Markov-tulajdonság bevezetésének éppen a kategoriális esetben van jelentősége, normális eloszlásnál ez felesleges: utóbbinál a (C5) teljesülése folytán a páronkénti $A \perp B$ és $A \perp C$ feltételes függetlenségek ekvivalensek az együttes függetlenséggel, $A \perp BC$ -vel.

Azt mondjuk, hogy a G gráf mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik a ...

13. definíció (P) páronkénti (*pairwise*) Markov-tulajdonsággal, ha bármely α, β nem-szomszédos pár feltételesen független az összes többi csúcsot feltéve:

$$\alpha \perp \beta \mid V \setminus \{\alpha, \beta\},$$

14. definíció (L) lokális (*local*) Markov-tulajdonsággal, ha bármely $\alpha \in V$ feltételesen független a vele össze nem kötött csúcsoktól a vele összekötött csúcsokat feltéve:

$$\alpha \perp V \setminus cl(\alpha) \mid bd(\alpha),$$

15. definíció (G) globális (*global*) Markov-tulajdonsággal, ha a csúcsok bármely (A, B, S) diszjunkt halmazhármasára, ahol S elválasztja A-t B-től:

$$A \perp B \mid S.$$

E harmadik, globális tulajdonságnak kitüntetett szerepe van, mert ha teljesül, akkor segítségével bármely *A, B, S* diszjunkt csúcshalmazról eldönthető, hogy igaz-e A és B S-re vett feltételes függetlensége. Ez a kritérium nem szigorítható, abban az értelemben, hogy ha A és B nincs elválasztva S által, akkor található olyan eloszlás, ami mellett $A \perp B \mid S$ nem teljesül.

A Markov-tulajdonságok között az alábbi kapcsolat áll fenn (Lauritzen, 1996):

9. tétel Bármely irányítatlan G gráf és bármely, χ -en adott valószínűség-eloszlás mellett

$$(G) \Rightarrow (L) \Rightarrow (P).$$

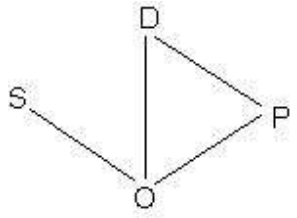
A tétel szerint a globális Markov-tulajdonság a legerősebb abban az értelemben, hogy a neki megfelelő feltételes függetlenségi állítások implikálják a másik két Markov-tulajdonságnak megfelelő állításokat is. Érdemes megjegyezni, hogy mivel e tétel bizonyításakor elég a (C1)-(C4) levezetési szabályokat felhasználni, az állítás igaz a tulajdonságok más modellje (pl. irrelevancia, merőlegesség) mellett is. Nem túl szigorú feltétel mellett a három tulajdonság ekvivalenciája is teljesül. Ez a feltétel az 3.3 fejezetben definiált (C5*). Igaz tehát a tétel (Lauritzen, 1996):

10. tétel Ha bármely páronként diszjunkt *A, B, C, D* halmazra fennáll (C5*), akkor:

$$(G) \Leftrightarrow (L) \Leftrightarrow (P).$$

A (C5*) például akkor teljesül, ha az eloszlás pozitív. Mint korábban említettem, a pozitivitást, ezt az eléggé természetes feltételt a korábbiakban is feltettük és legtöbbször a későbbiekben is feltesszük. A három tulajdonság ekvivalenciájának jelentőségét az adja, hogy eszerint elég a globális tulajdonság használata, ezzel a tulajdonsággal pedig bármely $A \perp B \mid S$ függetlenségi állításról eldönthető, hogy teljesül-e egy adott gráf mellett. Ugyanis elég azt megvizsgálni, hogy az A és B csúcshalmazokat az S csúcshalmaz elválasztja-e.

8. példa Az IG-re épülő Markov-modell példájaként tekintsük az alábbi ábra gráfját, ami néhány, a politikai közvélemény-kutatásokban gyakran használt változó egy lehetséges kapcsolatrendszerét mutatja be. A változók a következők: az Ország gazdasági helyzetével (O), a Saját gazdasági helyzettel (S), a Demokráciával való elégedettség (D) és a kormányon levő Párt megítélése (P).



Ha a gráfot a páronkénti Markov-tulajdonság mellett értelmezzük, akkor a modell szerint az össze nem kötött csúcsok feltételesen függetlenek, vagyis a két hiányzó élhez tartozó két alábbi állítás teljesül:

$$P \perp S \mid OD,$$
$$D \perp S \mid OP.$$

Ezek szerint a politikai attitűddel kapcsolatos változók (P ill. D) nincsenek közvetlen kapcsolatban a saját gazdasági helyzet megítélésével (S). A saját helyzet megítélése csak az ország helyzetének megítélésén keresztül gyakorol hatást a politikai attitűdökre. A PS ill. DS kétdimenziós marginális kereszttáblában esetleg megfigyelhető marginális kapcsolatuk annak következménye, hogy az O-val mind P, mind S, mind D össze van kötve, tehát kapcsolatban van. Pl. akik jobb helyzetben érzik magukat, azok hajlamosabbak az ország helyzetét is pozitívabban értékelni, az ország helyzetével elégedettek pedig hajlamosak a demokrácia állapotát és a kormányt is pozitívabban megítélni. De a feltételes kereszttáblában már nem figyelhető meg kapcsolat: pl. az OD-t egy értékén rögzítve (csak azokat nézve, akik elégedettek az ország helyzetével és a demokráciával is) a saját helyzet megítélése független a pártpreferenciától.

Ha a globális Markov-tulajdonság mellett értelmezzük a gráfot, akkor, mivel például az S-ből a DP halmazba csak az O-n keresztül vezet út, a modell szerint az alábbi állítás is igaz:

$$DP \perp S \mid O.$$

Eszerint a politikai attitűdváltozók együttes eloszlása sincs közvetlen kapcsolatban a saját gazdasági helyzet megítélésével, S-sel. Itt természetes feltevés a változók pozitív eloszlása (nincs okunk üresnek feltételezni a kontingenciatábla egyik celláját sem), ezért a 10. tétel szerint az utolsó feltételes függetlenség nem jelent újabb megszorítást, következik a korábbiakból.

3.5 Irányított körmentes gráfok

Az irányítatlan gráfok esetében a változók között egyenrangú kapcsolatot tételeztünk fel. Előfordulnak ugyanakkor olyan esetek is, amikor a változók szereposztása nem szimmetrikus, ami annak felel meg, hogy X hathat Y-ra, de Y nem hathat X-re[25]. Ilyen szituációban az irányítatlan gráfok használata természetellenes lenne, ez motiválta az irányított körmentes gráfok (IKG) bevezetését, ahol a változók közti nyilak az aszimmetrikus függő-független szereposztást jelölik. A 11. ábra modellje jól példázza, mi motiválja az aszimetrikus szerepek bevezetését. Itt egyetlen hatásirány tételezhető csak fel logikailag ill. a kialakulás időrendjét alapul véve pl. a társadalmi háttér és a felsőfokú továbbtanulási elhatározás között, vagy a nem és a szülői bíztatás között.

IKG esetén is az irányítatlan esetben már látott, ott ekvivalensnek bizonyult háromfajta Markov-tulajdonságot szokás megkülönböztetni. Azt mondjuk, hogy a G gráf mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik ...

16. definíció (IP) ... az *irányított páronkénti Markov-tulajdonság*gal, ha bármely α, β nem-szomszédos párra, ahol $\beta \in nd(\alpha)$:

$$\alpha \perp \beta \mid nd(\alpha) \setminus \{\beta\},$$

vagyis α feltételesen független nem-leszármazottaitól (kivéve a szüleit) a többi nem-leszármazottat feltéve. Itt érdemes megjegyezni, hogy a feltételhalmazban mindig benne van az összes szülő;

17. definíció (IL) ... az *irányított lokális Markov-tulajdonság*gal, ha bármely $\alpha \in V$ feltételesen független nem-leszármazottaitól a szüleire, mint feltételre nézve:

$$\alpha \perp \left(nd(\alpha) \setminus pa(\alpha)\right) \mid pa(\alpha).$$

IG-k esetében a globális Markov-tulajdonság az elválasztás fogalmára épült. Irányított esetben ennek megfelelőjeként az IKG-hoz tartozó irányítatlan erkölcsös gráfra vezetjük vissza a tulajdonságot:

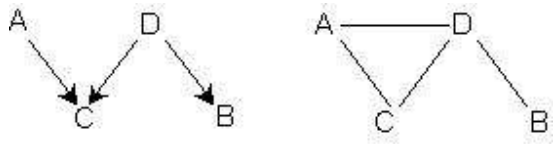
18. definíció (IG) A valószínűségi mérték rendelkezik az *irányított globális Markov-tulajdonság*gal, ha a csúcsok bármely (A, B, S) halmazhármasára, amelyek olyanok, hogy S elválasztja A-t B-től a $\left(G_{An(A \cup B \cup S)}\right)^m$-ben (vagyis az $A \cup B \cup S$-t tartalmazó legkisebb őshalmazhoz tartozó erkölcsös gráfban), fennáll:

$$A \perp B \mid S.$$

9. példa Mint már az IG-kel kapcsolatban említettük, a globális tulajdonság jelentőségét az adja, hogy segítségével dönteni tudunk tetszőleges csúcshalmazoknak egy harmadik csúcshalmazra vett feltételes függetlenségéről. Döntsük el a például a IG tulajdonságból

[25] Itt a hatás fogalmát nagyon általános értelemben használom, az ok-okozatiság problematikáját elkerülve (erről bővebben lásd majd a 6.1.1. fejezetet.

kiindulva, hogy az alábbi, baloldali IKG szerint A és B független-e a C-t feltéve? Az $A \cup B \cup C$ halmaz legkisebb őshalmazához a D is hozzátartozik, mert C és B szülője. Az így kapott legkisebb őshalmazhoz tartozó erkölcsös gráfhoz (alábbi jobb oldali gráf) az AD élet is hozzá kellett venni, mert közös gyerekük van. Az erkölcsös gráfban viszont ezért C nem választja el A-t és B-t, hiszen a D-n keresztül is vezet út köztük, tehát az eredeti IKG modellben a IG tulajdonság mellett A nem független B-től C-re, mint feltételre nézve.



A Markov-tulajdonságok között itt is tartalmazási reláció áll fent, mégpedig az irányítatlan esetnél több is elmondható (Lauritzen, 1996):

11. tétel Bármely IKG gráf és bármely, χ-en adott valószínűség-eloszlás mellett

$$(DG) \Leftrightarrow (DL) \Rightarrow (DP).$$

Az IG esetéhez hasonlóan a (C5*) itt is ekvivalenciát implikál (Lauritzen, 1996):

12. tétel Bármely IKG gráf és bármely, χ-en adott valószínűség-eloszlás mellett, ha (C5*) teljesül, akkor

$$(G) \Leftrightarrow (L) \Leftrightarrow (P).$$

Társadalomtudományi gráfmodelleket véve példának: ha az irányított lokális Markov-tulajdonság szerint értelmezzük a 70. oldalon található 11. ábra modelljét, akkor a modell szerint a felsőfokú továbbtanulási elhatározás feltételesen független nem-leszármazottjától, vagyis a nemtől, szüleire, azaz az összes többi változóra, mint feltételre nézve. A nem független továbbá az intelligenciától és a társadalmi háttértől, méghozzá feltétel nélkül, hiszen a nemnek nincs szülője. Belátható, hogy az irányított globális Markov-tulajdonság itt nem hoz új feltételes függetlenségeket.

3.6 Láncgráfok

3.6.1 Bevezető

Az irányított körmentes gráfok minden éle aszimmetrikus kapcsolatot reprezentál. Kézenfekvőnek látszik ugyanakkor olyan modellek bevezetése is, ahol ezekkel együtt megengedünk szimmetrikus, irány nélküli kapcsolatokat is. Ezek pl. olyan esetek lehetnek, ahol

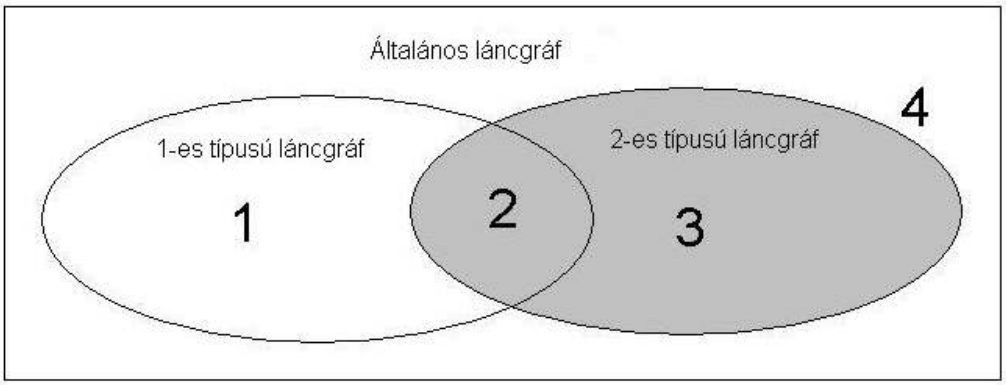
- nem ismert, hogy x okozza y-t vagy fordítva (nyíl lenne a megfelelő, de nem tudjuk, hogyan irányítsuk),
- vagy ahol y visszahat x-re (*feedback,* lásd az iskolázottság és a foglalkozás közötti nyíl vonalra cserélését a 7.1. fejezetben, abból a megfontolásból, hogy a szocialista rendszerben a politikailag megbízható, de aluliskolázott munkaerő kékgalléros vezető

pozíciókba való juttatásának utólagos legitimizációjaként felnőttkori beiskolázásokat hajtottak végre; vagy mai példával: a műszaki-technikai haladás révén egy adott szakmában is állandósulhat a változás, amellyel az érintetteknek lépést kell tartaniuk, ami viszont csak az ismeretek folyamatos gyarapításával lehetséges),

- vagy ahol az x és y közötti kapcsolatot egy harmadik, nem mért változó okozza (lásd a gólyák számának és a születésszámnak ismert esetét a 6.1.2. fejezetben).

Az irányított és irányítatlan éleket is tartalmazó gráfokat láncgráfoknak nevezzük. Mint definiálásuknál, a Gráfelméleti definíciók és jelölések c. fejezetben láttuk, két különböző és nem kompatibilis meghatározás is található velük kapcsolatban a szakirodalomban (1-es és 2-es típus). Ugyanakkor nem csak a gráfelméleti definíció, hanem a hozzájuk kapcsolódó feltételes függetlenségek kiolvasására szolgáló Markov-tulajdonságok meghatározása sem egységes. Az alábbiakban ezért mindig jelölni fogom, hogy mely szerzők megközelítése alapján dolgozom. Az IG-k és IKG-k esetén korábban látott három (páronkénti, lokális és globális) Markov-tulajdonság mellett a láncgráfok esetében blokk-rekurzív (block-recursive) Markov-tulajdonságot is megkülönböztetünk. Míg a páronkénti tulajdonság két csúcs, a lokális egy csúcs és egy csúcshalmaz közötti feltételes függetlenségről szól, addig a blokk-rekurzív és a globális tulajdonság két csúcshalmaz függetlenségéről. A blokk-rekurzív tulajdonság a komponensek (blokkok) rekurzitivitására épül. A globális tulajdonság itt is az elválasztás fogalmát használja, mint az IG-k és IKG-k esetén. Ezeket a tulajdonságokat veszem sorra a továbbiakban.

A fejezetben a tételek mindegyike az 1-es típusú láncgráfokra lesz kimondva, mert a kapcsolódó szerzők ebben a definíciós körben gondolkodtak. Nem ismert, hogy ezek a tételek igazak-e a 2-es típusú láncgráfokra is. Valójában az is kérdés, hogy az 1-es és 2-es láncgráf típusok a gráfelméleti különbségen túl a statisztikai modellt tekintve is különböznek-e. A szakirodalom ugyanis, ahogyan általában eltekint az 1-es és 2-es típusú láncgráf explicit megkülönböztetésétől, ugyanúgy nem vizsgálja azt a kérdést sem, hogy egy adott Markov-tulajdonság a két típusú gráfon definiálva azonos modellosztályt eredményez-e. A kérdés precízen így szól: adott Markov-tulajdonság mellett van-e olyan, 2-es típusú láncgráffal kapott feltételes függetlenségi állítás-halmaz, ami nem ekvivalens az 1-es típusú láncgráfokkal kapott állításhalmazok egyikével sem, és fordítva. Sőt az is kérdéses, hogy az „általános” láncgráfoknak mi a viszonya a két altípussal, ahol az általános láncgráfokat a két típus egyesítéseként úgy határozhatjuk meg, hogy komponenseik definiálása az 1-es típushoz hasonlóan, „előre” történik, de a komponensek a 2-es típushoz hasonlóan nem rendezettek. Vagyis kérdés, hogy van-e modell az alábbi halmazábra 1-essel, 3-assal és 4-essel jelölt részhalmazaiban.



6. ábra. Az általános, illetve az 1-es ill. 2-es típusú láncgráfokon, valamely Markov-tulajdonsággal definiált modellosztályok tartalmazási viszonya (sötét háttér: nemüres, fehér háttér: nem ismert, hogy üres-e)

Saját, a későbbiekben bemutatott eredményem szerint (lásd a 11. példa) a 3-as részhalmaz nemüres, tehát van olyan, 2-es típusú láncgráf-modell, ami nem állítható elő 1-es típusú láncgráffal. Ebből következően az 1-es típusú láncgráfokra kimondott, az alábbiakban bemutatott tételek nem feltétlenül teljesülnek a 2-es típusra. Ezt a tényt az általam ismert szakirodalom teljesen figyelmen kívül hagyja.

3.6.2 Páronkénti Markov-tulajdonságok

Néhány feltételes függetlenségi tulajdonság definiálásával kell kezdenünk.

Ha α és β ugyanazon K_i komponensbe tartozik, és az őket összekötő *vonal* hiányzik, akkor a két csúcs feltételesen független komponensükre és az összes olyan komponenst feltéve, amely a K_i-nek nem leszármazottja:

$$\alpha \perp \beta \mid \{K_{Nd}(i) \setminus \beta\}. \tag{24}$$

Ha α és β ugyanazon K_i komponensbe tartozik, és az őket összekötő *vonal* hiányzik, akkor a két csúcs feltételesen független az összes olyan komponenst feltéve, amely a K_i-nek nem leszármazottja:

$$\alpha \perp \beta \mid \{K_{nd}(i) \setminus \beta\}. \tag{25}$$

Ha β a $K_{nd}(i)$ eleme (azaz az α K_i komponensének nem-leszármazottjai közé tartozó komponensbe tartozik), és a β-ból az α-ba menő *nyíl* hiányzik, akkor a két csúcs feltételesen független α komponensére és az összes, nem-leszármazott komponenst feltéve:

$$\alpha \perp \beta \mid \{K_{Nd}(i) \setminus \beta\}, \tag{26}$$

Ha β a $K_{nd}(i)$ eleme (azaz az α K_i komponensének nem-leszármazottjai közé tartozó komponensbe tartozik), és a β-ból az α-ba menő *nyíl* hiányzik, akkor a két csúcs feltételesen független az összes, nem-leszármazott komponenst feltéve:

$$\alpha \perp \beta \mid \{K_{nd}(i) \setminus \beta\}. \tag{27}$$

Vegyük észre, hogy az első és harmadik tulajdonság csak abban különbözik a másodiktól és a negyediktől, hogy az utóbbiaknál az α-t tartalmazó komponens nem tartozik bele a feltételhalmazba. Továbbá: ha 2-es típusú láncgráfról van szó, ahol a komponensek teljesen rendezettek, akkor a feltételben levő $K_{nd}(i)$ és $K_{nd}(i)$ halmazok a $K_{an}(i)$ és $K_{An}(i)$ halmazokká egyszerűsödnek.

Mint korábban, a láncgráfok definiálásakor az 3.2. fejezetben említettem, Wermuth és Cox (pl. 2004) a 2-es típusú láncgráfok további specifikálásaként megengednek nem csak teli, hanem szaggatott vonalakat és nyilakat is. A fenti négy tulajdonság felsorolásuk sorrendjében az ő megfogalmazásukban a hiányzó teli vonal, hiányzó szaggatott vonal, hiányzó teli nyíl ill. hiányzó szaggatott nyíl esetéhez tartoznak.

E tulajdonságok kombinálásával négy fajta páronkénti Markov-tulajdonság definiálható.

19. definíció (LP1) Azt mondjuk, hogy a G gráf mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik a láncgráfokra definiált 1-es típusú *páronkénti* Markov-tulajdonsággal, ha (24) és (26) teljesül.

Az ezzel a Markov-tulajdonsággal definiált modelleket elsőként Lauritzen és Wermuth (1989) illetve Frydenberg (1990a) tanulmányozta, ezért a szakirodalom gyakran (pl. Andersson et al., 2001, Drton, 2008) páronkénti LWF Markov-tulajdonságként utal rá. A modellt Wermuth és Cox (2004) teli vonalakkal és teli nyilakkal ábrázolja, és blokk-regressziós (*block-regression model*) vagy blokkolt-koncentrációs (*blocked-concentration*) modellnek nevezi[26].

20. definíció (LP2) A valószínűségi mérték rendelkezik a láncgráfokra definiált 2-es típusú *páronkénti* Markov-tulajdonsággal, ha (24) és (27) teljesül.

A szakirodalom (Andersson et al., 2001, Drton, 2008) páronkénti alternatív Markov-tulajdonságként, röviden AMP Markov-tulajdonságként utal erre a tulajdonságra. A tulajdonsággal definiált modellt Wermuth és Cox (2004) teli vonallal és szaggatott nyíllal ábrázolja, és koncentráció-regressziós (*concentration-regression*) modellnek nevezi.

21. definíció (LP3) A valószínűségi mérték rendelkezik a láncgráfokra definiált 3-es típusú *páronkénti* Markov-tulajdonsággal, ha (25) és (26) teljesül.

A (Wermuth és Cox jelölését követve) szaggatott vonallal és teli nyíllal ábrázolható modelleket Wermuth és Cox (2004) röviden említi, Rudas, Bergsma és Németh (2009) pedig részletesebben is tárgyalja.

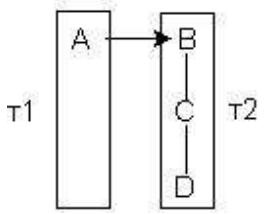
22. definíció (LP4) A valószínűségi mérték rendelkezik a láncgráfokra definiált 4-es típusú *páronkénti* Markov-tulajdonsággal, ha (25) és (27) teljesül.

Ezeket, a szaggatott vonallal és szaggatott nyíllal ábrázolható modelleket Wermuth és Cox (2004) többváltozós regressziós modelleknek nevezi. Richardson (pl. 2003) említendő itt még meg, aki az IKG-knak egy sajátos általánosításával, az irányított körmentes keverék gráfokkal

[26] Megjegyezném, hogy ők ezt és a következő tulajdonságokat is 2-es típusú gráfon definiálják.

(*acyclic directed mixed graphs*) foglalkozik. Ezek a gráfok a szokásos egyirányú nyilakon kívül kétirányú (*bi-directed*) nyilakat (↔) is tartalmaznak. Richardson ezeken a gráfokon speciális Markov-tulajdonságokat vezet be, pl. az m-szeparáció kritériumát. Belátható, hogy a 4-es típusú páronkénti Markov-tulajdonsággal adott modellek Richardsonnak az m-szeparációval az irányított körmentes keverék gráfon definiált modelljeivel tartalmazási kapcsolatban vannak. Precízen: az m-szeparáció implikálja a 4-es típusú páronkénti Markov-tulajdonságot, de a fordított implikáció nem áll általánosságban, csak a folytonos esetben, normális eloszlást feltételezve.[27]

10. példa A négy Markov-tulajdonság használatára példaként tekintsük az alábbi láncgráfot. (Az egyszerűség kedvéért olyan gráfot választottam, ami az 1-es és 2-es típusú láncgráf definíciójának is megfelel.)



7. ábra. Egy láncgráf

LP1 mellet a gráfmodell az alábbi feltételes függetlenségek teljesülésével ekvivalens:

$$B \perp D \mid CA,$$
$$C \perp A \mid BD,$$
$$D \perp A \mid BC.$$

LP2 mellett hiányzó nyíl esetén a saját komponensbeliek nem szerepelnek a feltételben, ezért az utolsó kettő állítás feltételhalmaza szűkül:

$$B \perp D \mid CA,$$
$$C \perp A,$$
$$D \perp A.$$

LP3 az LP1-hez képest a hiányzó vonal esetén jelent szűkebb feltételhalmazt:

$$B \perp D \mid A,$$
$$C \perp A \mid BD,$$
$$D \perp A \mid BC.$$

[27] Richardson említését azért tartottam fontosnak, mert megközelítését több helyen (pl. Wermuth, Cox, 2004) Wermuth és Cox szaggatott nyílt és szaggatott vonalat tartalmazó modelljével azonosítják, pedig, mint azt a tétel mutatja, nem áll fenn azonosság. A tartalmazási tétel bizonyítását Richardson egy nekem írt levelében vázolta, erre itt nem térnék ki.

LP4 pedig mind a hiányzó vonal, mind a hiányzó nyíl esetére szűkebb feltételhalmazt jelent:

$$B \perp D \mid A,$$
$$C \perp A,$$
$$D \perp A.$$

Tehát az LP1 implikálja a „leginkább feltételes” függetlenséget, LP4 a „leginkább marginális” függetlenséget, LP2 és LP3 pedig a kettő között van.

11. példa Ebben a példában mutatom meg, hogy a 6. ábra 3-as részhalmaza nem üres. Definiáljuk a *G* 2-es típusú gráfot a következőképpen: legyen *A, B* és *C* a *G* gráf három csúcsa, tartozzanak mindhárman ugyanahhoz a komponenshez, közülük szaggatott vonalak hiányozzanak (azaz értelmezzük a gráfot az LP3 vagy LP4 mellett). Ekkor három marginális függetlenség áll elő:

$$A \perp B,\ A \perp C,\ B \perp C.$$

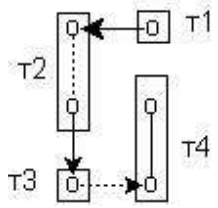
Könnyen belátható, hogy ez a három állítás nem állítható elő 1-es típusú láncgráf mellett. Ugyanis a csúcsok között nem húzódhat él (hogy teljesüljenek az állítások), ezért csak külön komponensbe vehetjük fel őket, tehát hiányzó nyilakat kell definiálnunk. Ez a konstelláció mind a négy páronkénti Markov-tulajdonság mellett három feltételes (és nem marginális!) függetlenséget jelent:

$$A \perp B \mid C,\ A \perp C \mid B,\ B \perp C \mid A.$$

A marginális ill. a feltételes függetlenségekkel definiált modellek ténylegesen különböznek egymástól, hiszen az 1. tétel szerint az előbbi három hatáshoz tartozó paraméter (λ_{AB}^{AB}, λ_{AC}^{AC}, λ_{BC}^{BC}), az utóbbi négy hatáshoz tartozó paraméter (λ_{AB}^{ABC}, λ_{AC}^{ABC}, λ_{BC}^{ABC}, λ_{ABC}^{ABC}) 0-ra állításával definiálható.

Vegyük észre, hogy az IG-ken (az IKG-ken), mint speciális 1-es típusú láncgráfokon az irányítatlan páronkénti (irányított páronkénti) Markov-tulajdonságok ugyanazokat a feltételes függetlenségeket implikálják, mint az LP1-LP4 bármelyike. Tehát ezek a modellek a korábbiak általánosításai.

Mint azt a 3.2. fejezetben már megjegyeztem, a többi szerzővel ellentétben Wermuth és Cox (pl. 1992) a legáltalánosabb esetben ugyanazon a gráfon belül megenged nem csak egyféle vonalat és egyféle befelé mutató nyilat tartalmazó komponenst, hanem ezek keverékét is. Ilyen pl. az alábbi gráf, ahol a második komponensen belül szaggatott vonalak vannak, és teli nyilak mutatnak bele, míg a negyedik komponensen belül teli vonalak vannak, és szaggatott nyilak mutatnak bele. Ezek a modellek nem tartoznak a fenti négy típus egyikébe sem, és esetükben valóban szükséges a teli/szaggatott megkülönböztetés a megfelelő Markov-tulajdonság megadásához. Ezeket a gráfmodelleket Rudas, Bergsma, Németh (2009) tárgyalja részletesebben.



3.6.3 Lokális Markov-tulajdonságok

Frydenberg (1990a) vezeti be az LP1, Andersson et al. (2001) az LP2 lokális megfelelőit.

23. definíció (LL1) Azt mondjuk, hogy a G gráf mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik a láncgráfokra definiált 1-es típusú *lokális* Markov-tulajdonsággal, ha

$$\alpha \perp \left(K_{Nd}(i) \setminus cl(\alpha)\right) | \, bd(\alpha), \tag{28}$$

minden α csúcsra, ami az i-vel indexelt komponenshez tartozik.

Azaz α feltételesen független minden vele össze nem kötött csúcstól, ami a saját komponensében vagy egy azt nem követő komponensben van, a vele összekötött csúcsokra, mint feltételre nézve.

Ezt a tulajdonságot LFW lokális Markov-tulajdonságnak nevezik, és valóban az LWF páronkénti tulajdonság megfelelője, hiszen adott csúcs az LP1 által tőle páronként függetlennek nyilvánított csúcsok uniójától független, és a feltételben saját komponensbeliek is szerepelnek.

24. definíció (LL2) Azt mondjuk, hogy a G gráf mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik a láncgráfokra definiált 2-es típusú *lokális* Markov-tulajdonsággal, ha

$$\alpha \perp \left(K_i(i) \setminus cl(\alpha)\right) | \left(K_{nd}(i) \bigcup ne(\alpha)\right), \tag{29}$$

illetve ha

$$\alpha \perp \left(K_{nd}(i) \setminus pa(\alpha)\right) | \, pa(\alpha), \tag{30}$$

minden α csúcsra, ami az i-vel indexelt komponenshez tartozik. Azaz az első feltétel szerint (hiányzó vonal esete) α feltételesen független minden vele össze nem kötött, saját komponensében levő csúcstól, a komponensének nem-leszármazottjaihoz tartozó ill. a vele összekötött saját komponensbeli csúcsokra, mint feltételre nézve. A második feltétel szerint pedig (hiányzó nyíl esete) α feltételesen független a saját komponensének nem-leszármazottjaihoz tartozó, de vele össze nem kötött csúcsoktól, a szüleire, mint feltételre nézve.

Ezt a tulajdonságot AMP lokális Markov-tulajdonságnak nevezik, és valóban az AMP páronkénti tulajdonság megjelelője, hiszen adott csúcs az LP2 által tőle páronként függetlennek

nyilvánított csúcsok uniójától független, és hiányzó vonal esetén a feltételben saját komponensbeliek is szerepelnek, hiányzó nyíl esetén nem.

Definiálható lenne az LP3 és LP4 lokális megfelelője is, ezeket a szakirodalom azonban nem említi.

A páronkénti esethez hasonlóan itt is könnyen belátható, hogy az IG-ken (az IKG-ken), mint speciális 1-es típusú láncgráfokon az irányítatlan lokális (irányított lokális) Markov-tulajdonságok ugyanazokat a feltételes függetlenségeket implikálják, mint az LL1-LL2 bármelyike. Tehát ezek a modellek is a korábbiak általánosításai.

Itt is felmerül az a még megválaszolatlan kérdés, hogy vajon az 1-es illetve 2-es típusú láncgráfokon lokális tulajdonságokkal definiált modellek halmaza megegyezik-e.

A páronkénti esetben vizsgált példára (7. ábra) visszatérve, a gráf az LL1 mellett az alábbi feltételes függetlenségekkel ekvivalens:

$$B \perp D \mid AC,$$
$$C \perp A \mid BD,$$
$$D \perp AB \mid C,$$

míg az LL2 mellett az alábbiakkal:

$$B \perp D \mid AC,$$
$$C \perp A,$$
$$D \perp A.$$

Ha megnézzük, milyen modelleket kaptunk ugyanezen a gráfon LP1 és LP2 mellett, akkor azt látjuk, hogy az LP2 és az LL1 modellje megegyeztek, az LP1 az LL1-től pedig csak abban különbözött, hogy az előbbi $D \perp A \mid BC$-t implikálta, az utóbbi $D \perp AB \mid C$-t. A két feltételes függetlenség közül a második implikálja az elsőt (a 3.3 fejezetben felsorolt feltételes függetlenségi tulajdonságok közül a C3 és C2 egymás utáni alkalmazásával), tehát az LL1 melletti modell implikálja az LP1 mellettit, más szóval: az LL1 melletti szűkebb a másiknál, almodellje annak. Viszont ha (C5*) áll, akkor teljesül, hogy

$$\text{ha } \left(D \perp B \mid AC \text{ és } D \perp A \mid BC\right), \text{ akkor D} \perp BA \mid C.$$

Vagyis (C5*) esetén az LP1 és LL1 melletti modellek ekvivalensek. Ez általában is igaz (AMP, 2001):

13. tétel Bármely 1-es típusú láncgráf és bármely, χ-en adott valószínűség-eloszlás mellett

$$(LL1) \Rightarrow (LP1),$$
$$\text{és}$$
$$(LL2) \Rightarrow (LP2).$$

Az IG és IKG esetéhez hasonlóan a (C5*), vagyis speciális esetként a pozitív eloszlás itt is ekvivalenciát implikál:

14. tétel Bármely 1-es típusú láncgráf és bármely, χ-en adott valószínűség-eloszlás mellett, ha (C5*) teljesül, akkor (AMP, 2001)

$$(LL1) \Leftrightarrow (LP1),$$
és
$$(LL2) \Leftrightarrow (LP2).$$

Mivel ezek a tételek az 1-es láncgráfok kontextusában nyertek bizonyítást, nem ismert, hogy a 2-es típusú láncgráfokra is teljesülnek-e.

3.6.4 Blokk-rekurzív Markov-tulajdonságok

A blokk-rekurzív Markov-tulajdonságnak nincsen megfelelője az IG-k és IKG-k esetére, hiszen ez a tulajdonság éppen a láncgráfok komponenseinek (másnéven blokkjainak) a rekurzív struktúrájára épülnek. Az alábbiakban Drton (2008) munkájára támaszkodom, ahol a szerző négyféle blokk-rekurzív Markov-tulajdonságot különböztet meg, a négyféle páronkénti tulajdonság analógiájára. Az első és a második tulajdonság megegyezik azzal, ahogyan Andersson et al. (2003), aki csak ezt a két tulajdonságot tárgyalja, bevezeti azokat.

Öt feltételes függetlenségi tulajdonság definiálásával kell kezdenünk.

A komponensek, mint csúcsok által definiált IKG-ra alkalmazzuk az irányított lokális Markov-tulajdonságot, eszerint minden K_i komponensre:

$$K_i \perp \left(K_{nd}(i) \setminus K_{pa}(i)\right) | K_{pa}(i). \tag{31}$$

Adott komponensen, mint irányítatlan gráfon alkalmazzuk az irányítatlan lokális Markov-tulajdonságot azon a feltételes eloszláson, amit a komponenst közvetlenül megelőző komponensekre (azaz a komponens szüleire), mint feltételre veszünk. Eszerint minden K_i komponensre és annak minden σ részhalmazára:

$$\sigma \perp \left(K_i \setminus \sigma \setminus ne(\sigma)\right) | \left(K_{pa}(i) \bigcup ne(\sigma)\right). \tag{32}$$

Ha az előbbi irányítatlan lokális Markov-tulajdonság helyett annak azt a variánsát vesszük, amikor a feltételben nem szerepelnek a szomszédok, akkor minden K_i komponensre és annak minden σ részhalmazára:

$$\sigma \perp \left(K_i \setminus \sigma \setminus ne(\sigma)\right) | K_{pa}(i). \tag{33}$$

Az előző két tulajdonság vonallal össze nem kötött csúcsok függetlenségét állította. Nyíllal össze nem kötött csúcsokra ismét kétfajta tulajdonságot definiálunk, ezek között megint az lesz csak a különbség, hogy a saját komponensbeli szomszédok szerepelnek-e a feltételben. Tehát minden K_i komponensre és annak minden σ részhalmazára:

$$\sigma \perp \left(K_{pa}(i) \setminus pa(\sigma)\right) | \left(pa(\sigma) \bigcup ne(\sigma)\right), \tag{34}$$

vagy

$$\sigma \perp \left(K_{pa}(i) \setminus pa(\sigma)\right) | \, pa(\sigma). \tag{35}$$

Most már definiálhatjuk a blokk-rekurzív Markov-tulajdonság négy típusát.

25. definíció (LB1) Azt mondjuk, hogy a G láncgráf mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik a láncgráfokra definiált 1-es típusú *blokk-rekurzív* Markov-tulajdonsággal, ha (31), (32) és (34) teljesül.

Ez a tulajdonság az LWF páronkénti tulajdonság megfelelője, hiszen a feltételhalmaz mind hiányzó nyilak, mind hiányzó vonalak esetén tartalmazza a saját komponensbeli szomszédokat. A tulajdonságot ezért LWF blokk-rekurzív tulajdonságnak nevezik.

26. definíció (LB2) A 2-es es típusú *blokk-rekurzív* Markov-tulajdonság (31), (32) és (35) teljesülésével határozható meg.

Ez a tulajdonság az AMP páronkénti tulajdonság megfelelője, hiszen a feltételhalmaz hiányzó nyilak esetén nem tartalmazza a saját komponensbeli szomszédokat. A tulajdonságot ezért AMP blokk-rekurzív tulajdonságnak nevezik.

27. definíció (LB3) A 3-es es típusú *blokk-rekurzív* Markov-tulajdonság (31), (33) és (34) teljesülésével határozható meg.

Ez a tulajdonság az LP3 páronkénti tulajdonság megfelelője, hiszen a feltételhalmaz hiányzó vonalak esetén nem tartalmazza a saját komponensbeli szomszédokat.

28. definíció (LB4) A 4-es es típusú *blokk-rekurzív* Markov-tulajdonság (31), (33) és (35) teljesülésével határozható meg.

Ez a tulajdonság az LP4 páronkénti tulajdonság megfelelője, hiszen a feltételhalmaz sem hiányzó vonalak, sem hiányzó nyilak esetén nem tartalmazza a saját komponensbeli szomszédokat.

Az első két tulajdonságot Andersson et al. (2001) definiálta, a második kettőt Drton (2008) vezette be. A négy tulajdonság mellett kapott modelleket Rudas, Bergsma és Németh (2009) is vizsgálja. Akárcsak a páronkénti esetben, itt is megjegyezhető, hogy az első tulajdonság implikál „leginkább feltételes”, a negyedik „leginkább marginális” feltételes függetlenségeket.

A 7. ábra visszatérve, nézzük meg most, hogy milyen modelleket definiál blokk-rekurzív tulajdonságok mellett. A (31)-es állítás egyik tulajdonság mellett sem ad semmit. A többi állítás az LB1 esetén az alábbi feltételes függetlenségeket adja:

$$B \perp D \,|\, AC,$$
$$CD \perp A \,|\, B,$$
$$C \perp A \,|\, BD,$$
$$D \perp A \,|\, C.$$

Az LB2 pedig a következőket:

$$B \perp D \mid AC,$$
$$CD \perp A,$$
$$C \perp A,$$
$$D \perp A,$$

Az LB3 a következőket:

$$B \perp D \mid A,$$
$$CD \perp A \mid B,$$
$$C \perp A \mid BD,$$
$$D \perp A \mid C.$$

Az LB4 pedig az alábbiakat:

$$B \perp D \mid A,$$
$$CD \perp A,$$
$$C \perp A,$$
$$D \perp A.$$

Megjegyzés: az implikált feltételes függetlenségek jó része redundáns, mert a többiekből következnek. Pl. $CD \perp A \rightarrow C \perp A$.

Ha megnézzük, milyen modellt definiált ugyanezen a gráfon LP1-LP4 és LL1-LL2, akkor a (C1)-(C4) feltételes függetlenségi tulajdonságok kihasználásával belátható, hogy a blokk-rekurzív tulajdonság mind a négy esetben implikálja a páronkéntit. Továbbá az LB1 implikálja az LL1-et és az LB2 az LL2-t. A fordított irányú tartalmazás pl. az LB4 és LP4 között azért nem áll fenn, mert az előbbihez $CD \perp A$ tartozik, az utóbbihoz $C \perp A$ és $D \perp A$, és az utóbbiak általános esetben nem implikálják az előbbit. Ez az implikáció az 3.3. fejezetben definiált (C6) tulajdonság alesete, amiről ott már megemlítettük, hogy általános esetben nem teljesül. Diszkrét változók esetén még pozitivitás esetén sem áll, teljesül viszont a folytonos esetben együttes normalitás esetén.

A példák az általános eset lényegére mutatnak rá. Ugyanis a páronkénti, lokális és blokk-rekurzív tulajdonságok tartalmazási relációiról az alábbiak ismertek általános esetben:

15. tétel Az LWF Markov-tulajdonságok esetén (AMP et al., 2001) 1-es típusú láncgráfon

$$(LB1) \Rightarrow (LL1),$$

és (C5*) teljesülésekor, tehát pl. pozitív eloszlásra az ekvivalencia is áll:

$$(LB1) \Leftrightarrow (LL1).$$

16. tétel Az AMP Markov-tulajdonságok esetén (AMP et al., 2001) 1-es típusú láncgráfon

$$(LB2) \Rightarrow (LL2),$$

és (C5*) és (C6) együttes teljesülésekor, tehát nagyon speciális esetben áll csak az ekvivalencia:

$$(LB2) \Leftrightarrow (LL2).$$

17. tétel A 3-es és 4-es Markov-tulajdonságok esetén (Drton, 2008) 1-es típusú láncgráfon

$$(LB) \Rightarrow (LP),$$

és (C5*) és (C6) együttes teljesülésekor teljesül csak az ekvivalencia.

Tehát pozitív diszkrét eloszláson az LWF blokk-rekurzív, lokális és páronkénti modellek ekvivalensek, de az AMP modell ill. a 3-es, 4-es blokk-rekurzív modellek valódi részhalmazai a megfelelő páronkénti modelleknek. A páronkénti modellek kevesebb megkötéssel adhatók meg, tehát több szabad paraméterrel rendelkeznek, így a modell dimenziója (ahogyan azt a 12. definícióban, a 29. oldalon megadtuk) a páronkénti esetben nagyobb, mint a blokk-rekurzív esetben.

Megjegyzés: folytonos esetben együttes normalitás esetén (C5*) és (C6) teljesül, tehát ilyen eloszlásokon elég a legegyszerűbb páronkénti modelleket tanulmányozni.

3.6.5 Globális Markov-tulajdonságok

A lokális tulajdonsághoz hasonlóan a globális tulajdonság is csak az LWF és AMP esetben van definiálva (Frydenberg, 1990a, Andersson et al., 2001), bár elképzelhető lenne hasonló általánosítás a másik kettő esetben is. IG-k esetében a globális Markov-tulajdonság az elválasztás fogalmára épült. Irányított esetben ennek megfelelőjeként, továbbra is az elválasztást használva az IKG-hoz tartozó irányítatlan erkölcsös gráfra vezettük vissza a tulajdonságot. Most a láncgráfhoz tartozó irányítatlan erkölcsös vagy kibővített gráfra fogunk hivatkozni.

29. definíció (LG1) Azt mondjuk, hogy a G láncgráf mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik az *1-es típusú (vagy LWF) globális Markov-tulajdonság*gal, ha a csúcsok bármely (A,B,S) halmazhármasára, amelyek olyanok, hogy S elválasztja A-t B-től a $\left(G_{An(A\cup B\cup S)}\right)^m$-ben (vagyis az $A\cup B\cup S$-t tartalmazó legkisebb őshalmazhoz tartozó erkölcsös gráfban), fennáll:

$$A \perp B \mid S.$$

30. definíció (LG2) Azt mondjuk, hogy a G láncgráf mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik az *2-es típusú (vagy AMP) globális Markov-tulajdonság*gal, ha a csúcsok bármely (A,B,S) halmazhármasára, amelyek olyanok, hogy S elválasztja A-t B-től a $\left(G_{An(A\cup B\cup S)}\right)^a$-ben (vagyis az $A\cup B\cup S$-t tartalmazó legkisebb őshalmazhoz tartozó kibővített gráfban), fennáll:

$$A \perp B \mid S.$$

A páronkénti és a lokális esethez hasonlóan itt is könnyen belátható, hogy az IG-ken (az IKG-ken), mint speciális 1-es típusú láncgráfokon az irányítatlan globális (irányított globális) Markov-tulajdonságok ugyanazokat a feltételes függetlenségeket implikálják, mint az LG1-LG2 bármelyike. Tehát ezek a modellek is a korábbiak általánosításai.

A két globális tulajdonság 1-es típusú láncgráfon ekvivalens a megfelelő blokk-rekurzív tulajdonsággal, még általános esetben is:

18. tétel Az LWF ill. AMP Markov-tulajdonságok esetén (AMP et al., 2001) 1-es típusú láncgráfon

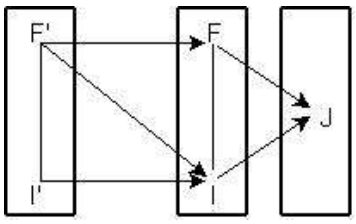
$$(LG1) \Leftrightarrow (LB1),$$
$$(LG2) \Leftrightarrow (LB2).$$

Összefoglalva az eddigi ekvivalenciatételeket elmondható, hogy *pozitív diszkrét eloszlás* esetén a tizenhat elvileg különböző modellből tizenkettőt ismertet az irodalom, s ezek között az ekvivalenciák miatt valójában csak hét különböző van, ahogyan az alábbi táblázat mutatja:

2. táblázat. A láncgráf-modellek típusai, kategoriális változók és pozitív eloszlás esetén (=: ekvivalencia, ⇒: implikáció)

Típus	*Markov-tulajdonság*						
1-es (LWF, leginkább feltételes)	Globális	=	Blokk-rekurzív	=	Lokális	=	Páronkénti
2-es (AMP)	Globális	=	Blokk-rekurzív	⇒	Lokális	=	Páronkénti
3-as			Blokk-rekurzív	⇒			Páronkénti
4-es (leginkább marginális)			Blokk-rekurzív	⇒			Páronkénti

A láncgráf-modellek definiálásának lezárásaként nézzünk most egy mobilitás-kutatási példát, a 7.1. fejezetben később szereplő modellt, ami úgy készült, hogy a klasszikus Treiman-modell gráfján végrehajtottam a Boguszak és társai (1990) által Csehszlovákia esetére javasolt két módosítás egyikét. A gráf eredeti alkalmazási kontextusában nem grafikus modellként volt értelmezve, hiszen folytonos változókkal és útmodellel közelítették meg. Itt azt mutatom meg, milyen jelentésbeli módosulást eredményezne, ha gráfmodellként értelmeznénk. A gráf indoklását, a módosítás lényegét részletesen majd a 7.1. fejezetben írom le, a modell evolúcióját és egyéb variánsait pedig a 3.8.-ben. Itt csak a gráfhoz tartozó feltételes függetlenségi állításokat vizsgálom meg.



