Látens változós modellezés

Diplomamunka

Írta: Tánczos Ervin

Alkalmazott matematikus szak

Témavezetők:

Pröhle Tamás, egyetemi tanársegéd
Valószínűségelméleti és Statisztika Tanszék
Eötvös Loránd Tudományegyetem, Természettudományi Kar

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2009

Tartalomjegyzék

1. Bevezetés

A többváltozós adatelemzés statisztikai módszereit előszeretettel alkalmazzák a legkülönbözőbb kutatási területeken. Ezek közül a látens (vagy rejtett) változós modellezés a társadalomtudományokban örvend nagy népszerűségnek. Ennek oka, hogy gyakran olyan változók értékeit kívánják meghatározni, amelyek közvetlenül nem mérhetők, mint például a boldogság, stressz vagy addikcióra való hajlam. Minden ilyen modell alapötlete, hogy a közvetlenül mérhető adatok értékeit háttérben meghúzódó, rejtett változók határozzák meg valamilyen módon. Mivel a magyarázó változók közvetlenül nem figyelhetők meg, több nehézség lép fel ezen modellek alkalmazásában egészen a modell felírásától a számításokon át a kapott eredmények interpretálásáig. Még az egyik leggyakrabban használt modellben, a faktoranalízisben, is számos kérdés nincs teljesen tisztázva, például az, hogy mikor egyértelmű a megoldás, és mely feltételek mellett fogja a konkrét számítási eljárásunk azt a megoldást adni.

A nehézségek ellenére azért használják gyakran az ilyen modelleket, mert igen tetszetős eredményt szolgáltatnak. A rengeteg megfigyelés információját, legalábbis jó részét, lényegesen kisebb számú változóba aggregálják, amelyek ráadásul könnyen áttekinthető kapcsolatban állnak egymással, illetve a mért változókkal is. Ez csábítóvá teszi az említett eljárásokat, hiszen lehetőséget nyújt arra, hogy a komplikált sokváltozós megfigyeléseinket egyszerű módon interpretáljuk.

A dolgozat témája egy általános látens változós modell, a Structural Equation Modeling (SEM) bemutatása. A második fejezet tartalmazza a modell felírását. Ez a fejezet a [2] *The Relationship Between Software Developement, Theory, and Education in Structural Equation Modeling* című cikk alapján készült. A harmadik fejezetben találhatóak a SEM illesztésére használt leggyakoribb módszerek. A fejezet elkészítéséhez az alábbi irodalmat használtam: [1] *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*, [3] *Bayesian Structural Equation Modelling* és [9] *Markov Chain Monte Carlo and Gibbs Sampling*. A negyedik fejezet foglalkozik a modell illeszkedésének minőségét mérő leggyakrabban használt mutatókkal. A negyedik fejezet a [1] *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis* című könyv második fejezetéből íródott. Az ötödik rész az útdiagramok SEM modellezésbeni szerepét tárgyalja. Az itt található tételek bizonyításai a [4] *Using Path Diagrams as a Structural Equation Modeling Tool* című cikkben olvashatók. Az utolsó részben pedig egy példán keresztül mutatom be a SEM modell használatát.

2. Structural Equation Modeling

2.1. A modell felírása

A SEM modell gyökerei a path analízisig nyúlnak vissza. A path analízis lényege, hogy a különböző változók között ok-okozati viszonyt tételezünk fel, és ez alapján írunk fel regressziós egyenleteket, amelyek összekapcsolják őket. A név onnan ered, hogy a változók közötti kapcsolatok szemléltetésére egy irányított gráfot rajzolunk fel, melyben a csúcsok a változók és a köztük futó irányított élek a regressziós együtthatók. A SEM ennek a modellnek az egyenletekkel felírt rejtett változókat is tartalmazó továbbfejlesztése.

Amiben ez a módszer többet nyújt a szokásos látens változós eljárásoknál az, hogy a rejtett változók közötti strukturális viszonyt is felírhatunk, és egyidejűleg ezt is figyelembe vesszük a modell illesztésekor. Ez alapján a látens változókat és az egyenleteket is két csoportra oszthatjuk. A változók lehetnek külső (vagy exogén) és belső (vagy endogén) változók. Azokat a változókat, melyekre nincs másik olyan látens változó, amely rájuk közvetlen hatással lenne külső változóknak nevezzük. Más szóval ezek azok a változók, amik magyarázzák a többit. A fennmaradó változók a másik csoportba esnek. Tehát a belső változók azok, amelyeket más látens változók magyaráznak. Az egyenletek azon csoportját, amelyek a látens változók közötti viszonyt írják le strukturális egyenleteknek nevezzük, míg azokat, amelyek a mért és a látens változók kapcsolatát írják le, mérési egyenleteknek nevezzük.