8. ábra. Ötváltozós státuszelérési modell, láncgráf modellel (I: iskolázottság, F: foglalkozás, I': apa iskolázottsága, F': apa foglalkozása, J: jövedelem)

Vizsgáljuk meg, hogy a fenti hét modell közül hány ténylegesen különböző modell definiálható itt. Először nézzük meg, van-e olyan csúcs, aminek a komponensében van más csúcs is, és hiányzik egy belé mutató nyíl vagy egy belefutó vonal. Az *F* ilyen (*I'F* nyíl hiányzik), de

hasonló szerepű hiányzó vonal nincsen. Tehát nem mindegy, hogy az LWF vagy az AMP tulajdonságot használjuk (ezek különböző feltételes függetlenséget adnak többcsúcsú komponensbe futó hiányzó nyíl esetére), de a 3-as és 4-es tulajdonság ezekkel megegyező feltételes függetlenségeket adna (mert ezek többcsúcsú komponensből hiányzó vonalra adnának az LWF ill. AMP típushoz képest más feltételes függetlenséget, de ilyen nincs). Tudjuk, hogy LWF esetben a páronkénti, lokális, blokk-rekurzív és globális modellek megegyeznek, ezért vegyük most a legegyszerűbb páronkénti tulajdonságot. Ezzel az alábbi feltételes függetlenségek olvashatók ki a gráfból:

$$\begin{aligned} &J \perp I' \mid F'FI, \\ &J \perp F' \mid I'FI, \\ &F \perp I' \mid F'I. \end{aligned} \tag{36}$$

Az AMP tulajdonságokról tudjuk, hogy közöttük valójában csak két különböző van, a páronkénti és a blokk-rekurzív. Az előbbi a következő függetlenségeket adja:

$$\begin{aligned} &J \perp I' \mid F'FI, \\ &J \perp F' \mid I'FI, \\ &F \perp I' \mid F'. \end{aligned} \tag{37}$$

Vegyük végül az AMP blokk-rekurzív tulajdonsággal megkapható feltételes függetlenségeket:

$$\begin{aligned} &J \perp F'I' \mid FI, \\ &F \perp I' \mid F'. \end{aligned}$$

Az utóbbi állításokról belátható, hogy megegyeznek a gráfból az AMP lokális tulajdonsággal kapott állításokkal. Az AMP lokális modellek viszont ekvivalensek az AMP páronkénti modellekkel, amiből az következik, hogy a gráfhoz tartozó AMP modellek is mind megegyeznek. Vagyis a gráfhoz csak két különböző modell tartozik: a (36) és a (37). Ezek pedig csak abban különböznek, hogy a foglalkozás és az apa iskolázottságának függetlenségéhez elég-e az apa foglalkozását feltenni (AMP), vagy az iskolázottságot is be kell-e vonni a feltételbe (LWF). Ha grafikus modellel közelítjük meg tehát Boguszak et al. gráfját, akkor e közül a két modell közül választhatunk.

3.7 A gráfok és a feltételes függetlenségek megfeleltetésének problémái: ekvivalens gráfok, ekvivalens Markov-tulajdonságok

A gráfok és a feltételes függetlenségi állítások megfeleltetésének kérdései közé tartozik az az egyszerűnek látszó probléma, hogy egy adott F függetlenség igaz-e az adott gráf és az adott Markov-tulajdonság mellett. Itt a nehézséget az adja, hogy F nem feltétlenül adódik közvetlenül a gráfból a Markov-tulajdonsággal kiolvasható függetlenségek L listájának tagjaként – lehet, hogy nincs köztük felsorolva, de azokkal ekvivalens, vagy azok következménye. Ennek belátásához fenti példáinknál, a páronkénti, lokális, blokk-rekurzív és globális tulajdonságok felhasználásával adódó modellek összehasonlításánál nem-triviális esetben a (C1)-(C4) levezetési szabályok használata is szükséges volt. Mivel azonban a feltételes függetlenség, mint logikai rendszer nem teljes (lásd a 3.3. fejezetet), előfordulhat olyan eset, amikor az $L \rightarrow F$ belátásához a (C1)-(C4) közé nem tartozó, de egyébként igaz szabályt kellene felhasználni. Az $L \rightarrow F$ összefüggés tehát

bizonyos esetekben bizonyítható a négy levezetési szabállyal, de ha segítségükkel nem sikerül igazolnunk az összefüggést, az még nem feltétlenül nyer cáfolatot. A nem-teljesség meghatározásából adódóan ez a probléma sem a négy szabály kibővítésével, sem azok kicserélésével nem oldható meg.

Ezzel a nem-teljességből fakadó problémával kapcsolatban az utóbbi két évtizedben fontos eredmények születettek. IG-kre az LWF globális tulajdonság mellett (Frydenberg, 1990b), IKG-kre szintén az LWF globális tulajdonság mellett (Geiger, Pearl, 1988), LG-kre az LWF (Studeny and Bouckaert, 1998) és az AMP (Andersson et al., 2001b) globális tulajdonság mellett belátták a teljességet, vagyis ezekben az esetekben a gráfból minden igaz feltételes függetlenség kiolvasható bizonyos szabályokkal. Ám ezek az eredmények nem feltétlenül igazak diszkrét eloszlásokra megszorítva. A bizonyításokban ugyanis legtöbbször azt mutatják meg (pl. Andersson et al., 2001b, Geiger, Pearl, 1988), hogy minden gráfhoz van legalább egy olyan eloszlás, ahol a felsorolt állítások és csak azok igazak. Ezt konstruktív módon, az eloszlás előállításával bizonyítják, a konstruált eloszlás viszont normális. Diszkrét eloszlású változókra Meek (1990) bizonyította be, hogy a Pearl által bevezetett d-szeparációval (*d-separation*, a feltételes függetlenségek kiolvasására szolgáló szabály) az LWF globális Markov-tulajdonság IKG-kon teljes. Másképpen fogalmazva: a d-szeparáció segítségével az IKG-ból a globális Markov-tulajdonsággal kiolvasott állításokból logikailag következő valamennyi feltételes függetlenség felsorolható. Az általam ismert irodalom nem nyújt választ diszkrét változókon értelmezett más gráfok és más Markov-tulajdonságok teljességének a kérdésére.

A következőkben ennek a problémának további általános megfogalmazásait tárgyalom. Elsősorban a problémák bemutatása a célom, a kapcsolódó eredményeket, tételeket inkább csak illusztrációként közlöm. A következő négy problémát vázolom:

1. Két Markov-tulajdonság ekvivalenciája adott gráfon
2. Két Markov-tulajdonság ekvivalenciája adott gráfosztályon
3. Két gráf Markov-ekvivalenciája
4. Két gráfosztály Markov-elvivalenciája

Két Markov-tulajdonság ekvivalenciája adott gráfon. Mint azt korábban, a különböző Markov-tulajdonságok felhasználásával adódó modellek összehasonlításánál láttuk, ugyanahhoz a gráfhoz tartozó két különböző Markov-tulajdonsággal definiált modell ekvivalenciájának eldöntése nem mindig egyszerű feladat. A modellek ekvivalenciájának eldöntését a gráfból a Markov-tulajdonságokkal kiolvasható két feltételes függetlenség-lista ekvivalenciájának eldöntésével végeztük el, és ehhez nem-triviális esetben a (C1)-(C4) levezetési szabályok használata kellett. Itt is igaz az, amit fent írtam: a feltételes függetlenség rendszerének nem-teljessége miatt ez a módszer nem alkalmazható általánosan. A modellek ekvivalenciája tehát néha ugyan bizonyítható, de nem cáfolható a négy levezetési szabállyal.

Két Markov-tulajdonság ekvivalenciája adott gráfosztályon. A fenti kérdésnek a gráfosztályokra általánosított változata az, amikor adott gráfosztályon két különböző Markov-tulajdonság ekvivalenciája a probléma: van-e minden *G* gráfhoz egy másik, ugyanebbe a gráfosztályba tartozó *G'* gráf, hogy *G* az első Markov-tulajdonság mellett ugyanazt a modellt adja, mint *G'* a második Markov-tulajdonság mellett? Andersson et al. (2001) szerint a

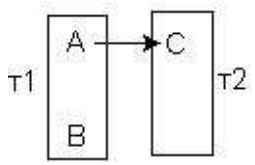
láncgráfokon az LWF- és AMP-tulajdonságok nem ekvivalensek. Ugyanők a gráfból kiolvasható (a gráf csúcsainak és éleinek bizonyos konfigurációival megadható) feltételekkel megadják az 1-es típusú láncgráfoknak azt a részhalmazát, amin az AMP - ill. LWF-Markov tulajdonságok már ekvivalensek.

Két gráf Markov-ekvivalenciája. Az első esetben ugyanahhoz a gráfhoz tartozó két különböző Markov-tulajdonság ekvivalenciájának eldöntéséről írtam. Egy fordított probléma két különböző gráf ugyanazon tulajdonság melletti ekvivalenciájának a kérdése.

31. definíció Két gráfot adott Markov-tulajdonság mellett Markov-ekvivalensnek nevezünk, ha ugyanazt a modellt indukálják.

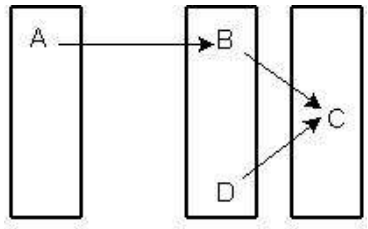
Míg az IG-k Markov-ekvivalenciájának szükséges és elégséges feltétele, hogy a csúcs- és élhalmazaik azonosak legyenek, addig az IKG-kre és LG-kre ez nem teljesül: akkor is egybeeshetnek a hozzájuk tartozó modellek, ha az élek halmaza különbözik (Andersson et al., 1997). Frydenberg (1990a) adott a gráfokból kiolvasható szükséges és elégséges feltételt két láncgráf LWF globális Markov tulajdonság melletti ekvivalenciájára. Hasonló eredményt közöl Verma és Pearl (1990) IKG-kre.

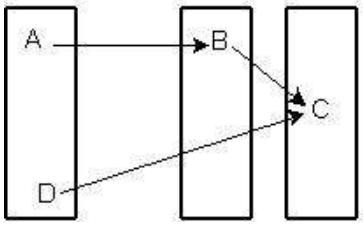
Gráfok Markov-ekvivalenciájának a körébe tartozik az 1-es és 2-es típusú láncgráfok problémája is. Láttuk, hogy a két gráfosztály gráfelméletileg valóban különbözik: van olyan él- és csúcs-halmaz, ami az egyiknek eleme, a másiknak nem. Azonban ettől független probléma az, hogy ugyanazon az él- és csúcshalmazon ugyanazzal a Markov-tulajdonsággal definiált két különböző típusú láncgráf által indukált modell különbözik-e. Tekintsük pl. az alábbi, 2-es típusú láncgráfot.



A gráfból az LP1-LP4 tulajdonság bármelyike mellett ezeket a feltételes függetlenségeket kapjuk: $B \perp A$, $C \perp B \mid A$. Ha viszont 1-es típusú láncgráfként kezeljük az adott csúcsokat és éleket, akkor a három csúcs három külön komponenst alkot. Ez az utóbbi modell bármelyik páronkénti tulajdonság mellett a következő függetlenségekkel azonos: $A \perp B$, $B \perp A \mid C$, $C \perp B \mid A$. Ez több megkötés, a modell ezért látszólag szűkebb, de a (21)-es levezetési szabály alkalmazásával kijön, hogy a plusz-megkötés a másik kettő implikáltja. Tehát a két típusú gráf ebben az esetben Markov-ekvivalens.

Egy másik ide tartozó, a gráfelméleti bevezetőben már említett probléma, hogy a 2-es típusú láncgráfok nem egyértelműen feleltethetők meg IKG-knak. Az alábbi két gráf ugyanannak az IKG-nak különböző 2-es típusú láncgráfjai.





Az LWF páronkénti Markov-tulajdonság mellett az első gráf az alábbiakat adja:

$$C \perp A \mid BD,\ D \perp A \mid B, D \perp B \mid A,$$

míg a második ezeket:

$$C \perp A \mid BD,\ D \perp A, D \perp B \mid A.$$

A két modell látszólag különbözik, de a (21)-es levezetési szabály alkalmazásával a második listából

$$(D \perp A,\ B \perp D \mid A) \rightarrow\ D \perp AB$$

áll elő, amiből (C3) és (C2) egymás utáni alkalmazásával

$$D \perp A \mid B$$

következik. Tehát a két modell ekvivalens, így a két gráf is Markov-ekvivalens az LWF páronkénti tulajdonság mellett.

Két gráfosztály Markov-elvivalenciája. Két gráf Markov-ekvivalenciája mellett beszélhetünk két gráfosztály Markov-ekvivalenciájáról is. A 3.6.1. fejezetben kitértünk az 1-es és 2-es típusú láncgráfok osztályának Markov-ekvivalenciájának kérdésére. Ebben az esetben nem két adott, csúcs- és élhalmazuk alapján összetartozó láncgráf modelljének páronkénti ekvivalenciája kérdéses, hanem a két típusba tartozó láncgráfok által indukált modellek halmazának azonossága. Elképzelhető, hogy az első kérdésre „Nem” a válasz, míg a másodikra „Igen”. Ugyanis lehet, hogy van olyan 1-es típusú gráf (G_1), amihez olyan 2-es típusú G_2 láncgráf tartozik, hogy G_2 modellje nem ekvivalens G_1 modelljével, de található egy hozzá egy másik 2-es típusú láncgráf (G_3), olyan, hogy G_3 modellje már ekvivalens G_1-ével.

Két gráfosztály ekvivalenciájának kérdéséhez tartoznak Andersson et al. (1997, 2001) eredményei: a gráfból kiolvasható szükséges és elégséges feltételeket adnak arra, hogy az IKG-k által adott modellosztály egybeessen az IG-k által adott modellosztállyal – általános esetben

ugyanis egyik sem tartalmazza a másikat, van modell, ami csak IG-vel, és van, amelyik csak IKG-vel fejezhető ki. A láncgráfok modellosztálya viszont (LWF vagy AMP-Markov tulajdonság mellett) tartalmazza az IKG-két és IG-két is. Andersson et al. megmutatja azokat a feltételeket, amiknek eleget tevő láncgráfok modellosztálya egybeesik az IKG-k ill. IG-k modellosztályával. Ez utóbbi bizonyos esetekben fontos kérdés lehet, hiszen ha kiderül, hogy egy láncgráfhoz van vele ekvivalens IKG, akkor a láncgráf-modell statisztikai elemzése jelentősen egyszerűsödhet.

Ugyanakkor ezek a hétköznapi gyakorlatban is fontos problémák, hiszen az adatainkhoz illeszkedő gráf kiválasztásánál megkerülhetetlen kérdés, hogy vajon két különböző gráf tényleg különböző modellt definiál-e adatainkon, és ha igen, akkor melyik feltevésében különbözik valójában a kéznél levő két modell. Ugyancsak fontos kérdés, hogy különbözik-e a választott gráf mellett két különböző Markov-tulajdonsággal definiált modell. Az alábbi szociológiai példáknál is látni fogjuk, hogyan merülnek fel ezek a problémák az alkalmazásokban.

3.8 Szociológiai példák grafikus modellekre

Az alábbiakban a két talán legismertebb szociológiai gráfmodellt, a Blau-Duncan és a Wisconsin modellt, illetve variánsaikat vizsgálom majd részletesebben. A modellek eredeti alkalmazási kontextusukban nem minden esetben voltak grafikus Markov-modellként értelmezve, ehelyett legtöbbször folytonos változókkal és útmodellel közelítették meg őket. Az útmodellek lineáris regressziós egyenletek rendszereként definiálhatók, ahol a függő változók azok, ahova él mutat a gráfban, a független változók pedig azok, ahonnét az élek kiindulnak. A függő és a független változók közötti kapcsolatot magyarázzuk meg direkt és indirekt hatásokra történő felbontással, úgy, hogy a függő és a független változók közötti korrelációt bontjuk fel a köztük húzódó gráfbeli utaknak megfeleltethető tagokra. Ez a megközelítés azonban, mint ahogyan azt majd a 6.1.4. fejezetben részletesebben is kifejtem, nem követi a grafikus modellek alapgondolatát. Grafikus modellekben az él törlése ugyanis feltételes függetlenséget implikál, de ez általános esetben nem áll az útmodellekre: össze nem kötött változópár feltételes függetlensége útmodellek esetében nem kell, hogy teljesüljön egyetlen feltételhalmaz mellett sem

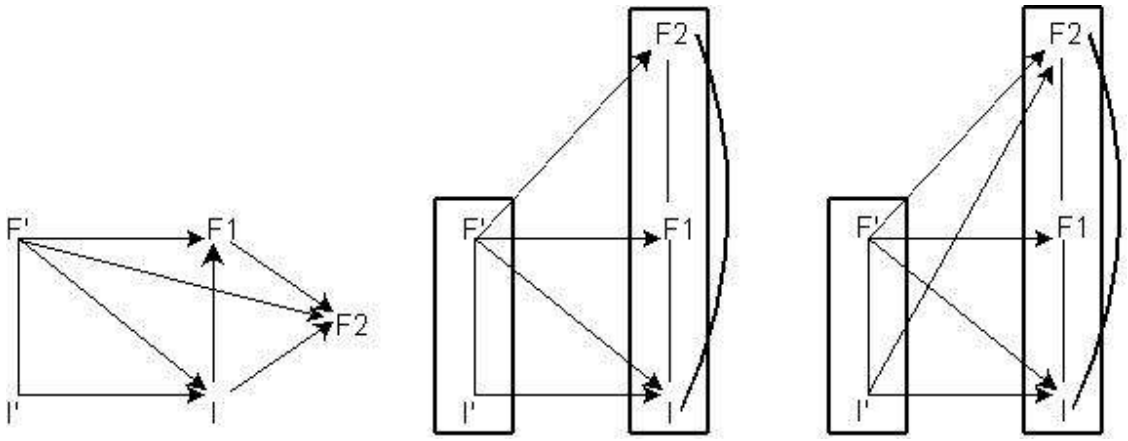
Az alábbiakban azt mutatom meg, milyen jelentésbeli módosulásokat eredményezne, ha grafikus modellként értelmeznénk az említett klasszikus gráfokat. Ennek haszna abban áll, hogy minden statisztikai modell esetében, akár illeszkedik az adatokhoz, akár nem, érdemes kissé eltérő modellek illeszkedését is megvizsgálni. Ugyanennek az elvnek a követésével bemutatom azt is, milyen következményekkel járnak átalakításaik (új élek bevonása, élek törlése, nyíl vonalra vagy vonal nyílra cserélése). Látunk olyan példát, amikor az átalakítással szűkebb vagy tágabb modellt kapunk, és látunk olyan példát is, amikor a gráf átrajzolása ellenére a modell változatlan marad. Megvizsgálom azt is, hogy különböző Markov-tulajdonságok alkalmazása esetén különbözik-e a kapcsolódó modell is. A két modell vizsgálata saját alkalmazásaim között a disszertáció végén is szerepel.

A következőkben a talán legismertebb szociológiai gráfmodellnek, a klasszikus Blau-Duncan modellnek az irodalomban előforduló különböző variánsait mutatom be. Az eredeti modell (Blau, Duncan, 1967) az alábbi 9. ábra baloldali gráfjával ábrázolható. Blau és Duncan célja a státuszelérési folyamat, ezen belül az iskolázottság közvetítő szerepének leírása volt. A modell változói a következők: I': apa iskolázottsága, F': apa foglalkozása, I: iskolázottság, F_1: első foglalkozás, F_2: jelenlegi foglalkozás. Az F' és az I' között nem nyíl, hanem kettősnyíl szerepelt az eredetiben. Ennek indoklása az volt, hogy a két változó nem csak azért korrelál, mert az egyik

hat a másikra, hanem azért is, mert közös okok húzódnak meg mögöttük: az apa első foglalkozása és az apa apjának foglalkozása ill. iskolai végzettsége. Itt ennek megfelelően vonalat húztam közéjük, de könnyen belátható, hogy a modellt definiáló feltételes függetlenségekre nincs hatása annak, hogy itt nyíl vagy vonal szerepel[28], ezért nyíl feltételezésével gráf tulajdonképpen IKG-ként is kezelhető. Tegyük most ezt, fogjuk fel a gráfot IKG-ként, és értelmezzük a globális Markov-tulajdonság mellett! Az alábbi feltételes függetlenséget kapjuk:

$$F_1F_2 \perp I' \mid F'I \tag{38}$$

Vagyis a modell alapvető mondanivalója, hogy az apa iskolázottsága közvetlenül nem befolyásolja az utód foglalkozását, csak közvetve, az iskolázottságon keresztül. Blau és Duncan funkcionalista magyarázata erre az volt, hogy a modern iparosodás a hatékonyságon alapul, ami egyfajta univerzalismust kíván meg: a legképzettebb embereket ülteti a legfontosabb pozíciókba. A társadalmi pozíció elérésénél a megszerzett (*achieved*) tulajdonságok tehát fontosabbak az örökölteknél (*ascribed*).



9. ábra. Az eredeti Blau-Duncan modell (1967), és annak Boguszak et al. (1990) általi módosítása két lépésben

Boguszak et al. (1990) Csehszlovákia esetére vonatkozóan két ponton javasolta módosítani a fenti modellt, a lépésenkénti változtatásról lásd a fenti 9. ábra középső és jobb oldali gráfját. Az első változtatás a jelenlegi foglalkozás és az iskolázottság közötti egyirányú kapcsolat kétirányúsítása, ennek indoka az, hogy a pozitívan diszkriminált, vezető pozíciókba juttatott politikailag megbízható, de iskolázatlan munkaerőt később gyakran mégis beiskolázták. (Boguszak et al. érveiről bővebben az alkalmazások között, az 7.1. fejezetben.) Grafikus modellként értelmezve a gráfot, az így láncgráffá alakul, mert az I és F_2 egy komponensbe kell, hogy tartozzon. Viszont az F_1-et is be kell vonnunk ebbe a komponensbe, hogy ne alakuljon ki félig-irányított kör. Mivel a komponensek teljesen rendezettek és egyenként összefüggő részgráfot alkotnak, a gráf 2-es típusú láncgráfként is egyértelműen áll elő, vagyis nem számít, hogy a gráfot 1-es vagy 2-es típusú láncgráfként kezeljük. Az ábra középső gráfjából az

[28] Ismét hangsúlyoznám, hogy ez a gráf grafikus modellként történő értelmezése mellett igaz. Nem érinti tehát azt a kérdést, hogy a gráfot útmodellként értelmező Blau-nak és Duncan-nek az *F'I'* él irányítatlanságára vonatkozó döntése is semmitmondó volt-e. Zárójelben megjegyzem azért: az útmodell együttható-egyenleteinek felírásából könnyen belátható, hogy az volt.

ekvivalencia-tételek használatával a különböző Markov-tulajdonságok mellett adódó modellek köre háromra szűkül. Ha a gráfot LWF modellként értelmezem, akkor a blokk-rekurzív Markov-tulajdonság mellett az a

$$F_1F_2 \perp I' \mid F'I \tag{39}$$

állítást adja, a páronkénti Markov-tulajdonság mellett pedig az ezzel a 2. táblázatba foglalt ekvivalencia-tételek szerint ekvivalens

$$\begin{aligned} F_1 \perp I' \mid F'IF_2 \\ F_2 \perp I' \mid F'IF_1 \end{aligned} \tag{40}$$

állításokkal adható meg. Ez a modell az eredeti Blau-Duncan modelltől abban különbözik, hogy itt az első foglalkozás és az apa iskolázottságának függetlenségéhez a jelenlegi foglalkozást is be kell vonni a feltételbe, ami nem biztos, hogy szerencsés feltevés. Ha AMP blokk-rekurzív modellként értelmezem a gráfot, akkor az

$$F_1F_2 \perp I' \mid F' \tag{41}$$

állítást kapom. A (39)-es és a (41)-es modell közötti különbség lényegi: a második azt mondja, hogy az apa iskolázottságának nincs direkt hatása az első és jelenlegi foglalkozásra, ha az apa foglalkozásán keresztül gyakorolt közvetett hatást kiszűrjük (értsd: ha az apa foglalkozását egy konkrét értéken rögzítjük). Az első modell szerint ugyanez akkor igaz, ha az utód iskolázottságán és az apai foglalkozáson keresztül gyakorolt közvetett hatásokat szűrjük ki (azaz ha a két utóbbi változót egyszerre rögzítjük egy-egy konkrét értéken). Az utóbbi modell nem szigorúbb az elsőnél, hanem mást mond, tehát egyik modell sem almodellje a másiknak.

Végül az AMP páronkénti Markov-tulajdonság szerint az alábbi függetlenségek olvashatók ki a gráfból:

$$\begin{aligned} F_1 \perp I' \mid F', \\ F_2 \perp I' \mid F'. \end{aligned} \tag{42}$$

Ennek a modellnek az előző almodellje, mert erősebb megszorításokat tartalmazott.

Ha (38)-at összevetjük (39)-cel, (41)-gyel ill. (42)-vel, láthatjuk, hogy az eredeti Blau-Duncan modell ekvivalens az első, LWF láncgráf-modellel (az LWF Markov-tulajdonság mellett nincs hatása a modellre a gráf változtatásának), viszont a másik két modellel nem összehasonlítható – sem nem szűkebb, sem nem bővebb azoknál. Ez azért előnyös, mert illeszkedő modell keresésekor rosszul illeszkedő modell esetén érdemes vele össze nem hasonlíthatóra térni, ahelyett, hogy automatikusan bővítenénk a modellt.

Buguszak et al. második változtatási javaslata szerint az apa iskolázottságának direkt hatása is van a jelenlegi foglalkozásra, mivel a magas iskolázottságú apáktól származó, egyébként negatívan diszkriminált utódok munkába állásukat követően mégis sikerrel érvényesültek a munkaerőpiacon (a kulturális tőke és az érvényesülési törekvés átörökítése folytán). A gráfból (9. ábra jobb oldali gráfja) egyetlen nyíl hiányzik, az I' és F_1 közötti. A különböző Markov-tulajdonságok mellett adódó modellek köre itt kettőre szűkül; ha LWF modellként értelmezem, akkor az

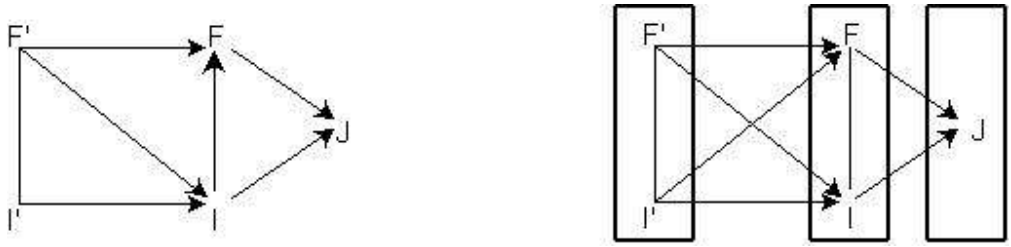
$$F_1 \perp I' \mid F' I F_2 \tag{43}$$

állítással adható meg, ha AMP modellként, akkor az

$$F_1 \perp I' \mid F' \tag{44}$$

állítással. A nyíl behúzásának a (39)-re, (42)-re ill. (44)-re gyakorolt következménye egyaránt az volt, hogy bővebb modellt kaptunk: az eredeti Blau-Duncan modell a módosított modellnek almodelljét képezi.

A Blau-Duncan modell egy variánsa a 10. ábra bal oldali modellje is. Eredeti forrása Duncan et al. (1968), de Donald Treiman (1970) tette ismertté, akinek modernizációs hipotézise szintén ezen a modellen belül fogalmazódott meg (erről bővebben lásd a 7.1. fejezetet). A változók köre nem tartalmazza az első foglalkozást, viszont a jövedelmet (*J*) igen. A jobb oldali gráf a Boguszak et al.-féle módosítások bevezetéséből adódik. Most nem mutatom be lépésenként a két módosítást, mert azt már megtettem korábban: az *IF* nyíl vonalra cseréléséből adódó gráfot a 3.6.5.-ben vizsgáltam. Az 7.1. alkalmazás e két modell illeszkedését veti össze 1992-es nemzetközi adatokon.



10. ábra. Az eredeti Blau-Duncan modell egy variánsa (Duncan et al., 1968, Treiman, 1970), és a modellre alkalmazva Boguszak et al. (1990) módosításait

A bal oldali gráf ismét tekinthető IKG-nek, a lokális Markov-tulajdonság alapján a következő feltételes függetlenségi állításokkal:

$$\begin{aligned} &J \perp I' F' \mid FI, \\ &F \perp I' \mid F' I. \end{aligned} \tag{45}$$

A a 3.6.5. fejezetben láttuk, hogy az *IF* nyíl vonalra cseréléséből adódó gráf LWF modellként (36) megegyezik ezzel, AMP modellként (37) azonban különbözik tőle. A nyíl-vonal csere után az *I'F* nyíl behúzását is elvégezve kapjuk a jobb oldali gráfot, ami bármely Markov-tulajdonság mellett ezt adja:

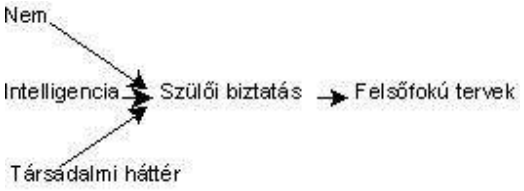
$$J \perp F' I' \mid FI,$$

tehát ennek az eredeti Treiman-modell almodellje.

A Blau-Duncan modell a státuszelérési folyamatot és benne az iskolázottságnak a családi háttér átörökítésében játszott közvetítő szerepét írja le, a folyamat leglényegesebb faktorainak kiemelésével. Nem jeleníti meg viszont a modell azokat a mechanizmusokat, melyek révén a

státuszelérési folyamatban a családi háttér hatása ténylegesen működik. Ezeknek a mechanizmusoknak a feltárása volt Sewell és Shah (1968) célja, amikor az örökölt erőforrások genetikai és szociálpszichológiai tényezőit is bevonták modelljükbe. A gyermekek szemében fontos személyek hatásának (az alábbi modellben a szülői biztatásnak) a megjelenítésével modelljük az iskolai teljesítmény mögött álló motivációk fontosságát hangsúlyozza.

Az alábbi példa Wisconsin középiskolásainak körében 1957-ben folytatott klasszikus vizsgálatából származik, azóta sokan mások újraelemezték (Fienberg, 1980, Whittaker, 1990, Spirtes et al., 1993, Heckerman et al., 1999). Sewell és Shah vizsgálata később longitudinális, több évtizedes kutatássá bővült, az itt bemutatott modellnek az elért foglalkozással is kibővített változata Wisconsin modellként vált ismertté. A Wisconsin modell a Blau-Duncan modellhez hasonlóan közismert a szociológiában, eredetileg ugyancsak folytonos változókkal, útelemzéssel közelítették meg, és szintén sok variánsa született. A most bemutatott első modellváltozatban a felsőfokú továbbtanulás (*F*) meghatározó tényezői a nem (*N*), intelligencia hányados (*I*), szülői bíztatás (*S*) és társadalmi háttér (*T*). Az adatok megtalálhatók az alkalmazások között, a 7.3. fejezetben. Sewell és Shah az alábbi ok-okozati sémát rendelte az adatokhoz[29]:



Fienberg (1980) újraelemezte az adatokat, mégpedig két logit-modellként tekintve azokra, ahol az *S* ill. az *F* adják a függő változókat, magyarázó változóik pedig a fenti sémában őket megelőző változók. Az általa preferált modell nem grafikus modell.

Grafikus modellként is elemezték már az adatokat. A legjobban illeszkedő IKG gráfot kereső algoritmusok alkalmazási példái között gyakran szerepel ugyanis ez a példa, az alább bemutatott gráfot (11. ábra) Spirtes et al. (1993) is egy ilyen algoritmussal kapta. Az általuk kereső algoritmussal végignézett gráfoknak meg kellett felelniük annak a kikötésnek, hogy olyan sorrendjét adják a változóknak, mely nem mond ellent a kézenfekvő ({*T, I, N*}, *S, F*) sorrendnek (itt az első három változó között nincs rendezés). Az így kapott „legjobb" modell a felsőfokú továbbtanulással kapcsolatos elhatározás kialakulását reprezentálja, objektív tényezőktől kiindulva az azoktól függő pszichológiai faktorokon át, mélyebb betekintést engedve ebbe a folyamatba annál, mint ha csak a szokott módon (pl. logisztikus regresszió alkalmazásával) az elhatározást, mint függő változót, az összes többi változót, mint azonos szintű magyarázó változót tekintenénk. Ezt a mélyebb betekintést segíti például, hogy a szülők biztatásának meghatározó faktorairól is képet kapunk, vagy az intelligenciának a továbbtanulási tervekre gyakorolt direkt (minden más faktortól független) és indirekt (a szülői biztatáson keresztül érvényesülő) hatását is el tudjuk különíteni.

[29] Ez eredeti értelmezésében persze nem grafikus modell, nem alkalmazható rá Markov-tulajdonság, inkább csak egy logikai séma.

11. ábra. Felsőfokú továbbtanulási tervek kialakulása, IKG (Spirtes et al., 2003)

A modell IKG modellként értelmezhető lokális (ez a 11. tétel szerint a globálissal ekvivalens) vagy páronkénti Markov-tulajdonságok mellett is, de mivel pozitív eloszlást feltételezhetünk, a két tulajdonság a 12. tétel szerint ekvivalens. A következő feltételes függetlenségek állnak elő a lokális esetben (csak a nem-redundánsakat felsorolva):

$$N \perp IT,\ F \perp N \mid TIS. \tag{46}$$

A második állítás adja a modell lényegi mondanivalóját: a felsőfokú továbbtanulási szándék független a nemtől, ha a többi tényező hatását kiszűrjük. A lányok és a fiúk azonos mértékű ambíciókkal bírnak, de esetleg a lányokat kevésbé bíztatják, mint az azonos képességű és azonos társadalmi háttérrel bíró fiúkat, ezért tűnhet úgy az *FN* marginális kapcsolatból, hogy a lányok kevésbé ambiciózusak. A modell illeszkedésének vizsgálata megtalálható az alkalmazások között, a 7.3. fejezetben.

A gráf átalakításainak következményeit bemutatandó: ha törölnénk a *TI* nyílt, feltételezve, hogy a társadalmi háttér nem befolyásolja a gyerekek képességeit, akkor egy szűkebb modellt kapnánk a (46)-nak a $T \perp I$ megszorítással történő kiegészítésével. Ha viszont a jelenlegi modellhez még hozzáadnánk az $N \rightarrow F$ nyilat, feltételezve, hogy a nemnek a szülői bíztatástól függetlenül is van hatása a továbbtanulásra, akkor bővebb modellt kapnánk, a (46)-ból a második állítást törölve. A jelenlegi modellben az $S \rightarrow F$ nyilat vonalra is cserélhetnénk, egyfajta visszacsatolást feltételezve, miszerint a felsőfokú tervek megjelenése fokozza a szülői bíztatást. Ekkor láncgráfmodellt kapunk, négy komponenssel, ahol az *S* és az *F* ugyanabba a komponensbe tartozik. Válasszuk ennek a modellnek a megadásához a láncgráfoknak az 1-es típusát[30], hogy alkalmazhassuk a 18. tétel szerinti ekvivalenciákat! Mivel a gráfban nincsen hiányzó vonal, ezért a Markov-tulajdonságok négy fajtájából csak kettő különbözik ténylegesen, az LWF és az AMP. Az eloszlás pozitivitásának feltevése mellett az LWF tulajdonságról tudjuk, hogy páronkénti, lokális, blokk-rekurzív és globális típusa ekvivalens. Válasszuk pl. a lokális típust. (28) szerint éppen a (46)-ban felsorolt feltételes függetlenségek olvashatók ki a gráfból. Tehát hiába cseréltük a nyilat vonalra, a kapott modell az LWF tulajdonság mellett valójában ugyanaz. Nézzük most az AMP tulajdonságot! Ennek két nem-ekvivalens típusa van, de könnyen belátható, hogy ezen a láncgráfon e két típus is ugyanazt a modellt indukálja. Nézzük tehát pl. a blokk-rekurzív típust! Az alábbi nem-redundáns függetlenségeket kapjuk a 26. definíció alapján:

$$N \perp IT,\ F \perp N \mid TI. \tag{47}$$

[30] A 2-es típus mellett három különböző gráfot kaphatnánk, attól függően, hogy az N-t, mint komponenst a T elé, a T és I közé, vagy az I és S közé tesszük.

Az IKG-hoz tartozó (46) és a láncgráfhoz tartozó (47) a második állítás feltételhalmazában különbözik egymástól. Tehát az AMP tulajdonság mellett a nyíl vonalra cserélésének a modellt tekintve ez a következménye.

Említettem, hogy Spirtes et al. a ($\{T, I, N\}$, S, F) sorrendet előfeltevésként használta. Vele szemben Heckerman et al. (1999, szintén grafikus modell-megközelítésben) ezt a rendezést előfeltételezte: ($\{T, N\}$, $\{I, S\}$, F). Bayes-féle megközelítést használva, a paraméterek egy konkrét a priori eloszlását feltételezve kerestek az IKG-k közül legjobb modellt. A második legjobban illeszkedő modelljük megegyezik Spirtes és társaiéval, de a legjobban illeszkedő modelljük olyan, amit Spirtes et al. nem enged meg: az *IS* nyíl irányítása náluk megfordul: a szülői bíztatás határozza meg az intelligenciát. Szerintük ez az ok-okozati viszony is elképzelhető, én azonban a láncgráf-modell eszközének birtokában azt gondolom, hogy itt nyíl helyett éppen a kétféle irány egyforma plauziblitása miatt inkább vonal lenne a megfelelő megoldás. Így az alábbi modellt kapnánk LWF Markov-tulajdonság mellett:

$$N \perp T,\ I \perp N \mid TS,\ F \perp N \mid TIS,$$

és az alábbit AMP Markov-tulajdonság mellett (ezen a gráfon az LP2, LL2, LB2 és LG2 tulajdonságok ekvivalensek):

$$N \perp T,\ I \perp N \mid T,\ F \perp N \mid TIS.$$

Az alkalmazások között, a 7.3. fejezetben Spirtes et al. eredeti modelljének és annak egy láncgráf-variánsának illeszkedését hasonlítom majd össze.

4 GRAFIKUS MODELL MARGINÁLIS MODELLKÉNT KATEGORIÁLIS ADATOKON

Az előző fejezetben tárgyalt grafikus modellek paraméterezése képezi ennek az új eredményeket közlő, ezért talán a legfontosabb fejezetnek a tárgyát. Mint a 2.3. fejezetben láttuk, a paraméterek jó megválasztása lényegi kérdés, ezen múlik például a modellek interpretálhatósága, sőt létezése is. Vannak a speciálisan a grafikus modellekre érvényes paraméterezési szempontok is: (1) egy grafikus modell ezeknek a paramétereknek bizonyos egyszerű megszorításaival (általában néhány paraméter nullára állításával) adható meg. Továbbá (2) azokban a gyakori esetekben, amikor a cél egy jól illeszkedő gráf (azaz az ezáltal prezentált modell) élei (a modellben szereplő hatások) elhagyhatóságának a tesztelése, akkor ezt a paraméter-megszorítások fokozatos feloldásával (*stepwise* - lépegetve) érhetjük el. Ugyanennek a az esetnek a megfelelője, amikor rosszul illeszkedő gráf javítása a cél, és bizonyos élek behúzásának hatását teszteljük a paraméter-megszorítások fokozatos feloldásával. Végül (3) a modellhez tartozó eloszlás értelmezése a becsült paraméterek interpretációjával történik. Mindhárom esetben azon múlik az eredményünk, hogy a tábla paraméterezése milyen paramétereket tartalmaz. Nem megfelelő paraméterezés esetén előfordulhat, hogy (1) nem világos, a modellt hogyan, a paraméterek milyen megszorításával definiálhatjuk, (2) a lépegetős módszer egyes lépéseinél az él elhagyását/behúzását nem tudjuk, hogyan feleltethetjük meg a paraméterek megszorításának/feloldásának, (3) a modellhez tartozó eloszlás paraméterei nem a modellezett jelenség minket érdeklő jegyeihez tartoznak.

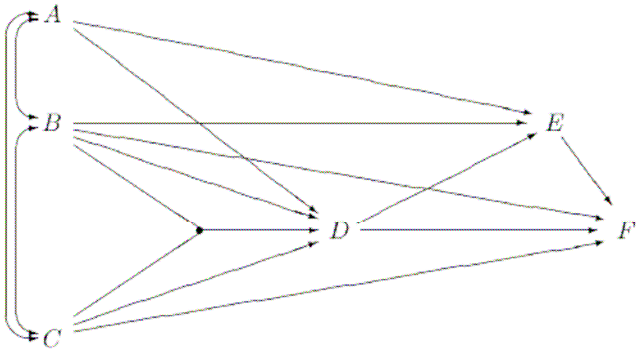
A grafikus modellek feltételes függetlenségi állításokkal adhatók meg. Ezek a feltételes függetlenségek a változóknak csak bizonyos részhalmazait érintik, ezért a modellek nem a teljes együttes eloszlásra, hanem csak a megfelelő marginális eloszlásokra vonatkozó megkötésekkel adhatók meg. Ha loglineáris paraméterezésben gondolkodunk, ezek a megszorítások egyszerre többfajta marginálishoz tartozó paraméter megszorítását igénylik, a hagyományos loglineáris keretben tehát nem fejezhetők ki. Természetesen adódik hát, hogy marginális loglineáris paraméterekkel paraméterezzük őket. Az alábbi fejezet tehát azon a kiindulóponton nyugszik, hogy a kategoriális változókon értelmezett grafikus modellek marginális loglineáris modellek. Az előző két, marginális ill. grafikus modellekkel foglalkozó fejezet eredményeit ezért közvetlenül fel fogjuk használni.

4.1 Egy előzmény: Goodman módosított útmodellje

A probléma megoldásának előzményeként szokás hivatkozni a Goodman (1973) által bevezetett módosított útmodellt (*modified path model*), ami ugyan nem grafikus modell, de valóban, sok szempontból rokonítható a grafikus modellekkel. Az alábbiakban Vermuntra (1996) támaszkodom a Goodman-modell interpretálásában, mert az ő megközelítése kicsit általánosabb, mint az eredeti.

Vermunt is megemlíti, hogy a Goodman-féle módosított útmodell segítségével a láncgráf-modellek is paraméterezhetők. (Vermunt, az általa hivatkozott források alapján itt vélhetően csak az LWF Markov-tulajdonság mellett értelmezett láncgráfokra gondol.) Goodman a feltételes függetlenségeknél általánosabb megszorítások paraméterezhetőségét célozta modelljével, ezt az alábbi, Vermunt-tól vett példán mutatom be.

Tekintsük az alábbi gráfot.



12. ábra. Módosított útmodell

Az ábra értelmezése a következő. A nyíllal összekötött változók között közvetlen kapcsolat van, a nyíl hiánya (páronkénti) feltételes függetlenséget jelent. A nyilak által meghatározott sorrend kauzális sort jelöl. A *BD* és *CD* nyilak ponttal összekötése azt jelzi, hogy a *B* és a *C D*-re gyakorolt hatása interakcióban van. Az *AE BE* és *DE* nyilak ugyanakkor pl. nincsenek ponttal összekötve, ami az interakció hiányát jelzi: az *A, B* és *D* mindegyikének van hatása *E*-re, de ezek a hatások egyenként függetlenek attól, hogy a másik két változó melyik kategóriájában vagyunk. Tehát a *D* az *A*-tól, a *B*-től, a *C*-től és a *BC* interakciótól függ, az *E* az *A*-tól, a *B*-től és a *D*-től, az *F* pedig a *B*-től, a *C*-től, a *D*-től és az *E*-től. Vegyük észre, hogy grafikus Markov-modellekkel a ponttal összekötött és össze nem kötött nyilak közti különbségtétel nem volt kifejezhető. Az *A*, *B* és *C* közötti kettős nyíl azt jelzi, hogy sem kauzális sorrendjük, sem együttes eloszlásuk nincs specifikálva.

Ha π_{abcdef} jelöli annak valószínűségét, hogy $A=a,\ B=b,\ C=c,\ D=d,\ E=e,\ F=f,$ akkor a változók kauzális sorrendje felhasználható arra, hogy ezt az együttes valószínűséget marginális és feltételes valószínűségek szorzatára bontsuk (Goodman, 1973):

$$\pi_{abcdef} = \pi_{abc}\pi_{d|abc}\pi_{e|abcd}\pi_{f|abcde}. \tag{48}$$

Ezen a módon közvetlenül is kifejezhető, hogy egy változó értéke csak az őt megelőző változók értékeitől függ. Ám ez még csak az első lépés a változók kapcsolatának specifikálásában. A következő lépésben az is meg kellene jelenítenünk, hogy az adott változó melyik korábbi változóval van tényleges kapcsolatban, és melyikkel nem. Ehhez fel kell használnunk azt a tételt[31], hogy változók feltételes függetlensége ekvivalensen definiálható a sűrűségfüggvény alábbi felbonthatóságával (pl. Lauritzen, 1990):

$$X \perp Y \mid Z \Leftrightarrow f(x \mid y,z) = f(x \mid z).$$

Ezt a gráfból kiolvasható

$$E \perp C \mid ABD,\ F \perp A \mid BCDE \tag{49}$$

[31] Ez a tétel az összevonhatósági elmélet (*collapsibility theorem*) következménye (Bishop et al., 1975). Az elmélet szerint ha *A* és *B* feltételesen független *C*-re nézve, akkor az *AC* hatás egyaránt becsülhető az *AC* és az *ABC* táblából is, vagyis az *ABC* tábla marginalizálható (összevonható) a *B*-re nézve.

páronkénti feltételes függetlenségek felhasználásával (48)–ra alkalmazva az alábbi kifejtést kapjuk:

$$\pi_{abcdef} = \pi_{abc}\pi_{d|abc}\pi_{e|abd}\pi_{f|bcde}. \tag{50}$$

Ha csak feltételes függetlenségi állításaink lennének, a modell így már megadható lenne. Azonban további specifikálás szükséges amiatt, hogy magasabbrendű interakciókat a modell nem enged meg. Így a kifejtés egyes tagjai tovább egyszerűsíthetők. Ez a feltételes valószínűségek logisztikus modellként történő paraméterezésével oldható meg:

$$\pi_{abc} = \frac{\exp\left(\sum_{L\subseteq ABC} \lambda_L^{ABC}(\mathbf{i}_L)\right)}{\sum_{abc}\exp\left(\sum_{L\subseteq ABC} \lambda_L^{ABC}(\mathbf{i}_L)\right)},\ L \neq \emptyset,\ \mathbf{i}_{ABC} = (a,b,c), \tag{51}$$

$$\pi_{d|abc} = \frac{\exp\left(\lambda_D^{ABCD}(d) + \lambda_{AD}^{ABCD}(a,d) + \lambda_{BD}^{ABCD}(b,d) + \lambda_{CD}^{ABCD}(c,d) + \lambda_{BCD}^{ABCD}(b,c,d)\right)}{\sum_d \exp\left(\lambda_D^{ABCD}(d) + \lambda_{AD}^{ABCD}(a,d) + \lambda_{BD}^{ABCD}(b,d) + \lambda_{CD}^{ABCD}(c,d) + \lambda_{BCD}^{ABCD}(b,c,d)\right)}, \tag{52}$$

$$\pi_{e|abd} = \frac{\exp\left(\lambda_E^{ABDE}(e) + \lambda_{AE}^{ABDE}(a,e) + \lambda_{BE}^{ABDE}(b,e) + \lambda_{DE}^{ABDE}(d,e)\right)}{\sum_e \exp\left(\lambda_E^{ABDE}(e) + \lambda_{AE}^{ABDE}(a,e) + \lambda_{BE}^{ABDE}(b,e) + \lambda_{DE}^{ABDE}(d,e)\right)}, \tag{53}$$

$$\pi_{f|bcde} = \frac{\exp\left(\lambda_F^{BCDEF}(f) + \lambda_{BF}^{BCDEF}(b,f) + \lambda_{CF}^{BCDEF}(c,f) + \lambda_{DF}^{BCDEF}(d,f) + \lambda_{EF}^{BCDEF}(e,f)\right)}{\sum_f \exp\left(\lambda_F^{BCDEF}(f) + \lambda_{BF}^{BCDEF}(b,f) + \lambda_{CF}^{BCDEF}(c,f) + \lambda_{DF}^{BCDEF}(d,f) + \lambda_{EF}^{BCDEF}(e,f)\right)}. \tag{54}$$

Látható, hogy azok a $\lambda_{XY}^{Xpa(X)}(xy),\ X \in \{D,E,F\},\ Y \subseteq pa(X)$ loglineáris paraméterek vannak a kifejtésekből elhagyva, amiknek az értéke nulla a magasabbrendű interakcióknak a modellből történt eltávolítása miatt. A gráfra visszapillantva látható, hogy egyedül *B*-nek és *C*-nek a *D*-re gyakorolt együttes hatásánál enged meg a modell másodfokúnál magasabb fokú hatást, ez a $\lambda_{BCD}^{ABCD}(b,c,d)$ (52)–ben történő szerepeltetésével egyenértékű. Az *E*-re és *F*-re gyakorolt hatások között nincs interakció, ezért az (53) és az (54) csak másodfokú hatásokat tartalmaz.

Ezek a logisztikus modellek könnyen átírhatók hagyományos loglineáris modellekké. Tehát Goodman-nél a modell specifikálása több különböző marginális táblához tartozó marginális loglineáris modell együttes specifikálásával oldható meg. Ebben az esetben ezek a táblák az *ABC, ABCD, ABDE* és *BCDEF* marginálisokhoz tartoznak.

Goodman megoldásában tehát több „kicsi” almodellből épül fel a „nagy”. A log-lineáris paraméterek és a várt cellagyakoriságok maximum likelihood-becslésekor ezután Goodman bizonyítása szerint elegendő az almodellekre külön becsléseket végezni. Ez abból adódik, hogy a likelihood almodell-specifikus tényezőkre bontható, melyek külön-külön maximalizálhatók.

Goodman tehát az irányított körmentes gráfmodelleknél általánosabb modellosztályt tárgyal, de megközelítése valóban jó kiindulópont lehet a grafikus modellek paraméterezéséhez.

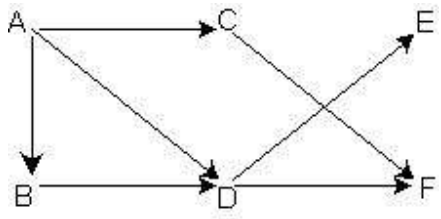
Megoldását érdemes lenne a disszertáció korábbi fejezeteinek keretei közé helyezni, ezt kísérlem meg a továbbiakban.

A sűrűségfüggvény (50)–ként való előállíthatósága, azaz a

$$P(V) = \prod_{\alpha \in V} P(\alpha \mid pa(\alpha)), \text{ ahol } V \text{ az összes változó halmaza} \tag{55}$$

kifejtés, másnéven faktorizáció létezése irányított körmentes gráfokon maga is Markov-tulajdonságként definiálható (pl. Lauritzen, 1990). Ez a faktorizációs tulajdonság IKG-k és kategoriális változók esetén a globális és lokális Markov-tulajdonságokkal ekvivalens. Tehát ha modellünket az adott gráf mellett ezen Markov-tulajdonságok valamelyikével értelmezzük, akkor valóban előállítható a sűrűségfüggvény az (55)–ben jelölt szorzat alakban. Tudjuk, hogy a páronkénti Markov-tulajdonság a lokálisnál bővebb modellosztályt határoz meg, de (C5*) mellett egybeesik a két modellosztály. Tehát a sűrűségfüggvény csak akkor faktorizálható, ha a (C5*), vagyis pl. a pozitivitás teljesül az eloszlásra. A faktorizáció megfelelője IG-k és LG-k esetén is létezik, mindkét esetben[32] implikálja a globális Markov-tulajdonságot, és (C5*) mellett egybeesik vele. Tehát a faktorizáció alapján a fenti példát talán általánosítani lehetne IG-kre és LG-kre is.

Ám a fenti heurisztika megvalósításakor több problémába ütközünk. Még IKG-k esetén és a (C5*) mellett sem olyan egyszerű a paraméterezés, mint azt a példa láttatja. A változók rendezése ugyanis nem teljes. Ezért kérdéses, hogy két össze nem hasonlítható csúcs közül melyiket válasszuk a (48) egyenlet felírásakor. Pl. az alábbi irányított körmentes gráf esetén



13. ábra. IKG (Rudas, Bergsma, 2004)

páronkénti Markov-tulajdonság mellett többek között ez a két állítás teljesül a modellre:

$$\begin{aligned} &C \perp E \mid ABD, \\ &E \perp C \mid ABDF. \end{aligned} \tag{56}$$

Hogy az első állítást beleépíthessük egy, az (50)-esnek megfelelő faktorizációba, az alábbi két valószínűség valamelyikének szerepelnie kellene egy, a (48)-asnak megfelelő faktorizációban:

$$\pi_{c|abde}, \ \pi_{e|abcd}.$$

[32] Láncgráfokra a faktorizációs tulajdonság csak LWF Markov-tulajdonság mellett volt vizsgálva (Lauritzen, 1996), és az ekvivalencia érvényessége is csak az LWF globális tulajdonság mellett ismert.

Hasonlóan, a második állítás beépíthetőségéhez az alábbi két tag valamelyikének kellene szerepelnie a faktorizációban:

$$\pi_{e|abcdf}, \ \pi_{c|abdef}.$$

Nyilvánvaló, hogy egyszerre nem tudjuk a két kívánalmat teljesíteni. Pl. $\pi_{c|abde}$ és $\pi_{e|abcdf}$ nem szerepelhet egyidejűleg a (48)-asnak megfelelő faktorizációban, mert az első tag szerint az E a C előtt van a csúcsok rendezésében, a második szerint éppen fordított a helyzetük. A problémát az okozza, hogy C és E nem összehasonlítható csúcsok a gráfból adódó részbenrendezés szerint. Megjegyezném, hogy ugyanez a probléma áll fenn 1-es típusú LG-k esetén is, ahol a faktorizáció a komponenseken, mint csúcsokon definiált IKG alapján működhetne, de a komponensek rendezése nem teljes.

További komoly problémák is előállnak. Ha a módosított útmodell példájában, a (49)-ben szerepelt feltételes függetlenségek paraméterezését közvetlenül az (50)-es faktorizációra építjük, akkor az alábbi marginális loglineáris paraméterek paraméterezik a modellhez tartozó eloszlást, vagyis ezek a szabad paraméterek:

$$\begin{aligned} &\lambda_X^{ABC}, \lambda_{DX}^{ABCD}, X \subseteq ABC, \\ &\lambda_{EV}^{ABDE}, V \subseteq ABD, \\ &\lambda_{FW}^{BCDEF}, W \subseteq BCDE. \end{aligned} \tag{57}$$

Ám kérdés, hogy melyek a rögzített paraméterek. Adódna, hogy további paraméterekkel kiegészítve teljessé tehetnénk a paraméterezést, és e kiegészítő paramétereket kellene megkötni, mégpedig minél egyszerűbb módon, pl. 0-ban rögzíteni. Csakhogy a teljessé tétel nem csak egy módon képzelhető el. Nem feltétlenül igaz, hogy bármelyik teljessé tételből adódó kiegészítő paramétereken valóban a 0-ban rögzítés adja a kellő modellt. Továbbá az is kérdés, hogy a többfajta elképzelhető mód közül van-e akár egy is, amely a 2.3. fejezetben felsorolt jó tulajdonságokkal (pl. rendezett dekomponálhatóság) rendelkezik.

Illusztráljuk ezeket az alapvető problémákat egy egyszerű IKG modellel:

$$A \longrightarrow B \longrightarrow C$$

A (48) megfelelője ez lenne:

$$\pi_{abc} = \pi_a \pi_{b|a} \pi_{c|ab},$$

amibe a modellt definiáló $C \perp A \mid B$ állítást beépítve a (50)-es megfelelőjét kapjuk:

$$\pi_{abc} = \pi_a \pi_{b|a} \pi_{c|b}.$$

A faktorizáció által közvetlenül adott szabad paraméterek ezek:

$$\lambda_{\varnothing}^{A}, \lambda_{A}^{A}, \lambda_{B}^{AB}, \lambda_{AB}^{AB}, \lambda_{C}^{CB}, \lambda_{CB}^{CB}.$$

Ha a teljessé tételhez az AC és ABC marginálisokat vonjuk be, akkor az alábbiak a kötött paraméterek:

$$\lambda_{AC}^{AC}, \lambda_{ABC}^{ABC},$$

míg ha csak az ABC marginálist használom a teljessé tételnél, akkor a következők:

$$\lambda_{AC}^{ABC}, \lambda_{ABC}^{ABC}.$$

A második megoldásnál valóban azok a paraméterek állnak elő kötött paraméterként, amiknek 0-ként való előírását a $C \perp A \mid B$ teljesüléséhez az 1. tétel is megkívánja. Ráadásul az így kapott teljes paraméterezés hierarchikus és rendezett módon dekomponálható is, ami, mint láttuk, Bergsma és Rudas (2002) szerint a paraméterezés jó tulajdonságait implikálja.

Az első megoldás viszont, ha a kötött paramétereket 0-ban rögzítettnek tekintjük, nem adja a $C \perp A \mid B$-t, mert a λ_{AC}^{ABC} paraméter nincs 0-ra állítva. Viszont az λ_{AC}^{AC} igen, tehát A és C marginális függetlensége (szükségtelenül!) része a modellspecifikációnak. Továbbá a paraméterezés nem rendezett módon dekomponálható. Tehát ha a λ_{AC}^{AC}, λ_{ABC}^{ABC} paraméterekkel definiáljuk a modellt, akkor nem 0-ra állítást, hanem valamely más, bonyolultabb megszorítást kell alkalmaznunk, ráadásul a paraméterek (mivel dekomponálhatók rendezett módon) nem lesznek variációsan függetlenek.

Vagyis a fenti heurisztika nem elég – nem csak a szabad paraméterek megadására, hanem arra is kell szabály, hogy mik a kötött paraméterek, és hogy azokra milyen megkötések érvényesek.

Végül még egy lényeges, már érintett problémát említenék. Kézenfekvő lenne, és a modell interpretálhatóságát szolgálná, ha a kötött paraméterek és a modell által implikált feltételes függetlenségek közvetlenül megfeleltethetők lennének egymásnak. Ez azonban ezzel a módszerrel, mint a legutolsó példánál láttuk, nem mindig teljesül.

A fentiekben a Goodman-féle, faktorizációra épülő paraméterezéssel kapcsolatos megkerülhetetlen, lényegi problémákra mutattam példákat. Összefoglalóan, ezek a problémák a következők:

1. a (48)-as faktorizáció megfelelő előállítása az IKG csúcsainak részbenrendezettsége miatt nem mindig kivitelezhető közvetlenül,
2. a kötött paraméterek előállítása nem egyértelmű, és nem mindegyik előállítás olyan, hogy a kötött paraméterek 0-ban való rögzítése adná a modellt, illetve
3. nem mindegyik előállítás mellett feleltethetők meg a modellt definiáló feltételes függetlenségek és a kötött paraméterek közvetlenül egymásnak, végül
4. a paraméterezés előnyös tulajdonságainak (pl. rendezett dekomponálhatóság) teljesülése nincsen garantálva.

Az utóbbi probléma fontosságát külön kiemelném. Korábban láttuk, hogy adott paraméterezés kapcsán alapvető kérdés, hogy létezik-e egyáltalán a paraméterekkel definiált modell / sima-e a paraméterezés / variációsan függetlenek-e a paraméterek / érvényesek-e a standard aszimptotikus tételek? A 2.3. fejezetben láttuk, hogy a marginális loglineáris paraméterezés, amilyen a Goodman által ajánlott is, általános esetben nem rendelkezik ezekkel a tulajdonságokkal. Egy megfelelő paraméterezés konstruálásakor tehát explicit választ kellene adni ezekre a kérdésekre.

4.2 Az általános megoldás: grafikus modellek marginális loglineáris modellekként

Bergsma és Rudas (2002, lásd a 2.3. fejezetet) jól használható általános feltételeket adott a legutóbb felsorolt kérdések (simaság, létezés stb.) megválaszolására. A továbbiakban vázolt megközelítés Goodman megoldásának olyan átalakítása, ami ezeknek az általános feltételeknek megfelelő paraméterezést állít elő. Rudas és Bergsma (2004) említi ezt a lehetőséget a marginális modellek néhány gyakorlati alkalmazását bemutató cikkében, irányított körmentes gráfokra és a belőlük kapott útmodellre Rudas, Bergsma és Németh (2006) ill. Németh, Rudas és Bergsma (2006) alkalmazza, láncgráfokra Rudas, Bergsma és Németh (2009) általánosítja.

Az alább bemutatott paraméterezések a Goodman-féle faktorizációból kapott paraméterek többi problémájától is mentesek lesznek, tételesen

1. mint látni fogjuk, létezik megoldás az irányított körmentes gráfok paraméterezésének az előző fejezetben bemutatott, a csúcsok részbenrendezettségéből fakadó problémájára,
2. a grafikus modell mindig a kötött paraméterek 0-ban való rögzítésével áll majd elő,
3. a modellt definiáló feltételes függetlenségek és a kötött paraméterek közvetlenül megfeleltethetők lesznek egymásnak, sőt, a paraméterezés a faktorizációs módszerrel szemben nem az eloszlás paraméterezéséből, hanem éppen ellenkezőleg, a függetlenségek paraméterezéséből indul majd ki.

Bergsma és Rudas (2002) írásának a következők szempontjából releváns tételeit összefoglalva: a paraméterezés simaságához, az aszimptotikus tételek teljesüléséhez és a modell dimenziójának (a valószínűséghányados próba szabadságfokának) az ismeretéhez hierarchikus és teljes paraméterezés kell. A modell létezéshez pedig elég, ha az üres halmaz csak egy marginális mellett szerepel hatásként. Tehát minimálisan ezeknek az előírásoknak megfelelő paraméterezést próbálunk a továbbiakban konstruálni. Ha találunk ilyent, akkor a paraméterek akkor és csak akkor függetlenek variációsan (vagyis nagyságuk akkor értelmezhető jól egymáshoz képest), ha rendezett módon dekomponálhatók.

Általános eljárást egy rendezett módon dekomponálható, hierarchikus és teljes paraméterezés megkonstruálására csak az IG-k és IKG-k esetén tudok majd mutatni. A láncgráfoknál néhány problémát mutatok majd, mely a kézenfekvő módon adódó hierarchikus paraméterezés útjában áll, és jelzem majd, mely almodellek esetén nem lehetséges elvileg sem hierarchikus paraméterezés létrehozása. Egyedi esetekben persze a láncgráfokra is található jó paraméterezés, erre is látunk majd példát.

Láttuk, hogy Goodman módosított útmodellje a magasabbrendű hatások elhagyását is lehetővé tette, tehát az IKG-nál általánosabb modell volt. E fejezet Útmodellek c. alfejezetében a grafikus modelleknek erre a módosítására (IKG útmodellek), és a kapott modellosztály paraméterezésére is látunk majd megoldást.

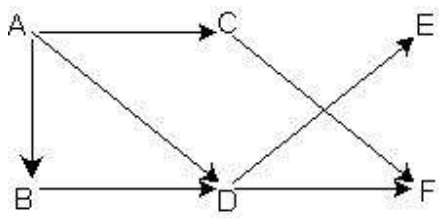
A fejezetben tárgyalt, valamely teljes P halmazhoz tartozó paraméterezésről mindig feltesszük, hogy az a marginális loglineáris modellek tárgyalásánál $\tilde{\lambda}_P$-vel jelölt paraméterezés, vagyis az ott látott módon a redundáns paraméterek elhagyásával jött létre.

4.3 A feltételes függetlenségek és a paraméterek megfeleltetésének problémái

Az előző fejezetekben láttuk, hogy a modellből adódó feltételes függetlenségek és a kötött paraméterek közvetlen megfeleltethetősége kívánatos lenne. E cél érdekében a grafikus modellek marginális loglineáris paraméterekkel történő felparaméterezésekor kiindulópontunk az 1. tétel lehet, amely az $X \perp Y \mid Z$ feltételes függetlenségnek az $X \cup Y \cup Z$ marginálison belüli paraméterezését adja meg. Kézenfekvő lenne a tétel alkalmazásával a grafikus modellt úgy paraméterezni, hogy az általa implikált $F_1, F_2, \ldots, F_n$ feltételes függetlenségeket a nekik megfelelő $M_1, M_2, \ldots, M_n$ marginálison belül paraméterezzük. Azonban ha - Bergsma és Rudas tételeit alkalmazva – hierarchikus és rendezett módon dekomponálható paraméterezést szeretnénk elérni, az ezen a módon nem mindig lehetséges.

Egy példával szemléltetném ezt a problémát. Az alábbi, korábban már látott IKG lokális Markov-tulajdonság mellett a következő feltételes függetlenségeket implikálja:

$$\begin{aligned} &C \perp BDE \mid A, \\ &D \perp C \mid AB, \\ &E \perp ABCF \mid D, \\ &F \perp ABE \mid CD, \\ &B \perp C \mid A. \end{aligned} \tag{58}$$



14. ábra. IKG (Rudas, Bergsma, 2004)

Az első függetlenség eléréséhez a *CE* hatást az *ABCDE* marginálison belül, a harmadik függetlenség eléréséhez pedig az *ABCDEF* marginálison belül kellene 0-vá tenni. Vagyis az így adódó paraméterezés nem hierarchikus, hiszen ugyanannak a hatásnak két különböző marginális mellett is szerepelnie kellene.

Ugyanakkor az is igaz, hogy a (21)-es levezetési szabály használatával belátható, hogy az első és a harmadik feltételes függetlenség átalakítható úgy ekvivalens módon, hogy a *CE* hatás már ugyanazon marginális mellett szerepeljen mindkét esetben:

$$C \perp BDE \mid A \Leftrightarrow \left(C \perp E \mid ABD \text{ és } C \perp BD \mid A\right),$$
$$F \perp ABE \mid CD \Leftrightarrow \left(C \perp E \mid ABD \text{ és } E \perp AB \mid D \text{ és } F \perp E \mid D\right).$$

tehát a *CE* szerepetetése megoldható. Arra azonban nincs garancia, hogy ez minden esetben alkalmazható megoldás. Nem feltétlenül igaz, hogy ha két feltételes függetlenség ilyen módon ütközik, akkor levezetési szabályokkal megfelelő módon átalakíthatók lennének (erre látunk majd példát a nem sima láncgráf-modellekkel kapcsolatban). Sőt meg kell említeni a feltételes függetlenségi rendszer nem-teljességét is: a (C1)-(C4) levezetési szabályok nem minden esetben elegendőek két, egyébként ekvivalens állítás ekvivalenciájának bebizonyítására.

Ugyanez a gráf alkalmas egy másik problémának az illusztrálására is. Az *AE* hatás szerepel az első feltételes függetlenséghez tartozó *ABCDE* marginálisban és a harmadik feltételes függetlenséghez tartozó *ABCDEF* marginálisban is. Ha hierarchikus paraméterezést szeretnénk előállítani, akkor a marginálisok olyan rendezésére lenne szükség, ahol az *ABCDE* korábban szerepel, mint az *ABCDEF*. Ekkor azonban, megint a hierarchikusság érdekében, az *AE* hatást a korábbi marginális mellett kell szerepeltetnünk (λ_{AE}^{ABCDE}), szabad paraméterként, tehát az *AE* nem szerepelhetne a későbbi marginális mellett hatásként λ_{AE}^{ABCDEF} formában, pedig a λ_{AE}^{ABCDEF} paraméterre szükségünk van, hiszen 0-ra kellene állítanunk az 1. tétel miatt.

Egy, a fentihez hasonló ad hoc megoldás ezt a problémát is megoldja. A harmadik függetlenség a (21)-es levezetési szabállyal felbontható:

$$E \perp ABCF \mid D \Leftrightarrow \left(E \perp AB \mid D \text{ és } E \perp CF \mid ABD\right),$$

az *EA* itt már az első függetlenség marginálisánál kisebb marginális mellett szerepel, vagyis az elé helyezendő a hierarchikus rendezésben. Ennek a megoldásnak az általánosításával kapcsolatban természetesen a fent leírt problémák szintén fennállnak.

E problémákra is megoldást fogunk mutatni a következőkben IKG esetére. Mint majd látni fogjuk, a megoldás azon alapszik, hogy a feltételes függetlenségek nem csak egyenként, az 1. tétellel paraméterezhetők, hanem van mód a gráf struktúráját kihasználó közvetlen paraméterezésre is.

Van azonban olyan probléma is, amire jelenleg még nincs megoldás. Ide tartoznak a láncgráfoknak az előbb mutatottakhoz hasonló, a hierarchikus paraméterezést akadályozó problémái. Problémás a láncgráfok rendezett módon dekomponálható paraméterezésének a megtalálása is, sőt azoknak a modelleknek a megadása is, amelyek paraméterezhetők így. A fenti módon, a feltételes függetlenségek egyenként történő paraméterezésével pl. a 2-es típusú láncgráfból, ahol egyetlen komponens van három össze nem kötött csúccsal, és a modellt a 3-as típusú páronkénti Markov-tulajdonság mellett definiáljuk, a következő három marginális függetlenség kapható:

$$A \perp B,\ B \perp C,\ A \perp C.$$

Ezek az *AB, BC* és *AC* marginálisokon belül paramétereződnének, amiknek nincsen rendezett módon dekomponálható rendezése. Nyitott kérdés, hogy van-e a modellhez más, problémamentes paraméterezés, illetve, hogy fellép-e ugyanez a probléma 1-es típusú láncgráfnál is.

4.4 Irányítatlan gráfok

Az előbb említett problémák egyike sem lép fel az irányítatlan gráfra épülő, pozitív eloszlást feltételező modellek esetén. Ilyenkor ugyanis az a fent vázolt ötlet, miszerint az 1. tétel segítségével egyenként paraméterezzük a modell által implikált F_1, F_2, ..., F_n feltételes függetlenségeket, közvetlenül alkalmazható, hiszen a nekik megfelelő marginálisok megegyeznek.

Ezt mutatjuk meg részletesebben is a következőkben. A 10. tétel alapján pozitív eloszlás mellett (vagy általánosabb esetben a (C5*) teljesülésekor) a lokális, páronkénti és globális Markov-tulajdonságok ekvivalensek. Tehát használhatjuk pl. a páronkénti Markov-tulajdonságot, ami szerint (lásd a 13. definíciót) bármely α, β össze nem kötött csúcspár feltételesen független az összes többi csúcsot feltéve, azaz:

$$\alpha \perp \beta \mid V \setminus \{\alpha, \beta\}.$$

Ez a függetlenség a V marginálison belül defniált loglineáris paraméterezésben az 1. tétel szerint akkor és csak akkor teljesül, ha

$$\lambda_L^V = 0, \ \forall L \in \mathbf{P}(V) \setminus \left(\mathbf{P}(V \setminus \alpha) \cup \mathbf{P}(V \setminus \beta)\right).$$

Most

$$\mathbf{P}(V) \setminus \left(\mathbf{P}(V \setminus \alpha) \cup \mathbf{P}(V \setminus \beta)\right) = \{\alpha \cup \beta \cup \gamma \mid \gamma \in V \setminus \alpha \setminus \beta\},$$

tehát minden olyan hatás értéke 0, amely a V α-t és β-t tartalmazó részhalmazaihoz tartozik. Bármely más páronkénti feltételes függetlenség hasonló módon, ugyancsak a V marginálison belül paraméterezhető. Így a függetlenségek uniója egyszerűen a külön-külön kapott 0-paraméterek uniójával adható meg, ezek között a közös marginális miatt nem léphet fel az előző fejezetben mutatott inkompatibilitás. Mivel a paraméterezésben egyetlen marginálist használunk, a V-t, tulajdonképpen hagyományos loglineáris modellt kapunk.

Ebből már közvetlenül adódik a jól ismert tétel (pl. Lauritzen, 1996):

19. tétel Pozitív eloszlás feltételezése mellett bármely, irányítatlan gráffal definiált modell megadható az alábbi loglineáris paraméterekre tett megszorításokkal:

$$\lambda_L^V = 0, \text{ minden } L \text{ csúcshalmazra, ami nem teljes részgráfhoz tartozik.}$$

A nem-nulla paraméterek, amelyek természetes módon a nullparaméterek teljessé tételével adódnak, paraméterezik a modellhez tartozó eloszlást. A redundáns paraméterek elhagyásával kapott

$$\tilde{\lambda}_P = \left\{\tilde{\lambda}_L^V : L \subseteq V\right\}$$

hagyományos loglineáris paraméterezésen belül tehát egyszerűen megadhatók az IG modellek. Ez a paraméterezés, mint a 2.1. fejezetben már láttuk, variációsan független, sima, a modell létezése és a standard aszimptotikus tételek alkalmazhatósága mindig garantált.

4.5 Irányított körmentes gráfok

Az alábbiak, hacsak nem jelzem másként, Rudas, Bergsma és Németh (2006) munkájára épülnek.

A 4.3. fejezetben felvillantott paraméterezési problémák jó része IKG modellekre vonatkozott, ami azt mutatja, hogy az IG-kre jól működő megoldás, miszerint az 1. tétel segítségével egyenként paraméterezzük a modell által implikált $F_1, F_2, \ldots, F_n$ feltételes függetlenségeket, itt közvetlenül nem alkalmazható. A probléma forrása az, hogy az IKG-kra érvényes Markov-tulajdonságok különböző marginálisok mellett adódnak. A megoldást Lauritzen et al. (1990) egy tétele nyújtja, amely az irányított lokális Markov-tulajdonságnak egy ekvivalens, és a paraméterezésben jól használható alternatíváját adja.

Ehhez definiáljuk az IKG csúcsainak egy rendezését:

32. definíció A G irányított körmentes gráf csúcsainak egy rendezését jólrendezésnek (*well-numbering*) nevezzük, ha bármely α, β csúcsra:

$$\alpha < \beta \Rightarrow \alpha \in nd(\beta),$$

azaz a nyilak mindig a sorban korábban levő csúcsokból mutatnak a később levő csúcsokba.

Könnyen belátható, hogy bármely IKG-hoz adható ilyen jólrendezés. Vegyük észre, hogy egy α csúcsot megelőző csúcsok között nem minden nem-leszármazott szerepel, de α szüleinek mindegyike szerepel köztük.

Jelöljük V_i-vel a V csúcshalmaz jólrendezésében az i. csúcsot, $V_{i<}$-vel az őt megelőző csúcsokat, $V_{i\leq}$-vel a sorban nála nem későbbi csúcsokat. A jólrendezéssel a lokális Markov-tulajdonság egy variánsa definiálható:

33. definíció (IJL) Azt mondjuk, hogy a G IKG mellett, az χ halmazon értelmezett P valószínűségi mérték rendelkezik az *irányított jólrendezési lokális Markov-tulajdonsá*ggal, ha bármely V_i csúcs feltételesen független a jólrendezésben őt megelőző csúcsoktól a szüleit feltéve:

$$V_i \perp V_{<i} \setminus pa(V_i) \mid pa(V_i).$$

Vagyis a lokális tulajdonságban, a

$$V_i \perp \left(nd\left(V_i\right) \setminus pa\left(V_i\right)\right) \mid pa\left(V_i\right)$$

formulában szereplő nem-leszármazottak helyett a jólrendezési tulajdonság csak a rendezésben korábbi nem-leszármazottakat veszi. Lauritzen et al. (1990) tétele szerint ez nem jelent tényleges változtatást:

20. tétel IKG gráfon az irányított lokális és az irányított jólrendezési lokális Markov-tulajdonság ekvivelens:

$$(IL) \Leftrightarrow (IJL).$$

Az 12. tétel kimondta, hogy pozitív eloszlás mellett IKG-ken a páronkénti, lokális és globális Markov-tulajdonságok ekvivalensek, tehát a mostani tételünk felhasználásával adódik, hogy pozitív eloszlás esetén mindezek helyett a jólrendezési lokális tulajdonság is használható.

Most már megadhatjuk az IKG-modell P paraméterezését. A marginálisok **M** halmazát a jólrendezés szeletei adják, n számú csúcs mellett:

$$\mathbf{M} = \{V_{i\leq},\ i = 1...n\}.$$

A marginálisok rendezése ebből közvetlenül adódik: $V_{1\leq}$, $V_{2\leq}$, ..., $V_{n\leq}$. Minden hatást pontosan egyszer, a lehető legkorábban szerepeltetünk. Ebben a paraméterezésben a $V_{i\leq}$ marginális mellett szereplő hatásokat az alábbi formában is felírhatjuk:

$$\varepsilon = \{V_i\} \cup F \cup G,\ \mathrm{F} \subseteq V_{<i} \setminus pa(V_i),\ \mathrm{G} \subseteq pa(V_i). \tag{59}$$

Ha az 1. tételt használjuk a paraméterezés megadásához, akkor a $V_i \perp V_{<i} \setminus pa(V_i) \mid pa(V_i)$ feltételes függetlenség a $\{V_i\} \cup \{V_{<i} \setminus pa(V_i)\} \cup \{pa(V_i)\} = V_{\leq i}$ marginálison belül értelmezett $\varepsilon = \{V_i\} \cup F \cup G,\ F \subseteq V_{<i} \setminus pa(V_i),\ G \subseteq pa(V_i),\ F \neq \varnothing$ hatások 0-ra állításával ekvivalens. Ezek a hatások az (59) alapján a most definiált P paraméterezésben is szerepelnek, vagyis a modellhez tartozó valamennyi feltételes függetlenség megadható P-ben.

A többi hatás paraméterezi az IKG modellbe tartozó eloszlásokat. Ezeknél $F = \varnothing$, tehát csak a V_i változó és annak szülei tartoznak a hatáshoz. Belátható, hogy ezek a paraméterek éppen a (55)-beli faktorizációval kapott faktorokat paraméterezik. Ugyanis a sűrűségfüggvény (49)–ben látott, a feltételes függetlenséggel ekvivalens átírása szerint a

$$V_i \perp V_{<i} \setminus pa(V_i) \mid pa(V_i)$$

feltételes függetlenség esetén

$$P(V) = \prod_{V_i \in V} P(V_i \mid pa(V_i)) = \prod_{V_i \in V} P(V_i \mid pa(V_i) \cup F),\ F \subseteq V_{<i} \setminus pa(V_i).$$

Tehát a modellbe tartozó eloszlás faktorizációjának tagjait paraméterező $\lambda_{V_i G}^{V_i \cup pa(V_i)}$, $G \subseteq pa(V_i)$ paraméterek azonos értéket vesznek fel, mint az azonos hatásnál de nagyobb marginálison belül adott társaik:

$$\lambda_{V_i G}^{V_i \cup pa(V_i)} = \lambda_{V_i G}^{V_i \cup pa(V_i) \cup F},\ G \subseteq pa(V_i),\ F \subseteq V_{<i} \setminus pa(V_i).$$

Ugyanez maximális F esetén a

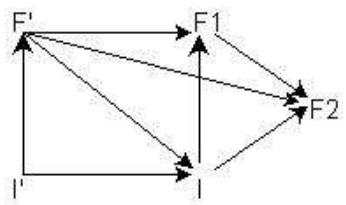
$$\lambda_{V_i G}^{V_i \cup pa(V_i)} = \lambda_{V_i G}^{V_i \cup V_{<i}} = \lambda_{V_i G}^{V_{\leq i}},\ G \subseteq pa(V_i) \tag{60}$$

egyenlőséget adja, ami éppen a bebizonyítandó állítás: a P paraméterezés szabad paraméterei (az egyenlőség jobb oldali tagja) megegyeznek a modellhez tartozó, faktorizációval definiált eloszlás paramétereivel (az egyenlőség bal oldali tagja). Ez fontos állítás, és - ahogyan az alábbi, 12. példa példa mutatja majd - a paraméterbecslések interpretálásánál érdemes kihasználni, hiszen a becslések kézenfekvőbb értelmezését teszi lehetővé. Segítségével egy adott nyílhoz tartozó hatás mérése csak a nyíl végpontjába befutó többi nyilat veszi figyelembe, mintha többszörös

regressziót definiálnánk a végponton, mint függő változón, és a szülőkön, mint független változókon.

E paraméterezésben a korábban, a 4.1. és 4.3. fejezetben felsorolt problémák már nem fordulnak elő. A jólrendezés megoldotta az IKG csúcsainak részbenrendezettsége miatt ott előálló technikai problémákat: a kötött paraméterek előállítása egyértelmű, a feltételes függetlenségek közvetlenül megfeleltethetők nekik, és nincs inkompatibilitás az egyes feltételes függetlenségeknek megfelelő nullparaméterek között (nem fordul elő ugyanaz a null-hatás két különböző marginális mellett, és korábban nem-nulla hatás sem lehet későbbi marginális mellett nulla). Végül: a paraméterezés előnyös tulajdonságainak teljesülése is garantálva van, ugyanis a paraméterezés hierarchikus, teljes és rendezett módon dekomponálható. Ezért a 2.3. fejezetben említett tételekből adódóan az IKG modellek mindig léteznek, paramétereik értékei egymástól függetlenül interpretálhatók, továbbá a modellek illeszkedése valószínűség-hányados próbával tesztelhető, a modelleknek megfelelő maximum likelihood-becslés egyértelműen létezik és előállítható, és konfidencia-intervallum is konstruálható hozzá.

12. példa Paraméterezzük fel a 3.8. fejezetben a klasszikus szociológiai modellek között már bemutatott, alábbi Blau-Duncan modellt!



15. ábra. Az eredeti Blau-Duncan modell (1967)

A csúcsok jólrendezése: *I', F', I, F_1, F_2*. A marginálisok **M** halmaza: *I', I'F', I'F'I, I'F'IF_1, I'F'IF_1F_2*. A lokális Markov-tulajdonság implikálta $F_2 \perp I' \mid F'IF_1$ állítás miatt az alábbi paraméterek értéke 0:

$$\lambda_{I'F_2*}^{I'F'IF_1F_2},$$

ahol a * bármilyen, akár üres részhalmazt is jelölheti a marginális szóba jöhető változóinak, IF_1F_2-nek. A $F_1 \perp I' \mid F'I$ feltételes függetlenség a

$$\lambda_{I'F_1*}^{I'F'IF_1}$$

paraméterek 0-ra állítását kívánja meg. A szabad paraméterek a következők:

$$\lambda_*^{I'},\ \lambda_{F'*}^{I'F'},\ \lambda_{I*}^{I'F'I},\ \lambda_{F_1}^{I'F'IF_1},\ \lambda_{F'F_1}^{I'F'IF_1},\ \lambda_{IF_1}^{I'F'IF_1},\ \lambda_{F'IF_1}^{I'F'IF_1},\ \lambda_{F_2}^{I'F'IF_1F_2},\ \lambda_{F'F_2}^{I'F'IF_1F_2},\ \lambda_{IF_2}^{I'F'IF_1F_2},\ \lambda_{F_1F_2}^{I'F'IF_1F_2},$$
$$\lambda_{F'F_1F_2}^{I'F'IF_1F_2},\ \lambda_{IF_1F_2}^{I'F'IF_1F_2},\ \lambda_{F'IF_2}^{I'F'IF_1F_2},\ \lambda_{F'IF_1F_2}^{I'F'IF_1F_2}.$$

Megfigyelhető, hogy a szabad paraméterek éppen azok, amiket szerettünk volna: bármely változó és a szülei közötti hatások jellemezhetők segítségükkel. Ennek belátásához az (60)-et kell felhasználnunk, ez alapján pl. a paraméterezésben szereplő $\lambda_{F_1F_2}^{I'F'IF_1F_2}$ megegyezik $\lambda_{F_1F_2}^{F'IF_1F_2}$-vel, tehát az előbbi paraméter értelmezhető úgy, mint az F_1-nek az F_2-re gyakorolt, a többi szülőre feltételes hatása.

4.6 Láncgráfok

A láncgráf[33] az irányított körmentes gráf általánosítása. Mint ilyen, annál lényegesen komplikáltabb, s mint látni fogjuk, néha irreguláris modelleket definiál. A pozitív eloszlás mellett előálló hétféle[34] modellosztály (lásd az alábbi, 3. táblázatot) paraméterezése, mint a 4.1. és a 4.3. fejezetben láttuk, számos problémát vet fel. Ezekre a problémákra jelenleg még nincs ismert megoldás. Vannak eredmények, melyek a hét modell közül kettőnek a simaságát bizonyítják (Lauritzen, 1996, Drton, 2008), illetve a maradék öt közül háromról megmutatják, hogy nem minden esetben simák (Drton, 2008). Ezeket az eredményeket 3. táblázatban foglaltam össze. A modell simasága, mint azt a 2.3.3. fejezetben láttuk, azért lényeges, mert az aszimptotikus eloszlások teljesüléséhez szükséges, másrészt sima modellhez létezik sima paraméterezés, ami szükséges feltétele a hierarchikus paraméterezés létezésének, a paraméterek interpretálhatóságának.

A 2.3.3. fejezetben azt is láttuk, hogy a sima paraméterezés létezésére elégséges feltétel ismert, de szükséges feltétel nem. Egy modellosztály simaságának cáfolata ezért csak konkrét esetekre támaszkodva (adott modell mellett, a változók kategóriaszámának rögzítésével) végezhető el. Drton ennek megfelelően egy modellosztály nem-simaságának bizonyítását úgy végzi el, hogy albegrai statisztikai módszerrel, a Singular szoftver használatával egy, az osztályhoz tartozó konkrét modellről látja be, hogy nem sima. Magam szerzőtársaimmal (Rudas, Bergsma, Németh, 2009) jelenleg fejezzük be azt a munkát, mely előzetes várakozásaink szerint megmutatja, hogy a hét közül melyek a simák, ezekre hierarchikus paraméterezést fog adni, a többiről pedig belátja, hogy nincs hierarchikus paraméterezése, vagyis az alábbi táblázat minden cellája kiszínezhető lesz már. Eredményeink közvetlenül következnek Bergsma és Rudas (2002) marginális loglineáris paraméterezésekkel kapcsolatos tételeiből, elsősorban a disszertáció 5. tételéből, amely alkalmas a paraméterezés, így a modell simaságának bizonyítására is (a tétel szerint egy hierarchikus és teljes P halmazhoz tartozó $\tilde{\lambda}_P$ paraméterezés sima).

3. táblázat. A láncgráf-modellek típusai, kategoriális változók és pozitív eloszlás esetén. Sötét háttér = mindegyik sima modell, világos = van köztük nem sima, fehér = kérdéses.

Típus	*Markov-tulajdonság*						
1-es (LWF, leginkább feltételes)	Globális	=	Blokk-rekurzív	=	Lokális	=	Páronkénti
2-es (AMP)	Globális	=	Blokk-rekurzív	⇒	Lokális	=	Páronkénti
3-as			Blokk-rekurzív	⇒			Páronkénti
4-es (leginkább marginális)			Blokk-rekurzív	⇒			Páronkénti

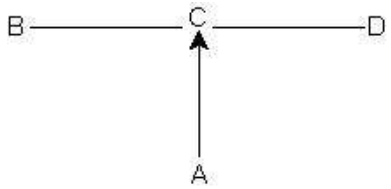
[33] Ebben a fejezetben végig 1-es típusú láncgráfot értek láncgráf alatt, mivel a létező tételek szinte kizárólag velük kapcsolatosak.

[34] Láttuk: az elvi tizenhatból csak tizenkettőt ismertet az irodalom, ezek közül öt ekvivalens a maradék hét valamelyikével.

Az alábbiakban a fenti, a láncgráf-modellek simaságát bemutató táblázat egyes celláit veszem sorra.

Az LWF-modellekből, ahogyan a táblázat is mutatja, tulajdonképpen csak egy van, hiszen pozitív eloszlás mellett valamennyien ekvivalensek. A modell Lauritzen (1996, 4.6.1 fejezet) által adott faktorizációja Drton szerint közvetlenül implikálja a modell simaságát. A modell paraméterezésére az irodalom nem ismertet módszert. A Lauritzen által mutatott faktorizáció alkalmas kiindulópont lenne, de a a 4.1. és a 4.3. fejezetben ismertetett paraméterezési problémák megoldása nem ismeretes.

AMP modellekből az ekvivalenciák miatt csak kettő van. Az AMP páronkénti modellnek sem a tulajdonságait sem a paraméterezését nem ismeri az általam feldolgozott irodalom. Az AMP blokk-rekurzív modellekről Drton (2008) megmutatja, hogy van köztük nem sima. Példája az alábbi gráf:



16. ábra. Láncgráf-modell (Drton, 2008)

A gráfból az AMP blokk-rekurzív Markov-tulajdonság mellett az alábbi feltételes függetlenségek olvashatók ki:

$$B \perp D \mid AC \text{ és } A \perp BD. \tag{61}$$

Drton a Singular segítségével látja be erről a modellről, hogy ha *A, B, C* és *D* binárisak, akkor a modell nem sima. Ilyenkor persze sima paraméterezése sincs a modellnek, vagyis (az 5. tétel alapján) hierarchikus paraméterezése sem létezik. A 4.3. fejezetben láttunk példákat arra, hogy a feltételes függetlenségek közvetlen paraméterezése hogyan okozhat inkompatibilitási problémákat, azáltal, hogy két különböző marginális mellett kellene ugyanazt a hatást 0-ra állítani, illetve azáltal, hogy a rendezésben korábbi marginális melletti nem-nulla hatás kellene későbbi marginális mellett nullára állítani. A probléma bemutatásakor mutattam egy ad hoc, egyedi esetekben működő megoldást is az inkompatibilitás kiküszöbölésére (a függetlenségek (C1)-(C4)-re épülő ekvivalens átalakításaival), de hangsúlyoztam, hogy a megoldás általánosíthatóságára nincsen garancia. A jelenlegi példa éppen ezt az általánosíthatóságot cáfolja: az *ABD* hatást az *ABD* és *ABCD* marginálison belül is 0-ban kellene rögzíteni, illetve a kisebb *ABD* marginálison belüli nem-nulla *BD* paramétert a nagyobb *ABCD* marginálison belül 0-ra kellene állítani. Mivel a modellnek nincs hierarchikus paraméterezése, tudható, hogy e problémák nem oldhatók meg a függetlenségek átalakításaival sem.

A 3. táblázat következő sora a 3-as típusú modelleké. Az előbbi gráf a 3-as típusú blokk-rekurzív Markov-tulajdonság mellett az alábbiakat implikálja:

$$B \perp D \mid A \text{ és } A \perp BD \mid C. \tag{62}$$

Erről a modellről is Drton látta be algebrai úton, hogy bináris változók mellett nem sima. Itt is megtalálhatók a hierarchikus paraméterezés létrehozását akadályozó konfigurációk: az *ABD* hatást az *ABD* és *ABCD* marginálison belül is 0-ban kellene rögzíteni, illetve az *AB, AD* hatások a kisebb marginálison belül nem-nullák, a nagyobb mellett viszont azok.