Ezek után a modell a következő formába írható: A struktúrális modell

$$\eta = B\eta + \Gamma\xi + \zeta$$

A belső változókat η , a külsőket pedig ξ jelöli. Mindkettőhöz tartozik egy-egy mérési modell:

$$y = \Lambda_y\eta + \varepsilon$$
$$x = \Lambda_x\xi + \delta$$

Ez a három egyenlet együttesen a SEM modell. A következőket tesszük fel:

- Minden változó 0 várható értékű
- ζ független ξ-től
- ε független η-tól
- δ független ξ-től

- ε független ξ-től
- ζ,ε és δ függetlenek
- $diag(B) = 0$

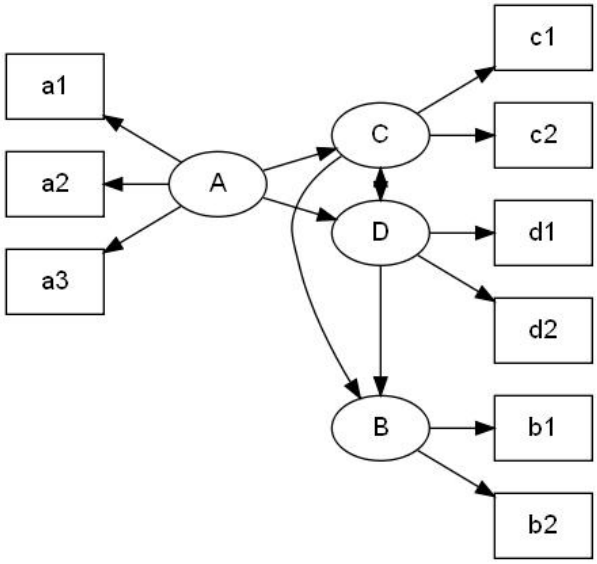
Az első feltevés csak technikai jellegű, és nem veszítjük el vele az általánosságot. Ha át akarunk térni a nem 0 várható értékű esetre, egyszerűen minden egyenletben hozzá kell adnunk a jobb oldalhoz bal oldal várható értékét. A következő öt feltétel a hibák egymástól és a látens változóktól való függetlenségét fejezi ki, míg az utolsó azt jelenti, hogy a belső változók regressziójában mindegyik legfeljebb az egyik oldalon szerepelhet.

Ez a modell felírás még nagyon általános, hiszen ezekből az egyenletekből csak x-et és y-t , és az ő tapasztalati kovarianciamátrixukat ismerjük, és rengeteg ismeretlen változónk van. Viszont a Λ_y,Λ_x,B és Γ együttható mátrixoknak speciális alakja van, amit mindig az adott modell felírás határoz meg. Ezen mátrixok elemeinek nagy része zérus, mivel előre meghatározzuk azt, hogy az egyes mért változókra mely látens változók hatnak, és azt is, hogy az egyes belső látens változókra mely másik látens változók hatnak. Ha olyan modellt írunk fel, amiben még mindig több ismeretlenünk van, mint egyenletünk, megtehetjük, hogy egyes együtthatók értékeit előre rögzítjük, vagy egyéb feltételeket teszünk rájuk. Ezzel azonban óvatosan kell bánni, mert könnyen előfordulhat, hogy nagyon rosszul illeszkedő modell lesz a végeredmény, ezért lehet, hogy ekkor érdemesebb újabb modellt felírni.

Ez a modell általános esetként magában foglal több eljárást is, mint például a faktoranalízist, vagy a többváltozós regressziót. A faktoranalízis például a következőképpen néz ki a SEM terminológiában: A látens változók mindegyike külső, hiszen nincs köztük semmilyen magyarázó viszony, így nincsenek strukturális egyenletek, és csak a külső változókra felírt mérési modell marad meg. Ebben a modellben pedig nem írunk elő semmit az együttható mátrix alakjára, azaz bármelyik látens változó bármelyik mérésre hatással lehet. Amit a faktoranalízisben saját faktornak neveztünk, most hibának hívunk.