A 3-as típusú páronkénti Markov-tulajdonság szerepel a táblázat következő cellájában. Itt is Drton példája lesz alkalmazható a nem-simaság bizonyítására. Az előbbi gráf mellett a Markov-tulajdonság által implikált állítások a következők:

$$B \perp D \mid A,\ A \perp D \mid BC \text{ és } A \perp B \mid DC. \tag{63}$$

Pozitív eloszlás mellett azonban a modell ekvivalens a blokk-rekurzív párjával, mert ilyenkor alkalmazható a (C5*) levezetési szabály, amivel belátható, hogy az (63) két utolsó állítása ekvivalens az (62) utolsó állításával. Vagyis a páronkénti modellek között is van nem-sima.

A táblázat utolsó sora a 4-es típusú modelleké. A páronkénti modellekről itt sem tudunk semmit. A blokk-rekurzív modellekről viszont Drton belátja, hogy simák. Bizonyításában a feltételes valószínűségi koordinátákra olyan lineáris transzformációt alkalmaz, ami leegyszerűsíti a modellt definiáló feltételes függetlenségi állításokat.

Három modellosztályról láttuk be, hogy tartozik hozzájuk nem sima modell is. Ha viszont a modellosztályon belül van nem sima modell, akkor nem adható minden modellre sima paraméterezés, tehát az 5. tétel alapján elmondható, hogy általános esetben nincs hierarchikus paraméterezése az ehhez a három modellosztályhoz tartozó modelleknek.

A rendezett módon dekomponálható paraméterezés létezése továbbra is kérdéses mindegyik modellosztály esetén.

13. példa Bár nincsen általános válasz az AMP páronkénti modellek paraméterezhetőségéről, egyedi esetekben adható jó paraméterezés. Ilyen eset a 9. ábra középső gráfjához tartozó modell, amely a Blau-Duncan modell egyetlen helyen (a jelenlegi foglalkozás és az iskolázottság közötti nyíl vonalra cserélésével) módosított változata. Láttuk, hogy a modell az

$$\begin{aligned} F_1 &\perp I' \mid F', \\ F_2 &\perp I' \mid F' \end{aligned} \tag{64}$$

állításokkal adható meg. Azt az eljárást használjuk, hogy az 1. tétellel egyenként paraméterezzük a feltételes függetlenségeket. Az első állítás így az $I'F'F_1$, a második az $I'F'F_2$ marginálison belül paraméterezhető, mégpedig az alábbi paraméterek nullára állításával:

$$\lambda_{I'F_1*}^{I'F'F_1},\ \lambda_{I'F_2*}^{I'F'F_2}\ .$$

A paraméterezés hierarchikus és rendezett módon dekomponálható. A szabad paraméterek megválasztásánál arra törekszünk, hogy a hozzájuk tartozó marginálisok a gráf komponens-struktúráját és a komponensek rendezését is kövessék. Vagyis pl. ha K_1 az első, K_2 a második komponense egy kétkomponensű láncgráfnak, akkor a K_1 lesz az első marginális (hogy a K_1-en

belül hiányzó vonalaknak megfelelő függetlenségek paraméterezhetők legyenek[35]), $K_1 \cup K_2$ az utolsó, és minden, $k \in K_2$-ből hiányzó nyíl miatt még a $K_1 \cup k$ is marginális kell, hogy szerepeljen[36]. Példánkra visszatérve: ha az első komponens miatt az *I'F'*-et, a második miatt az $I'F'IF_1F_2$-t is hozzávesszük az eddigi marginálisokhoz, akkor a paraméterezés teljessé tehető az alábbi szabad paraméterekkel:

$$\lambda_{\varnothing}^{I'F'},\ \lambda_{F'}^{I'F'},\ \lambda_{I'}^{I'F'},\ \lambda_{I'F'}^{I'F'},\ \lambda_{F_1}^{I'F'F_1},\ \lambda_{F'F_1}^{I'F'F_1},\ \lambda_{F_2}^{I'F'F_2},\ \lambda_{F'F_2}^{I'F'F_2},\lambda_{I'F'IF_1}^{I'F'IF_1F_2},\ \lambda_{I}^{I'F'IF_1F_2},\ \lambda_{I'I}^{I'F'IF_1F_2},\ \lambda_{F'I}^{I'F'IF_1F_2},$$
$$\lambda_{I'F'I}^{I'F'IF_1F_2},\ \lambda_{I'IF_1}^{I'F'IF_1F_2},\ \lambda_{F'IF_1}^{I'F'IF_1F_2},\ \lambda_{IF_2}^{I'F'IF_1F_2},\ \lambda_{F_1F_2}^{I'F'IF_1F_2},\ \lambda_{IF_1F_2}^{I'F'IF_1F_2},\ \lambda_{F'F_1F_2}^{I'F'IF_1F_2},\ \lambda_{F'IF_2}^{I'F'IF_1F_2},\ \lambda_{F'IF_1F_2}^{I'F'IF_1F_2}.$$

Az így előálló paraméterezés hierarchikus, a négy marginális *I'F'*, $I'F'F_1$, $I'F'F_2$, $I'F'IF_1F_2$ rendezése pedig rendezett módon dekomponálható.

A szabad paraméterekhez tartozó becslések interpretálásánál érdemes kihasználni, hogy a (60) megfelelője itt is alkalmazható, így a szülő-leszármazott feltételes asszociációk feltételhalmazából a nem-szülők elhagyhatók. Pl. a $\lambda_{F'F_2}^{I'F'F_2}$ paraméter a $\lambda_{F'F_2}^{F'F_2}$-vel egyezik meg, ami sokkal kézenfekvőbb értelmezést tesz lehetővé.

Vegyük azt is észre, hogy ez nem az egyetlen lehetséges megválasztása a szabad paramétereknek. Egyedül az $I'F'IF_1F_2$ új marginálisként történő bevonásával is megoldható a teljessé tétel, az alábbi, szintén hierarchikus és rendezett módon dekomponálható paraméterekkel, de ezek nem követik a gráf komponens-struktúráját, így nem azokat a hatásokat mérik, amiket szeretnénk (pl. az apát jellemző *I'F'* hatást az utód első foglalkozására feltételesen veszi):

$$\lambda_{\varnothing}^{I'F'F_1},\ \lambda_{F'}^{I'F'F_1},\ \lambda_{I'}^{I'F'F_1},\ \lambda_{I'F'}^{I'F'F_1},\ \lambda_{F_1}^{I'F'F_1},\ \lambda_{F'F_1}^{I'F'F_1},\ \lambda_{F_2}^{I'F'F_2},\ \lambda_{F'F_2}^{I'F'F_2},\lambda_{I'F'IF_1}^{I'F'IF_1F_2},\ \lambda_{I}^{I'F'IF_1F_2},\ \lambda_{I'I}^{I'F'IF_1F_2},\ \lambda_{F'I}^{I'F'IF_1F_2},$$
$$\lambda_{I'F'I}^{I'F'IF_1F_2},\ \lambda_{I'IF_1}^{I'F'IF_1F_2},\ \lambda_{F'IF_1}^{I'F'IF_1F_2},\ \lambda_{IF_2}^{I'F'IF_1F_2},\ \lambda_{F_1F_2}^{I'F'IF_1F_2},\ \lambda_{IF_1F_2}^{I'F'IF_1F_2},\ \lambda_{F'F_1F_2}^{I'F'IF_1F_2},\ \lambda_{F'IF_2}^{I'F'IF_1F_2},\ \lambda_{F'IF_1F_2}^{I'F'IF_1F_2}.$$

Az alkalmazások között további példákat találhatunk láncgráfoknak ezen az úton történő paraméterezésére. Fontos hangsúlyoznom ugyanakkor, hogy ez az eljárás, amikor az 1. tétellel paraméterezzük egyenként a feltételes függetlenségeket, csak egyedi megoldásként használható. Általánosítása több problémába, köztük a 4.3. fejezetben korábban már felsoroltakba ütközik.

4.7 Útmodellek

Láttuk, hogy Goodman módosított útmodellje a feltételes függetlenségi megszorításokon kívül további magasabbrendű hatások elhagyását is lehetővé tette, tehát a grafikus modellnél általánosabb modell volt. Ebben az alfejezetben Rudas, Bergsma, Németh (2006) munkáját[37] felhasználva de azt általánosítva az IKG modelleknek erre a módosítására és a kapott modellosztály paraméterezésére mutatok megoldást. A kapott modellosztályt irányított körmentes útmodelleknek (IKG útmodelleknek) nevezem majd.

[35] Ez az itt használt AMP-tulajdonság mellett érvényes. Ha helyette 3-as típusú Markov-tulajdonságot használnák, akkor a K_1 minden hiányzó vonalához tartozó csúcspár külön marginális kellene, hogy legyen.

[36] Ez az itt használt AMP-tulajdonság mellett érvényes. Ha helyette LWF Markov-tulajdonságot használnák, akkor elég lenne a $K_1 \cup K_2$ szerepeltetése a marginálisok között.

[37] A hivatkozott munka csak olyan útmodelleket tárgyal, ahol minden, másodfokúnál nagyobb hatás törlődik a modellből. Ezzel szemben itt a magasabbrendű hatások csak részben is törölhetők.

Az IKG útmodellek tehát Goodmant követve úgy kaphatók az IKG modellekből, hogy adott α csúcs és szülei (β_1 .... β_n) közötti kapcsolatok interakciója szabadon szűkíthető, tehát pl. a β_1 .és β_2 α-ra gyakorolt hatásával kapcsolatban feltehetjük, hogy azok nincsenek interakcióban. Ez loglineáris paraméterezésben a $\lambda_{\alpha\beta_1\beta_2}^{\alpha\cup pa(\alpha)}$ paraméter 0-ra állításával érhető el. További tetszőleges megszorítások tehetők a

$$\lambda_{\alpha L}^{\alpha\cup pa(\alpha)},\ L\subseteq pa(\alpha),\ |L|\geq 2 \tag{65}$$

paraméterek bármelyikének 0-ra állításával. Mivel tudjuk, hogy az IKG modell paraméterei hierarchikusak és variációsan függetlenek, ezért ezek a megkötések szabadon megtehetők, mindig létező modellt fognak definiálni, és a megkötések nem befolyásolják a megmaradó szabad paraméterek interpretációját.

Ezzel sikeresen parametrizáltuk Goodman modelljét. Két potenciálisan jól használható és az IKG útmodellből levezetett modellt szeretnék még ebben a fejezetben bemutatni. Az első azzal a megkötéssel adódik, hogy ha az IKG-modellre alkalmazott interakciós szűkítéseknél egy L hatáshoz tartozó paramétert 0-ra állítunk, akkor minden, nála nagyobb hatáshoz tartozó paramétert is 0-ra kell állítanunk. A hagyományos loglineáris esetben az ennek a megkötésnek megfelelő modellt nevezik hierarchikus modellnek, ezért ezekre a modellekre *hierarchikus útmodell*ként fogok hivatkozni. A megkötést az motiválhatja, hogy így a modell a 12. ábra látott módon, a nyilak ponttal való összekötésével reprezentálható, és a paramétermegkötések is jobban értelmezhetők. Pl. nem fordulhat elő az a nehezen interpretálható eset, hogy a D-re gyakorolt hatások vizsgálatában az A és B interakcióját eltávolítjuk, de meghagyjuk az A, B és C interakcióját.

Az utolsóként bemutatott modell szintén az IKG útmodellek almodellje, és Rudas, Bergsma, Németh (2006) ezt vezeti be *útmodell*ként (*path model*) tanulmányában. A modell az IKG modellből minden, a (65)-ben felsorolt lehetséges interakciós hatást eltöröl, tehát csak a következő szabad paramétereket engedi meg: $\lambda_{\alpha\beta}^{\alpha\cup pa(\alpha)},\ \beta\in\{pa(\alpha),\varnothing\}$. Ezek azok a paraméterek, amelyek (a $\beta=\varnothing$ esetben) vagy a leszármazottnak a szülőkre feltételes eloszlását, vagy (a $\beta\in pa(\alpha)$ esetben) a többi szülőre feltételes szülő-leszármazott kapcsolatokat paraméterezik. Az utóbbi paraméterek becslései könnyen bemutathatók az IKG nyilaira írt paraméterekkel. Egy ilyen példát mutatunk is a már említett 2006-os cikkünkben.

E két utóbbi modell megint közvetlenül volt megadható az IKG modell paraméterezésén, jól példázva a marginális loglineáris modellek flexibilitását. A láncgráf-modellek paraméterezésének tisztázása után vélhetően a láncgráfokon definiált útmodellek általános paraméterezése is tisztázódik majd. Ilyen láncgráfon definiált útmodellekre látunk majd példát az alkalmazások között, a 7.4. fejezetben.

4.8 MLH becslés, illeszkedésvizsgálat

A 2.3. fejezetben láttuk, hogy ha a modellünk paraméterezése hierarchikus és teljes (az IG-knál és IKG-knál mindig van ilyen, a láncgráfoknál nem mindig), akkor a szabad paraméterek maximum likelihood-becslése egyértelműen létezik és előállítható. A becslés előállíthatósága abból következik, hogy a fenti feltételek esetén a maximum likelihood-becslés a likelihood egyenlet stacionárius pontja. Ilyenkor az un. gradiens módszer alkalmazható a maximum

megtalálására. Több szerző is közölt ilyen algoritmust, én Wicher Bergsma (1997) algoritmusát használom az alkalmazásokban.

A hierarchikus és teljes paraméterezés további, már említett előnye, hogy a modellek illeszkedése valószínűség-hányados próbával tesztelhető, és a hozzá tartozó szabadságfok a kötött paraméterek számával egyezik meg. A valószínűség-hányados próba alkalmazhatóságának az is előnye, hogy ily módon két egymásbaágyazott modell összehasonlítása is lehetséges[38]. A szűkebb modell ilyenkor a másik továbbspecifikálásával jön létre. Ez a továbbspecifikálás definíció szerint a szabadságfok növekedésével jár, és általában a tesztstatisztika növekedését is eredményezi, azonban nem biztos, hogy olyan mértékben, ami ne lenne elfogadható a modell takarékosságáért cserébe. Erre a kérdésre ad objektív választ a valószínűség-hányados próba, kihasználva, hogy a két modell tesztstatisztika-értékének különbsége aszimptotikusan maga is khi-négyzet eloszlást követ, szabadságfoka a továbbspecifikációval bevont új kötött paraméterek száma. Grafikus modellek vizsgálatánál gyakran adódnak egymásbaágyazott modellek, mert, mint korábban többször láttuk, a gráfból egy-egy él törlése szűkebb modellhez vezet. Ilyen esetre látunk majd alkalmazási példát a 7.1. fejezetben.

Nézzünk most egy példát a modell szabadságfokának kiszámítására. Tekintsük az eredeti Blau-Duncan modellt. A modell által adott feltételes függetlenségeket az 3.8. fejezetben olvastuk ki a gráfból, a modell felparaméterezése pedig a 4.5. fejezetben történt meg. A modell az *I', F', I, F_1, F_2* változók ötdimenziós kontingencia-tábláján van definiálva. A szokott módon tegyük fel, hogy az *I* és az *I'* bináris változók, míg a többi háromértékű. Ez 2*3*2*3*3=108 cellát jelent, ami éppen ennyi független paraméterrel paraméterezhető fel.

A 4.5. fejezetben adtuk meg az alábbi szabad és kötött paramétereket.

Kötött: $\lambda_{I'F_2*}^{I'F'IF_1F_2}$, $\lambda_{I'F_1*}^{I'F'IF_1}$,

szabad: $\lambda_{*}^{I'}$, $\lambda_{F'*}^{I'F'}$, $\lambda_{I*}^{I'F'I}$, $\lambda_{F_1}^{I'F'IF_1}$, $\lambda_{F'F_1}^{I'F'IF_1}$, $\lambda_{IF_1}^{I'F'IF_1}$, $\lambda_{F'IF_1}^{I'F'IF_1}$, $\lambda_{F_2}^{I'F'IF_1F_2}$, $\lambda_{F'F_2}^{I'F'IF_1F_2}$, $\lambda_{IF_2}^{I'F'IF_1F_2}$, $\lambda_{F_1F_2}^{I'F'IF_1F_2}$, $\lambda_{F'F_1F_2}^{I'F'IF_1F_2}$, $\lambda_{IF_1F_2}^{I'F'IF_1F_2}$, $\lambda_{F'IF_2}^{I'F'IF_1F_2}$, $\lambda_{F'IF_1F_2}^{I'F'IF_1F_2}$,

ahol a * a marginális bármely, akár üres részhalmazát jelöli. Számoljuk ki a modell szabadságfokát, amire a modell illeszkedésének vizsgálatánál lenne szükségünk! A szabadságfok a kötött paraméterek számával egyezik meg. Itt a $\lambda_{I'F_2*}^{I'F'IF_1F_2}$ több paramétert jelöl, a * helyén az *I, F_1* és *F'* bármely részhalmaza állhat (összesen 2*2*2 ilyen részhalmaz van). Ez lehet pl. az üres halmaz, ekkor a paraméterhez az *I'F_2* hatás tartozik, amihez 2*3 paraméterérték tartozik (hiszen az *I'* kétértékű, az F_2 háromértékű). A nem-redundáns paraméterek, mint a $\tilde{\lambda}_P$-t meghatározó 7. definícióban láttuk, változónként egy érték elhagyásával állnak elő. Így a 2*3 paraméterérték közül (2-1)*(3-1)=2 nem-redundáns van. A $\lambda_{I'F_2*}^{I'F'IF_1F_2}$ és $\lambda_{I'F_1*}^{I'F'IF_1}$ paraméterekhez tartozó valamennyi hatásra hasonlóan számolható ki a nem-redundáns paraméterek száma:

[38] Természetesen az, hogy az egyik modell a másiknál jobban illeszkedik, nem jelenti azt, hogy ténylegesen jól illeszkedik az adatokhoz; a bővebb modell jó illeszkedése feltétele a jobban illeszkedő szűkebb modell elfogadásának.

marginális	**hatás**	**nem-redundáns paraméterek száma**
$I'F'IF_1F_2$	$I'F_2$	2
	$I'F'F_2$	4
	$I'IF_2$	2
	$I' F_1F_2$	4
	$I'F'IF_2$	4
	$I'F'F_1F_2$	8
	$I'IF_1F_2$	4
	$I'F'IF_1F_2$	8
$I'F'IF_1$	$I'F_1$	2
	$I'F'F_1$	4
	$I'IF_1$	2
	$I'F'IF_1$	4

Ez 48 paraméter, ha tehát a Blau-Duncan modell adatainkhoz való illeszkedését vizsgálnák, a valószínűséghányados-próbát 48-as szabadságfok mellett kellene értékelni. A szabad paraméterek száma pedig 108-48=60.

A 7. fejezetben bemutatott alkalmazások konkrét példáit adják majd a grafikus modellek illesztésének és a paraméterek becslésének. Az egyik alkalmazás teljes számítástechnikai megvalósítása pedig a 8. fejezetben található meg.

5 NYITVA MARADT MÓDSZERTANI KÉRDÉSEK

Néhány helyen már említettem, mely pontok kérdésesek jelenleg is a grafikus modellek kutatásában. Ezeket foglalnám röviden össze az alábbiakban.

A láncgráfokon pozitív eloszlás mellett hét modell volt definiálható, lásd az alábbi, korábban már szerepelt összefoglaló táblázatot. (A hét modell úgy áll elő, hogy az ekvivalens modelleket egyetlen modellnek vettem. A táblázat üres celláinak megfelelő modelleket nem definiálja az irodalom.) Ezek közül kettőről belátható, hogy van hozzá sima paraméterezés. Kérdéses, hogy ez a sima paraméterezés általánosan hogyan adható meg. Az is kérdés, adható-e hozzájuk a modell létezését és a becslések standard aszimptotikus tulajdonságait garantáló hierarchikus (ezenkívül esetleg még a variációs függetlenséget garantáló rendezett módon való dekomponálhatósággal is rendelkező) paraméterezés. Két további modellről még az sem tudható jelenleg, hogy létezik-e hozzá sima paraméterezés. Szerzőtársaimmal, Rudas Tamással és Wicher Bergsmával jelenleg befejezéséhez közeledő (?) kutatásunk reményeink szerint választ ad ezekre a kérdésekre.

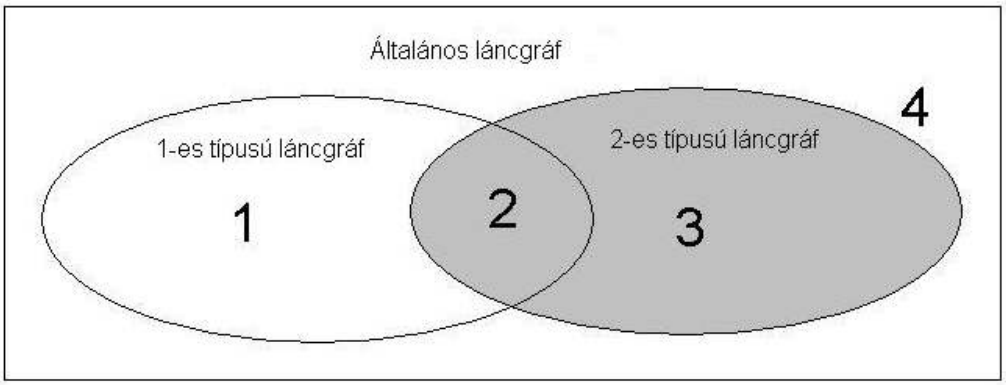
Drton (2008) is megemlíti cikkének végén, hogy a láncgráfok simaságának végleges tisztázását szolgálná, ha a nem sima modelleket biztosan tartalmazó 2-es és 3-as típusok sima modelljeit sikerülne karakterizálni. Kérdés az is, van-e olyan sima modell a 2-es és 3-as típuson belül, ami a típus valódi tagja, vagyis amihez nincs vele Markov-ekvivalens 1-es vagy 4-es típusú modell.

4. táblázat. A láncgráf-modellek típusai, kategoriális változók és pozitív eloszlás esetén. Sötét háttér = mindegyik sima modell, világos = van köztük nem sima, fehér = kérdéses.

Típus	*Markov-tulajdonság*							
1-es (LWF, leginkább feltételes)	Globális	=	Blokk-rekurzív	=	Lokális	=	Páronkénti	
2-es (AMP)	Globális	=	Blokk-rekurzív	⇒	Lokális	=	Páronkénti	
3-as			Blokk-rekurzív	⇒			Páronkénti	
4-es (leginkább marginális)			Blokk-rekurzív	⇒			Páronkénti	

A láncgráfok tárgyalásának egységesítését szolgálná, ha a táblázat jelenleg üres négy cellája is ki lenne töltve, tehát ha a 3-as és 4-es páronkénti és blokk-rekurzív Markov-tulajdonságoknak is lenne lokális és globális megfelelője.

A párhuzamosan futó kutatások eredményeinek egyesítését és a láncgráfok megközelítésének egységesítését szolgálná, ha tisztázódna az 1-es ill 2-es típusú láncgráfokhoz tartozó modellosztályok viszonya (lásd az alábbi, a disszertációban már szintén szerepelt ábrát). Sikerült belátnom, hogy a két típus nem Markov-ekvivalens: van olyan 2-es típusú gráfhoz tartozó modell, ami nem adható meg 1-es típusúval. Kérdés maradt viszont, hogy teljesülnek-e a 2-es típusú láncgráfokra az 1-es típusra már belátott tételek (elsősorban a fenti táblázatban feltüntetett ekvivalenciák). Szintén kérdés maradt, hogy az ábrán 1-essel és 4-essel jelölt halmazok üresek-e.



17. ábra. Az általános, illetve az 1-es ill. 2-es típusú láncgráfokon definiált modellosztályok tartalmazási viszonya (sötét háttér: nemüres, fehér háttér: nem ismert, hogy üres-e)

Láttuk, hogy milyen kézenfekvő módon volt bevezethető útmodell az irányított körmentes gráfokon. Az előbbi kérdések megválaszolása után vélhetően a láncgráfokon definiált útmodellek paraméterezése is megoldható majd.

Végül az elemzési módszer társadalomtudományi ismertségét és alkalmazhatóságát támogatná, ha felhasználóbarát statisztikai program íródna hozzá.

Remélem, a jövőben lehetőségem nyílik ezeket a kérdéseket legalább részben tisztázni.

6 ALKALMAZÁS

6.1 A grafikus modellek általános alkalmazási problémái

6.1.1 A gráf felállítása

Egy, a vizsgálatunk tárgyát képező gráf alapvetően két módon állhat elő. Az első, Bayes-félének is nevezhető megközelítésben (pl. a 3.1.-ben említett szakértői rendszerek esetében) tipikusan nagyszámú változónk van, és az adott eloszlást legjobban reprezentáló (értsd: ahhoz statisztikai értelemben legjobban illeszkedő modellt adó) gráfot keressük a lehetséges összes gráf közül, a „lehetséges összes gráf”-ra esetleg olyan a priori megkötéseket téve, mint a csúcsok időben adott rendezése. Ilyenkor hatékony kereső algoritmus konstruálása a feladat (pl. Spirtes et al., 1993). A másik megközelítésben, és a disszertációban ilyen esetekre látunk majd alkalmazási példákat, a gráf felállítása ismereteinken vagy hipotéziseinken alapszik, és a hipotézis elfogadását vagy elutasítását az adott eloszlás dönti el. Ezek a fentinél könnyebben kezelhető esetek, csak néhány (négy-tíz) változóval. Természetesen ebben az esetben is lehetséges több szóba jöhető modell összehasonlítása, pl. modellépítésnél újabb megszorítások bevonásával (egy ilyen alkalmazás a 7.5. fejezetben szerepel), vagy ha két hipotézist szeretnénk összevetni (pl. 7.1. fejezet).

Kérdés lehet, hogy a gráf egy adott csúcspárja között nyilat vagy vonalat használjunk-e. Lauritzen és Richardson (2003) azokat az eseteket tárgyalja, ahol IKG-k helyett LG-k használata a megfelelő, vagyis ahol egy nyilat vonalra érdemesebb cserélni. Az alábbiakban a fordított esetről írok röviden, vagyis arról, amikor vonal helyett nyíl használata lehet indokolt.

Az irányított élekkel reprezentált asszociációk oksági viszonyként való interpretálhatósága fontos kérdés. Éppen e kauzális megközelítés képezi a grafikus modellekkel foglalkozó mai statisztika egyik fő problémáját (pl. Pearl 1995, Freedman 1997, 2001, Cox, Wermuth, 2001, 2004). Itt Cox és Wermuth írásai (2001, 2004) alapján e problémára, illetve általában a kauzalitás három elterjedt definíciójának statisztikai szempontú összevetésére térnék ki röviden.

A gráf két pontja közötti irányított él használata, vagyis az általuk reprezentált változó-pár magyarázó ill. függő változó szereposztása nem feltételezi feltétlenül a kauzalitás meglétét. A nyíl használata általában két lehetséges alapon indokolható. Az egyik érvelés szerint a korábbi időponttal összefüggésbe hozható változó (iskolázottság) magyarázó, a későbbi időponthoz kapcsolható (foglalkozás) függő változó lehet. Egy másik érvelés a változók tényleges tartalmával kapcsolatos (elméleti kiindulású vagy empirikus adatokkal alátámasztott) munkahipotézisre épül. Például a politikai attitűd függő és a jövedelmi helyzet magyarázó változóként történő kezelése egy komplexebb mechanizmus részeként ugyan, de plauzibilisnek tűnik.

Az irányított él irányának ily módon történő eldöntése után az egyszerű asszociáció helyett a tényleges kauzalitás feltételezése a közkeletű statisztikai megközelítés szerint (kauzalitás *1. definíciója*) akkor engedhető meg, ha a magyarázó (M) és függő (F) változó között kimutatható statisztikai összefüggés és annak előjele megmarad bármely más magyarázó változóra (M’) történő korrigálás után is.

A kauzalitás egy másik, hipotetikus „intervencióval” történő meghatározása (*2. definíció*) ennél erősebb interpretációját adja a fogalomnak. Eszerint az M (az egyszerűség kedvéért igen / nem értékűként, az intervenció végrehajtása / elmaradása) oka az F-nek, ha az M „igen” értéke esetén az F értéke a populációban szisztematikusan magasabb (vagy alacsonyabb), mint az M „nem” értéke esetén, az F-fel potenciális kapcsolatban álló minden M’ változatlan volta mellett.

Végül – a tudományos fogalomhasználatot leginkább lefedő – *3. definíció* az M-től F-ig vezető folyamat tényleges megértését feltételezi. E megértés tipikusan elméleti alapokon áll, vagy a szóban forgó adatoknál alacsonyabb szintre vonatkozó ismereteinkre vezethető vissza. Erre a visszavezetésre példa a szociológia makroszintű jelenségekre vonatkozó magyarázatainak az egyéni cselekvések szintjén működő racionális döntéselméletre történő hivatkozása.

6.1.2 Konfigurációk, melyek klasszikus asszociációs problémákat fednek

Az alábbi, háromcsúcsú gráfra érvényes példák alkalmasak arra, hogy nagyobb gráfokban is hasonló összefüggéseket segítsenek felismerni. Irányított körmentes gráfokban vagy láncgráfokban fordulhatnak elő, a szakirodalomban hamis függőség (*spurious dependency*), hamis asszociáció (*spurious association*), illetve kiválasztási torzítás (*selection bias*) néven ismertek (Wermuth, 2003).

18. ábra. Klasszikus problémákat fedő konfigurációk három csúcson

Mindhárom eset három változón értelmezett, és egy háromcsúcsú IKG olyan konfigurációit reprezentálják, ahol az egyik csúcspár között nincsen nyíl. Az első gráf szerint

$$S \perp N \mid K,\ \cancel{S \perp N},$$

vagyis a sikeres elhelyezkedés feltételesen független a nemtől, de a képzés típusának figyelembe nem vételekor (a szerinte történő marginalizálás esetén) nem függetlenek, ilyenkor *hamis függőség* jön létre köztük. Hasonlóan, a második gráfból a

$$S \perp G \mid H,\ \cancel{S \perp G}$$

állítások olvashatók ki, tehát gólyák száma feltételesen független a születésszámtól, de a háztetők száma szerinti marginalizálás *hamis asszociáció*t eredményezne. A harmadik gráf szerint

$$E \perp N,\ \cancel{E \perp N \mid J},$$

vagyis a nem és az elvégzett osztályok száma (marginálisan) független egymástól, de a jövedelmet feltéve összefüggést mutathatnak. Pl. a férfiak magasabb keresete és a keresetnek az iskolai osztályokkal fennálló pozitív kapcsolata esetén adott jövedelmi kategórián belül a nők magasabb iskolai végzettséget mutatnának. *Kiválasztási torzítás* akkor lépne fel, ha pl. csak magas jövedelmű személyeket válogatnánk abba a mintába, amin a nem és az iskolázottság kapcsolatát vizsgálnánk.

Vegyük észre, hogy mindhárom példa a feltételes, ill. marginális kapcsolat közti különbségtételre vezethető vissza.

Nagyobb gráfra alkalmazva a fentieket, vegyük példának a 11. ábra látott IKG modellt, amit, mint említettem, az adott adatbázishoz egy kereső algoritmus választott ki legjobban illeszkedő gráfként. A gráf értelmezésekor érdemes megkeresni a fenti konfigurációkat, mégpedig úgy, hogy a három csúcs helyén csúcshalmazok is állhatnak. Az első és a harmadik konfiguráció fedezhető fel. Az első konfiguráció megvalósulása: nem→(szülői bíztatás, intelligencia, társadalmi helyzet)→felsőfokú továbbtanulási tervek. A nem a többi tényezőre nézve feltételesen független a felsőfokú továbbtanulási tervektől, ám ha őket nem vennénk figyelembe, hamis függőség alakulna ki köztük. Valószínűsíthető pl., hogy rögzített társadalmi helyzet és rögzített intelligenciaszint mellett a fiús szülők hajlamosabbak gyerekeiket továbbtanulásra bíztatni, és mivel a jobban bíztatott gyerekek nagyobb arányban terveznek továbbtanulást, kialakulhat az a marginális összefüggés, hogy a fiúk ambiciózusabbak a továbbtanulást illetően. A harmadik konfigurációnak két megvalósulása is van, a nem→szülői bíztatás←intelligencia, illetve a nem→szülői bíztatás←társadalmi helyzet kapcsolatok. Itt is kiválasztási torzítás léphet fel, mert, akárcsak a háromcsúcsú gráfban, az össze nem kötött csúcsok (NI ill. NT) itt is marginálisan függetlenek. Eszerint ha a szülők hajlamosak a fiúkat a lányokkal összevetve már egy alacsonyabb intelligenciaszint esetén is továbbtanulásra bíztatni, akkor, bár a nem és az intelligencia függetlenek, a szülői bíztatást feltéve összefüggés mutatkozik köztük (a lányok intelligensebbek). Hasonlóan, ha a fiúkat már egy alacsonyabb társadalmi helyzet esetén is továbbtanulásra bíztatják szüleik, a nem és a társadalmi helyzet a szülői bíztatást feltéve összefügg (a lányoknak kedvezőbb a társadalmi háttere).

6.1.3 Direkt és indirekt hatás

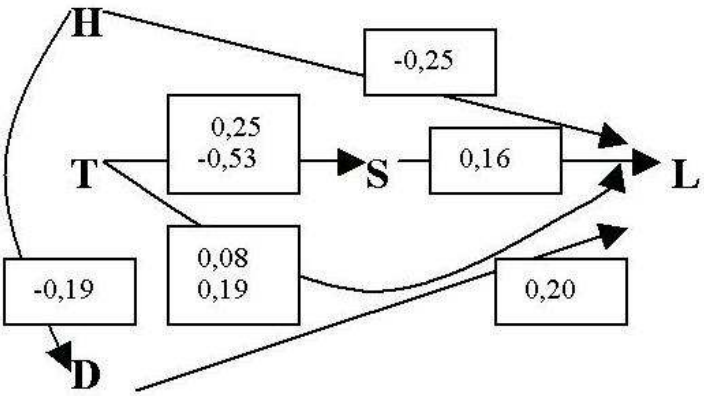
A klasszikus Wright-féle útelemzéshez hasonlóan a grafikus modellek esetén is megkülönböztethetünk direkt és indirekt hatásokat, ha két változó között egy ill. több lépés hosszú utat is tudunk találni. Jó példa erre Rudas, Bergsma, Németh (1996) egy alkalmazása, ahol egy, a kategoriális adatelemzés irodalmában sokszor elemzett (pl. Fienberg, 1970, Bishop et al., 1975) biológiai adatbázist vizsgáltak újra grafikus modellezéssel. Az adatok két gyíkfaj, a grahami és az opalinus napközben megfigyelt viselkedését írják le, és a gyíkok tartózkodási helyével kapcsolatban (*perching site*) a megfigyelt gyík faját (L), a megfigyelés időpontját (T), a tartózkodási hely magasságát (H), átmérőjét (D), és benapozottságát (S) rögzítik. A változók kategóriái a következők: faj (L): grahami/opalinus, időpont (T): délelőtt/dél/délután, magasság (H): 5 láb alatt/felett, átmérő (D): 2 ujjnál nagyobb/kisebb, benapozottság (S): napos/árnyékos.

Ha célunk egy adott helyszínen talált gyík faját meghatározó tényezőknek és a köztük levő kapcsolatoknak az azonosítása, akkor pl. a 19. ábra IKG modellje, ami a lokális Markov-tulajdonság mellett értelmezendő, elfogadható lehet. A lokális Markov-tulajdonság az alábbi feltételes függetlenségeket implikálja:

$$HD \perp T, D \perp S \mid TH,$$

amelyek szerint (1) a tartózkodási hely fizikai jellemzői (magassága és átmérője) függetlenek a megfigyelés idejétől, és (2) a tartózkodási hely átmérője nincs közvetlen hatással arra, hogy a hely napos-e, vagyis a két változó a tartózkodási hely többi adottságát feltéve független.

19. ábra. Paraméterbecslések, DAG útmodell



Az ábrán jelölt paraméterbecslések ennél az IKG-nél szűkebb modellhez tartoznak. Ez a szűkebb modell minden, a gráf nyilainak közvetlenül meg nem feleltethető hatás elhagyásával (lásd a 4.7. fejezet végén az IKG útmodellek utolsó variánsát) és minden további, nyílhoz kapcsolódó, de nem szignifikáns hatás elhagyásával adódott. A modell illeszkedése a valószínűséghányados-próba szerint elfogadható (p=0,81), és a kiinduló, egyébként jól illeszkedő (p=0,35) IKG modellhez képest sem szignifikáns az illeszkedésromlás (p=0,93).

Az ábrán bemutatott paraméterek marginális loglineáris paraméterek, a modell egy hierarchikus és rendezett módon dekomponálható paraméterezéséhez tartoznak (az IKG-khez tartozó hierarchikus paraméterezés megadási módját lásd a 4.5. fejezetben), tehát rendelkeznek a 2.3. fejezetben felsorolt valamennyi előnyös tulajdonsággal. A paraméterezés marginálisai: *H, HD, HDT, HDTS, HDTSL.* A nullparaméterek a következők:

$$\lambda_{HT}^{HDT}, \lambda_{DT}^{HDT}, \lambda_{HDT}^{HDT}, \lambda_{HS}^{HDTS}, \lambda_{DS}^{HDTS}, \lambda_{HDS}^{HDTS}, \lambda_{HTS}^{HDTS}, \lambda_{DTS}^{HDTS}, \lambda_{HDTS}^{HDTS}, \lambda_{HDL}^{HDTSL}, \lambda_{HTL}^{HDTSL}, \lambda_{HSL}^{HDTSL}, \lambda_{DTL}^{HDTSL}, \lambda_{TSL}^{HDTSL},$$
$$\lambda_{HDTL}^{HDTSL}, \lambda_{HDSL}^{HDTSL}, \lambda_{HTSL}^{HDTSL}, \lambda_{DTSL}^{HDTSL}, \lambda_{HDTSL}^{HDTSL},$$

míg a szabad paraméterek az alábbiak:

$$\lambda_{\varnothing}^{H}, \lambda_{H}^{H}, \lambda_{D}^{HD}, \lambda_{HD}^{HD}, \lambda_{T}^{HDT}, \lambda_{S}^{HDTS}, \lambda_{TS}^{HDTS}, \lambda_{L}^{HDTSL}, \lambda_{HL}^{HDTSL}, \lambda_{DL}^{HDTSL}, \lambda_{TL}^{HDTSL}, \lambda_{SL}^{HDTSL}.$$

A szabad paraméterek közül a másodfokú hatáshoz tartozó, a nyilakhoz kapcsolható paramétereket tüntettük csak fel. Közülük is csak a nem-redundánsakat mutatjuk, szokott módon a változók utolsó értékéhez tartozó redundáns paramétereket hagyva el. Látható, hogy a *T* és az *L* között két út van, tehát az időpontnak a gyíkfajra gyakorolt direkt hatása és a benapozottságon keresztül gyakorolt indirekt hatása elkülöníthető. A direkt hatás paraméterei alapján elmondható,

hogy az egyes időpontokban a gyíkok nem számarányuknak megfelelően figyelhetők meg. A délben megfigyelt gyíkok inkább grahamik, mint opalinusok ($\lambda_{TL}^{HDTSL}(\text{délben, grahami}) = 0{,}19$), de délelőtt is inkább grahamik láthatók, még ha kisebb különbséggel is ($\lambda_{TL}^{HDTSL}(\text{délelőtt, grahami}) = 0{,}08$). A harmadik, redundáns paraméter ezekből kiszámolható, és az opalinusok délutáni fölényét mutatja $\lambda_{TL}^{HDTSL}(\text{délután, grahami}) = 0 - 0{,}19 - 0{,}08 = -0{,}27$. Az indirekt hatás útvonalát követve, annak első lépésében $\lambda_{TS}^{HDTS}(\text{délelőtt, napos}) = 0{,}25$, míg $\lambda_{TS}^{HDTS}(\text{délben, napos}) = -0{,}53$, amiből $\lambda_{TS}^{HDTS}(\text{délután, napos}) = 0{,}28$, vagyis délben inkább árnyékos helyen találhatók a gyíkok, délelőtt és este pedig napos helyeken. Az indirekt út második lépésében, az *S-L* nyílon pedig $\lambda_{SL}^{HDTSL}(\text{napos, grahami}) = 0{,}16$ szerepel, ami azt jelzi, hogy napos helyeken inkább grahami figyelhető meg. Tehát a direkt és az indirekt hatás ellenkező előjelű: délben, minden egyéb tényezőre, így a benapozásra is korrigálva inkább grahamik láthatók, ugyanakkor a déli órákban megfigyelt tartózkodási helyek inkább árnyékosak, és árnyékos helyeken inkább opalinusok találhatók.

6.1.4 A grafikus modell és a strukturális egyenletek modell (SEM) összevetése

A strukturális egyenletek modell (*structural equation model*, SEM, a szociológiában gyakran inkább *útmodell*) az eredeti Wright-féle útelemzésnek egyfajta általánosítása, akárcsak az itt tárgyalt grafikus Markov-modell, így bizonyos értelemben vetélytársaknak tekinthetők. Ezért tartottam fontosnak külön fejezetet szánni összevetésüknek. A következő, speciálisan a mobilitás-kutatási alkalmazásokról szóló fejezetben a szociológiában útmodellnek nevezett SEM-mel kapcsolatban még vissza fogok térni erre a témára.

A SEM a grafikus modellektől eltérő matematikai alapokon, de szintén gráfokkal jelenít meg asszociációs struktúrákat, a nyilakhoz számokat rendelve, ahol a számok szintén a vonatkozó hatás erősségét jellemzik. A kategoriális adatokra alkalmazott SEM-nek a ’80-as években még nem volt más alternatívája, mint a hagyományos loglineáris modell, utóbbival szembeni előnyeként elsősorban azt jelölték meg (pl. Xie, 1989, Heckman, 1978), hogy alkalmasabb oksági viszonyok elemzésére, és nem csak szimmetrikus kapcsolatok reprezentálását teszi lehetővé.

Az utóbbi évtizedekben kidolgozott grafikus modellre (mely általánosan nem csak kategoriális, hanem folytonos változókra is értelmezhető) ezek a kritikai észrevételek már nem teljesülnek. Láttuk, hogy a grafikus modellekben szimmetrikus és asszimmetrikus kapcsolatok egyaránt megjeleníthetők. A változók közötti oksági kapcsolatok útvonalai (*path*) is ugyanúgy követhetők, mint a SEM esetében, így a direkt és indirekt hatások is ugyanúgy megfoghatók, ahogyan azt az 6.1.3. fejezetben láttuk. Kategoriális változók esetén a teljes hatás direkt és indirekt hatásokra való *számszerű* felbontása ugyanakkor nem megoldott, de ez a kategoriális SEM esetén is problematikus, és csak az utóbbi években születtek rá javaslatok (Kuha, Goldthorpe, 2007).

Ugyanakkor a grafikus modelleknek bizonyos előnyei is vannak a SEM-mel szemben. A fentiekben láttuk, hogy grafikus modellekben az élek feltételes kapcsolatoknak feleltethetők meg, az él törlése pedig feltételes függetlenséget implikál. Ez általános esetben nem teljesül a SEM-ekre. Egy meglévő él közvetlenül kapcsolható a SEM-et definiáló egyenletek valamelyikének egy paraméteréhez, de a megfelelő változópár feltételes vagy marginális kapcsolatához csak nagyon nehézkesen köthető. Ugyanígy, össze nem kötött változópár feltételes függetlensége sem kell,

hogy teljesüljön egyetlen feltételhalmaz mellett sem (Cox, Wermuth, 2001, Wermuth, 2003). Hasonló kritikát fejt ki Whittaker (1990), megemlítve, hogy a SEM-ben nem szereplő interakciók feltételhalmaza nem mindig egyértelmű. Többváltozós normális esetre a probléma részletes kifejtését lásd pl. Cox és Wermuth 1993-as munkájában. Egy, a kritikákra született válasz Pearl 1993-as tanulmánya. A két megközelítés egy újabb összevetése Lauritzen és Richardson (2003) írásában található.

6.2 Mobilitás-kutatási alkalmazás: használati feltételek, alternatívák

6.2.1 Történeti áttekintés

A grafikus modellek kvantitatív empirikus szociológiai alkalmazhatóságát a mobilitás-kutatás példáján szeretném megmutatni. Ezzel, azt hiszem, nem szűkítem le túlságosan tárgyamat, hiszen, ahogyan majd az alábbi történeti összefoglalóból is kitűnik, a mobilitás-kutatás módszereinek bővülése általában szorosan követi az általános statisztikai módszerek fejlődését, sőt néha generálja is azt.

Az alábbiakban tehát az intergenerációs mobilitás kutatásának kontextusában szeretném a kategoriális változókra definiált grafikus modelleket elhelyezni. Ehhez Ganzeboom et al. (1991), Kuha, Goldthorpe (2007), Treiman, Ganzeboom (2000) és Kelley (1990) összefoglalói alapján röviden vázolni szeretném a kutatási módszerek és fogalmak változásának szempontunkból releváns pontjait, majd a grafikus modellek alkalmazási lehetőségeit és alternatíváit ezek tükrében megvizsgálni.

Elsőként azt a legegyszerűbb megközelítést említeném, amikor az apát és annak fiát azonos foglalkozási kategóriák szerint kereszttáblázzák, és ebben a táblázatban végeznek számításokat a mobilitás mértékére és irányára vonatkozóan. Történetileg ez a módszer a Featherman (1981) ill. Treiman és Ganzeboom (2000) által a mobilitás-kutatások első nemzedékének[39] nevezett kutatásokhoz, az un. **ISA paradigmához** kapcsolható. Ez a paradigma a Nemzetközi Szociológiai Társaság (*International Sociological Association*, ISA) 1951-ben megfogalmazott kezdeményezésével jött létre. Egységes elméleti alapon és egységes módszerekkel végeztek országos mobilitás-vizsgálatokat, azzal a céllal, hogy az eredményeket nemzetközileg összehasonlíthassák. Módszereiket tekintve elsősorban a belépési (pl. munkások között a szellemi származásúk aránya) és kilépési (munkás származásúak között a szellemiek aránya) arányszámokat vizsgálták. Ezeknek az arányszámoknak az összehasonlítására épült a közismert Lipset-Zetterberg tétel (Lipset, Zetterberg, 1959), ami szerint a mobilitás az iparosodottság fokával párhuzamosan nő, és mintázata a nyugati iparosodott országokban állandónak tekinthető. Ez a tétel az Egyesült Államok, mint „új nemzet” nagyobb mobilitásával kapcsolatos, akkoriban közkeletű vélekedést cáfolta.

Mások (Duncan, 1966, Rogoff-Romsoy, 1966) azonban felismerték, hogy ezek a mobilitási arányszámok a tábla marginálisainak is függvényei, ezért nem használhatók közvetlenül komparatív célokra. Ez a probléma alapvető kérdésnek bizonyult. A mobilitás-kutatás utóbbi évtizedeiben több javaslat is született a megfigyelt, „abszolút” mobilitásnak a „relatív” vagy

[39] Featherman használta először a „mobilitás-kutatás generációi” kifejezést, Treiman és Ganzeboom 2000-es tanulmánya pedig az első, ahol a mobilitás-kutatások negyedik, a korábbi szerzők által még nem tárgyalt nemzedéke is megjelent.

„nettó” mobilitási esélyektől való megkülönböztetésére. A különféleképpen operacionalizált nettó mobilitási esélyek egyenlősége vagy egyenlőtlensége alapján jellemeztek ezután egyes társadalmakat "nyitottnak" vagy "zártnak", és e fogalmak mentén vizsgálták felül újra és újra a klasszikus Lipset-Zetterberg tételt. A legkorábbi ilyen céllal született módszer a strukturális (a foglalkozási struktúra változása, vagyis a tábla marginálisainak különbsége által kikényszerített) és a cirkuláris („tiszta”) mobilitási arány szétválasztása volt. Ezeket használta később a szintén a Lipset-Zetterberg tétel módosításának tekinthető FJH-tétel is (Featherman, Jones, Hauser, 1975). A tétel szerint a mobilitás tényleges nagysága ugyan különböző a piacgazdasággal és nukleáris családtípussal jellemezhető iparosodott országokban, mivel azokban a strukturális változások üteme eltérő nagyságú mobilitást kényszerít ki, ám a strukturális változásoktól független cirkuláris mobilitás minden iparosodott társadalomban egyforma.

A két foglalkozási változóhoz később másokat is hozzávettek (pl. a fiú vagy az apa iskolázottságát), és a háromdimenziós tábla elemzését a kétváltozós feltételes táblákkal próbálták megoldani, de nehézséget okozott a marginálisok változása által okozott hatásoknak és a három változó tényleges harmadfokú kapcsolatának az elkülönítése (Kuha, Goldthorpe, 2007).

Az egyszerű kereszttábla-elemzések után a történetileg is őket követő útmodelleket kell megemlítenem, mint a grafikus modellek alternatíváját/előzményét. Ezeknek a modelleknek a kereszttábla-elemzés előbb említett nehézségeit sikerült megoldaniuk. Az „útmodell” elnevezés az előző fejezetben említett strukturális egyenletek modellezéshez tartozó módszert jelöl, és eredeti formájában folytonos változókból álló lineáris regressziós egyenletek rendszerére épült. A szociológiában az 1960-as évek elején született és az 1970-es évek végéig[40] a mobilitás-kutatásban egyeduralkodó mondható **Blau-Duncan paradigma** honosította meg[41]. Ez a módszer ebben a tudományágban addig gyakorlatilag ismeretlennek számított. Sikerét annak köszönhette, hogy a mobilitási, sőt általában a kvantitatív empirikus szociológiai problémáknál nagyon jól alkalmazhatónak bizonyult. Sikerült felnyitni a mobilitásnak azt a „fekete dobozát” (idézet Kelley-től), amit az első nemzedék addig az apa foglalkozása – fiú foglalkozása kereszttábla egyszerű elemzéseivel próbáltak megközelíteni. A módszer alkalmas volt kettőnél több változó kezelésére, így lehetőség volt az apa és a fiú iskolázottságát, ill. az utód első foglalkozását is figyelembe venni. Ennek köszönhetően megmutatták, hogy az apa iskolázottságának önálló, a foglalkozásától független hatása van. Feltárták a mobilitásnak az oktatáson keresztül érvényesülő csatornáját is, és bebizonyították, hogy ez a kulcsa a szülői háttér és a foglalkozási státusz közötti kapcsolatnak. A paradigmát később más problémákra is kiterjesztették, vizsgálták a jövedelemmel kibővített modellt (pl. Treiman, 1970), mások a családi erőforrások genetikai és pszichológiai tényezőit is megjelenítették, mint pl. az utód intelligencia-szintjét vagy a szülői biztatást (lásd 3.8. fejezet, Wisconsin modell), megint mások az intragenerációs mobilitást elemezték a karrier több állomásának bevonásával. A módszer különösen gyümölcsözőnek bizonyult országok közötti összevetésre ill. ugyanazon ország időbeli változásának vizsgálatára. A Lipset-Zetterberg tételhez visszatérve újra felmerült a kérdés, hogy a státuszelérési folyamat

[40] Treiman és Ganzeboom ezt a korszakot nevezi a mobilitás-kutatás második nemzedékének.

[41] A szóhasználat nem csak hétköznapi értelmében helytálló: ezt az iskolát kuhn-i értelemben vett paradigmaként aposztrofálja Colclough és Horan (1983) és Kelley (1990) is.

mely aspektusai állandóak, melyek változnak a gazdasági vagy politikai rendszer változásával, és melyek azok, amik ezekkel a változásokkal nem magyarázható, országspecifikus vonások.[42]

Az útmodellekhez a foglalkozási státusz és az iskolázottság megfelelő operacionalizálására volt szükség. Blau és Duncan, ill. Treiman a foglalkozási státusz folytonos változóját a foglalkozások presztízshierarchiája alapján határozták meg. Ez a megközelítés elméleti megközelítésükből is fakadt, hiszen a társadalmi helyzetet az amerikai iskola tradicionálisan a társadalmi presztízsből kiindulva határozza meg, szemben az európai osztályhelyzet-alapú megközelítéssel. Az iskolázottságot az elvégzett iskolai osztályok számával mérték.

A ’70-es években felmerülő kritikák a paradigma több előfeltevését is kifogásolták. A legfőbb módszertani kritika az volt, hogy a modell nem enged meg sem nem-lineáris kapcsolatokat, sem a magyarázó változók közötti interakciókat. A státuszelérési folyamat egyéni szintű megközelítését is kifogásolták, ami így nem mond semmit a kollektív, makroszintű mobilitási arányokról, ami a mobilitási témakör eredeti motivációja volt. A szociológiaelméleti kritikák a változók folytonos voltát kifogásolták – a foglalkozásét azért, mert koncepcionális alapon a társadalmi státusz helyett inkább a társadalmi osztályt használták volna; az iskolázottságét pedig azért, mert az elvégzett osztályok számával szemben a megszerzett képesítés szintjét és típusát megfelelőbbnek tartották.

Vagyis a kritikusok ismét kategoriális változók elemzésére alkalmas módszert kerestek. E harmadik nemzedék legnagyobb része az ISA-paradigma kétváltozós kereszttáblájához tért vissza, de a kifinomultabb loglineáris modelleket használva (a korábban említett Goodman is ehhez a nemzedékhez tartozott). A második nemzedék útmodelljeivel szemben támasztott kritikáik a **loglineáris paradigma** (Kelly kifejezése, 1990) módszereire már nem voltak érvényesek, ráadásul az első nemzedék problémája, a nettó mobilitás és a marginális hatás elkülönítése is megoldódott.

Az utóbbi problémának a megoldása az esélyhányadossal történt. Emlékeztetnék arra az 2.1. fejezetben már említett problémára, hogy két változó (itt: apa és fia foglalkozása) közötti kapcsolat erősségének (itt: az átöröklődés, immobilitás erősségének) méréséhez olyan asszociációs mérőszámot érdemes rendelni, ami variációsan független a marginálisoktól, és csak az együttes eloszlástól függ. Hiszen bármilyen is az egy generációval ezelőtti és az aktuális foglalkozási struktúra, a társadalom elvileg tetszőlegesen zárt vagy tetszőlegesen nyitott is lehet. A 2.1. fejezetben említettem, hogy az esélyhányados, amely a loglineáris paraméterek függvényeként előálló mutató, a marginálisoktól variációsan független. Így az *F’* (apa foglalkozása) - *F* (utód foglalkozása) mobilitási táblának a marginálisai és a $\lambda_{F'F}^{F'F}$ paraméterek is variációsan függetlenek. Ebből adódóan az apa generációjának *1...I* foglalkozási kategóriáinak arányához és az utód generációjának *1...J* kategóriáinak arányához képest az (*i,j*) cellához tartozó $\lambda_{F'F}^{F'F}(i,j)$ paraméter értéke bármilyen pozitív vagy negatív szám lehet. Pozitivitás esetén a függetlenséghez, vagyis a teljes nyitottsághoz képest többen vannak a cellában (az *i*. foglalkozást űző apák fiai másokhoz képest nagyobb arányban kerülnek a *j*. kategóriába), negativitás esetén kevesebben.

[42] A Szociológiai Szemlében megjelent tanulmányomban, melynek kivonatát a 7.1. fejezetben közlöm, részletesen is foglalkoztam ezzel a kérdéskörrel a magyar rendszerváltásnak a státuszelérési folyamatra gyakorolt hatásával kapcsolatban.

A loglineáris modellek lehetőséget adtak továbbá az *F'F* kétváltozós összefüggés komplikáltabb leírására is, pl. azáltal, hogy az osztály-átöröklődést kifejező, átlóban levő cellák a többitől külön kezelhetők, a $\lambda_{F'F}^{F'F}(i,i)$ paraméterek bármilyen megkötésével.

A harmadik nemzedék a CASMIN (*Comparative Analysis of Social Mobility in Industrial Nations*) projektben intézményesült a '70-es évek végén és a '80-as években. A projekt egyik eredménye a nemzetközi összehasonlításokban elfogadottá vált osztálykategóriák létrehozása volt (EGP kategóriák, Erikson, Goldthorpe és Portocarero után), az alkalmazások között, a 7. fejezetben magam is ezt a kategorizációt használom majd.

A harmadik nemzedék igazi központi kérdése azonban ismét a Lipset-Zetterberg tételhez kapcsolódott. A loglineáris keretben újra megvizsgálták, hogy a társadalmi mobilitás valóban növekszik-e az iparosodottság fokával. A mobilitási mintázatot (itt: *social fluidity*) a $\lambda_{F'F}^{F'F}(i,j)$ loglineáris paraméterekkel definiálva azonban nem ezt a tételt, hanem Sorokin (1964) trendnélküli fluktuáció elméletét látták alátámaszthatónak, ez a konklúziója a nemzedék egyik nagy összefoglaló munkájának, a *The Constant Flux*-nak (Erikson, Goldthorpe, 1992). Erikson és Goldthorpe országközi és időbeli összevetéseket is végzett, utóbbiakat nem valódi időeltéréssel felvett adatokon, hanem a keresztmetszeti adatok kohorszokra osztásával. Közös fluiditás (*common fluidity*) modelljük az országonként vett $\lambda_{F'F}^{F'F}(i,j)$ paraméterek, állandó fluiditás (*constant fluidity*) modelljük pedig a kohorszonként vett $\lambda_{F'F}^{F'F}(i,j)$ paraméterek országtól/kohorsztól való függetlenségét tesztelte. Eredményeik szerint (1) az ipari társadalmak közös mobilitási mintázattal rendelkeznek; ennek apró különbségei inkább történelmi és politikai körülményekre, mint az iparosodottság fokára vezethetők vissza, és (2) nem bizonyítható, hogy a mobilitás nőne a történelem folyamán.

A harmadik nemzedék tehát megoldotta az első nemzedék egyik legfőbb problémáját, ugyanakkor a második nemzedék többváltozós módszerével ellentétben az iskolai végzettség modellbe vonása továbbra sem valósult meg. Történt ugyan kísérlet a kétváltozós elemzés kiszélesítésére, és a Blau-Duncan paradigmának megfelelően történő általánosítására (lásd a 4.1. fejezetben már elemzett Goodman-féle módosított útmodellt, a grafikus modellek közvetlen előzményét), ez azonban nem terjedt el. A harmadik nemzedék megközelítésével szemben felhozott további kritikák között szerepelt az is, hogy a relatív mobilitást a tábla sok (Erikson és Goldthorpe szám szerint nyolc) paraméterével jellemzik, és nem képesek egyetlen mutatóba sűríteni a zártság/nyitottság mértékét. A kifogás statisztikai hátterében az áll, hogy így a diszkrimináló erő csökken, hiszen két tábla közötti különbség több szabadságfok között oszlik meg.

A harmadik nemzedéket a '90-es évek óta követő kutatások, bár Treiman és Ganzeboom (2000) negyedik nemzedékként hivatkozik rájuk, nem rendelkeznek közös kutatási programmal. Sem módszereikben, sem kutatási kérdéseikben nem egységesek. Néhány olyan példát említek itt csak közülük, ami a grafikus modellek alkalmazhatósági kérdéséhez kapcsolódik. Az egyik ilyen módszertani fejlesztés az útmodellek általánosításának, a folytonos és kategoriális változókat egyaránt kezelni képes strukturális egyenletek modelleknek a megjelenése volt (Muthén, 1983, 1984, Winship, Mare, 1983, Xie, 1989). Speciális esetként ide tartozik a kategoriális kimeneti változós regressziós modellek (logisztikus és probit regresszió, ordinális vagy multinomiális logisztikus regresszió) elterjedése a szociológiában (pl. az iskolából a munkába történő átmenet

modellezése, Shavit, Müller, 1997). Ez utóbbiakba beépíthető a magyarázó változók nemlineáris hatása vagy a köztük levő interakció is, amit a Blau-Duncan paradigmából kritikusai hiányoltak.

Egy másik újdonság az időbeli folyamatok előtérbe kerülése. Az egyik jellemző megközelítés ismételt keresztmetszeti vizsgálatokat használ, és az időt, mint önálló változót vonja be az elemzésbe, a társadalmi rétegződés évenkénti változásait makro jellegű változásokra, pl. társadalompolitikai beavatkozásokra vezetve vissza (pl. DiPrete, Grusky, 1990). Egy másik ide tartozó megközelítés eseménytörténeti (ugyanazokról a személyekről több különböző időpontból rendelkezésre álló) adatokkal dolgozik.

A harmadik nemzedék által használt loglineáris modellekkel szemben felhozott egyik, már említett kifogás az volt, hogy nem tudják egyetlen mutatóval jellemezni az adott társadalom nyitottságát. Erre a kihívásra született válaszként a unidiff modell (Erikson, Goldthorpe, 1992, másnéven *log-multiplikatív layer effect model,* Xie, 1992). Ez a modell elfogadottnak mondható a kortárs mobilitás-kutatásban, egy friss nemzetközi munkában is egyezményes módszerként használják (Breen, 2004). A modell nevét onnét, kapta, hogy időbeli összehasonlításoknál két időpont között egyenletesen változó mobilitási esélyeket (*uniform difference- unidiff*) tételez fel. Vagyis a modell szerint a különböző időpontokban mért apa-fiú mobilitási táblák időpontra nézve feltételes log-esélyhányadosai csak egy, az adott időpontra jellemző konstans-szorzóval térnek el a feltétel nélküli log-esélyhányadosoktól.

Ez a megszorítás az apa foglalkozása (*F'*) – utód foglalkozása (*F*) feltételes asszociációnak egy φ_k konstanssal való szorzástól eltekintve állandó voltával azonosítható, ahol a feltétel a *T* (időpont), a *k* pedig a *T* egy értékéhez tartozik. A tábla felparaméterezésével az apák *i.*, az utódok *j.* foglalkozásához és a *k.* időponthoz tartozó cella log-gyakorisága a modell szerint tehát így áll elő:

$$\log \mu(i,j,k) = \lambda_{\varnothing}^{FF'T} + \lambda_{F}^{FF'T}(i) + \lambda_{F'}^{FF'T}(j) + \lambda_{T}^{FF'T}(k) + \lambda_{FT}^{FF'T}(i,k) + \lambda_{F'T}^{FF'T}(j,k) + \phi_k \lambda_{FF'}^{FF'T}(i,j),$$

ami az alábbi hagyományos loglineáris paraméterezést

$$\log \mu(i,j,k) = \lambda_{\varnothing}^{FF'T} + \lambda_{F}^{FF'T}(i) + \lambda_{F'}^{FF'T}(j) + \lambda_{T}^{FF'T}(k) + \lambda_{FT}^{FF'T}(i,k) + \lambda_{F'T}^{FF'T}(j,k) + \lambda_{F'F}^{F'FT}(i,j) + \\ + \lambda_{F'FT}^{F'FT}(i,j,k)$$

annyiban szorítja meg, hogy annak utolsó két tagját egyetlen, a *k*-tól és az *(i,j)*-től csak külön-külön (interakciójuk nélkül) függő taggal teszi egyenlővé:

$$\lambda_{F'F}^{F'FT}(i,j) + \lambda_{F'FT}^{F'FT}(i,j,k) = \varphi_k \psi_{ij}.$$

Itt a ψ_{ij} mutatja az apa-utód kereszttábla *(i,j)* cellájában a kapcsolat minden időpontra közös értékét, φ_k pedig ennek a kapcsolatnak az intenzitásának időbeli változását. Ez, az adott időpontra jellemző φ_k érték a unidiff együttható. Csökkenő (növekvő) φ_k–k az asszociáció gyengülését (erősödését), vagyis társadalmi nyitódást (záródást) jeleznek.

Ahogyan máskor, az időre feltételes táblák helyett itt is nézhetünk országokra feltételes táblákat, a unidiff együttható ilyenkor az országok nyitottságának összevetésére alkalmazható, országonként változó intenzitású, de azonos mintázató mobilitási esélyek feltételezésével. Ez a módszer tehát egyetlen mutatóba sűríti a zártság/nyitottság mértékét; erre a második és harmadik nemzedék módszerei, ahogyan már említettem, nem voltak alkalmasak. Így a mobilitás-kutatást eredetileg motiváló, az első nemzedék által vizsgált kérdés az általános mobilitás nagyságával

kapcsolatban ismét előtérbe került. A unidiff modell alkalmazható komplikáltabb felállásban is, egy friss cikk pl. (Beller, Hout, 2006) a unidiff mutató értékének országközi különbségeit a jóléti rendszer és az oktatáspolitika típusára (a felsőoktatás hozzáférhetősége) vezeti vissza. Technikailag pl. a jóléti rendszerre való visszavezetés a φ_k értékének a jóléti rendszer típusától való függésének megengedésével valósítható meg.

Itt említenék meg egy, a unidiff modellel kapcsolatos módszertani problémát, ami ismét a klasszikus kérdéshez, a mobilitási mutató strukturális változásoktól való függetlenségéhez tartozik. Firth (2005) ill. Breen és Luijkx (2004, a már említett nemzetközi összehasonlító munka zárótanulmánya) alapján ugyanis a unidiff modell paraméterei nem rendelkeznek a marginális változásoktól való függetlenség tulajdonságával. Breen és Luijkx szerint könnyen konstruálható pl. olyan tábla, ahol csupán az egyik, nagyon egyenlőtlen kilépési arányszámokkal rendelkező származási osztály arányának csökkentése csökkenti a unidiff paraméter értékét is. Vagyis a társadalom látszólag nyitottabbá válhat, pedig a mobilitási mintázat (az esélyhányadosokkal jellemzett együttes eloszlás) valójában változatlan. Firth (2005) szerint ugyanakkor ez a probléma csak rosszul illeszkedő unidiff modell esetén állhat elő, vagyis ha nem teljesül az az előfeltevés, hogy a vizsgált mobilitási táblák log-esélyhányadosai csak egy országspecifikus szorzótényezőben térnek el egymástól.[43]

6.2.2 A történeti áttekintés tanulságai

A fenti történeti összefoglaló alapján, azt gondolom, jól látszik, hogy az egymást követő kutatási paradigmák módszereit legtöbbször nem lehet sem szociológiaelméletileg „megfelelő”, sem matematikailag „helyes” voltuk alapján rangsorolni. Bár úgy tűnhet, mintha egy adott paradigmát a felmerülő problémák miatt leváltó és a problémákat megoldó új módszer „jobb” lenne a korábbinál, észrevehető, hogy a legtöbbször nem matematikai vagy technikai jellegű problémák lépnek fel, hanem inkább az új elméleti koncepciók inkompatibilisek a régi módszerekkel. Mivel a paradigmák középpontjában álló kutatási kérdések lényeges pontokban különböznek egymástól, a használt módszerek sem csak statisztikailag térnek el egymástól, hanem koncepcionális hátterüket tekintve is. Az előző fejezetben felsorolt módszertani alternatívák ennek megfelelően tehát véleményem szerint nem jobbak vagy rosszabbak egymásnál, hanem más-más kérdések megválaszolására alkalmasak. Ugyanez igaz a grafikus modellekkel való összevetésükre is, ezt részletezném a továbbiakban.

[43] A már említett 2004-es nemzetközi tanulmánykötet (Breen, 2004) néhány eredménye Breen és Luijkx szerint a unidiff modell ezen hátrányos tulajdonságával magyarázható. Ugyanis az apa „önálló” foglalkozási kategóriája éppen ilyen származási kategória: a többitől nagyon különböző, az általános esélyegyenlőtlenséget inkább növelő mobilitási jegyekkel bír. Magyarországon ez a származási kategória az 1973-1983 közötti intervallumon erősen csökkenő arányt mutatott, ami Breenék szerint önmagában oka lehetett a unidiff modell paraméterében ezen az intervallumon tapasztalt csökkenésnek, tehát az erősen növekvőnek ítélt mobilitásnak. Róbert és Bukodi (2004) jegyezte a Breen-kötet Magyarországról szóló fejezetét. Vizsgálatukban a Unidiff modell a disszimilaritási index szerint jól magyarázta az adatokat, de a valószínűség-hányados statisztikát tekintve rosszul illeszkedett, tehát elvileg valóban felmerülhetett ez a probléma. Kíváncsi lévén arra, hogy Breen és Luijkx feltevése helytálló-e, néhány éve megvizsgáltam a Magyarországra vonatkozó eredmények érzékenységét a marginálisok változtatására, s tapasztalataim szerint azok stabilitást mutattak (Németh, 2006b). Vagyis Róbert és Bukodi eredményeit vélhetően nem befolyásolta az önálló apák számarány-csökkenése. David Firth tanácsa alapján az érzékenységtesztet a következőképpen végeztem: a három évhez tartozó három tábla marginálisait a lokális esélyhányadosok rögzítése mellett homogenizáltam az un. iteratív arányos illesztés algoritmusának segítségével, majd megismételtem a unidiff modell illesztését.

Talán legalapvetőbb különbség a fenti módszerek között, hogy **folytonos vagy kategoriális változók** elemzésére (esetleg mindkettőre) alkalmasak-e. Az előbbiek közé a lineáris regressziós módszerek és általánosításaik (pl. útelemzés) tartoznak, az utóbbiak közé a mobilitási arányszámokat vizsgáló, a loglineáris modellekkel ill. azok variánsaival (pl. unidiff modellel) dolgozó megközelítések. A strukturális egyenletek modellezés mind folytonos, mind kategoriális változó kezelésére képes. A disszertációban kategoriális változókra alkalmazható grafikus modelleket ismertettem, de a grafikus modellezés folytonos változókon is értelmezhető (pl. Cox, Wermuth, 1996).

A folytonos vagy kategoriális típus közötti választás elméleti megalapozottságú: a foglalkozás vagy az iskolázottság operacionalizálása elsősorban koncepcionális, s nem technikai kérdés. Ezzel kapcsolatban említettem a foglalkozási státuszt osztály-alapon ill. presztízs-alapon megközelítő európai ill. amerikai iskola példáját. Hibás az a (pl. Kelley 1990-es írásában olvasható) érvelés, hogy a folytonos változó, mivel sok különböző értéket vehet fel, pontosabb mérőeszköz, mint a kategoriális változó. A kevesebb kategóriás társadalmi osztályok használata mellett is állnak mind szociológia-elméleti, mind statisztikai érvek. A két típus közötti elméleti különbség elsősorban az, hogy az osztályok strukturális alakulatok, a presztízs-hierarchia viszont az egyéni szereplőhöz kapcsolódik. Xie és Goyette (2003) szerint a foglalkozás kategoriális megközelítése a mobilitás kontextusában azért megfelelőbb, mert így konkrét foglalkozási osztályokról is tudunk kijelentéseket tenni. Valóban, folytonos esetben foglalkozások közötti távolságot mérünk, és a vertikális mobilitás van a figyelem középpontjában: „a magasabb iskolai végzettség magasabb presztízsű foglalkozással jár". Kategoriális esetben pozíciók között teszünk különbséget: „az Egyesült Államokban bizonyos foglalkozások »toborzási« kritériumaikban univerzalisztikusabbak másoknál, ezért az ázsiai amerikaiak kulturális tőke híján inkább ezeket választják" (Xie és Goyette példája, 2003). Az iskolázottság kategoriális használata mellett is vannak elméleti érvek, pl. ha középiskola-típusok vagy egyetemi szakok mobilitásban való szerepe a kérdés, és nem az iskolai osztályok száma – társadalmi státusz lineáris kapcsolat. Általában elmondható, hogy a kategoriális megközelítés a mobilitási mintázat, a tipikus mobilitási csatornák és sorompók (*channels, barriers*, Blau és Duncan szóhasználata, 1967) megtalálását teszi könnyebbé (Yamaguchi, 1983). Yamaguchi példáival élve: azzal, hogy különbséget teszünk pozíciók között, felfedezhetjük pl. az önállók és vállalkozók közötti, a fizikai javak átörökítése és annak konverziója révén működtetett csatornát, vagy az önálló gazdálkodók és a mezőgazdasági munkások közötti, technológiai, információs és földrajzi izoláltságuk miatt létrejött csatornát. Ugyanezen osztályok és a többi osztály között ugyanakkor sorompót találhatunk; az utóbbiakat éppen izoláltságuk választja el másoktól, az előbbieket kulturális és anyagi tőkéjük: nem feltétlenül elég a közéjük való belépéshez a megfelelő iskolai végzettség megszerzése. E csatornák és sorompók nem definiálhatók folytonos változók mellett.

A kategoriális változók ill. a folytonos változók használata közötti különbséget kitűnően példázza Xie (1989) vizsgálata, aki a Blau-Duncan modellt az eredeti 1962-es adatokkal elemezte újra, azzal a céllal, hogy a különböző foglalkozási kategóriák eltérő „toborzási" mintázatát felfedje. Strukturális egyenletek modellt használt, az iskolázottságot kategóriákkal definiálta, a foglalkozást pedig bináris módon határozta meg (két különböző változóval: tudományos vagy mérnöki foglalkozás vs. egyéb, ill. hivatalnok/vezető/önálló vs. egyéb). Eredményei szerint az eredeti, folytonos változókkal működő Blau-Duncan modell eredményei összemossák a foglalkozásonként eltérő toborzási mintákat, pedig a foglalkozási struktúra nemlineáris és különböző dimenziókban megfogható kapcsolatokban termelődik újra. A tudományos vagy

mérnöki foglalkozásba való belépést például csaknem teljes egészében az iskolai végzettség határozza meg, a szülői hatás csak ezen keresztül érvényesül, az iskola elérését megkönnyítve. Ezzel szemben a hivatalnokok/vezetők/önállók csoportjába való belépés sokkal inkább múlik az apa társadalmi státuszán, két azonos iskolai végzettségű utód közül a kedvezőbb háttérrel rendelkezőnek van a belépésre nagyobb esélye.

A folytonos ill. kategoriális változók alkalmazhatóságával kapcsolatban az elméleti szempontok után a technikai szempontok közül is megemlítenék néhányat. A hozzájuk rendelt módszerek abban is különböznek, hogy a regresszióra épülő módszerek alapvetően lineáris kapcsolatot képzelnek el két változó között, míg a keresztábla-alapú módszerek, mint a grafikus modellek is, bármilyen nemlineáris kapcsolatot is képesek leírni. A keresztábla-típusú elemzési módszerek hátránya viszont, hogy általában nem tudnak megbirkózni a túl sok üres vagy majdnem üres cella problémájával. Ezért nem vezethetünk be túl sok kategóriát egy adott változóhoz – kettő-öt kategória általában az átlagos. Hasonló problémát okoz, ha túl sok változót vonunk az elemzésbe, vagy ha a bevont változók között nagyon erős az összefüggés.

A mobilitás **mikro- ill. makroszintű megközelítés**e alapján is különbséget tehetünk a mobilitás-kutatás módszerei között. Ahogyan az összefoglalóból kitűnik, alapvetően ez a kétfajta megközelítés is egymást váltogatta az elmúlt évtizedekben. A mikroszintű megközelítés az egyéni státuszelérési folyamatot próbálja követni (útmodellek, Goodman módosított útmodellje, élettörténeti modellek). A makroszintű megközelítés tárgya a társadalom mint egész, aggregált táblákkal foglalkozik, és lehetőleg egyetlen mérőszámmal próbálja jellemezni egy társadalom nyitottságát (első nemzedék táblaelemzései, unidiff modellezés). Egy ilyen összevont mutató nagyban megkönnyíti a nemzetközi/időbeli összevetéseket, statisztikai szempontból is, hiszen egyetlen paramétert használ, így nagyobb statisztikai erővel észleli a nem túl erős trendeket is. A harmadik nemzedék loglineáris modelljei egyik kategóriába sem sorolhatók be: aggregált szinten jellemzik ugyan a mobilitást, de sok paraméterrel írják le azt. A grafikus modellek inkább az első megközelítéshez tartoznak. Az egyéni szintű státuszelérési folyamat egyes állomásainak egyenlőtlenségei jellemezhetők vele, és nem ad makroszintű egyenlőtlenségi mutatót.

Az egyetlen mérőszámmal jellemzett társadalmi mobilitás ugyan könnyű összevethetőséget kínál, de a különbségek magyarázatára nem ad lehetőséget. Magyarázatokat inkább a státuszelérési folyamatot követő mikroszintű megközelítés kínálhat. Véleményem szerint éppen ezért lenne jól kombinálható egymással a két megközelítés: az aggregált mérőszámmal jelzett különbségek mögötti mechanizmusok a folyamatmodellekkel lennének megtámogathatók. Jó példa erre a megtámogatási igényre Beller és Hout (2006) már említett tanulmánya, melyben a mobilitási esélyek (a unidiff paraméter) országközi különbségeit makroszintű mutatók, mint pl. a felsőoktatás hozzáférhetősége alapján kísérelik megmagyarázni. Az eredmények diszkussziójában, ahol több hipotézist állítanak fel az oktatáspolitikának a mobilitási esélyek meghatározódásánál betöltött szerepével kapcsolatban, a Blau-Duncan modellt hivatkozzák, ahol az apai foglalkozás és a foglalkozás közötti, az iskolai végzettséget is érintő utak jól követhetők. Egyik hipotézisük szerint pl. az iskolai végzettségnek az apa iskolai végzettségével és/vagy az apa foglalkozásával való kapcsolatának országközi különbségei magyarázhatják a unidiff paraméter különbségeit. Hipotéziseik támogatást kapnának, ha grafikus modellként illesztenék a Blau-Duncan modellt, és konkrét paraméterbecslésekre hivatkoznának.

Mint láttuk, a mobilitási módszertan klasszikus problémája a mobilitási esélyek olyan mérőszámának létrehozása, ami **független a társadalmi struktúra változásaitól**. A probléma

felismerése óta gyakorlatilag mindegyik újonnan bevezetett módszer képes volt megoldani ezt a problémát, és a grafikus modellek módszere is alkalmas erre. A státuszelérési folyamatot marginális loglineáris paraméterekkel írja le, és ezek a paraméterek, mint a 2.3.1-ben láttuk, megfelelő paraméterezés esetén variációsan függetlenek egymástól, így az egyváltozós hatásoktól is.

A mobilitás-kutatások negyedik nemzedékének egyik általam említett problémája az idő kezelése, az egyik ide tartozó megközelítés eseménytörténeti modellekkel dolgozik. Léteznek loglineáris modellek is az eseménytörténetek feldolgozására (Vermunt, 1997). A marginális modelleknek is létezik az idővel is számoló, ismételt mérések vagy paneladatok esetén használható alkalmazása (Rudas, Bergsma, 2004). Bár disszertációmnak ez közvetlenül nem képezi tárgyát, de a 7.5. fejezetben mutatok néhány példát paneladatok elemzésére.

A strukturális egyenletek modell, ami az eredeti Blau-Duncan-féle útelemzés általánosítása, a grafikus modellhez sok szempontból hasonlít: mindketten gráffal szemléltetik az asszociációs struktúrát, a változók közötti **oksági kapcsolatok útvonalai** ugyanúgy követhetők, és a **direkt és indirekt hatások** is ugyanúgy megfoghatók mindkét esetben. Így bizonyos értelemben vetélytársaknak tekinthetők, ezért szenteltem korábban egy önálló fejezetet összevetésüknek (6.1.4. fejezet). Összefoglalva az ott írtakat: a grafikus modellek hátránya a SEM-el szemben, hogy a teljes hatás direkt és indirekt hatásokra való *számszerű* felbontása itt nem lehetséges, vagy legalábbis még nem született rá javaslat, de ez a kategoriális SEM esetén is problematikus, és csak az utóbbi években látszik megoldódni (Kuha, Goldthorpe, 2007). Ugyanakkor a grafikus modelleknek is vannak előnyei a SEM-mel szemben. Grafikus modellekben az élek feltételes kapcsolatoknak feleltethetők meg, az él törlése pedig feltételes függetlenséget implikál. Így a modell paraméterei kézenfekvő módon, a feltételes kapcsolatok erősségeként interpretálhatók. Ez általános esetben nem teljesül a SEM-re. A grafikus modell tehát akkor megfelelő választás, ha **feltételes függetlenségek**kel szeretnénk leírni a változók struktúráját. A társadalmi mobilitás kutatása ilyen terület lehet, ha az a kutatási kérdés, hogy a státuszelérési folyamat egyes állomásain mely korábban megszerzett erőforrások hagyhatók már figyelmen kívül más erőforrások birtokában.

Összefoglalóan: a módszer megválasztása nem redukálható statisztikai kérdéssé. Az elméleti háttérkoncepció és a kutatási kérdés alapján lehet eldönteni, hogy a grafikus modellek mely esetekben nyújthatnak a többi alternatívához képest jobb megoldást, vagy mikor egészíthetik ki egy másik szempont bevonásával azokat.

7 SAJÁT ALKALMAZÁSOK

7.1 Klasszikus modell, friss adatok: Treiman hipotézise a magyar rendszerváltásban

A további fejezetekben néhány saját alkalmazást részletesebben is bemutatok, ezek nagy része a mobilitás-kutatási alkalmazás, egy részüket korábban már publikáltam. Az első, legnagyobb terjedelmű alfejezet a Szociológiai Szemlében (Németh 2006b, angolul Németh 2007) megjelent tanulmányom kivonata. Az ide történő beemelésével elsődleges célom az volt, hogy megmutassam, a grafikus modellek módszere beilleszthető az adott keretek közé, eredményei összevethetők más kutatásokéval. A kutatás alábbi összefoglalásánál erre a célra koncentráltam. A további, módszertanilag bonyolultabb alkalmazásokban inkább a statisztikai munka lesz majd hangsúlyos.

7.1.1 Kutatási motivációk, kutatási kérdések

Treiman gyakran hivatkozott modernizációs hipotézise szerint a gazdasági-technológiai fejlődés a mobilitás szempontjából nyitottabb társadalom kialakulásához vezet, mivel a munka fokozott bürokratizálódása megnehezíti a pozíciók közvetlen átörökítését, ugyanakkor a fejlett gazdaság munkaerőpiaca a formális oktatással megszerezhető képességeket jutalmazza. A hipotézis szerint az oktatás elterjedtsége, az átfogóbb tömegkommunikáció, a nagyobb urbanizáció és fokozott földrajzi mobilitás egyaránt a merev osztályszerkezet lebontásának irányába hat (Treiman, 1970). Andorka és társai 1973-as, 1983-as és 1992-es mobilitási táblázatok összevetésekor Magyarországon az esélyegyenlőtlenségek bizonytalan és lassú csökkenését detektálják, amit a tétel gyenge alátámasztásaként értékelnek (Andorka és társai, 1994). Luijkx-nak és társainak (1995) magyarországi, 1973 és 1993 között felvett, férfiakra vonatkozó adatokon az iskolai végzettség hatásának növekvő, ill. a származás hatásának csökkenő trendjét sikerült kimutatniuk, ugyanakkor egyik tendencia sem volt stabil, bizonyos visszafordulást tapasztaltak a 80-as évek közepétől. Interpretációjuk szerint ezek az eredmények nem cáfolják, inkább támogatják Treiman állítását, hiszen a nyolcvanas évek második felétől megindult gazdasági teljesítmény-romlás az iparosodás, a modernizációs folyamatok visszaéseséként értékelhető.

Felmerül a kérdés, hogy ezeknek az éveknek az összevetése valóban alkalmas-e az iparosodás tézisének tesztelésére. A rendszerváltás sokdimenziós folyamat, elválaszthatatlanul összefonódó politikai, gazdasági, társadalmi és kulturális összetevőkkel. A változásoknak az iparosodásra történő egyoldalú visszavezetése a különböző okok összemosását jelentené.

Ugyancsak megfontolandó az elemzés időkeretének kiszélesítése. Ahogyan Luijkx és társai is említik, 1992-1993 több szempontból mélypontja volt a rendszerváltást követő gazdasági átalakulás negatív kísérőjelenségeinek. Mivel ma már frissebb adatok is rendelkezésre állnak, referenciapontként későbbi, a gazdasági-társadalmi átalakulás stabilizálódását követő időpont is kijelölhető.

További kérdéses pont a makrotársadalmi folyamatok reakcióidejének kérdése. Ha a rendszerváltás hozta politikai változásokat tételezzük fel hatótényezőként, vajon mikorra várható

hatásuknak megjelenése a mobilitási mintákban? Ez a probléma ritkán kerül expliciten említésre, bár hasonló kérdések felmerülnek; pl. Breen és Luijkx (2004) a kövezető problémát említi: „A »szociáldemokrácia« példáját véve, az aktuálisan szociáldemokrata kormányzatú országokban várnánk nagyobb fluiditást, vagy plauzibilisebb lenne csupán a szociáldemokrata kormányzás hosszabb (még specifikálandó) idejű fennállását relevánsnak venni?”[44] A klasszikus nemzetközi vizsgálatok általában egy-két évtizedes időhosszakat fognak közre, impliciten feltételezve, hogy a társadalmi folyamatok ilyen távon reagálják le a gazdasági/politikai változásokat. Ahogyan fenti példáim mutatják, a hazai munkák az iparosodási hipotézist tesztelve hasonló feltételezéssel élnek.

Végül, ettől nem független probléma az, hogy a rendszerváltás vizsgálatakor tulajdonképpen jórészt a rendszerváltást megelőzően beindult folyamatokról beszélhetünk – mind a gazdasági dimenziókat (a kvázipiaci formák, ill. a makroszintű válságtünetek, a foglalkoztatás csökkenésének megjelenése), mind a politikai, oktatáspolitikai dimenziókat tekintve. Pl. Gazsó, Laki (2004) szerint a származáson alapuló szelekciós mechanizmusok erősödése, az állam kihátrálása esélykiegyenlítő szerepéből nem a rendszerváltás időszakában kezdődött, csak az államszocializmus utolsó évtizedének mechanizmusai folytatódtak tovább.

Az alábbiakban 1983-as, 1992-es és 2000-es felvételek státusmegszerzésre vonatkozó adatait vetem össze. *Kutatási kérdésem a klasszikus modernizációs hipotézisből indul ki: vajon milyen irányban változott a vizsgált időszakban az apa státusza (iskolázottsága, foglalkozása) és a kérdezett státusza közötti kapcsolat?* A fent említett problémáknak megfelelően a detektált változásokat nem tisztán a gazdasági fejlődés, hanem (már a ’80-as években megindult) soktényezős rendszerváltozási folyamat következményeként értékelem majd. Kerülném azt a megfogalmazást, hogy a rendszerváltás társadalmi hatásainak végigkísérése lenne a cél – a gazdasági hanyatlás hatása talán már felfedezhető ezekben az adatokban, a növekvése valószínűleg még nem.

Az elemzés során a Központi Statisztikai Hivatal 1983-as és 1992-es Társadalmi mobilitás kutatásának ill. 2000-es Életmód és időmérleg-vizsgálatának adatait használom. Ez az időszak nem kontinuus fejlődésnek, inkább U-görbét leíró, két ellentétes irányú periódusból álló folyamatnak tekinthető. Munkaerőpiaci szempontból Kézdi (2002) veti fel az 1986-1999 közötti időszak két fázisú voltát, 1995-ös osztóponttal. Elemzése szerint az első fázisra a munkahelyek nagyarányú megszűnése és az ágazatok közötti tömeges reallokáció, míg a második fázisra a foglalkoztatás csökkenésének megállása, a magasabb képzettségűek aránynövekedése és az iskolázottság bérekben mért hozamának gyorsabb növekedése jellemző. Szociológiai megközelítésben Kolosi és Róbert (2004) tárgyalják a fenti időszak ellentétes irányt mutató átalakulási fázisait.

7.1.2 Elméleti háttér, korábbi vizsgálatok

A rendszerváltozás *longitudinális* vizsgálatakor általában alkalmazott elméletek a korábbi nemzetközi, elsősorban *keresztmetszeti* összehasonlító vizsgálatok szokásosan alkalmazott magyarázó elméleteire támaszkodnak. Így az 6.2.1. fejezetben már említett, Lipsettől és Zetterbergtől származó, Blau és Duncan ill. Treiman által pontosított, az ipari társadalmak feltételezetten egyetemes szerkezetére épülő, alapvetően funkcionalista **gazdasági** magyarázatok

[44] Saját fordításom.

merülnek fel - ezek a gazdasági növekedést követő társadalmi nyitódást jósolnának[45]. Megemlítendők továbbá a politikai szervezetek rétegződésre gyakorolt hatását vizsgáló **politikai** magyarázatok. A politikai megközelítésre koncentráló vizsgálatok a szocialista blokkon belül is lényeges különbségeket találtak, illetve - elsősorban nemzetspecifikus történelmi jellegzetességekre visszavezethetően - nem látták megalapozottnak a szocialista mobilitási minta, mint speciális típus meglétét (Simkus, 1981, Kurz-Müller, 1987). Más vizsgálatok szerint mivel néhány nem-baloldali kormányzatú országban a baloldaliakéhoz hasonló a mobilitás szintje, a politikai intervenció, bár feltételezhető, de nem az egyedüli társadalmi meghatározó tényező (pl. Breen-Luijkx, 2004). Ezért a gazdasági megközelítéssel szemben a politikai elmélet esetében nem beszélhetünk a rendszerváltozással kapcsolatos egyértelmű „előrejelzésekről".

E két, a mobilitási mintázatok változását szisztematikus (és megmagyarázható) folyamatnak látó megközelítéssel szemben áll az 6.2.1. már említett harmadik tézis, mely Sorokin (1964: 142.o) nevéhez fűződő változatában szélesebb történelmi perspektívából trendmentes ingadozásnak tekinti a mobilitás abszolút rátáinak változásait. Az elmélet újabb, a mobilitás-kutatások harmadik nemzedékének CASMIN-projektjéhez kapcsolható változatában a relatív mobilitási esélyeknek (*social fluidity*) a társadalmi rétegződés konkrét megoszlásától független mintázatának és szintjének állandóságát tételezi fel az iparosodott országokban.

Sorokinhoz hasonlóan, igaz, más elméleti alapon, de Boudon (1974) is cáfolja az egyenlőség növekedésének tendenciáját. Véleménye szerint még ha gyengül is a származás és az iskolai végzettség kapcsolata, a hátrányos származás akkor is megjelenik a munkába álláskor, vagyis az elért státuszban – így az egyenlőtlenségek egy szinttel feljebb tolódnak. Mare (1981) klasszikus elemzésében szintén az oktatás demokratizálódásához fűzött optimista várakozásokat cáfolja. Kimutatja, hogy az Egyesült Államokban a származási esélykülönbségek az expanzió ellenére növekedtek a magasabb iskolai szinteken.

Az elméletekhez kapcsolódó korábbi vizsgálatoknak a jelen kutatási kérdést érintő eredményeire térve: Treiman és Yip (1989) nemzetközi összehasonlító vizsgálatát a modernizációs tézis empirikus igazolásaként szokás citálni. Elemzésükben huszonegy, az iparosodottság különböző fokán álló ország keresztmetszeti kutatásból származó adatain az apa foglalkozásának és a kérdezett iskolázottságának a kérdezett foglalkozására gyakorolt hatását vizsgálják. Fontos megjegyezni, hogy Treiman hipotézisében a mobilitás-kutatások második nemzedékének elméleti és módszertani irányvonalát követve direkt és teljes hatásokat különböztet meg, ennek megfelelően ebben a vizsgálatban is többszörös regressziós modellek együtthatóival, mint a direkt hatások megfelelőjével dolgozik. Eredménye szerint az iparosodottabb országokban az iskolázottságnak a foglalkozásra gyakorolt hatása erősebb és az apa foglalkozásának hatása gyengébb, azaz sikerül bizonyítania az iparosodottság – társadalmi nyitottság kapcsolatot.

Egy másik gyakran hivatkozott, újabb munkában Goldthorpe (1996) arra a következtetésre jut, hogy bár a '60-as, '70-es években végzett vizsgálatok a foglalkozás, az iskolázottság és az apa foglalkozása közötti direkt hatásokat tekintve többé-kevésbé valóban alátámasztják az Erősödő Meritokratikus Kiválasztódás elvét (lásd a 45. lábjegyzetet), a '90-es évekből származó

[45] Hasonló az implikációja az újabb meritokratikus elméleteknek is, így például Jonssonénak (1992), aki bevezette az Erősödő Meritokratikus Kiválasztódás hipotézisét, mely szerint a posztindusztriális társadalmakban azok hatékonyságvezérelt volta következtében az egyéni érdemek válnak az oktatáshoz való hozzáférésnek, majd a munkaerőpiaci pozíció megszerzésének kulcstényezőivé.

munkák Anglia, Wales, Svédország adatain éppen ellentétes következtetésekre jutnak. Goldthorpe interpretációja szerint a vizsgálatok eltérő metodikáján kívül valódi változások is okozhatják ezeket az eredményeket: a modern társadalmak az elterjedt vélekedéssel szemben talán sokkal kevésbé működnek „hatékonyan”.

A Magyarországon végzett vizsgálatok közül Luijkx és társai 1995-ös eredményeit a bevezetőben már említettem; fontos még megjegyezni a munkáról, hogy az alkalmazott többszörös logisztikus regresszió folytán a fentiekhez hasonlóan szintén iskolázottság – foglalkozás ill. apa foglalkozása – foglalkozás direkt hatásokat számszerűsít. E tanulmány továbbfejlesztett változata (Luijkx és társai, 2002) szerint tényleges elmozdulás figyelhető meg a származástól a teljesítmény irányába, időintervalluma (1989 a végpont) azonban nem illeszkedik jelen kérdésünkhöz. Mégis érdemes megemlíteni, mert nem csak gazdasági, hanem politikai magyarázatokat is figyelembe vesz.

Az *intergenerációs foglalkozási mobilitás*sal kapcsolatos újabb hazai publikációk közé tartozik Bukodi (2001) munkája. Eredményei alapján a nemzedékek közötti mobilitás 1983 és 1992 között alig változott, 2000-re viszont lecsökkent. A szerző a mobilitás mértékét a teljes mobilitási arányszámmal definiálja. Későbbi munkájában (2003) ugyanezen adatok alapján a teljes mobilitási arányszámot felbontja vertikális és horizontális mobilitási arányszámokra[46], amiből kiderül, hogy az eltelt időszakban a vertikális (tehát ténylegesen presztízs-változtató) mobilitás nem változott, mert a csökkenő teljes mobilitással párhuzamosan a horizontális mobilitás aránya is csökkent. Ugyancsak az apa foglalkozása –foglalkozás kétdimenziós kapcsolatot elemzi egy újabb tanulmányban Róbert és Bukodi (2004), egy nemzetközi vizsgálat részeként, ahol Erikson és Goldthorpe által a Constant Flux c. munkában kialakított modelleket használják. Az 1973-as, 1983-as, 1992-es és 2000-es évek összevetésével úgy találják, hogy a korábbi mobilitás-növekedés a 90-es években a nők esetén rendkívül lelassult, a férfiak esetén pedig visszafordult.

A *származásnak az iskolai végzettségre* gyakorolt hatására vonatkozóan a hazai szociológia eredményei közül elsősorban módszertani szempontból érdemes megemlíteni Andorka és Simkus (1983) Boudon és Mare eredményeire reagáló kutatását. A szerzők explicit szándéka volt, hogy a mobilitás megváltozását az oktatási expanzió hatásától függetlenül mérhessék, ezért loglineáris elemzést végeztek.

A téma oktatáspolitikai fontosságánál fogva több új hazai vizsgálat is készült e kérdéskörben. Székelyi és társai (1998) a származás hatásának időbeli változását vizsgálják egy 1997-ben, felsőoktatási hallgatók körében végzett kutatásban, ahol elsősök és végzősök családi hátterét vetik össze. Eredményük szerint az elsősök között nagyobb az alacsonyabb státuszú családból származók aránya. Gazsó és Laki (2004) szerint ez az összevetés nem releváns, hiszen tárgya az elsősök és a végzősök közötti különbség, de nem az esélyek változása. Valóban, a módszerrel

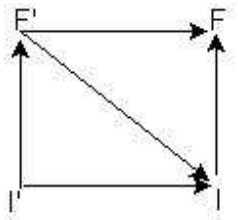
[46] Az arányszámok pontos definíciójához: mindhárom arányszám adott foglalkozási osztályozás mellett értelmezhető. A teljes mobilitási arányszámot az apjuktól eltérő osztályba tartozó válaszadók arányaként határozhatjuk meg. A teljes mobilitási arányszámot ésszerű tovább bontani vertikális és horizontális mobilitási arányszámokra, ahol a vertikális mobilitás eltérő presztízsű foglalkozási osztályok között, a horizontális mobilitás azonos presztízsű foglalkozási osztályok között valósul meg. Ugyanis a foglalkozási osztályozások mögött nem áll feltétlenül presztízsrangsor, pl. az eredeti EGP skálán külön kezelt önálló kereskedők, iparosok és szolgáltatók, vagy közvetlen termelésirányítók és szakmunkások a presztízsrangsorokban leggyakrabban azonos szinten helyezkednek el.

szemben felhozható, hogy a végzősök között már nincsenek ott az időközben lemorzsolódottak, akik esetleg inkább alacsonyabb státuszú szülők gyerekei. Gazsó és Laki 1973-as, 1997-es és 2000-es adatok alapján veti össze az egyetemi/főiskolai hallgatók származási hátterét és alapvetően változatlanságot tapasztal; interpretációjuk szerint a felsőoktatás expanziója nem kapcsolódott össze esélykiegyenlítődéssel[47]. Bukodi (1995) 1992-es adatokon, többszörös regressziós modellel vizsgálja az iskolázottság és az apa foglalkozása ill. iskolázottsága közötti direkt kapcsolatot. Az időbeli változásokat a minta életkori csoportokra bontásával, mintegy keresztmetszeti kohorszelemzést végezve számszerűsíti. Azonban a tényleges változások ezzel a módszerrel sem detektálhatók teljes megbízhatósággal. E kohorszelemzés problémája, hogy az idősebb kohorszok a természetes elhalás miatt nem teljesek, ami azért okozhat gondot, mert hazánkban igen erős a várható élettartam társadalmi meghatározottsága (pl. Kovács, Hablicsek, 2006).

További vizsgálati eredmények ’90-es évek végi adatokra támaszkodva erős származási hatást mutatnak az iskolai teljesítmény tekintetében, így Andor és Liskó (2000), vagy Róbert (2004). Ezek a vizsgálatok azonban nem végeznek összevetést korábbi évek adataival.

7.1.3 Módszertan

Kutatási kérdéseim az alábbi ábrán látható gráfhoz tartozó modell hatásaira vonatkoznak. A gráf Donald Treimannak a 21. ábra látható modelljére épül, ami tulajdonképpen az eredeti Blau-Duncan modell egy variánsa, eredeti forrása Duncan et al. (1968), de igazán ismertté valóban Treiman (1970) tette. Az eredeti Treiman-modell tehát az adatbázisaimban nem szereplő jövedelmet is tartalmazza, de ez a különbség nem érinti az itt vizsgált kapcsolatokat, azok azonosak. A gráf alapján látható, hogy a modellből az *I’F* hatás került eltávolításra.



20. ábra. A vizsgált modell (I: iskolázottság, F: foglalkozás, I’: apa iskolázottsága, F’: apa foglalkozása)

A modell irányított körmentes gráf-modell, a lokális Markov-tulajdonság[48] az alábbi feltételes függetlenségi állítást implikálja:

$$F \perp I' \mid F'I. \qquad (66)$$

[47] Ez a kijelentés első hallásra talán paradoxonnak tűnhet, ám matematikailag nem az: mint Mare (1974) korábban említett eredménye is mutatja, az adott végzettséget megszerzők arányának változása független a társadalmi esélykülönbségek változásától.

[48] A 3.5. fejezetben láttuk, hogy IKG-n és pozitív eloszlás mellett minden Markov-tulajdonság ekvivalens, ezért nem jelent megszorítást a lokális tulajdonság választása.

Treiman a modernizációs hipotézisben konkrét alhipotéziseket fogalmaz meg, eszerint a modernizációval gyengül

1. az apa foglalkozása - foglalkozás teljes hatás,
2. az apa foglalkozása - foglalkozás direkt hatás,
3. az apa foglalkozása - iskolázottság direkt hatás,

ugyanakkor erősödik az

4. iskolázottság – foglalkozás direkt hatás.

A modernizációs hipotézis egyes pontjait követve az alábbi elemzés célja a marginális loglineáris paraméterek értékeiből kiindulva a kapcsolatok időbeli változásának vizsgálata lesz. Az elemzést nemenként külön végezem, hiszen a tárgyalt kérdésekben több korábbi vizsgálat is nemi különbségeket mutatott ki (hazai friss példa Róbert, Bukodi, 2004). Fontos ismét hangsúlyozni azt a 2.1. fejezetben már tárgyalt tényt, hogy a loglineáris paraméterek függetlenek a táblák marginálisaitól, ezért pl. az apa-kérdezett összevetésekben előnyös módon a foglalkozási struktúra változásaitól független változást jelenítik csak meg, vagy - másik példával élve - kiküszöbölik az egyetemi képzés expanziójának hatását. Természetesen ugyanezzel az előnyös tulajdonsággal bír a klasszikus loglineáris elemzés is, ahogyan azt pl. Andorka és Simkus (1983) a módszertan megválasztásának indoklásakor hangsúlyozza.

Az 4.5. fejezetben részletezett eljárást követve a modellnek megfelelő eloszlás az alábbi marginális loglineáris paraméterekkel adható meg:

$$\lambda_{\varnothing}^{I'}, \lambda_{I'}^{I'}, \lambda_{F'}^{I'F'}, \lambda_{I'F'}^{I'F'}, \lambda_{I}^{I'F'I}, \lambda_{I'I}^{I'F'I}, \lambda_{F'I}^{I'F'I}, \lambda_{I'F'I}^{I'F'I}, \lambda_{F}^{I'F'IF}, \lambda_{F'F}^{I'F'IF}, \lambda_{IF}^{I'F'IF}, \lambda_{F'IF}^{I'F'IF}. \quad (67)$$

Ugyanakkor a (66)-nak megfelelő paraméterek értéke nulla:

$$\lambda_{I'F}^{I'F'IF}, \lambda_{I'F'F}^{I'F'IF}, \lambda_{I'IF}^{I'F'IF}, \lambda_{I'F'IF}^{I'F'IF}. \quad (68)$$

A modernizációs hipotézis 2.-4. pontjához a $\lambda_{F'F}^{I'F'IF}$, $\lambda_{F'I}^{I'F'I}$, $\lambda_{IF}^{I'F'IF}$ paraméterek kapcsolódnak, míg 1. pontjához a paraméterezésben nem szereplő $\lambda_{F'F}^{F'F}$ paraméter tartozik. Ezek időbeli változását vizsgálom a továbbiakban.

Ugyanezen adatok felhasználásával több vizsgálat is készült már (pl. Bukodi 2001, Bukodi 2003, Róbert, Bukodi 2004). Az eredmények közvetlen összevethetősége érdekében Róbert és Bukodi (2004) döntését követve a vizsgálati populációt a 20-69 éves aktuálisan dolgozókban ill. a valaha dolgozó munkanélküliekben határoztam meg. A nemenként külön bontott együttes eloszlás celláit a szokásosan alkalmazott 0,5-ös elemszámmal töltöttem fel, az üres cellák elkerülése érdekében. A minta elemszámai a következőképpen alakultak:

1983:	férfiak - 9076 mintaelem (2 üres cella),	nők – 7834 (3 üres)
1992:	férfiak - 7045 (2 üres),	nők – 6364 (2 üres)
2000:	férfiak - 2463 (nincs üres),	nők – 2009 (5 üres)

Az elemzés során az adatbázisokhoz eredetileg hozzárendelt súlyokat használtam.

A modell viszonylagos összetettségét ellensúlyozandó[49] a változók képzésénél a kategóriák alacsony számára törekedtem. Az *iskolázottság*ot bináris változóként használtam, mégpedig az apa (érettségizett/érettségi nélkül) ill. a kérdezett (diplomás/diploma nélkül) esetére más kategóriákkal, ennek oka a megközelítőleg azonos eloszlású változók képzésére való törekvés volt.

Az *apa foglalkozását* a kérdezett 14 éves korában meglevő foglalkozással definiálták a kérdőívek. A *foglalkozás* meghatározásakor szerencsés lett volna a kérdezett első foglalkozásának használata, de ez az információ nem állt rendelkezésre mindegyik adatbázisban. Ezt a problémát megoldandó Andorka (1995) azzal próbálkozik, hogy külön vizsgálja a fiatalokat, ahol a "jelenlegi" foglalkozás jó proxy a pályakezdő foglalkozásra. Jelen vizsgálatban ezzel a céllal külön elvégzem majd a 20-29 éves korosztály vizsgálatát. A foglalkozási kategóriák kialakításánál Erikson és Goldthorpe (1992) ajánlását követve az EGP skálához illeszkedő hármas beosztást használtam:

1. felső osztály (szolgáltató osztály: vezető, szellemi szabadfoglalkozású, diplomás alkalmazott)
2. középosztály (egyéb szellemi, szakmunkás, iparos/kereskedő/szolgáltató, gazdálkodó)
3. alsó osztály (alacsonyabb szintű kereskedelmi és szolgáltató dolgozók, betanított munkás/segédmunkás/mezőgazdasági fizikai).

Ennek a megoldásnak a kategóriák alacsony számán kívül az is előnye, hogy az ordinalitás következtében vertikális mobilitást határoz meg a kategória-ugrás.

7.1.4 Eredmények

A valószínűség-hányados tesztstatisztika szerint a modell illeszkedése mindhárom évben, mindkét nemre rossz ($p<0,01$), kivéve a 2000-es évben a nők esetét, ahol a modell jól illeszkedik ($p=0,115$). Vagyis a modellt definiáló, (66)-ban meghatározott, az apa iskolázottságára és a kérdezett foglalkozására vonatkozó feltételes függetlenségi állítás nem fogadható el. Ezért a modellt elvetettem, és a továbbiakban a 2. ábra modelljének módosított, az *I'-F* nyilat is tartalmazó változatát alkalmaztam. Ez, mivel egyetlen hiányzó nyilat sem tartalmaz, telített modell. A paraméterezést tekintve ez a (68)-ban felsorolt paraméterek szabad, nem nullában rögzített paraméterként történő használatát jelenti. A következőkben e paramétereket fogom megbecsülni.

A kutatási kérdéseim szempontjából releváns négy paraméter értékének becslése az alábbi táblázatokban található. A cellákban a felső érték a férfiakra, az alsó a nőkre kapott becslés.

5. táblázat. Paraméterbecslések a modernizációs hipotézis 2. pontjához

[49] Visszautalnék az 6.2.2. fejezetre: sok kategória használatával esetleg túl sok üres vagy majdnem üres cella állhatna elő.

év	*apa foglalkozása- foglalkozás direkt hatás* $\lambda_{F'F}^{I'F'IF}$								
	apa: felső osztály			*apa: középosztály*			*apa: alsó osztály*		
	felső	*közép*	*alsó*	*felső*	*közép*	*alsó*	*felső*	*közép*	*alsó*
1983	0,06	-0,14	0,08	0,04	0,12	-0,16	-0,10	0,02	0,08
	0,36	-0,25	-0,11	-0,09	0,28	-0,19	-0,27	-0,03	0,3
1992	0,33	-0,13	-0,20	-0,02	0,17	-0,15	-0,31	-0,04	0,35
	0,33	-0,2	-0,13	-0,07	0,08	-0,01	-0,26	0,12	0,14
2000	0,30	-0,20	-0,10	-0,11	0,31	-0,20	-0,19	-0,11	0,30
	0,15	-0,10	-0,05	-0,10	0,06	0,04	-0,05	0,04	0,01

6. táblázat. Paraméterbecslések a modernizációs hipotézis 3. pontjához

év	*apa foglalkozása – iskolázottság direkt hatás* $\lambda_{F'I}^{I'F'I}$					
	apa: felső osztály		*apa: középosztály*		*apa: alsó osztály*	
	diplomás	*diploma nélküli*	*diplomás*	*diploma nélküli*	*diplomás*	*diploma nélküli*
1983	0,32	-0,05	-0,27	-0,32	0,05	0,27
	0,30	-0,13	-0,17	-0,30	0,13	0,17
1992	0,34	-0,06	-0,28	-0,34	0,06	0,28
	0,34	-0,09	-0,25	-0,34	0,09	0,25
2000	0,42	-0,13	-0,29	-0,42	0,13	0,29
	0,28	-0,07	-0,21	-0,28	0,07	0,21

7. táblázat. Paraméterbecslések a modernizációs hipotézis 4. pontjához

év	*iskolázottság - foglalkozás direkt hatás* $\lambda_{IF}^{I'F'IF}$					
	diplomás			*diploma nélküli*		
	felső	*közép*	*alsó*	*felső*	*közép*	*alsó*
1983	1,19	-1,19	-0,28	0,28	-0,91	0,91
	1,28	-1,28	-0,43	0,43	-0,85	0,85
1992	1,35	-1,35	-0,4	0,4	-0,95	0,95
	1,15	-1,15	-0,4	0,4	-0,75	0,75
2000	0,92	-0,92	-0,31	0,31	-0,61	0,61
	1,01	-1,01	-0,19	0,19	-0,82	0,82

8. táblázat. Paraméterbecslések a modernizációs hipotézis 1. pontjához

év	*apa foglalkozása – foglalkozás teljes hatás* $\lambda_{F'F}^{F'F}$								
	apa: felső osztály			*apa: középosztály*			*apa: alsó osztály*		
	felső	*közép*	*alsó*	*felső*	*közép*	*alsó*	*felső*	*közép*	*alsó*
1983	0,72	-0,20	-0,52	-0,23	0,09	0,14	-0,49	0,11	0,38
	0,73	0,02	-0,75	-0,25	-0,05	0,30	-0,48	0,03	0,45
1992	0,79	-0,23	-0,56	-0,15	0,08	0,07	-0,64	0,15	0,49
	0,71	-0,06	-0,65	-0,17	0,02	0,15	-0,54	0,04	0,50
2000	0,95	-0,41	-0,54	-0,32	0,27	0,05	-0,63	0,14	0,49
	0,66	-0,03	-0,63	-0,23	0,05	0,18	-0,43	-0,02	0,45

A fenti táblázatokban szereplő minden paraméter értéke szignifikáns[50]. A hatások időbeli változásának szignifikanciájára vonatkozó próba szerint pedig az

- apa foglalkozása - foglalkozás direkt hatás nem változott szignifikáns módon egyik nem esetén sem,
- az apa foglalkozása- iskolázottság direkt hatás a férfiak esetében 1983 és 1992 között nem változott szignifikánsan, de 1992-2000 között (és a teljes 1983-2000 közötti intervallumot tekintve is) szignifikánsan erősödött, míg a nők esetében sehol nem változott szignifikánsan,
- az iskolázottság – foglalkozás direkt hatás nem változott szignifikáns módon sem a nők, sem a férfiak esetében, míg
- az apa foglalkozása – foglalkozás teljes hatás a férfiak esetében mind 1983-1992 között, mind 1992-2000 között (és a teljes 1983-2000 intervallumon is) szignifikánsan változott. Az egyes paraméterértékek változására külön elvégzett szignifikancia-teszt alapján elmondható, hogy ezt a változást a záródás irányába mutató (abszolút értékükben növekedő) paraméterek szignifikáns változása hozta. A nők esetében a változás csak az 1983-1992 intervallumon szignifikáns. Itt a részletes teszt alapján elmondható, hogy a középosztálybeli apák utódainak esélyei változtak szignifikánsan, mégpedig a nagyobb nyitottság irányába. 1992-2000, sőt 1983-2000 sem találunk viszont szignifikáns változást.

Már említett okok miatt a 20-29 éves korosztályt külön is vizsgáltam, bár az esetszámok így erősen lecsökkentek (2000-ben csak 703 férfi és 530 nő tartozott a mintába), tehát óvatosan értelmezendők az eredmények. A tapasztalatok jórészt a fentiekhez hasonló tendenciát– a nők esetén változatlan esélyegyenlőtlenségeket, a férfiaknál bizonyos mutatók esetén már növekvő egyenlőtlenségeket - mutatnak. Az utóbbi megfigyelést részletezve: az apa foglalkozásának és a fiú iskolázottságának 1983-1992 között felerősödő összefüggését, illetve az iskolázottság – foglalkozás hatás 1983 és 2000 közötti szignifikáns csökkenését találtuk.

7.1.5 Összegzés

Eredményeim szerint (1) *az iskolázottságnak lényegi szerepe van a társadalmi státusz elérésében*, ám ez a hatás nem erősödött, sőt, a fiatal férfiaknál inkább csökkent. Ugyanakkor (2) bár az apa foglalkozásának a kérdezett foglalkozására vonatkozó teljes hatása mindhárom évben mindkét nemre igen erős, (3) a direkt hatás jóval gyengébb, de azért szignifikáns. Figyelembe véve, hogy (4) az apa foglalkozása erős hatást gyakorol az iskolázottságra, elmondható, hogy a társadalmi újratermelődés egy része az iskolai végzettségen keresztül valósul meg, vagyis *az iskolának a társadalmi reprodukcióban is lényeges szerepe van, az újratermelődés fontos csatornájának tekinthető*.

Az (1)-(4) eredmények mind egybevágnak az ipari társadalmakra vonatkozó klasszikus megfigyelésekkel, pl. a múlt század közepének Amerikáját vizsgáló, a mobilitás-kutatások második nemzedékéhez tartozó, így a szokásos útelemzést használó Blau és Duncan eredményeivel (Ganzeboom, Treiman, Ultee, 1991). Ugyanakkor Treiman és Yip (1989) 21

[50] Ha másként nem jelzem, 0,05 szintű szignifikanciát számolok.

országnak a ’70-es évek elejéről származó adatait összevetve az iskolázottságot lényegében a származástól függetlennek találják (szemben a (4) eredménnyel), ezzel arra a következtetésre jutva, hogy a foglalkozási státusz sokkal inkább a végzettség, mint a származás függvénye. Ezzel szemben egy friss nemzetközi vizsgálat angol, francia, ír, svéd és holland adatokból kiindulva a származásnak a státusmegszerzésre vonatkozó hatását az iskolázottságra korrigálva is erősnek találta (Breen-Luijkx, 2004). Ezekben az országokban, Magyarországhoz hasonlóan, a társadalmi újratermelődésnek az iskolázottságon kívül más jelentős csatornái is feltételezhetők.

Az *apa foglalkozásának az iskolázottságra* vonatkozó hatása eredményeim szerint mindhárom évben szignifikáns, ez megfelel az iskolai teljesítmény tekintetében erős származási hatást kimutató hazai vizsgálatoknak (Andor, Liskó 2000, Róbert 2004). A hatás férfiak esetén tapasztalt erősödése, nők esetén talált változatlansága az Elméleti háttér, korábbi vizsgálatok c. részben említett – igaz, nem teljesen azonos kérdésfeltevéssel, nem azonos intervallumon és nem azonos módszertannal készült – hazai vizsgálatok legtöbbjével egybevág. Székelyi és társai (1998) írása ugyan ellentmondani látszik, ők a mobilitás növekedését detektálják, ám az eredmények összevethetőségével kapcsolatban korábban már jeleztük a lehetséges problémákat.

Eredményeim lehetséges magyarázata elsősorban az oktatás demokratizálódásához fűzött optimista várakozásokat cáfoló vizsgálatokhoz (Mare, Boudon) köthető. Míg ezek a vizsgálatok a pusztán a felsőoktatási férőhelyek számának növelését célzó társadalompolitikai beavatkozások szükségszerű sikertelenségét hangsúlyozzák, addig hazai szerzők a felsőoktatás expanziójának végrehajtási módját is kritizálják, pl. Gazsó (1997) szerint a szelektív felsőoktatást a spontán makrotársadalmi folyamatokat felerősítve a közoktatás-politika is támogatja.

Az Elméleti háttér, korábbi vizsgálatok c. részben említett hazai tanulmányok közül az *apa foglalkozásának az utód foglalkozására* vonatkozó *direkt*, iskolázottságra kontrollált hatását az adott periódusra egyedül Luijkx és társai (1995) vizsgálták férfiakra vonatkozóan, 1973-as, 1983-as és 1992-1993-as adatokat összevetésével. Megközelítésük – az elemzési módszer eltérésén kívül - abban különbözik a jelen tanulmányétól, hogy nem korrigáltak az apa iskolázottságára, illetve külön vizsgálták a le- ill. felfelé mutató mobilitás esélyeit. A származás felfelé mutató mobilitásra vonatkozó hatásával kapcsolatban U-alakú görbét detektáltak, míg a lefelé mutató mobilitással kapcsolatban változatlan hatást. Bár eredményük a módszertani eltéréseket figyelembe véve csak óvatosan vethető össze közvetlenül az enyémmel, de nem is feltétlenül mond ellent az általam a le- ill. felfelé mutató mobilitásra összességében talált 1983-1992 közötti változatlanságnak.

Hasonlóan, az *iskolázottságnak a foglalkozására* vonatkozó *direkt*, származásra kontrollált hatásával az említett hazai források közül Luijkx és társai foglalkoztak. Ők 1973-as, 1983-as és 1992-1993-as adatok alapján fordított U-alakú görbét detektáltak a végzettség lefelé mutató mobilitásra gyakorolt hatására vonatkozóan, a felfelé mutató mobilitásra vonatkozó hatás változatlanságát tapasztalták, ami újfent nem mond ellent az én (eltérő módszerrel kapott) eredményemnek, azaz az 1983-1992 közötti, bármilyen irányba mutató mobilitás változatlanságának.

Több tanulmány foglalkozik ugyanakkor a korábban már idézettek közül az *apa foglalkozása – foglalkozás teljes hatással*. Bukodi (2001) és (2003) mobilitási arányszámot használ a kérdés

megválaszolására, így a két eredmény nem vethető össze közvetlenül[51]. Róbert és Bukodi (2004) eredményei, bár más módszerrel, unidiff modellekkel készültek (ezekről már írtam az 6.2.1. fejezetben), jórészt alátámasztják tapasztalataimat. A unidiff modell alapján az 1983-as, 1992-es és 2000-es évek összevetésében a nőknél a mobilitás-növekedés lassulására, míg a férfiaknál már záródásra utaló jeleket találtak.

A *munkaerőpiac működésére*, Treiman hipotézisének kiindulási pontjára vonatkozó eredményeim magyarázatakor Goldthorpe meritokratikus elméletekkel szembeni szisztematikus kritikája (1996) citálható. Szerinte *nem feltétlenül a hatékonyság* működteti a munkaerő allokálást: az iskolázottság egyfajta életstílust, kulturális hátteret is jelent, gyakran ezért és nem a tényleges tudásért keresnek a munkaadók diplomás alkalmazottat. A *posztindusztriális*, tudás-alapú gazdaságban ugyanakkor nem is feltétlenül a kognitív képességek kerülnek előtérbe: olyan, tipikusan expanzióban levő poszt-indusztriális területeken, mint a kereskedelem, PR vagy a promóció inkább a megjelenés, a beszédstílus kap nagy szerepet. Ezek viszont a családból öröklődnek, s nem az iskolán keresztül szerezhetők meg. A modern társadalmakban ugyancsak erősödő kis- és középvállalkozói szektorban hasonló a helyzet: a szülők nem csak a tárgyi javakat örökíthetik át utódaikra, hanem a szintén szükséges vállalkozói kultúrát, individualisztikus munkaerkölcsöt is.

A fentieket összefoglalva: a vizsgált időszakban Magyarországon Treiman hipotézisével ellentétes (férfiaknál bizonyos egyenlőtlenségek növekedése), vagy azt legalábbis nem alátámasztó (változatlan esélyegyenlőtlenségek) tendenciák figyelhetők meg. A szülői háttér hatása a fiúk iskolázottságára folyamatosan erősödik, sőt a származás az iskolarendszert megkerülve, más csatornákon keresztül is befolyásolja az elért státuszt. Eközben a diploma munkaerőpiaci értéke, vagyis a foglalkozási státusz elérésében betöltött szerepe a hipotézisben várttal szemben nem növekedik, sőt, a fiataloknál inkább csökken.

Az eredmények ér vényességéről még két megjegyzést szeretnék tenni. Az egyik, hogy esetleg előfordulhat, hogy a használt változók egyre kevésbé képesek megragadni a foglalkozási státuszt, tehát az itt használt kategóriák esetleg átdefiniálásra szorulhatnak. Továbbá a bevezetőben említett reakcióidő-problémára hivatkozva az is elképzelhető, hogy a rendszerváltás és a vizsgálat időpontja között eltelt idő túl rövid a gazdasági/politikai változások tényleges társadalmi leképeződéséhez, bár a jelenlegi folyamatok kevéssé valószínűsítik az itt kimutatottal ellentétes irányú tendencia megjelenését. Pl. az általános iskolai teljesítményben megmutatkozó társadalmi egyenlőtlenségek növekedését kimutató vizsgálatok (Kertesi, Kézdi, 2004) alapján inkább ennek a felsőoktatásra való továbbgyűrűzését jelezhetnénk.

7.2 A Treiman-modellnek és a Boguszak és társai (1990) által javasolt módosításának összevetése nemzetközi adatokon

Az alábbiakban egy korábban már szintén közölt (Németh, 2006a, és Németh, Rudas, Bergsma, 2006) vizsgálat kivonatát mutatom be. A Treiman-modell Boguszak és társai (1990) által javasolt módosításának érvényességét vizsgálom majd. Az előző fejezethez képest módszertani szempontból az jelent újdonságot, hogy irányított körmentes gráf mellett láncgráfot

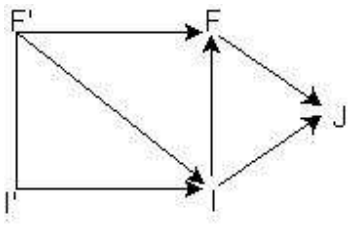
[51] A mobilitási arányszám ugyanis, mint azt a 6.2.1. fejezetben kiemeltem, az itt használt marginális loglineáris paraméterekkel szemben változással reagálhat csupán a táblázat marginálisainak, tehát a struktúra megváltozásának a hatására is.

is vizsgálok majd. Az első modellhez tartozó gyakoriságok és paraméterek Mathematica-ban ill. SPSS-ben elvégzett becslése, ill. illeszkedésvizsgálata a Függelékben, a 8. fejezetben található meg teljes részletességgel.

A Treiman-modell a 21. ábra látható. Mint már említettem, ez az eredeti Blau-Duncan modell egy variánsa, eredeti forrása Duncan et al. (1968), de igazán ismertté Treiman (1970) tette. Az előző alfejezetben ismertetett modernizációs hipotézis eredetileg ezen a modellen volt értelmezve. Az *I'F'* hatás irányítatlan, ezért a modellt láncgráf modellként kezelem majd, bár ahogy az 3.8. fejezetben említettem már, a láncgráf-modell Markov-ekvivalens az IKG párjával (amelyben ez a hatás irányítottra van cserélve). A láncgráfhoz négy komponens tartozik: $\{I'F'\}$, $\{I\}$, $\{F\}$, $\{J\}$.

Az 3.8. fejezetben láttuk, hogy ezen a gráfon a láncgráfokra definiált valamennyi Markov-tulajdonság ekvivalens. Nem teszek hát igazi megszorítást, ha a lokális tulajdonságot veszem, amiből a következő feltételes függetlenségi állításokat kapjuk:

$$\begin{aligned} &J \perp I'F' \mid FI, \\ &F \perp I' \mid F'I. \end{aligned} \tag{69}$$



21. ábra. Treiman-modell (Treiman, 1970, eredetileg Duncen et al., 1968). I': apa iskolázottsága, F': apa foglalkozási státusza, F: foglalkozási státusz, I: iskolázottság, J: jövedelem

Bár, mint láttuk, még nincsen kidolgozott általános eljárás a láncgráfok paraméterezhetőségére, egyedi esetekben azonban jó paraméterezés adható azzal a kézenfekvő eljárással, hogy az 1. tétellel paraméterezzük egyenként a feltételes függetlenségeket. A marginálisok M halmazát a komponensek rendezése definiálja: *I'F', I'F'I, I'F'IF, I'F'IFJ*. A feltételes függetlenségek miatt az alábbi paraméterek értéke 0:

$$\lambda_{I'F*}^{I'F'IF},\ \lambda_{I'J*}^{I'F'IFJ},\ \lambda_{F'J*}^{I'F'IFJ}.$$

A modellhez tartozó eloszlás szabad paraméterei:

$$\begin{aligned} &\lambda_{*}^{I'F'},\ \lambda_{I*}^{I'F'I}, \\ &\lambda_{F*}^{I'F'IF}, \text{ ha } I' \notin *, \\ &\lambda_{J*}^{I'F'IFJ}, \text{ ha } I', F' \notin *. \end{aligned}$$

Mivel a paraméterezés rendezett módon dekomponálható, hierarchikus és teljes, ezért minden, a 2.3. fejezetben említett előnyös tulajdonsággal rendelkezik.

Boguszak és társai (1990) Csehszlovákiát vizsgálva a modell két hatásra kiterjedő módosítását javasolták[52], mindkettőt a „korábbi diszkrimináció kompenzálására" vonatkozó elméletük alapján. (1) Modelljükben az iskolázottságra is hatása van a foglalkozásnak. Eszerint a politikailag megbízható, de aluliskolázott munkaerő kékgalléros vezető pozíciókba való juttatását (e pozitív diszkriminációt) gyakran legitimizálták utólag felnőttkori beiskolázásokkal. (2) A foglalkozásra direkt hatása van az apa iskolázottságának. Ez a „szocialista anomáliának" nevezett jelenség a negatív iskolai diszkrimináció hozadéka: a magas státuszú családokból származó utódok elől bár az iskola, mint érvényesülési csatorna el volt zárva, a kulturális tőke és az érvényesülési törekvés átörökítése folytán más csatornákon ugyan, de mégis nagy számban sikerült érvényesülniük.

A 22. ábra mutatja a két módosítás bevezetésével kapott láncgráf-modellt. Az 3.8. fejezetben kiderült, hogy a gráf bármely Markov-tulajdonság mellett ezt adja:

$$J \perp F'I' \mid FI,$$

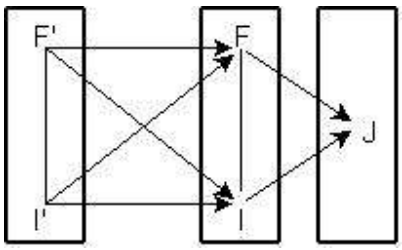
tehát az eredeti Treiman-modellnél tágabb modell. A paraméterezéséhez szükséges, a gráf komponenseihez illeszkedő marginálisok a következők: *I'F', I'F'IF, I'F'IFJ.* Az alábbi paraméterek értéke 0:

$$\lambda_{I'J*}^{I'F'IFJ},\ \lambda_{F'J*}^{I'F'IFJ}.$$

míg a szabad paraméterek a következők:

$$\lambda_{*}^{I'F'},\ \lambda_{I*}^{I'F'IF},\ \lambda_{F*}^{I'F'IF},$$
$$\lambda_{J*}^{I'F'IFJ},\ \text{ha}\ I', F' \notin *.$$

Ismét elmondható, hogy sikerült rendezett módon dekomponálható, hierarchikus és teljes paraméterezést adnunk, ezért az minden, a 2.3. fejezetben említett előnyös tulajdonsággal rendelkezik.



22. ábra. A Treiman-modell módosítása Boguszak et al. (1990) alapján

[52] Ők az eredeti Blau-Duncan modell módosításában gondolkodtak, de értelemszerűen a Treiman-modellen is bevezethetők javaslataik.

Látható, hogy az első modell valóban tartalmazza a második modellt, hiszen az utóbbi kötött paraméterei az előbbiben is 0-ra állítva szerepelnek, tehát egymásbaágyazott a két modell. Mint a 4.8. fejezetben láttuk, a valószínűség-hányados statisztika (L^2) alkalmas eszköz egymásbaágyazott modellek összevetésére. Bár ismert, hogy az aszimptotikus *p* érték kevésbé megbízható akkor, ha a táblában kis elemszámú cellák találhatók (mint ahogy a jelen esetben is), az L^2 a modellek *összevetésére* azonban ilyenkor is alkalmas eszköz (Agresti, 2002, pp. 361.).

A modellek összevetésére az ISSP 1992-es adatait használom, ebben az évben az eredeti modell mind az öt változójáról rendelkezésre állnak az adatok. Boguszak és társai eredeti vizsgálati tárgya, Csehszlovákia mellé kontrollként egy másik kelet-európai ex-szocialista államot (Magyarország), illetve kontrollként egy klasszikus liberális államot, az Egyesült Államokat vontam be az elemzésbe.

A vizsgálati populációt a munkaerőpiacon jelen levő, havi jövedelemmel rendelkező legalább 25 évesek alkotják. Az együttes eloszlásnak a statisztikai elemzésben problémát okozó üres cellái miatt a szokásosan alkalmazott 0,5-ös elemszámmal történő feltöltést végeztem (ezt az értéket adtam hozzá minden cellához, nem csak az üresekhez). Az elemzés során az adatbázisokhoz eredetileg hozzárendelt súlyokat használtam.

A változók képzését az előző alfejezethez hasonlóan végeztem. Az *iskolázottság* kategóriái: 1 „diplomás”, 2 „diploma nélküli”. A *jövedelem* kategóriánál az egyéni havi jövedelmet vettem számításba, mégpedig az eloszlás változásának és az inflációnak a hatását kiküszöbölendő a mediánt, mint küszöbértéket alkalmaztam; a két kategória: 1 ”medián felett”, 2 „medián alatt”. A kérdezett és az apa *foglalkozás*ának definiálása az előző fejezetben látott módon, az EGP foglalkozási presztízs skálához illeszkedő hármas foglalkozás-kategorizációval történt.

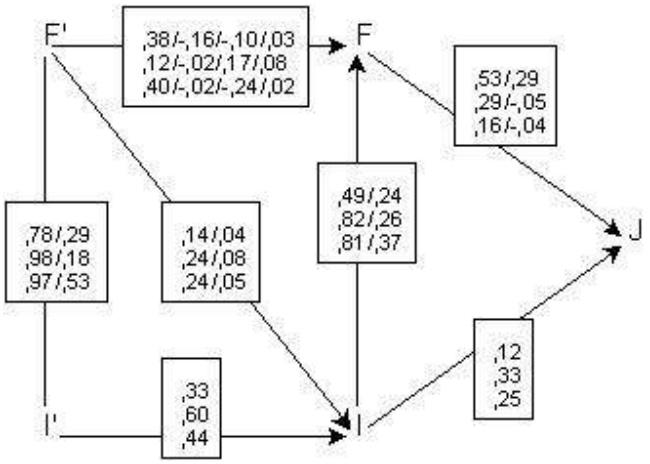
A 9. táblázat tartalmazza az illeszkedéstesztek eredményét. Mindkét modell jól illeszkedik mindhárom ország adataihoz. Az 1. modell 2. modellel való összevetésében az illeszkedésromlás nem szignifikáns, még a két kelet-európai országban sem. Vagyis érdemes több megkötést alkalmazni, tehát az eredeti Treiman modell jobbnak tűnik a Boguszak és társai által ajánlottnál.

9. táblázat. Az ötváltozós modellek illeszkedéstesztjeinek eredményei

		1. modell	*2. modell*	*1. modell a 2.-hoz*
	Szabadságfok	42	30	42-30=12
USA	L^2	25,6	14,7	25,6-14,7=10,9
	p	0,978	0,991	0,538
Magyarország	L^2	38,1	23,3	38,1-23,3=14,8
	p	0,643	0,802	0,253
Csehszlovákia	L^2	38,8	25,5	38,8-25,5=13,3
	p	0,614	0,700	0,348

Érdemes lenne a jobbnak bizonyult modell paramétereinek becslését is megvizsgálni, ezzel az előző fejezet kontextusába visszahelyezkedve Treiman modernizációs elmélete alapján

egymáshoz képest értékelni a három ország mobilitási jegyeit. Az alábbi ábra az 1. modellhez tartozó, az élekhez rendelhető (tehát kételemű hatáshoz tartozó) becsült paraméterértékeket mutatja be.



23. ábra. Az 1. modell paraméterbecslései, a dobozokban soronként felülről lefelé: USA, Magyarország, Csehszlovákia

A becslések jobb értelmezését teszi lehetővé, ha kihasználjuk, hogy a (60) megfelelője itt is alkalmazható, így a szülő-leszármazott feltételes asszociációk feltételhalmazából a nem-szülők elhagyhatók. Pl. az *IF* nyílon levő érték az *IF* hatáshoz tartozó paraméter, vagyis $\lambda_{IF}^{I'F'IF}$ becsült értékét mutatja, ami az apa jellemzőire (*I'F'*) korrigált feltételes *IF* kapcsolat becsült nagysága. Ám mivel értéke megegyezik $\lambda_{IF}^{F'IF}$ -vel, az F_2-vel ténylegesen kapcsolatban álló, csupán az *F'*-re korrigált hatásként is értelmezhető. A paraméterek tendenciaszerű mintázatait interpretálva elmondható, hogy

1. a családi háttér az ex-szocialista országokban erősebben befolyásolja az iskolai végzettséget, mint Amerikában (*I'I* és *F'I* hatás előbbiekben nagyobb), ugyanakkor
2. az iskolázottság ezekben az országokban fontosabb meghatározója a foglalkozásnak (*IF* hatás), mint Amerikában,
3. a jövedelemet Amerikában a foglalkozás erősebben meghatározza (*FJ* hatás), mint a kelet-európai országokban,
4. végül: utóbbi államokban Amerikához képest kissé erősebb az iskolázottság közvetlen jövedelem-meghatározó szerepe (*IJ* hatás).

Az 1., 3. és 4. szempont alapján Amerika a treimani értelemben modernizáltabb, mint a két kelet-európai ország, és csak a 2. szempont billenti utóbbiak felé a mérleget.

7.3 Klasszikus adatok, új modell: Sewell és Shah (1968) vizsgálata a felsőfokú továbbtanulásról

Az alábbi fejezetben egy irányított körmentes gráf-modellt, egy 2-es típusú láncgráfhoz tartozó modellt, és az előbbihez tartozó útmodellt vizsgálunk majd. Az adatok a mobilitáskutatás

egyik alapmodelljéhez kapcsolódnak, és gyakran szerepelnek IKG-k példájaként, az itt bemutatott láncgráf-illesztés azonban újdonságot jelent.

Az alábbi adatok Sewell és Shah (1968) klasszikus vizsgálatából származnak, azóta is gyakran használt példák kategoriális adatok elemzésében (Fienberg, 1980, Whittaker, 1990, Spirtes et al., 1993, Heckerman et al., 1999).[53] A Wisconsin végzős középiskolásainak egy mintáján felvett adatok a felsőfokú továbbtanulással kapcsolatos elhatározást és annak befolyásoló tényezőit írják le. A 3.8. fejezetben már említettem ezt a vizsgálatot, a hozzá kapcsolódó kutatási paradigmát, és a belőle később kifejlődő, a Blau-Duncan modellhez hasonlóan közismert Wisconsin-modellt. Említettem már a vizsgálat motivációját is: az előző alkalmazásokban tárgyalt Blau-Duncan modell és variánsai, így a Treiman-modell is a státuszelérési folyamatot és benne az iskolázottságnak a családi háttér átörökítésében játszott közvetítő szerepét írják le, de nem jelenítenek meg olyan mechanizmusokat, melyek révén ebben a folyamatban a családi háttér hatása ténylegesen működik. Ezeknek a mechanizmusoknak a feltárása volt Sewell és Shah célja, amikor a családi erőforrások genetikai és szociálpszichológiai tényezőit is bevontak modelljükbe. A gyermekek számára fontos személyek hatásának (az alábbi modellben a szülői biztatásnak) a megjelenítésével modelljük az iskolai teljesítmény mögött álló motivációk fontosságát hangsúlyozza.

10. táblázat. Sewell és Shah adatai (n=10.318). A: alacsony, AK: alsó-közép, FK: felső-közép, F: felső

				Társadalmi háttér			
Nem	*IQ*	*Szülői biztatás*	*Felsőfokú tervek*	*A*	*AK*	*FK*	*F*
Férfi	*A*	*Kisfokú*	*igen*	4	2	8	4
			nem	349	232	166	48
		Nagyfokú	*igen*	13	27	47	39
			nem	64	84	91	57
	AK	*Kisfokú*	*igen*	9	7	6	5
			nem	207	201	120	47
		Nagyfokú	*igen*	33	64	74	123
			nem	72	95	110	90
	FK	*Kisfokú*	*igen*	12	12	17	9
			nem	126	115	92	41
		Nagyfokú	*igen*	38	93	148	224
			nem	54	92	100	65
	F	*Kisfokú*	*igen*	10	17	6	8
			nem	67	79	42	17
		Nagyfokú	*igen*	49	119	198	414
			nem	43	59	73	54
Nő	*A*	*Kisfokú*	*igen*	5	11	7	6
			nem	454	285	163	50
		Nagyfokú	*igen*	9	29	36	36
			nem	44	61	72	58

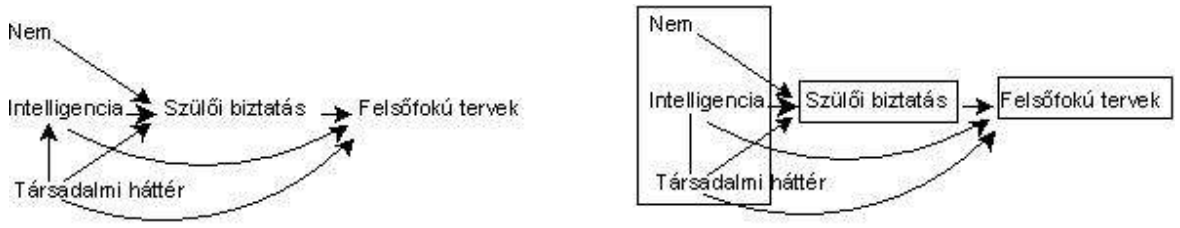
[53] Az itt tárgyalt első modell ill. a modell variánsai már szerepeltek a 3.8. fejezetben. Ugyanott kitértem a korábbi szerzők megközelítésmódjára is.

AK	*Kisfokú*	*igen*	5	19	13	5
		nem	312	236	193	70
	Nagyfokú	*igen*	14	47	75	110
		nem	47	88	90	76
FK	*Kisfokú*	*igen*	8	12	12	12
		nem	216	164	174	48
	Nagyfokú	*igen*	20	62	91	230
		nem	35	85	100	81
F	*Kisfokú*	*igen*	13	15	20	13
		nem	96	113	81	49
	Nagyfokú	*igen*	28	72	142	360
		nem	24	50	77	98

Az első most vizsgált modellt (a 24. ábra bal oldali gráfját) kereső algoritmussal kapta Spirtes et al. (1990), az algoritmussal az összes szóba jöhető IKG-t végignézve, a legjobban illeszkedő modellt keresve. Egyetlen megszorításuk az volt, hogy a változók időbeli sorrendje a következő legyen: ($\{N, I, T\}, S, F$). A gráfot páronkénti Markov-tulajdonság mellett értelmezve az alábbi modellt kapjuk:

$$\text{1.modell: } N \perp T,\ N \perp I \mid T,\ F \perp N \mid TIS. \tag{70}$$

A Spirtesék által használt algoritmus nem tudott dönteni a *TI* él irányáról, mert az él bármely irányítás mellett ugyanazokat a feltételes függetlenségeket (=ugyanazt a modellt) definiálja. Ezért a kézenfekvőbb *T-I* irányítás mellett döntöttek. Elképzelhető az is ugyanakkor (ezt ők is megemlítik), hogy a köztük levő kapcsolatot egy közös, itt nem mért ok hozza létre. Én a láncgráf-modellek eszközének birtokában ezért azt javaslom, hogy az élt irányítatlannak vegyük, és láncgráfra térjünk át. Az ábra jobb oldali gráfja így jön létre. Ez 2-es típusú láncgráf, hiszen az *N* nincs összekötve egyik komponensbelijével sem. A modell megfelel Fienberg (1980) koncepciójának is, aki a problémát két regressziós modellként fogta fel, az elsőben *S* a függő és *NIT* a magyarázó változók halmaza, a másodikban *F* a függő és a maradék négy a magyarázó változók halmaza.



24. ábra. Két modell a Sewell és Shah-féle adatokhoz

A gráfot 3-as típusú páronkénti Markov-tulajdonság mellett értelmezem majd, mert az első komponensben a feltételes függetlenségeket nem kívánom a komponens többi tagját feltéve értelmezni. Mivel a láncgráf-modellek ekvivalencia-tételei 1-es típusú gráfon voltak értelmezve, ezért nem tudható, hogy ez a páronkénti tulajdonság általános esetben ekvivalens-e a blokk-

rekurzív tulajdonsággal. Egyedi esetben azonban a kérdés megválaszolható. Itt pl. a 3-as típusú blokk-rekurzív tulajdonság mellett a modell nem ekvivalens a páronkénti megfelelőjével, viszont ekvivalens az LWF-tulajdonság mellett értelmezett megfelelőjével, ezért nem érdemes külön vizsgálni. A 3-as típusú páronkénti tulajdonság mellett az alábbi modellt adja a gráf:

$$\text{2. modell: } N \perp T,\ N \perp I,\ F \perp N \mid TIS. \tag{71}$$

Mint látható, a két modell csak a középső feltételes függetlenség feltételhalmazában különbözik. Mivel

$$\left(N \perp T \text{ és } N \perp I \mid T\right) \Leftrightarrow N \perp TI \Rightarrow N \perp I,$$

az első modellhez tartozó állítások implikálják a másodikhoz tartozókat. Tehát a láncgráf-modellnek az IKG-modell almodellje.

Az 1. modell paraméterezéséhez a 4.5. fejezet alapján pl. a következő marginálisok használhatók: *N, NT, NTI, NTIS, NTISF*. A 0-paraméterek az alábbiak:

$$\lambda_{NT}^{NT},\ \lambda_{NI*}^{NTI},\ \lambda_{NF*}^{NTISF},$$

míg a modellhez tartozó eloszlást paraméterező szabad paraméterek:

$$\lambda_{*}^{N},\ \lambda_{T}^{NT},\ \lambda_{I}^{NTI},\ \lambda_{TI}^{NTI},\ \lambda_{S*}^{NTIS},$$
$$\lambda_{F*}^{NTISF},\ N \notin *.$$

A 2. modell láncgráf modell, amiknek a paraméterezése általánosan még nem megoldott, de egyedi esetekben, mint ez a mostani is, adható jó paraméterezés. Az alábbi paraméterezés rendezett módon dekomponálható, hierarchikus és teljes paraméterezés, tehát minden, a 2.3. fejezetben említett előnyös tulajdonsággal bír. A felhasznált marginálisok: *NT, NI, NTI, NTIS, NTISF*. A null-paraméterek:

$$\lambda_{NT}^{NT},\ \lambda_{NI}^{NI},\ \lambda_{NF*}^{NTISF},$$

míg a modellhez tartozó eloszlást paraméterező szabad paraméterek:

$$\lambda_{\varnothing}^{NT},\ \lambda_{T}^{NT},\ \lambda_{N}^{NT},\ \lambda_{I}^{NI},\ \lambda_{TI}^{NTI},\ \lambda_{NTI}^{NTI},\ \lambda_{S*}^{NTIS},$$
$$\lambda_{F*}^{NTISF},\ N \notin *.$$

Vizsgáljuk meg végül azt is, valóban szükséges-e interakciókat feltételezni az első modell hatásai között, azaz vezessünk be útmodellt az 1. modellen! Ez egyszerűen a háromnál nagyobb számosságú hatáshalmazt tartalmazó szabad paraméterek nullára állítását jelenti. Legyen ez a 3. modell. A modell értelmezhető úgy (mint azt Goodman módosított útmodelljénél láttuk), hogy a nyilak hegyénél levő változók (szülői biztatás ill. felsőfokú tervek), mint függő változók alkotta regressziós modellben a független (a nyilak kiindulásánál levő) változók között nincsen interakció.

A modellek illeszkedését valószínűség-hányados próbával teszteljük. A 2. és az 1., ill. az 1. és a 3. modell közvetlenül összehasonlítható, mivel egymásbaágyazott modellek. Az eredményeket az alábbi táblázat mutatja:

11. táblázat. Az illeszkedéstesztek eredményei

Modell	szabadságfok	L^2	p
1. (irányított körmentes)	47	85,8	0.000
2. (láncgráf)	38	74,3	0.000
1. modell | 2. modell	9	11,5	0.243
3. (útmodell)	95	165,6	0.000
3. modell | 1. modell	48	79,8	0.003

Mivel az 1. modell a 2.-hoz képest nem rontja szignifikánsan az illeszkedést, a *T* és *I* közötti nyíl a vonalnál jobb megoldásnak mutatkozik. Az útmodell az 1. modellhez képest szignifikánsan rosszabbul illeszkedik, ezért szintén nem fogadható el. Az 1. modell tűnik a leginkább elfogadhatónak a három közül, de, mint láthatjuk, az ő illeszkedése is rossz. Tudjuk, hogy ezt a modellt Spirtes et al. (bizonyos kézenfekvő előfeltételek mellett) a legjobb IKG modellként választotta ki, tehát az is elmondható, hogy IKG modellel ezek az adatok nem írhatók le jól.

7.4 Szerhasználat amerikai fiatalok körében

Ennek a fejezetnek a módszertani fontosságát az adja, hogy a nem-egymásbaágyazott modellek összevetésére alkalmas BIC mutató használatára mutatok majd példát.

Az alábbi táblázat (Agresti, 2002, pp. 361) egy 1992-ben, az Ohio államelő Dayton városának Wright Egyetemén végzett felmérés eredményét mutatja be. Az adatok a nem (*G*) és rassz (*R*) szerint kategorizált végzős hallgatók egy mintájának szerhasználati szokásaira vonatkoznak. A szerek között az alkohol (*A*), cigaretta (*C*) és marihuána (*M*) szerepel. Agresti (2002, pp. 362) irányítatlan gráffal modellezi az adatokat, azonban, mivel egyirányú kapcsolatok is feltehetők, a következőkben én irányított éleket is használok. Az alábbiakban az adatstruktúrát potenciálisan jól leíró modelleket veszem sorra, majd az adatokhoz való illeszkedésük alapján összehasonlítom őket. Végül a legjobban illeszkedő modell mellett kapott paraméterbecsléseket is megadom.

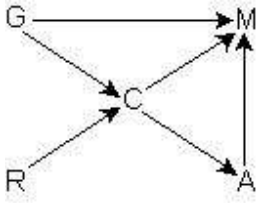
12. táblázat. A daytoni adatok, n=2276

		Marihána használat							
		Rassz=Fehér				*Rassz=Egyéb*			
		Nő		*Férfi*		*Nő*		*Férfi*	
Alkohol	*Cigaretta*	igen	nem	igen	nem	igen	nem	igen	nem
igen	igen	405	268	453	228	23	23	30	19
	nem	13	218	28	201	2	19	1	18
nem	igen	1	17	1	17	0	1	1	8
	nem	1	117	1	133	0	12	0	17

Egy lehetséges irányított körmentes gráfmodellt mutat be a 25. ábra (1. modell). Az egyes szerek fogyasztását azok "súlyossági foka" szerint rendezi, mintegy feltételezve, hogy a fogyasztó a könnyebbektől halad a súlyosabbak felé. A demográfiai változók (nem és kor) az

alkoholfogyasztást közvetlenül nem befolyásolják, csak a cigarettafogyasztáson keresztül. Hasonlóan, nincs közvetlen összefüggés a rassz és a marihuána-fogyasztás között.

25. ábra. Szerhasználat, 1. modell



Könnyű látni, hogy a gráfhoz a különböző Markov-tulajdonságok ugyanazt a modellt definiálják. Válasszuk ezért pl. a lokális tulajdonságot, ami szerint

$$G \perp R,\ A \perp RG \mid C,\ M \perp R \mid CAG.$$

A paraméterezés az alábbi marginálisok mellett hierarchikus és rendezett módon dekomponálható lesz: *R*, *RG, RGC, RGCA, RGCAM*, tehát minden, a 2.3. fejezetben említett előnyös tulajdonsággal bírni fog. A két feltételes függetlenségnek megfelelő null-paraméterek:

$$\lambda_{RG}^{RG},\ \lambda_{RA*}^{RGCA},\ \lambda_{GA*}^{RGCA},\ \lambda_{RM*}^{RGCAM},$$

míg a szabad paraméterek:

$$\lambda_{*}^{R},\ \lambda_{G}^{RG},\ \lambda_{C*}^{RGC},\ \lambda_{AC}^{RGCA},\ \lambda_{A}^{RGCA},$$
$$\lambda_{M*}^{RGCAM},\ R \notin *.$$

Egy másik szóba jöhető modellként (2. modell) definiáljunk útmodellt az 1. modellhez, elhagyva a modell szabad paraméterei közül minden legalább harmadfokú hatást. Ezzel az alábbi paraméterek értékét írjuk elő 0-ként:

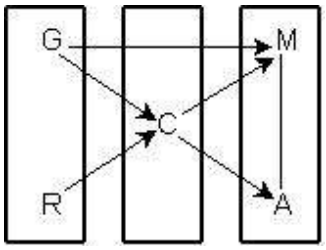
$$\lambda_{RGC}^{RGC},\ \lambda_{CAM}^{RGCAM},\ \lambda_{GAM}^{RGCAM},\ \lambda_{GCM}^{RGCAM},\ \lambda_{GCAM}^{RGCAM}.$$

A modell értelmezhető úgy (lásd Goodman módosított útmodelljét), hogy a nyilak hegyénél levő változók (pl. a marihuána), mint függő változók alkotta regressziós modellben a független (a nyilak kiindulásánál levő) változók között nincsen interakció.

A 3. modell (26. ábra) láncgráf-modell, az 1. modellhez képest azzal a változtatással, hogy irányítatlan kapcsolatot tételez fel az *M* és az *A* között, arra alapozva, hogy a marihuána fogyasztása is befolyásolhatja az alkoholfogyasztást. Itt az LWF és AMP Markov-tulajdonságok már más modellt definiálnak. Válasszuk az AMP tulajdonságot. Az adott gráf esetén az egyébként nem ekvivalens blokk-rekurzív és lokális Markov-tulajdonságok ugyanazt adják:

$$G \perp R,\ A \perp RG \mid C,\ M \perp R \mid CG.$$

Látható, hogy a modell az 1. modelltől csak az utolsó állítás feltételhalmazában különbözik.



26. ábra. Szerhasználat, 3. modell

A megfelelő marginálisokat a 4.6. fejezet végén leírt módon határozzuk meg: *RG, RGC, RGCM, RGCA, RGCAM*. A kötött paraméterek a következők:

$$\lambda_{RG}^{RG},\ \lambda_{RA*}^{RGCA},\ \lambda_{GA*}^{RGCA},\ \lambda_{RM*}^{RGCM},$$

míg a szabad paraméterek ezek:

$$\lambda_{\varnothing}^{RG},\ \lambda_{R}^{RG},\ \lambda_{G}^{RG},\ \lambda_{C*}^{RGC},\ \lambda_{AC}^{RGCA},\ \lambda_{A}^{RGCA},$$
$$\lambda_{M*}^{RGCM},\ R \notin *,$$
$$\lambda_{MA*}^{RGCAM}.$$

Könnyen belátható, hogy az utóbbi láncgráf-modell útmodell párja megegyezik a 2. modellel (ugyanazokat a kötött paramétereket tartalmazzák), így ezt külön nem kell definiálnunk: a 2. modellt a 3. modell útmodelljeként is tekinthetjük.

A kötött paraméterekből is látható, hogy az így definiált modellek páronként egymásbaágyazottak, az alábbi tartalmazási relációkkal:

$$1.\ \text{modell} \supseteq 2.\ \text{modell}.$$

és

$$3.\ \text{modell} \supseteq 2.\ \text{modell}.$$

Mint a 4.8. fejezetben láttuk, a valószínűség-hányados statisztika (L^2) alkalmas eszköz az egymásbaágyazott modellek összevetésére. Azonban pl. az 1. és a 3. modell, mivel nem egymásbaágyazottak, nem hasonlíthatók össze ezen a módon. Nem-egymásbaágyazott modellek összevetésére a Bayes-féle Információs Kritériumot (*Bayesian Information Criterion,* BIC) használják gyakran. A BIC képlete a következő:

$$BIC = L^2 - szabadságfok \times \ln(n),$$

ahol az n a mintanagyság. Két modell közül mindig az alacsonyabb BIC értékkel rendelkező preferálandó. Látható, hogy a nagy L^2 értékkel rendelkező modellekhez a kötött paraméterek számával arányosan csökkentett BIC-érték tartozik, vagyis a kritérium a sok kötött paraméterrel rendelkező, takarékos modelleket jutalmazza.

13. táblázat. Az illeszkedéstesztek eredményei

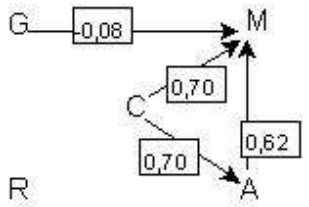
Modell	szabadságfok	L^2	p	BIC
1. (irányított gráf)	15	28,0	0,021	-87,9
2. (irányított útmodell)	20	31,3	0,051	-123,3
2. modell | 1. modell	5	3,3	0,654	-35,4
3. (láncgráf)	11	26,4	0,006	-58,6
2. modell | 3. modell	9	4,9	0,843	-64,7

Ha a 2. modellt az L^2 segítségével összevetjük az 1. modellel, látható, hogy az illeszkedésromlás nem szignifikáns. Ugyanez igaz a 4. és a 2. modell összehasonlítására is. Vagyis a gráfok útmodellé alakítása ésszerűnek tűnik.

A BIC alapján a 2. modell adja a legjobb modellt, és a valószínűséghányados próba szerint is egyedül ennek a modellnek az illeszkedése elfogadható.

Tehát az adatokhoz egy útmodell illeszkedik a legjobban. Tovább is mehetünk, az útmodell egyes nyilaihoz tartozó hatásokról egyenként eldöntve, hogy azok elhagyhatók-e. Un. visszalépegető elhagyást (*backward elimination)* használtam. Az első lépésben a 2. modellhez képest a λ_{GC}^{RGC} paraméter elhagyása okozta a legkisebb illeszkedésromlást (ΔL^2=0,03). Ennek elhagyása után a λ_{RC}^{RGC} paraméter hagyható el (ΔL^2=2,8). A további elhagyások már szignifikáns illeszkedésromlással járnak.

A végső modellhez tartozó szabad paraméterek: $\lambda_{CA}^{RGCA} = \lambda_{CA}^{CA}$, λ_{GM}^{GCAM}, λ_{CM}^{GCAM}, λ_{AM}^{GCAM}. Itt az első paraméter, mint látható, az *RGCA* helyett értelmezhető a *CA* marginális mellett is, a 4.5. fejezetbeli (60) tétel alapján (ott már említettük, hogy a tétel előnye éppen ez: a szülő-gyerek hatások feltételhalmazából elhagyhatjuk azokat a csúcsokat, amelyek a gyereknek nem szülei). Tehát a szabad paraméterek a többszörös regressziós paraméterekhez hasonlóak: a többi szülőre korrigált szülő-gyerek hatások nagyságát adják. Az alábbi ábra mutatja be a szabad paraméterek becsléseit. Az ábrán csak a nem-redundáns paramétereket tüntettem fel. Mint látható, a nemnek csak gyenge hatása van a marihuána fogyasztására. A paraméter értéke negatív, ami azt jelzi, hogy inkább a férfiak hajlamosak a marihuána fogyasztásra. A cigarettázásnak a marihuána-fogyasztásra gyakorolt direkt és indirekt (alkohol-fogyasztáson keresztül érvényesülő) hatása is erősnek tűnik, és ugyancsak erős az alkohol-marihuána kapcsolat is.



27. ábra. Szerhasználat, a legegyszerűbb modell, paraméterbecslések.

7.5 A demokráciával ill. az életszínvonallal való elégedettség kapcsolata, az MHP három hulláma alapján

Az alábbi fejezet fontosságát az adja, hogy a marginális loglineáris modellek paneladatok elemzésére való alkalmasságát szemlélteti. Olyan modellek paraméterezésére is látunk példát, amelyek nem grafikus modellek, de feltételes függetlenségi állításokkal definiálhatók, így a disszertációban bemutatok keretek között elemezhetők.

A panelvizsgálat több időpontban (hullámban) ugyanazon a mintán, ugyanazzal a tematikával megismételt felmérések sorozata. A szokásos keresztmetszeti vizsgálatokkal szembeni előnye, hogy a változási folyamatok ténylegesen követhetők, nem csak valószínűsíthetők. Pl. politikai preferenciavizsgálatok esetén az *X* párt szavazótáborának két különböző hónapban mért csökkenése keresztmetszeti vizsgálat esetén létrejöhet egyszerűen a véletlen folytán is - ha a mintába kevesebb *X*-párti került beválogatásra. Ezzel szemben panelvizsgálatnál, mivel ugyanazokat a személyeket kérdezzük minden hónapban, a szavazótábor csökkenése tényleges népszerűségvesztést jelent (a mintán belül biztosan, a populáción belül persze csak valószínűsíthetően). Ha viszont a keresztmetszeti vizsgálatban retrospektív (a múltra visszaemlékező) adatokat próbálunk gyűjteni, a kérdezettek a múltbeli pártpreferenciáikra nem emlékszenek jól, különösen, ha az nem is létezett. Ugyanakkor a panelvizsgálatok rendkívül költségesek, munkaigényesek és a minta folyamatos lemorzsolódásával is számolni kell. Az (egyébként az utóbbi okok miatt nagyon ritka) magyarországi szociológiai panelvizsgálatok közül elsősorban a TÁRKI Magyar Háztartás Panelje érdemel említést tematikája és mintaelemszáma folytán.

Rudas és Bergsma (2004) említi a marginális loglineáris modellek alkalmazási lehetőségei között a panelvizsgálatok példáját. Az alábbiakban grafikus modellel megfogható panelvizsgálati hipotézisekre is mutatok példát, ennek megfelelően a disszertáció korábbi fejezeteiben részletezett módszertan közvetlenül alkalmazható lesz. A grafikus modellekre történő visszavezetés lényege, hogy a különböző (t_i) időpontokban mért ugyanazon *V* változót időpontonként külön változóként (V_i-k) kezelem. Ez azért lehetséges, mert panelvizsgálat lévén ismert a V_i-k együttes eloszlása. Több időpontban ismételt keresztmetszeti vizsgálat nem adna erre lehetőséget, ott az együttes eloszlás ismeretlen, csak a V_i marginális eloszlások adottak. Az alábbiakban olyan modellek is szerepelnek majd, amelyek bár gráf-modellek továbbspecifikálásával adódnak, nem fejezhetők ki közvetlenül gráf-modellel. Ugyanakkor feltételes függetlenségi állításokkal adhatók meg, ezért paraméterezésük kézenfekvő módon, az IKG modelleknél látott módszerből kiindulva elvégezhető. Ilyen, nem-gráf modellt definiál az a gyakori panelvizsgálati hipotézis is, hogy a folyamat memóriája véges és csak a megelőző *k* hullámra emlékszik. Ilyenkor azt mondjuk, hogy k-rendű Markov-láncot tételezünk fel, miszerint a t_i időpontban mért változó(k) feltételes eloszlása csak a megelőző *k* időpontban mért értékektől függ. Ilyen hipotézisre is mutatok példát az alábbiakban. A modelleknek megfelelő eloszlás paraméterezésekor utalni fogok a „jó" paraméterezéssel szemben támasztott követelményekre (hierarchikusság, rendezett módon való dekomponálhatóság), látunk példát arra is, amikor ezek a követelmények nem teljesülnek. A Magyar Háztartási Panel 1993-as, 1995-ös és 1997-es hullámát használom fel, a minta azon részét kizárva, akiknél valamelyik vizsgált kérdésre valamelyik hullámban hiányzó válasz volt. A vizsgált változók a következők:

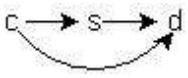
d: a demokráciával, konkrétan a politikai beleszólás lehetőségével való elégedettség, bináris változóként (0: elégedett, 1: nem elégedett)

s: a saját életszínvonallal való elégedettség, bináris változóként (0: elégedett, 1: nem elégedett)

c: kohorsz-hovatartozás, három kategóriában, az első hullámban mért terciliseekkel (1: idősek, 2: középkorúak, 3: fiatalok). Vagyis nem életkorcsoportot használok, így ennek a változónak az értéke nem változik hullámonként.

A változók kiválasztásának oka, hogy a politológusok a kelet-európai országokkal kapcsolatban a saját élethelyzettel való elégedettség és a demokráciapárti attitűd között általában szoros kapcsolatot tételeznek fel. Mivel a demokráciába/piacgazdaságba történő átmenet az egyes korosztályokat gazdaságilag másként érintette, vélhetően az átmenet megítélése is generációnként képet mutat.

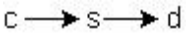
Összesen hét változót definiáltam: $d_1, d_2, d_3, s_1, s_2, s_3, c,$ az alsóindexben jelölve az 1-es, 2-es, 3-as hullámokat. Az együttes eloszlás kontingencia-táblája 192 cellás, 18 üres cellával. Az üres cellák megszűntetése céljával egy kis konstanst (0,5) adtam minden cellagyakorisághoz. Az alábbiakban a 28. ábra irányított körmentes gráfjából indulunk ki, vagyis az életszínvonallal való elégedettséget a demokrácia megítélése elé helyezzük kialakulási sorrendjét tekintve, és az ábra nyilaihoz rendelhető hatásokra teszünk megkötéseket.



28. ábra. Panelvizsgálat, telített modell

1. modell: látszólagos függőség

Az első vizsgált modell az 18. ábra látott első gráffal (nem - képzés típusa -elhelyezkedés sikeressége) bemutatott *látszólagos függőség* egy esete. A modell ezt a látszólagos függőséget teszteli a demokráciával való elégedettség és a kohorsz-hovatartozás kapcsolatában, egyformán mindhárom hullámban. A feltevés szerint a demokráciával való elégedettség generációs különbségei csak amiatt állnak elő, hogy az idősebbek elégedetlenebbek saját életszínvonalukkal, és az életszínvonallal való elégedetlenség beárnyékolja a demokrácia megélését is. A hipotézis a 28. ábra *c-d* nyilának elhagyásával kapható, és az alábbi irányított körmentes gráffal szemléltethető:



A modell az irányított páronkénti Markov-tulajdonságot[54] használva hullámonként így formalizálható:

$$d_1 \perp c \mid s_1,$$

$$d_2 \perp c \mid s_2,$$

$$d_3 \perp c \mid s_3.$$

[54] Ez pozitív eloszlás mellett ekvivalens a többi irányított Markov-tulajdonsággal, tehát nem tettem valódi megszorítást a páronkénti tulajdonság választásával.

A modellnek megfelelő eloszlás felparaméterezésekor használandó marginálisok az IKG-k paraméterezésénél látottak alapján:

$$c,\ cs_1,\ cs_1d_1,\ cs_2,\ cs_2d_2, cs_3,\ cs_3d_3,\ cs_1s_2s_3d_1d_2d_3 \tag{72}$$

Ezekkel a marginálisokkal könnyen adható hierarchikus, rendezett módon dekomponálható és teljes paraméterezés. Az 1. tétel alapján a következő paraméterek értéke 0:

$$\lambda_{cd_1}^{cs_1d_1}, \lambda_{cs_1d_1}^{cs_1d_1}, \lambda_{cd_2}^{cs_2d_2}, \lambda_{cs_2d_2}^{cs_2d_2}, \lambda_{cd_3}^{cs_3d_3}, \lambda_{cs_3d_3}^{cs_3d_3}.$$

A modell szabadságfoka 12, mert 6 paraméterre teszünk megkötést, és a hozzájuk tartozó hatás mindegyik esetben egyetlen háromértékű változót tartalmaz a binárisokon kívül, így mindegyikhez két nem-redundáns paraméterérték tartozik. A valószínűség-hányados tesztstatisztika szerint a modell illeszkedése rossz (L^2=27,07, p=0,008). Vagyis a demokráciához való viszony generációs különbségei valószínűsíthetően nem látszólagosak, nem magyarázhatóak az életszínvonal közbejövő hatásával.

2. modell: időben állandó *c-s* hatás

A második modell szerint az életszínvonallal való elégedettség generációs különbségei nem változtak ’93 és ’97 között, vagyis konstans a (28. ábra *c-s* nyilához rendelhető) marginális asszociáció. A (72) paraméterezést használhatjuk itt is, a modell formálisan az alábbi egyenlőséggel fejezhető ki:

$$\lambda_{cs_1}^{cs_1} = \lambda_{cs_2}^{cs_2} = \lambda_{cs_3}^{cs_3}.$$

Ezek a megszorítások nem definiálnak grafikus modellt, de, mivel az alábbi két lineáris megszorítással azonosíthatók:

$$\lambda_{cs_1}^{cs_1} - \lambda_{cs_2}^{cs_2} = 0,\ \lambda_{cs_2}^{cs_2} - \lambda_{cs_3}^{cs_3} = 0,$$

a grafikus modellekhez hasonlóan 0-paraméterekkel definiálhatók, így a modell a 4.8. fejezetben bemutatott eljárással becsülhető. A modell szabadságfoka 4, mert a két megkötéshez itt is két független paraméterérték tartozik. A valószínűség-hányados tesztstatisztika szerint a modell illeszkedése jó (L^2= 2,34, p=0,674). Ezek szerint elfogadható, hogy a három hullám során állandó erősségű az életszínvonal megítélése és a kohorsz-hovatartozás közötti kapcsolat. Specifikáljuk tovább a modellt!

3. modell: időben állandó *s-d* hatás

Vizsgáljuk meg a 28. ábra *s-d* nyilához rendelhető hatás időbeli állandóságát is! Most a 2-es modell továbbspecifikálásával, az életszínvonal ill. a demokrácia megítélése közötti, életkorra korrigált kapcsolat változatlanságának hipotézisével állítjuk fel modellünket. A paraméterezés itt is a (72) paraméterezéssel oldható meg, a modell a 2. modellhez tartozó megszorításokat az alábbiakkal kiegészítve adható meg:

$$\lambda_{s_1 d_1}^{d_1 c s_1} = \lambda_{s_2 d_2}^{d_2 c s_2} = \lambda_{s_3 d_3}^{d_3 c s_3},$$

azaz:

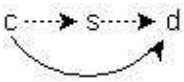
$$\lambda_{s_1 d_1}^{d_1 c s_1} - \lambda_{s_2 d_2}^{d_2 c s_2} = 0, \lambda_{s_2 d_2}^{d_2 c s_2} - \lambda_{s_3 d_3}^{d_3 c s_3} = 0.$$

A modell szabadságfoka 6, hiszen a 2. modellhez 4 szabadságfok tartozott, amihez még 2 szabadságfok jött a két új (csak bináris változókat tartalmazó hatásokkal definiált paraméterekre vonatkozó) megszorításból. A valószínűség-hányados tesztstatisztika szerint a modell illeszkedése jó (L^2=2,40, p=0,880), sőt a 2. modellhez képest csak egészen kismértékű illeszkedésromlást okozott az újabb megkötés:

L^2(3. model3| 2. modell)= 0,07, p(3. modell|2. modell)=0,968.

A 2. és 3. modellek egymásbaágyazott volta adott lehetőséget összehasonlításukkor a valószínűség-hányados próba használatára.

Tehát a *c-s* és az *s-d* nyilakhoz tartozó asszociációk konstans voltának feltételezése elfogadható. Az eddigiek szerint az alábbi ábra szaggatott nyilaihoz tartozó hatások időben állandó erősségűek:



Vizsgáljuk most meg a harmadik nyílhoz tartozó hatás időben állandó voltát!

4. modell: időben állandó *c-d* hatás

A 3. modell továbbspecifikálása a *c* és *d* közötti feltételes hatás konstans voltával. A paraméterezés ugyanaz mint (72)-ben, a modell a 3. modellhez képest az alábbi paraméter-megkötéseket is tartalmazza:

$$\lambda_{c_1 d_1}^{d_1 c s_1} = \lambda_{c_2 d_2}^{d_2 c s_2} = \lambda_{c_3 d_3}^{d_3 c s_3},$$

azaz:

$$\lambda_{c_1 d_1}^{d_1 c s_1} - \lambda_{c_2 d_2}^{d_2 c s_2} = 0, \lambda_{c_2 d_2}^{d_2 c s_2} - \lambda_{c_3 d_3}^{d_3 c s_3} = 0.$$

A modell szabadságfoka 10 (a 3. modellé 6 volt, ehhez jön még a két megszorítás, amelyek háromértékű és bináris változót tartalmazó hatásokra vonatkoznak). A valószínűség-hányados tesztstatisztika szerint a modell illeszkedése jó (L^2=8,93, p=0,539) és a 3. modellhez képest nem szignifikáns mértékű az illeszkedésromlás:

L^2(4. modell | 3. modell)=6,52, p(4. modell | 3. modell)=0,163.

Vagyis a harmadik hatás konstans volta is elfogadhatónak tűnik.

Ezek szerint a 28. ábra összes nyilához tartozó hatás időben állandónak tételezhető fel. Becsüljük meg a modellhez tartozó, kétváltozós hatásokhoz kapcsolódó szabad paramétereket!

$$\lambda_{cs_i}^{cs_i}(1,0)=-0{,}009,$$

$$\lambda_{cs_i}^{cs_i}(2,0)=-0{,}166.$$

Ezekből következik, hogy a c_i=3-hoz kapcsolódó paraméterérték 0,175, hiszen a c_i=1, c_i=2 és c_i=3-hoz tartozó paraméterérték összege 0 kell, hogy legyen (s_i bármely rögzített értéke esetén). Mindhárom hullámban a fiatalok hajlamosabbak leginkább életszínvonaluk pozitív megítélésére, míg a középkorúak a legelégedetlenebbek. A második hatás becslése:

$$\lambda_{s_i d_i}^{d_i cs_i}(0,0)=0{,}210,$$

vagyis a két attitűdváltozó között (az előzőnél kissé erősebb) pozitív kapcsolat van: akik elégedettek az életszínvonalukkal, azok nagyobb eséllyel elégedettek a demokráciával is. Végül a harmadik hatás becslése:

$$\lambda_{c_i d_i}^{d_i cs_i}(1,0)=-0{,}019,$$

$$\lambda_{c_i d_i}^{d_i cs_i}(2,0)=-0{,}041.$$

A c_i=3-hoz tartozó paraméterérték ismét könnyen adódik, értéke 0,060. Eszerint a fiatalok a legelégedettebbek a demokráciával, a középkorúak a legkevésbé, de ezek a generációs különbségek kisebbek, mint az életszínvonallal való elégedettség esetében.

Építsük tovább a modellt, most egy Markov-tulajdonság bevonásával!

5. modell: 1-rendű Markov-lánc

A modell a 4. modell megszorítása, annak feltételezésével, hogy az életszínvonallal való elégedettség 1-rendű Markov-lánc, memóriája csak a megelőző állapotra terjed ki:

$$s_3 \perp s_1 \mid s_2, \tag{73}$$

vagyis (az 1. tétel szerint)

$$0=\lambda_{s_3 s_1}^{s_3 s_2 s_1}=\lambda_{s_3 s_2 s_1}^{s_3 s_2 s_1}.$$

Hogy ez a paraméter a paraméterezésnek része legyen, a modellnek megfelelő eloszlás felparaméterezésekor használandó marginálisokat az $s_1s_2s_3$ marginálissal kell bővíteni. Viszont könnyen belátható, hogy a (72)-ban látott marginális-sorba bárhova szúrjuk is be ezt a marginálist, az eredmény nem lesz rendezett módon dekomponálható. Például az alábbi módon:

$$c,\ cs_1,\ cs_1d_1,\ cs_2,\ cs_2d_2,\ cs_3,\ cs_3d_3,\ s_1s_2s_3,\ cs_1s_2s_3d_1d_2d_3.$$

az $s_1s_2s_3$ metszete a korábbi maximális elemekkel $s_1s_2s_3$, ez a halmaz viszont egyik korábbi halmaznak sem részhalmaza. Nehéz kérdés, hogy a (72) marginálisaival vagy azok kibővítése által hogyan definiálható a (73)-as feltételes függetlenség. Nem jelentene ilyen problémát pl. a legbővebb feltétellel adott függetlenség megadása:

$$s_3 \perp s_1 \mid s_2 c d_1 d_2 d_3,$$

mert ehhez nem kellene az eredeti marginálisokat bővíteni

Ezért most önmagában a (73)-mal adott Markov-lánc tesztelését végezzük el. Jegyezzük itt meg, hogy a modell klasszikus loglineáris modell, hiszen egyetlen marginális ($s_1s_2s_3$) mellett definiálható. A modell szabadságfoka 2 (két null-paraméterünk van, bináris változókat tartalmazó hatásokkal). A valószínűség-hányados tesztstatisztika szerint a modell illeszkedése rossz (L^2=71,94, p=0,000).

Eszerint az életszínvonallal való elégedettséget nem csak a legutóbbi elégedettségi szint, hanem az az előtti is befolyásolja, vagyis a folyamatnak legalább két lépés hosszú memóriája van. Vizsgáljunk meg most egy hasonló kérdést a demokráciával való elégedettséggel kapcsolatban is!

6. modell: *d*-nek nincs kapcsolata a korábbi *s*-ekkel

Láttuk, hogy a demokráciával való elégedettség pozitív kapcsolatban van az életszínvonal azonos hullámban mért értékelésével. Nézzük meg, van-e szignifikáns befolyása rá a életszínvonal korábbi hullámokban mért értékelésének! A 6. modell[55] ennek megfelelően az alábbi feltételes függetlenségi állításokkal állítható elő:

$$d_3 \perp s_1\, s_2 \mid s_3,$$

$$d_2 \perp s_1 \mid s_2.$$

A paraméterezés hierarchikus és rendezett módon dekomponálható lesz az alábbi marginálisokkal:

$$s_1s_2d_2,\ s_1s_2s_3d_3,\ s_1s_2s_3d_2d_3.$$

A paraméterek közül az alábbiak értékét állítjuk 0-ra:

$$0 = \lambda_{d_3s_1}^{d_3s_1s_2s_3} = \lambda_{d_3s_2}^{d_3s_1s_2s_3} = \lambda_{d_3s_1s_2}^{d_3s_1s_2s_3} = \lambda_{d_3s_1s_2s_3}^{d_3s_1s_2s_3} = \lambda_{d_3s_1s_3}^{d_3s_1s_2s_3} = \lambda_{d_3s_2s_3}^{d_3s_1s_2s_3} = \lambda_{d_2s_1s_2}^{d_2s_1s_2} = \lambda_{d_2s_1s_2}^{d_2s_1s_2}.$$

A modell szabadságfoka 8. A valószínűség-hányados tesztstatisztika szerint a modell illeszkedése jó (L^2=8,00, p=0,434). Vagyis elfogadhatónak tűnik, hogy a demokráciával való elégedettség csak az egyidejű életszínvonal-értékeléstől függ. Szorítsuk meg most tovább a modellt az életszínvonal megítélésének kialakulására vonatkozóan!

7. modell: *s*-nek csak saját múltján keresztül van kapcsolata a korábbi *d*-ekkel

Eddig feltettük, hogy az életszínvonallal való elégedettség kialakulása megelőzi a demokrácia megítélését. Ugyanakkor ez a sorrend nem nyilvánvaló, hiszen a demokráciában való csalódás is

[55] Az 5. modellhez hasonlóan a 6. modellnek sem paraméterezhető kézenfekvő módon a korábbi, 4. modellel vett metszete (a marginálisok uniója nem rendezhető dekomponálható módon).

okozhat olyan pesszimista beállítódást, ami akár a saját helyzet megítélését is ronthatja. Az előbbi hipotézis tesztelése az alábbi modellel történhet[56]:

$$s_3 \perp d_1\, d_2 \mid s_1\, s_2,$$

$$s_2 \perp d_1 \mid s_1,$$

vagyis

$$0 = \lambda_{d_1 s_3}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_1 d_2 s_3}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_2 s_3}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_1 s_3 s_1}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_1 d_2 s_3 s_1}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_2 s_3 s_1}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_1 s_3 s_2}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_1 d_2 s_3 s_2}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_2 s_3 s_2}^{d_1 d_2 s_1 s_2 s_3} =$$
$$= \lambda_{d_1 s_3 s_1 s_2}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_1 d_2 s_3 s_1 s_2}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_2 s_3 s_1 s_2}^{d_1 d_2 s_1 s_2 s_3} = \lambda_{d_1 s_2}^{d_1 s_1 s_2} = \lambda_{d_1 s_1 s_2}^{d_1 s_1 s_2}.$$

A modell szabadságfoka 22 (14 új null-paraméter jött a 6. modellben adott 8-hoz). A valószínűség-hányados tesztstatisztika szerint a modell illeszkedése rossz (L^R=58,04, p=0.000), tehát a korábbi demokrácia-megítéléstől nem független a jelenlegi életszínvonal-megítélés, még a korábbi életszínvonal-megítéléseket figyelembe véve sem. Adjunk becslést ennek a kapcsolatnak az erősségére, a 6. modellnek megfelelő eloszlás alapján!

$$\lambda_{d_1 s_2}^{d_1 s_1 s_2}(0,0) = 0{,}116,$$

$$\lambda_{d_1 s_3}^{d_1 d_2 s_1 s_2 s_3}(0,0) = 0{,}06,$$

$$\lambda_{d_2 s_3}^{d_1 d_2 s_1 s_2 s_3}(0,0) = 0{,}09,$$

eszerint az életszínvonal megítélésére pozitív hatása van a demokráciával való korábbi elégedettségnek, de ez az emlék felejtődik: a korábban mért elégedettségnek (d_1) kisebb hatása van, mint a későbbinek (d_2).

[56] Az 5. és 6. modellhez hasonlóan a 7. modellnek sem paraméterezhető kézenfekvő módon a korábbi, 4. modellel vett metszete (a marginálisok uniója nem rendezhető dekomponálható módon).

8 FÜGGELÉK: SZÁMÍTÁSTECHNIKAI MEGVALÓSÍTÁS

Az alábbiakban a 7.3 fejezetben bemutatott, Sewell és Shah továbbtanulási adataira definiált 1. modell programnyelvi megvalósítása következik. A modell változói a felsőfokú továbbtanulás (*C*, mint *college plans*) és annak meghatározó tényezői, mint a nem (*G, gender*), intelligencia hányados (*I*), szülői bíztatás (*P, parental encouragement*), és társadalmi háttér (*S, social background*).

8.1 Modellbecslés

A modelleknek megfelelő eloszlás maximum-likelihood becslése Mathematica-ban íródott, Wicher Bergsma MLE függvényének[57] felhasználásával. A program az alábbi egyenlet kifejezéseit használja:

$$X = ct[0].Log\left(at[0].Mt.n\right),$$

ahol . a mátrixszorzás, *X* a fix paraméterek rögzített értékét tartalmazó, szabadságfok hosszú vektor (ez grafikus modellek esetén, mivel 0-ban rögzítünk paramétereket, mindig nullvektor). Az *n* a megfigyelt gyakoriságok vektora, ebből az *Mt* mátrixszal való szorzással kapjuk a megfigyelt marginálisokat. Az *at[0]* grafikus modellek esetén mindig identitásmátrix (Bergsma függvénye a grafikus modelleknél általánosabb marginális modellekre íródott, ahol ez a mátrix más is lehet). A *ct[0]* kontrasztmátrixszal történő szorzás adja meg a kívánt paraméterértékeket a marginálisok logaritmusának lineáris kombinációjaként. A program inputként ezt az öt kifejezést várja. A program outputja *est*, a becsült gyakoriságok vektora.

A 7.3 fejezet 1. modelljét definiáló feltételes függetlenségek az alábbi nullparaméterek rögzítését kívánják meg:

$$\lambda_{GS}^{GS}, \lambda_{GI*}^{GSI}, \lambda_{GC*}^{GSIPC}.$$

S és *I* négyértékű, a többi bináris változó, ennek megfelelően három-három ill. egy-egy független paraméter tartozik hozzájuk.

*A MEGFIGYELT GYAKORISÁGOK MEGADÁSA

m={{4}, {349}, {13}, {64}, {9}, {207}, {33}, {72}, {12}, {126}, {38}, {54}, {10}, {67}, {49}, {43}, {2}, {232}, {27}, {84}, {7}, {201}, {64}, {95}, {12}, {115}, {93}, {92}, {17}, {79}, {119}, {59}, {8}, {166}, {47}, {91}, {6}, {120}, {74}, {110}, {17}, {92}, {148}, {100}, {6}, {42}, {198}, {73}, {4}, {48}, {39}, {57}, {5}, {47}, {123}, {90}, {9}, {41}, {224}, {65}, {8}, {17}, {414}, {54}, {5}, {454}, {9}, {44}, {5}, {312}, {14}, {47}, {8}, {216}, {20}, {35}, {13}, {96}, {28}, {24}, {11}, {285}, {29}, {61}, {19}, {236}, {47}, {88}, {12}, {164}, {62}, {85}, {15}, {113}, {72}, {50}, {7}, {163}, {36}, {72}, {13}, {193}, {75}, {90}, {12}, {174}, {91},

[57] A program elérhető az interneten:
http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html.print

```
{100}, {20}, {81}, {142}, {77}, {6}, {50}, {36}, {58}, {5}, {70}, {110}, {76}, {12}, {48},
{230}, {81}, {13}, {49}, {360}, {98}};

n=Flatten[m];

*Wicher Bergsma MLE függvénye

Clear[MLE]

MLE[n_, start_, g_, zeta_, Zt_, Mt_, X_] := Module
	[{MaxStepSize=1,
			MinStepSize=0.1,
			MaxError=1/10.^10,
			MaxIterations=100,
			dg, G, z, Ut, v, m, step, error, iterate},
	Ut=NullSpace[Transpose[X]];
	v[m_, step_]:= v[m, step]=Module[{mu, hm, Htm, Hm, lambda},
		mu=Mt.m;
		hm=Ut.g[zeta[mu]];
		Htm=(m*#1&) /@ (Ut.g']['zeta[mu]]*Zt[mu].Mt));
		Hm=Transpose[Htm];
		lambda=-Inverse[Htm.(Hm/m)].(Htm.(n/m-1)+hm);
		Log[m]+(step*(n-m+Hm.lambda))/m];
	error[m_] := (v[m, 1]-Log[m]).(m*(v[m, 1]-Log[m]));
	iterate[m_, step_:MaxStepSize] := Module[{newm},
		newm=Exp[v[m, step]];
		Print[N[step], "   ", error[newm]];
		Which[
			step < MinStepSize, Exp[v[m, MaxStepSize]],
			error[newm] > error[m], iterate[m, step/2],
			True, newm]];
	Print["stepsize, error :"];
	FixedPoint[iterate, start, MaxIterations, SameTest -> (error[#1] < MaxError & )]]
t[pi_, {a_, c_, 0}] := pi;
t[pi_, {a_, c_, i_}] :=
	Exp[c[i-1].Log[a[i-1].t[pi, {a, c, i-1}]]];
T[pi_, {a_, c_, 0}] := IdentityMatrix[Length[pi]];
T[pi_, {a_, c_, i_}] := t[pi, {a, c, i}]*
c[i-1].(a[i-1].T[pi, {a, c, i-1}]/
a[i-1].t[pi, {a, c, i-1}])
Null

Clear[ct, at, zeta1, Zt1]

<< LinearAlgebra`MatrixManipulation`;
```

*AZ Mt MÁTRIX DEFINIÁLÁSA

*1. marginális: *GSIPC* (128 cella)

```
M1=IdentityMatrix[128];
```

*2. marginális: *GSIP* (64 cella)

```
M2=ZeroMatrix[64, 128];
For[i=0, i < 64, i=i+1; M2[[i,2*i-1]]=1];
For[i=0, i < 64, i=i+1; M2[[i,2*i]]=1];
```

*3. marginális: *GSP* (16 cella)

```
M3=ZeroMatrix[16, 128];
Do[M3[[i,12*Floor[(i-1)/2]+(i-1)*2+Mod[Ceiling[Mod[j, 4, 1]/2]+
      Floor[(j-1)/4]*4, 16, 1]]]=1, {i, 1, 16}, {j, 1, 128}];
```

*4. marginális: *GSI* (32 cella)

```
M4=ZeroMatrix[32, 128];
Do[M4[[i,j]]=1, {i, 1, 32}, {j, (i-1)*4+1, i*4}];
```

*5. marginális: *GI* (8 cella)

```
M5=ZeroMatrix[8, 128];
Do[M5[[i,48*Floor[(i-1)/4]+(i-1)*4+Mod[Ceiling[Mod[j, 16, 1]/4]+
      Floor[(j-1)/16]*16, 64, 1]]]=1, {i, 1, 8}, {j, 1, 128}];
```

*6. marginális: *GS* (8 cella)

```
M6=ZeroMatrix[8, 128];
Do[M6[[i,j]]=1, {i, 1, 8}, {j, (i-1)*16+1, i*16}];
Mt=BlockMatrix[{{M1}, {M2}, {M3}, {M4}, {M5}, {M6}}];
```

*AZ *at[0]* MÁTRIX DEFINIÁLÁSA

```
at[0]=IdentityMatrix[256];
```

*A paraméterek megadása, külön minden marginálisra és külön minden, a paramétert definiáló hatás-halmaz számosságra. Itt példaként csak a *GSI* marginálishoz tartozó paraméterezést ismertetem.

(v4-gal jelöljük a paramétereket, a 4-es a 4. marginálishoz tartozást rövidíti. A * a paraméterhez tartozó változók *GSIPC* sorban kapott sorszámát ill. a változók értékeit jelöli, pl. v41131 az 1. (*G*) változó 1. értékéhez és a 3. (*I*) változó 1. értékéhez tartozik, azaz $\lambda_{GI}^{GSI}(1,1)$-hez.

```
data4=ZeroMatrix[3, 32];
Do[data4[[1,i]]=Ceiling[i/16], {i, 1, 32}];
Do[data4[[2,i]]=Ceiling[Mod[i, 16, 1]/4], {i, 1, 32}];
Do[data4[[3,i]]=Ceiling[Mod[i, 4, 1]], {i, 1, 32}];

empty4=ZeroMatrix[1, 32];
For[i=0, i < 32, i=i+1; empty4[[1,i]]=1];
empty4=empty4/VectorNorm[N[empty4[[1]]], 1];

p411=ZeroMatrix[1, 32];
For[i=0, i < 32, i=i+1; If[data4[[1,i]]==1, p411[[1,i]]=1]];
p411=p411/VectorNorm[N[p411[[1]]], 1];

p421=ZeroMatrix[1, 32];
For[i=0, i < 32, i=i+1; If[data4[[2,i]]==1, p421[[1,i]]=1]];
p421=p421/VectorNorm[N[p421[[1]]], 1];

p422=ZeroMatrix[1, 32];
For[i=0, i < 32, i=i+1; If[data4[[2,i]]==2, p422[[1,i]]=1]];
p422=p422/VectorNorm[N[p422[[1]]], 1]; p423=ZeroMatrix[1, 32];

For[i=0, i < 32, i=i+1; If[data4[[2,i]]==3, p423[[1,i]]=1]];
p423=p423/VectorNorm[N[p423[[1]]], 1];

p431=ZeroMatrix[1, 32];
For[i=0, i < 32, i=i+1; If[data4[[3,i]]==1, p431[[1,i]]=1]];
p431=p431/VectorNorm[N[p431[[1]]], 1];

p432=ZeroMatrix[1, 32];
For[i=0, i < 32, i=i+1; If[data4[[3,i]]==2, p432[[1,i]]=1]];
p432=p432/VectorNorm[N[p432[[1]]], 1]; p433=ZeroMatrix[1, 32];

For[i=0, i < 32, i=i+1; If[data4[[3,i]]==3, p433[[1,i]]=1]];
p433=p433/VectorNorm[N[p433[[1]]], 1];

ketelemu[x_, y_, v_, w_] := Module[{s}, s=ZeroMatrix[1, 32];
        For[i=0, i < 32, i=i+1; If[data4[[x,i]]==y && data4[[v,i]]==w, s[[1,i]]=1]];
        s=s/VectorNorm[N[s[[1]]], 1];        Evaluate[ToExpression[StringJoin["p4",
ToString[x],    ToString[y], ToString[v], ToString[w]]]]=s]

Do[ketelemu[1, 1, k, j], {k, 2, 3}, {j, 1, 3}];
Do[ketelemu[2, m, k, j], {m, 1, 3}, {k, 3, 3}, {j, 1, 3}];

haromelemu[x_, y_, v_, w_, z_, q_] :=
        Module[{s}, s=ZeroMatrix[1, 32];
        For[i=0, i < 32, i=i+1;
```

```
                If[data4[[x,i]]==y && data4[[v,i]]==w &&
                data4[[z,i]]==q, s[[1,i]]=1]];
        s=s/VectorNorm[N[s[[1]]], 1];
        Evaluate[ToExpression[StringJoin["p4", ToString[x], ToString[y], ToString[v],
        ToString[w], ToString[z], ToString[q]]]]=s]

Do[haromelemu[1, 1, 2, j, 3, m], {j, 1, 3}, {m, 1, 3}]

*egyelemű halmazokhoz tartozó paraméterek

Do[s=StringJoin["p4", ToString[k], ToString[l], "-empty4"];
      Evaluate[ToExpression[StringJoin["v4", ToString[k],
      ToString[l]]]]=Evaluate[ToExpression[s]], {k, 1}, {l, 1}];

Do[s=StringJoin["p4", ToString[k], ToString[l], "-empty4"];
      Evaluate[ToExpression[StringJoin["v4", ToString[k],
      ToString[l]]]]=Evaluate[ToExpression[s]],
        {k, 2, 3}, {l, 1, 3}];

*kételemű halmazokhoz tartozó paraméterek

Do[s=StringJoin["p4", ToString[k], ToString[l], ToString[m],
        ToString[n], "-", "v4", ToString[k], ToString[l], "-", "v4",
        ToString[m], ToString[n], "-empty4"];
  Evaluate[ToExpression[StringJoin["v4", ToString[k],ToString[l],
        ToString[m], ToString[n]]]]=
        Evaluate[ToExpression[s]],{k, 1},{l, 1},{m, 2, 3},{n, 1, 3}];

Do[s=StringJoin["p4", ToString[k], ToString[l], ToString[m],
        ToString[n],"-", "v4", ToString[k], ToString[l], "-", "v4",
        ToString[m],ToString[n], "-empty4"];
        Evaluate[ToExpression[StringJoin["v4",     ToString[k],ToString[l],ToString[m],
        ToString[n]]]]=
        Evaluate[ToExpression[s]],{k,2,2},{l,1,3},{m,3,3},{n, 1, 3}];

*háromelemű halmazhoz tartozó paraméterek

Do[s=StringJoin["p4", ToString[k], ToString[l], ToString[m],
        ToString[n],ToString[r], ToString[t], "-", "v4", ToString[k],
        ToString[l], ToString[m], ToString[n], "-", "v4",
        ToString[k], ToString[l], ToString[r], ToString[t], "-",
        "v4", ToString[m], ToString[n], ToString[r], ToString[t],
        "-", "v4", ToString[k], ToString[l], "-","v4", ToString[m], ToString[n], "-", "v4",
        ToString[r], ToString[t], "-empty4"]; Evaluate[ToExpression[StringJoin["v4",
        ToString[k],
        ToString[l], ToString[m], ToString[n], ToString[r], ToString[t]]]]=
   Evaluate[ToExpression[s]],{k,1},{l,1},{m,2,2},{n,1,3},{r,3,3},
```

```
{t,1,3}];
```

*AZ *X* VEKTOR MEGADÁSA

```
X=ZeroMatrix[47, 1];
```

*A *ct[0]* MÁTRIX MEGADÁSA

*A *GSIPC* marginálishoz tartozó fix paraméterek megadása

```
ct[1]=AppendColumns[v11151, v1114151, v1112151, v1112251, v1112351, v1113151,
v1113251, v1113351, v111213151, v111213251, v111213351, v111223151, v111223251,
v111223351, v111233151, v111233251, v111233351, v111214151, v111224151, v111234151,
v111314151, v111324151, v111334151, v11121314151, v11122314151, v11123314151,
v11121324151, v11122324151, v11123324151, v11121334151, v11122334151, v11123334151];

null1=ZeroMatrix[32, 128];

ct[1]=AppendRows[ct[1], null1];
```

*A *GSI* marginálishoz tartozó fix paraméterek megadása

```
ct[2]=AppendColumns[v41131, v41132, v41133, v4112131, v4112132, v4112133, v4112231,
v4112232, v4112233, v4112331, v4112332, v4112333];

null21=ZeroMatrix[12, 208];
null22=ZeroMatrix[12, 16];

ct[2]=AppendRows[null21, ct[2], null22];
```

*A *GS* marginálishoz tartozó fix paraméterek megadása

```
ct[3]=AppendColumns[v61121, v61122, v61123];

null3=ZeroMatrix[3, 248];

ct[3]=AppendRows[null3, ct[3]];

ct[0]=AppendColumns[ct[1], ct[2], ct[3]];
```

*A BECSLÉST VÉGZŐ MLE FÜGGVÉNY MEGHÍVÁSA

```
zeta1[mu_] := t[mu, {at, ct, 1}];
Zt1[mu_] := T[mu, {at, ct, 1}];
est=MLE[n, n, Log, zeta1, Zt1, Mt, X];
```

stepsize, error :
1. 1.77868
1. 0.00564231
1. 3.595894875865832^-7
1. 4.2265604340836644^-10
1. 1.7263617907216004^-12
1. 7.629174052500964^-15

kk=Transpose[{n, est}];

Export["model1.csv", kk];

*ellenőrzés. kell: az *X* vektor jó közelítéssel nullvektor.

ct[1].Log[at[0].Mt.est]

{1.9082·10^-17, -8.67362·10^-18, 5.72459·10-^17, 1.26635·10^-16, -7.45931·10^-17, 6.41848·10^-17, 5.0307·10^-17
3.64292·10^-17, -2.94903·10^-17, -4.33681·10^-17, -5.72459·10^-17, 9.54098·10^-17, -1.96024·10^-16, 9.54098·10^-17,
2.13371·10^-16, 6.07153·10^-17, -3.64292·10^-17, -1.92554·10^-16, 7.11237·10^-17, -1.9082·10^-17, 9.19403·10^-17
7.80626·10^-17, -7.45931·10^-17, 2.2031·10^-16, 3.98986·10^-17, 4.68375·10^-17, 9.54098·10^-17, -2.94903·10^-17
5.20417·10^-18, -2.23779·10^-16, 6.76542·10^-17, 4.68375 10^-17}

8.2 Illeszkedésteszt

A modell valószínűséghányados próbával történő illeszkedéstesztje szintén Mathematica-ban készült, az előző program folytatásaként. Output: a tesztstatisztika értéke és a p érték.

**n* és *est* mint fent: a megfigyelt és becsült gyakoriságok.

```
<< Statistics`ContinuousDistributions`;
lr=2*n*Log[n/est];
one=Table[i/i,{i,128}];
test=lr.one
```

85.8192

*a modell szabadságfoka 47

1-CDF[ChiSquareDistribution[47],test]

0.000468856
*ez a p érték.

8.3 Paraméterbecslés

A modellnek megfelelő eloszláshoz tartozó szabad paraméterek kiszámítása SPSS 15.0-ban készült el, mert ez feladat ebben a programcsomagban nagyon könnyen megvalósítható. A

modell paraméterezése (lásd a 7.3. fejezetet) a G, GS, GSI, GSIP, GSIPC marginálisok mellett történik, az alábbi példában a GSI marginálishoz tartozó szabad paraméterek értékét számítjuk ki, azaz az alábbiakat:

$$\lambda_I^{GSI}, \lambda_{SI}^{GSI}.$$

Az SPSS Hierarchical Loglinear Model parancsa λ, General Loglinear Model parancsa μ paraméterezésnek megfelelően ad paraméterbecslést. Bár ezek a procedúrák marginális loglineáris paraméterezésre nem használhatók, marginálisonként külön-külön telített modellt definiálva kiszámíthatjuk segítségükkel az egyes paramétereket. A Mathematica-ból való áttérés után cellánként aggregált adatbázis használata a kézenfekvő, ezért az *est* változóval, azaz a becsült gyakoriságokkal való súlyozás szükséges. A GSI marginálisra a Hierarchical Loglinear Model procedúra az alábbi módon hívandó meg (a cellák δ–val, alapbeállításként 0,5-tel történő simítása felülírandó):

```
WEIGHT
  BY est .
HILOGLINEAR
  g(1 2) s(1 4) i(1 4)
  /CRITERIA DELTA(0)
  /PRINT=ESTIM .
```

A program csak a nem redundáns paramétereket jeleníti meg, mindig a változók utolsó értékét elhagyva. A többi marginálishoz tartozó szabad paraméter kiszámítása hasonlóan, marginálisonként külön történik.

9 IRODALOM

Agresti, A. (2002): Categorical data analysis. Wiley.

Andersson, S. A., Madigan, D., Perlman, M. D. (2001): Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics*, 28, 33-85.

Andersson, S. A., Madigan, D., Perlman, M. D. (1997): On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scandinavian Journal of Statistics*, 24, 81-102.

Andor, M., Liskó, I. (2000): *Iskolaválasztás és mobilitás*. Budapest: Iskolakultúra Kiadó.

Andorka, R., Bukodi, E. and Harcsa, I. (1994): Társadalmi mobilitás, 1992. In Andorka, R., Kolosi, T., Vukovich, Gy. (szerk.): *Társadalmi Riport.* Budapest: Tárki, 293-310.

Andorka, R. (1995): A társadalmi mobilitás félévszázados trendjei Magyarországon. In: Andorka, R., Hradil, Peschar: *Társadalmi rétegződés*. Budapest: Aula, 393-427.

Andorka, R., Simkus, A. (1983): Az iskolai végzettség és a szülői család társadalmi helyzete. *Statisztikai szemle*, 6: 592-611.

Beller, E., Michael Hout, M. (2006): Welfare states and social mobility: How educational and social policy may affect cross-national differences in the association between occupational origins and destinations. *Research in Social Stratification and Mobility*, 24: 353–365.

Bergsma, W. (1997): *Marginal models for categorical data*. Tilburg: Tilburg University Press.

Bergsma, W., Rudas, T. (2002): Marginal models for categorical data. *The Annals of Statistics*, (30/1), 140-159.

Bishop, Y. V. V., Fienberg, S. E., Holland, P. W. (1975): *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA.

Blau, P. M, Duncan, O. D. (1967): *The American Occupational Structure*. New York: Wiley and Sons.

Boudon, R. (1974): *Education, Opportunity and Social Inequality: Changing Prospects in Western Society*. New York: Wiley

Breen, R., Luijkx, R. (2004): Conclusions. In: Breen, R. (szerk.): *Social Mobility in Europe*. Oxford: Oxford University Press, 383-446.

Boguszak, M., Gabal, I., Matějů, P. (1990): Ke koncepcím vývoje sociální struktury v ČSSR. *Sociologický časopis* (26/3), 168- 186.

Bukodi E. (1995): *Az iskolázottsági esélyek alakulása*. Budapest: Központi Statisztikai Hivatal.

Bukodi E. (2001): *A nemzedékek közötti mobilitás alakulása 1983 és 2000 között*. Budapest: Központi Statisztikai Hivatal.

Bukodi, E. (2003): Társadalmi rétegződés, mobilitás. In: uő (szerk.): *Társadalmi helyzetkép 2002*. Budapest: Központi Statisztikai Hivatal.

Colclough, G., Horan, P. M. (1983): The Status Attainment Paradigm: An Application of a Kuhnian Perspective. In: *The Sociological Quarterly*, Vol. 24, 25-42.

Cox, D. R., Wermuth, N. (2001): Some statistical aspects of causality. *European Sociological Review*, 17(1): 65–74.

Cox, D.R., Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, 8, 204-218; 247-277.

Cox, D.R., Wermuth, N. (1996). *Multivariate Dependencies – Models, Analysis and Interpretation*. London: Chapman & Hall.

Cox, D. R., Wermuth, N. (2004): Causality: a statistical view. *International Statistical Review*, (72/3), 285-305.

Dawid, A. P. (1979): Conditional independence in statistical theory. *Journal of the Royal Statistical Society* 41B 1-31.

Dawid, A. P. (1980): Conditional independence for statistical operations. Annals of Statistics, 8: 598–617.

Dessens, J., Jansen, W., Ringdal, K. (1995): Log-linear models in Comparative Research: A Paradigm Lost? *European Sociological Review* 11:261-271.

DiPrete, Th. A., Grusky, D. B. (1990): Structure and Trend in the Process of Stratification for American Men and Women. *American Journal of Sociology*, 96:107-43.

Drton, M. (2008). Discrete chain graph models. *Bernoulli*, közlésre elfogava.

Drton, M., Sullivant, S. (2007): Algebraic statistical models. *Statistica Sinica*, 17: 1273–1297.

Duncan, O. D., Featherman, D. L., Duncan, B. (1968): *Socioeconomic Background and Occupational Achievement: Extensions of a Basic Model*. Washington, D. C. U. S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research.

Duncan, O. D. (1966). Methodological Issues in the Analysis of Social Mobility. In: Smelser, N. J., Lipset, S. M. (szerk): *Social Structure and Mobility in Economic Development.* Chicago: Aldine, 51-97.

Erikson, R.; Goldthorpe, J.H. (1992): *The Constant Flux*. Oxford: Calderon.

Featherman, D. L., Jones, F. L., Hauser, R. M. (1975): Assumptions of social mobility research in the U.S.: The case of occupational status. *Social Science Research*, 4:329-360.

Featherman, D. L. (1981): Social Stratification and Mobility. Two Decades of Cumulative Social Science. *The American Behavioral Scientist* 24, 3, 364-385

Fienberg, S. E. (1970): The analysis of multidimensional contingency tables. *Ecology*, 51, 419-433.

Fienberg, S. E. (1980): *The Analysis of Cross-classified Categorical Data.* 2nd Edition. Cambridge, MA.: M.I.T. Press.

Firth, D. (2005): Some Topics in Social Statistics. In: Davison, A. C., Dodge, Y., Wermuth, N. (szerk.): *Celebrating Statistics: Papers in Honour of Sir David Cox on his 80th Birthday* Oxford: Oxford University Press.

Freedman, D. (2001): *On Specifying Graphical Models for Causation, and the Identification Problem*. University of California, Berkeley, Statistical Technical Reports, 601.

Freedman, D. (1997): From association to causation via regression. *Advances in Applied Mathematics* (18/1), 59-110.

Frydenberg, M. (1990a): The chain graph Markov property. *Scandinavian Journal of Statistics*. 17, 333-353.

Frydenberg, M. (1990b) Marginalization and collapsibility in graphical interaction models. *Annals of Statistics*, 18, 790-805.

Ganzeboom, H. B. G.; Treiman, D. J.; Ultee, W. C. (1991): Comparative Intergenerational Stratification Research: Three Generations and Beyond. *Annual Review of Sociology*, 17: 277-302.

Gazsó, F. (1997): A társadalmi folyamatok és a felsőoktatás. *Eszmélet* 35. szám

Gazsó, F., Laki, L. (2004): *Fiatalok az újkapitalizmusban*. Budapest: Napvilág Kiadó.

Geiger, D. and Pearl, J. (1988). Logical and algorithmic properties of conditional independence. *Technical Report R-97*, Cognitive Systems Laboratory, University of California, Los Angeles.

Glonek, G. J. N., McCullagh, P. (1995): Multivariate logistic models. *Journal of the Royal Statistical Society*, Ser. B, 57, 533-546.

Goldthorpe, J. H. (1996): Problems of "Meritocracy". In: Erikson, R. és Jonsson, J.O. (szerk.): *Can Education Be Equalized? The Swedish Case in Comparative Perspective*. Colorado: WestviewPress, 255-88.

Goodman, L.A. (1973): The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika*, 60, 179-192.

Heckerman, D., Meek, C., Cooper, G. (1999): *A Bayesian Approach to Causal Discovery*. In Glymour, C., Cooper, G. (szerk.): Computation, Causation, and Discovery. p. 141-165. Cambridge, MA: MIT Press.

Heckman, J. J. (1978): Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, 46/4, 931-959.

Jonsson, J. O. (1992): *Towards the Merit-Selective Society?* Stockholm: Swedish Institute for Social Research.

Kelley, J. (1990): The failure of a paradigm: Log linear models of social mobility. In , Modgil, C. (szerk.): *John Goldthorpe: Consensus and Controversy* London, England: Falmer Press, 319-46.

Kertesi, G., Kézdi, G. (2004): *Általános iskolai szegregáció – okok és következmények*. Budapesti Munkagazdaságtani Füzetek, BWP. 2004/7.

Kézdi, G. (2002): Two Phases of Labor Market Transition in Hungary: Inter-Sectoral Reallocation and Skill-Based Technological Change. *Budapest Working Papers on the Labour Market* 3. szám. Budapest: MTA KTK és BKÁE

Kolosi, T., Róbert, P. (2004): A magyar társadalom szerkezeti átalakulásának és mobilitásának főbb folyamatai a rendszerváltás óta. In: Kolosi T., Tóth I. Gy. és Vukovich Gy. (szerk.): *Társadalmi Riport 2004*. Budapest: TÁRKI, 48-74.

Kovács K., Hablicsek L. (2006): *Iskolázottság és halandóság*. Budapest: Egészségügyi Stratégiai Kutatóintézet.

Kuha, J. and Goldthorpe, J. H. (2007). *Path analysis for discrete variables: Education as mediator of social mobility in Britain*. Research Report 144, Department of Statistics, London School of Economics.

Kurz, K., Müller, W. (1987): Class mobility in the industrial world. *Annual Review of Sociology*, 13, 417-442.

Lauritzen, S. L., Spiegelharter, D. J. (1988): Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B,* 50, 157-224.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., Leimer, H.-G. (1990): Independence properties of directed Markov fields. *Networks*, 20, 491-505.

Lauritzen, S. L. (1996): *Graphical Models*. Oxford: Clarendon Press.

Lauritzen, S. L., Richardson, T. S. (2002): Chain graph models and their causal interpretation (with discussion). *Journal of the Royal Statistical Society, Series B,* 64 , 321 - 361.

Lauritzen, S. L., Wermuth, N. (1989): Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17, 31-57.

Levitz, M., Madigan, D., Perlman, M. D. (2001): Separation and Completeness Properties for Amp Chain Graph Markov Models. *Annals of Statistics,* 29/6, 1751-1784.

Lipset, S. M., Zetterberg, H. L. (1959). Social mobility in industrial societies. In: Lipset, S. M., Bendix, R. (szerk.): *Social Mobility in Industrial Society* (pp. 11-75). Berkeley: University of California Press.

Luijkx, R., Róbert, P., de Graaf, P. M., Ganzeboom H. B. G. (1995): A származástól a teljesítményig: A státuszmegszerzés folyamata Magyarországon. *Szociológiai Szemle,* 4: 3-27.

Luijkx, R., Róbert, P., de Graaf, P. M., Ganzeboom, H. B. G. (2002): Changes in Status Attainment in Hungary between 1910-1989. *European Societies*, 4 (1): 107-140.

Mare, R. D. (1981): Change and Stability in Educational Stratification. *American Sociological Review,* 46: 72-87

Meek, C. (1995): Strong Completeness and Faithfulness in Bayesian Networks. In: Besnard, P., Hanks, S. (szerk.): *Uncertainty in Artificial Intelligence, 11.*, Morgan Kaufmann Publishers, Inc., San Mateo, 411-418.

Muthén, B. O. (1983): Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22: 43-65.

Muthén, B. O. (1984): A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, 49: 115-132.

Németh, R. (2004): An application of marginal log-linear models to examine changes in social mobility in Hungary during the transition period. In *Recent Developments and Applications In Social Research Methodology. Proceedings of the RC33 Sixth International Conference on Social Science Methodology*, (RC33), Amsterdam.

Németh, R. (2006a): Grafikus modellek kategoriális adatokon - társadalomtudományi alkalmazással. In: Némedi, D., Somlai, P., Szabari, V., Szikra, D. (szerk.) *Kötő-jelek.* Az Eötvös Loránd Tudományegyetem Szociológia Doktori Iskolájának Évkönyve.

Németh, R., Rudas, T., Bergsma, W. (2006): *Analyzing categorical data with graphical models - a social science application.* SMABS-EAM (Society for Multivariate Analysis in the Behavioural Sciences, European Association of Methodology) konferencia, Budapest.

Németh, R. (2006b): A társadalmi mobilitás változásai Magyarországon a rendszerváltás folyamán. *Szociológiai Szemle*, 2006/4, 19-35.

Németh, R. (2007): Changes in social mobility in Hungary during the transition period. *Review of Sociology*, Vol 13, No. 1, 49-66. Az előző cikk angol nyelvű változata.

Pachter, L., Sturmfels, B. (2005): *Algebraic statistics for computational biology*. Cambridge University Press, New York.

Pearl, J. (1988): *Probabilistic reasoning in expert systems*. Morgan Kaufman, San Mateo.

Pearl, J. (1993) *On the Statistical Interpretation of Structural Equations* UCLA Cognitive Systems Laboratory, Technical Report (R-200).

Pearl, J. (1995): Causal diagrams for empirical research. *Biometrika*, (82/4), 669-710.

Richardson, T. (2003): Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics* 30, Vol 30: 145-157.

Róbert, P. (2004): Iskolai teljesítmény és társadalmi háttér nemzetközi összehasonlításban. In.: Kolosi, T., Vukovich, Gy., Tóth, I. Gy. (szerk.): *Társadalmi Riport 2004*. Budapest: Tárki, 193-205.

Róbert, P., Bukodi, E. (2004): Changes in Intergenerational Class Mobility in Hungary, 1973–2000. In: Breen, R. (szerk.): *Social Mobility in Europe*. Oxford: Oxford University Press, 287-315.

Rogoff-Romsoy, N. (1966): Changes in rates and form of mobility. In: Smelser, N. J., Lipset, S. M. (szerk.): *Social Structure and Mobility in Economic Development*. Chicago: Aldine, 213-234.

Rudas, T., Bergsma, W., Németh, R. (2006): Parameterization and estimation of path models for categorical data. In: Rizzi, A., Vich, M. (szerk.) *COMPSTAT 2006 Proceedings in Computational Statistics*, Physica-Verlag, 383-394.

Rudas, T., Bergsma, W., Németh, R. (2009): Markov marginal models for categorical data. (kézirat)

Rudas, T., Bergsma, W. (2004): On application of marginal models for categorical data. *Metron*, (62/1), 1-23.

Rudas, T: (1998): *Odds Ratios in the Analysis of Contingency Tables*. Thousand Oaks: Sage.

Sewell, W., Shah, V. (1968): Social Class, Parental Encouragement and Educational Aspirations. *American Journal of Sociology*, 73: 559–572.

Shavit, Y., Müller, W. (szerk.) (1997): *From School to Work: A Comparative Study of Educational Qualifications and Occupational Destinations*. Oxford: Oxford University Press.

Simkus, A. (1981): Comparative stratification and mobility. International Journal of *Comparative Sociology* 22 (3): 213-236.

Sorokin, P.A. (1964): *Social and Cultural Mobility*. Chicago: Free Press.

Stillman, M. (2005): Tools for computing primary decompositions and applications to ideals associated to Bayesian networks. In: *Solving Polynomial Equations. Foundations, Algorithms and Applications*, Volume 14 of *Algorithms and Computation in Mathematics*, Springer-Verlag, 203-239.

Spirtes, P., Glymour, C., Scheines, R. (1993): *Causation, Prediction, and Search*. Springer-Verlag, New York, 149-150.

Studený, M. (1992): Conditional independence relations have no finite complete characterization. In: *Transactions of the 11th Prague conference on information theory, statistical decision functions and random process*. Academia: Prague, 377-96.

Studeny, M. and Bouckaert, R. R. (1998). On chain graph models for description of conditional independence structure. *Annals of Statistics*, 26, 1434-1495.

Székelyi, M., Csepeli, Gy., Örkény, A., Szabados, T. (1998): *Válaszúton*. Budapest: Új Mandátum.

Treiman, D. J. (1970): Industrialization and social stratification. In Laumann, E. O. (szerk.): *Social Stratification, Research and Theory for the 1970s*. Indianapolis: Bobbs-Merill, 207-34. Magyarul in Róbert, P. (szerk.): *Társadalmi mobilitás: hagyományos és új megközelítések*. Budapest: Új Mandátum Könyvkiadó, 1998, 86-111.

Treiman, D.J., Yip, K-B. (1989): Educational and Occupational Attainment in 21 Countries. In: Kohn, M.L. (szerk.): *Cross-national research in sociology*. Newbury Park: Sage, 373-394.

Treiman, D. J., Ganzeboom, H. B. G. (2000): The Fourth Generation of Comparative Stratification Research. In: Quah, S. és Sales, A. (szerk.): *The International Handbook of Sociology*. Thousand Oaks, CA: Sage, 123-150.

Verma, Th., Pearl, J. (1990): Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence*, Proceedings of the Sixth Conference, pp. 220-227. San Francisco: Morgan Kaufman.

Vermunt, J.K. (1997). *Log-linear models for event histories*. Thousand Oakes: Sage Publications.

Wermuth, N., Lauritzen, S. L. (1990). On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models. *Journal of the Royal Statistical Society. Series B*, Vol. 52, No. 1, pp. 21-50.

Wermuth, N. (2003). Analysing social science data with graphical Markov models. In: Green, P., Hjort, N. és Richardson, S. (szerk.): *Highly Structured Stochastic Systems*. Oxford: University Press, 47-52.

Wermuth, N. & Cox, D.R. (1992). Graphical models for dependencies and associations. In: Dodge, Y. & Whittaker, J. (szerk.): *Computational Statistics*. (Vol. 1), Heidelberg: Physica, 235-247.

Wermuth, N. & Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society B*, 66, 687-717.

Wermuth, N. (2005). Graphical chain models. In: Everitt, B. és Howell, D. C. (szerk.): *Encyclopedia of Behavioral Statistics, II.* Chichester: Wiley, 755-757.

Whittaker, J. (1990): *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons.

Winship, C., Mare, R. D. (1983): Structural equations and path analysis for discrete data. *American Journal of Sociology*, 89: 54-110.

Xie, Y. (1989). Structural Equation Models for Ordinal Variables. An Analysis of Occupational Destination. *Sociological Methods & Research*, Vol. 17, No. 4, 325-352.

Xie, Y. (1992): The log-multiplicative layer effect model for comparing mobility tables. *American Sociological Review*, 57: 380-95.

Xie, Y., Goyette, K. (2003). Social Mobility and Educational Choices of Asian Americans. *Social Science Research* 32:467-498.

Yamaguchi, K. (1983). The Structure of Intergenerational Occupational Mobility: Generality and Specificity in Resources, Channels, and Barriers. *American Journal of Sociology* Vol. 88, No. 4: 718-745.

10 FELHASZNÁLT ADATBÁZISOK

Életmód és időmérleg-vizsgálat Magyarországon 1999-2000. Budapest: TÁRKI Adatbank.

International Social Survey Programme, Social Inequality II, 1992. Köln: GESIS-ZA.

TÁRKI Magyar Háztartás Panel II. hullám, 1993. Budapest: TÁRKI Adatbank.

TÁRKI Magyar Háztartás Panel IV. hullám, 1995. Budapest: TÁRKI Adatbank.

TÁRKI Magyar Háztartás Panel VI. hullám, 1997. Budapest: TÁRKI Adatbank.

Társadalmi mobilitás Magyarországon 1983. Budapest: TÁRKI Adatbank.

Társadalmi mobilitás Magyarországon 1992. Budapest: TÁRKI Adatbank.

11 ÁBRÁK JEGYZÉKE

1. ábra. Egy gráf-modell (I: iskolázottság, F: foglalkozás, I': apa iskolázottsága, F': apa foglalkozása) ... 36
2. ábra. 1-es típusú láncgráf, és a gráf komponensei ... 38
3. ábra. 2-es típusú láncgráfok, $\tau_1 \ll \tau_2 \ll \tau_3 \ll \tau_4$ ... 39
4. ábra. Ugyanahhoz az IG-hez rendelhető 2-es típusú LG-k, $\tau_1 \ll \tau_2 \ll \tau_3$ ... 40
5. ábra. Ugyanahhoz az IKG-hez rendelhető 2-es típusú LG-k, $\tau_1 \ll \tau_2 \ll \tau_3$ ... 40
6. ábra. Az általános, illetve az 1-es ill. 2-es típusú láncgráfokon, valamely Markov-tulajdonsággal definiált modellosztályok tartalmazási viszonya (sötét háttér: nemüres, fehér háttér: nem ismert, hogy üres-e) ... 50
7. ábra. Egy láncgráf ... 52
8. ábra. Ötváltozós státuszelérési modell, láncgráf modellel (I: iskolázottság, F: foglalkozás, I': apa iskolázottsága, F': apa foglalkozása, J: jövedelem) ... 60
9. ábra. Az eredeti Blau-Duncan modell (1967), és annak Boguszak et al. (1990) általi módosítása két lépésben ... 66
10. ábra. Az eredeti Blau-Duncan modell egy variánsa (Duncan et al., 1968, Treiman, 1970), és a modellre alkalmazva Boguszak et al. (1990) módosításait ... 68
11. ábra. Felsőfokú továbbtanulási tervek kialakulása, IKG (Spirtes et al., 2003) ... 70
12. ábra. Módosított útmodell ... 73
13. ábra. IKG (Rudas, Bergsma, 2004) ... 75
14. ábra. IKG (Rudas, Bergsma, 2004) ... 79
15. ábra. Az eredeti Blau-Duncan modell (1967) ... 84
16. ábra. Láncgráf-modell (Drton, 2008) ... 86
17. ábra. Az általános, illetve az 1-es ill. 2-es típusú láncgráfokon definiált modellosztályok tartalmazási viszonya (sötét háttér: nemüres, fehér háttér: nem ismert, hogy üres-e) ... 93
18. ábra. Klasszikus problémákat fedő konfigurációk három csúcson ... 95
19. ábra. Paraméterbecslések, DAG útmodell ... 97
20. ábra. A vizsgált modell (I: iskolázottság, F: foglalkozás, I': apa iskolázottsága, F': apa foglalkozása) ... 112
21. ábra. Treiman-modell (Treiman, 1970, eredetileg Duncen et al., 1968). I': apa iskolázottsága, F': apa foglalkozási státusza, F: foglalkozási státusz, I: iskolázottság, J: jövedelem ... 119
22. ábra. A Treiman-modell módosítása Boguszak et al. (1990) alapján ... 120
23. ábra. Az 1. modell paraméterbecslései, a dobozokban soronként felülről lefelé: USA, Magyarország, Csehszlovákia ... 122
24. ábra. Két modell a Sewell és Shah-féle adatokhoz ... 124
25. ábra. Szerhasználat, 1. modell ... 127
26. ábra. Szerhasználat, 3. modell ... 128
27. ábra. Szerhasználat, a legegyszerűbb modell, paraméterbecslések. ... 129
28. ábra. Panelvizsgálat, telített modell ... 131

12 TÁBLÁZATOK JEGYZÉKE

1. táblázat. Hipotetikus példa egy *deliberative poll* végeredményéről. Zárójelben a vita maximális hatása esetén fennálló gyakoriságok 25
2. táblázat. A láncgráf-modellek típusai, kategoriális változók és pozitív eloszlás esetén (=: ekvivalencia, ⇒: implikáció) 60
3. táblázat. A láncgráf-modellek típusai, kategoriális változók és pozitív eloszlás esetén. Sötét háttér = mindegyik sima modell, világos = van köztük nem sima, fehér = kérdéses. 85
4. táblázat. A láncgráf-modellek típusai, kategoriális változók és pozitív eloszlás esetén. Sötét háttér = mindegyik sima modell, világos = van köztük nem sima, fehér = kérdéses. 92
5. táblázat. Paraméterbecslések a modernizációs hipotézis 2. pontjához 114
6. táblázat. Paraméterbecslések a modernizációs hipotézis 3. pontjához 115
7. táblázat. Paraméterbecslések a modernizációs hipotézis 4. pontjához 115
8. táblázat. Paraméterbecslések a modernizációs hipotézis 1. pontjához 115
9. táblázat. Az ötváltozós modellek illeszkedéstesztjeinek eredményei 121
10. táblázat. Sewell és Shah adatai (n=10.318). A: alacsony, AK: alsó-közép, FK: felső-közép, F: felső 123
11. táblázat. Az illeszkedéstesztek eredményei 126
12. táblázat. A daytoni adatok, n=2276 126
13. táblázat. Az illeszkedéstesztek eredményei 129

13 A SZERZŐ PUBLIKÁCIÓI

13.1 Folyóirat cikkek és könyvfejezetek

Rudas, T., Bergsma, W., Németh, R. (2009): Markov marginal models for categorical data. Kézirat.

Kósa, Zs. et al (köztük Németh, R.): Does socioeconomic status fully mediate the effect of ethnicity on the health of Roma people in Hungary? *Journal of Epidemiology and Community Health*, közlésre elfogadva

Vokó, Z. et al (köztük Németh, R.): Az agyérbetegségek epidemiológiája Magyarországon az ezredfordulót követően. *Lege Artis Medicinae*. 2008, 18. évf., 1. sz. , 31-38.

Zajkás, G. et al (köztük Németh, R.): Táplálkozási vizsgálat Magyarországon, 2003-2004. Mikrotápanyagok, vitaminok. *Orvosi Hetilap*, 2007, 148. évfolyam, 34. szám, 1593-1600.

Bíró, L. et al (köztük Németh, R.): Táplálkozási vizsgálat Magyarországon, 2003-2004. *Orvosi Hetilap*, 2007, 148. évfolyam, 15. szám, 703-8.

Németh, R.: Changes in social mobility in Hungary during the transition period. *Review of Sociology*, 2007, Vol 13, No. 1, 49-66.

Boros, J., Görög, K., Kaposvári, Cs., Németh, R.: Általános epidemiológia. In: Ember, I. (szerk.): *Népegészségügyi Orvostan* tankönyv, 2007, Dialog Campus, Pécs.

Kósa, Zs. et al (köztük Németh, R.): A Comparative Health Survey of the Inhabitants of Roma Settlements in Hungary. *American Journal of Public Health*, 2007, Vol 97, No. 5, 853-859.

Németh, R.: A társadalmi mobilitás változásai Magyarországon a rendszerváltás folyamán. *Szociológiai Szemle*, 2006/4, 19-35.

Németh, R.: Grafikus modellek kategoriális adatokon - társadalomtudományi alkalmazással. In: Némedi, D., Somlai, P., Szabari, V., Szikra, D. (szerk.) *Kötő-jelek*. Az Eötvös Loránd Tudományegyetem Szociológia Doktori Iskolájának Évkönyve, 2006.

Rodler Imre et al. (köztük Németh Renáta): Táplálkozási vizsgálat Magyarországon *Orvosi Hetilap*, 2005, 146/34.

Németh, R.: Representativeness problems inherent in address-based sampling and a modification of the Leslie Kish grid. *Bulletin of Sociological Methodology*, 2004, N. 83, 43-60.

Szende Á., Németh R.: A magyar lakosság egészségi állapothoz kapcsolódó életminősége. *Orvosi Hetilap*, 2003. 144. évfolyam, 34. szám, 1667-1674.

Borsos, K., Vitrai, J., Boros, J., Németh, R., Országh, S.: A felnőtt lakosság egészségügyi kiadásai a 2000. Évi országos felmérés adatai alapján. *Egészségügyi Gazdasági Szemle*, 2003, 41. évfolyam, 4. szám 25-33.

Németh, R., Rudas, T.: Mintavétel a Leslie Kish kulcs alkalmazásával. *Statisztikai Szemle*, 2002, 4.

13.2 Konferencia kötetek

Rudas, T., Bergsma, W., Németh, R. (2006): Parameterization and estimation of path models for categorical data. In: Rizzi, A., Vich, M. (eds.) *COMPSTAT 2006 Proceedings in Computational Statistics*, Physica-Verlag, 383-394.

Vokó Z. et al (köztük Németh R.) (2005): Health of the inhabitants of Roma settlements in Hungary – a comparative health survey. In: *European Journal of Public Health* 15 (6): 65, 2005 EUPHA (European Public Health Association) Annual Conference, Graz, Ausztria.

Csizmadia, P., Németh, R. (2004): Work stress factors and their associations with health in the contemporary Hungarian working population. In: *European Journal of Public Health* 14 (4): 25-26, Dec 2004. EUPHA (European Public Health Association) Annual Conference, Oslo, Norvégia.

Németh, R.(2004): An application of marginal log-linear models to examine changes in social mobility in Hungary during the transition period.In: *Proceedings of the RC33 Sixth International Conference on Social Science Methodology*, Amsterdam.

Zakariás, I., Németh, R. (2004): Comparing sampling frames. In: *Proceedings of the RC33 Sixth International Conference on Social Science Methodology*, Amsterdam.

Németh, R. (2001): Respondent selection within the household – A modification of the Kish grid. In: *Proceedings of the Sixth Austrian, Hungarian, Italian and Slovenian Meeting of Young Statisticians*, Ossiach, Austria.

13.3 Konferencia kivonatok

Vokó, Z. et al. (köztük Németh, R.): *The complex effect of ethnicity on the health of Roma people in Hungary*. XVIII World Congress of Epidemiology, Porto Alegre, Brazília, 2008.

Németh, R., Rudas, T., Bergsma, W.: *Analyzing categorical data with graphical models - a social science application*. SMABS-EAM (Society for Multivariate Analysis in the Behavioural Sciences, European Association of Methodology) konferencia, Budapest, 2006.

Németh, R.: *Grafikus modellek kategoriális adatokon*. Az ELTE Társadalomtudományi Kar Szociológia Doktori Iskolájának konferenciája. Budapest, 2006.

Lobmayer, P., Németh, R.: *Income inequality and health in Hungarian counties*. International Symposium, az Eighth International Congress of Behavioural Medicine szatellitje. Budapest, 2004.

Németh, R. *Időbeli trendek vizsgálata: az OLEF2000 és OLEF2003 összehasonlításának módszertana*. A Népegészségügyi Tudományos Társaság XIII. Nagygyűlése, Szekszárd, 2004.

Németh R.: *Sampling Design of Health Surveys: Household as a sampling unit*. Biannual Seminar, ROeS, International Biostatistics Society, St. Gallen, Svájc, 2003.

Lobmayer, P., Németh, R.: *Társadalmi egyenlőtlenség és egészség összefüggése Magyarországon az OLEF2000 adatai alapján*. Magyar Higiénikusok Társasága XXXIV. Vándorgyűlés, Siófok, 2003.

Németh, R., Szende, Á.: *A magyar lakosság egészségi állapothoz kapcsolódó életminősége: a demográfiai, társadalmi, és gazdasági tényezők hatása*. Méltányosság az egészségügyben konferencia, Budapest, 2002.

Németh, R.: *Miért célszemély?* Vitaindító. II. Országos Egészségstatisztikai Fórum, Budapest, 2002. Mintavétel szekció, kerekasztal beszélgetés.

Vizi, J. et al. (köztük Németh, R.): *Egészségre fordított kiadások az OLEF2000-ben.* Népegészségügyi Tudományos Társaság XI. Nagygyűlése, Nyíregyháza, 2002.

Németh R.: *OLEF2000 - mintavétel, elemzés.* Országos Egészségstatisztikai Fórum, Budapest, 2001.

Népegészségügyi Tudományos Társaság X. Nagygyűlése, Gyula, 2001:

Az OLEF2000 tervezésekor alkalmazott egészségmodell. Vokó Z., Boros J., Grajczjar I., Németh R., Országh S., Széles Gy., Vitrai J., Vizi J.

Az OLEF2000 tervezésének és megvalósulásának tapasztalatai. Vitrai J., Boros J., Grajczjar I., Németh R., Országh S., Széles Gy., Vizi J., Vokó Z.

Az OLEF2000 mintavételi és elemzési módszere. Németh R., Boros J., Grajczjar I., Országh S., Széles Gy., Vitrai J., Vizi J., Vokó Z.

Az OLEF2000 adatkezelési kérdései. Országh S., Boros J., Grajczjar I., Németh R., Széles Gy., Vitrai J., Vizi J., Vokó Z.

Társadalmi-gazdasági meghatározottság felmérése az OLEF2000-ben. Grajczjar I., Boros J., Németh R., Országh S., Széles Gy., Vitrai J., Vizi J., Vokó Z.

Az egészségmagatartás vizsgálata az OLEF2000-ben. Boros J., Grajczjar I., Németh R., Országh S., Széles Gy., Vitrai J., Vizi J., Vokó Z.

Krónikus betegségek, korlátozottság, mentális egészségi állapot az OLEF2000-ben. Széles Gy., Boros J., Grajczjar I., Németh R., Országh S., , Vitrai J., Vizi J., Vokó Z.

Az igénybevett egészségügyi szolgáltatások felmérése az OLEF2000-ben. Vizi J., Boros J., Grajczjar I., Németh R., Országh S., Széles Gy., Vitrai J., Vokó Z.

13.4 Munkaanyagok, kutatási jelentések

Deckovic-Dukres, V., Hrkal, J., Németh, R., Vitrai, J., Zach, H. (2007): *Inequalities in health system responsiveness. Joint World Health Survey Report Based on Data from Selected Central European Countries*, Jelentés a WHO megbízásából.

Remák, E., Gál, R. I., Németh, R. (2006): *Health and morbidity in the accession countries. Country report – Hungary*. ENEPRI Research Reports 28, Brussels: ENEPRI.

Albert, F., Dávid, B., Németh, R. (2005): *Társas támogatás, társadalmi kohézió*. In.: Országos Lakossági Egészségfelmérés OLEF2003, Kutatási Jelentés.

Németh, R., Csizmadia, P. (2005): *Anyagi helyzet*, ill. *A munka világa* c. fejezetek. In.: Országos Lakossági Egészségfelmérés OLEF2003, Kutatási Jelentés.

Boros J., Németh R., Vitrai J. (szerk.) (2002): *Országos Lakossági Egészségfelmérés OLEF2000, Kutatási Jelentés*. Országos Epidemiológiai Központ.

Németh R. (2002): *Elemzés* és *Mintavétel* fejezetek. In.: Országos Lakossági Egészségfelmérés OLEF2000, Kutatási Jelentés. Szerk.: Boros J., Németh R., Vitrai J. Az Országos Epidemiológiai Központ kiadványa.

Németh R. (2003): *Sampling Design of Health Surveys: Household as a sampling unit*. LIS Working Papers No. 358.

Doorslaer, E. van, Masseria, C. és az OECD Health Equity Research Group (köztük Németh, R.) (2004): *Income-related inequality in the use of medical care in 21 OECD countries*. In: Towards High-Performing, Health Systems: Policy Studies, OECD.

Németh, R. (szerk.) (1998): *Kategoriális adatok elemzése.* Egységes jegyzet az ELTE Szociológiai Intézet survey statisztikus szakirány számára.