Így ha újra megnézzük a modellt, a következő módon is interpretálhatjuk: A mért változókra felírunk egy vagy több faktoranalízist, majd a látens változókra felírunk egy többváltozós regressziót. Valójában megtehetnénk ezt úgy is, hogy először kiszámoljuk a faktoranalíziseket, kinyerjük a faktorokat, majd ezekre felírjuk a regressziót. Ez a módszer azonban más eredményt adna, hiszen a SEM számításakor a látens változók közti strukturális viszonyt már akkor figyelembe vesszük, mikor a mért és a látens változók közti kapcsolatot tárjuk fel.

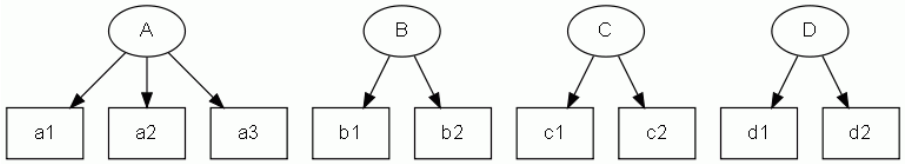
Gyakran a SEM modelleket ábrákkal adják meg. Ennek előnye, hogy könnyebben átlátható, mint a mátrixokkal történő felírás. Egy irányított gráfot szokás rajzolni, melynek csúcsai a változók, és az élek jelölik a regressziós együtthatókat.

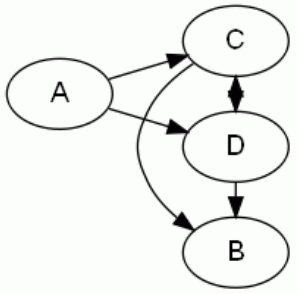


Néhány megállapodás teszi könnyen áttekinthetővé a gráfos felírást. Az első, hogy a csúcsokat úgy próbáljuk meg elrendezni, hogy köztük az irányított élek jobbról balra, vagy fentről lefelé fussanak. Előfordulhat, hogy ez nehezen áttekinthető ábrát eredményezne, ezért ez a szabály nem szigorú. A megfigyelt változókat téglalapba, a látens változókat pedig ellipszisbe szokás foglalni. Ez alapján az ábrán levő SEM-ben négy látens változó található, melyeket nagy betűkkel jelöltünk. Azok az irányított élek, melyeknek csak egy feje van, a regressziós együtthatókat jelölik. Az ábrán levő példában az egyetlen exogén változó az A, hiszen rá sehonnan nem mutat él, míg a másik három rejtett változó mind endogén, hiszen C-be és D-be A-ból vezet él, míg B-be C-ből és D-ből. Azokkal az élekkel, amelyeknek két feje van, kovarianciát jelölünk. Ezek az élek vagy mért-, vagy belső látens-, vagy külső látens változók változók között futnak, egyik csoportból a másikba nem. Egy ilyen nyíl jelenléte azt mutatja, hogy a két változó hibája korrelál. A példánkban a C és D csúcsok közti kétfejű nyíl azt jelenti, hogy a belső látens változók hibájának (korábbiakban ζ) kovarianciamátrixa nem diagonális, a főátló fölött egyetlen elem nem 0, mégpedig a C és D kovarianciájának megfelelő. Ezen felül néha szokás a mért változóinkhoz ún. reziduális éleket húzni. Ezek az élek a változókba mutatnak, azonban a farkuk nincs befejezve. A mért változók hibáit jelölik, vagy másképpen fogalmazva azon magyarázó tényezők hatásait, amelyek a mi modellünkön kívül esnek. A reziduális élek a későbbiekben nem fognak szerepelni az ábrákon.

A könnyebb áttekinthetőség érdekében az előzőekhez hasonlóan itt is szétbont-

hatjuk az ábránkat mérési és strukturális modellre:





Gyakran ez a felírás készül el előbb, amikor egy konkrét modellt szeretnénk illeszteni, mivel könnyebben átlátható és javítható, mint a mátrixfelírás. Az alkalmazók az irányított élek mentén oksági viszonyt feltételeznek. Úgy interpretálják a felírt modellt, hogy annak a csúcsnak megfelelő változó, amely az él végénél helyezkedik el okozza azt, ami az él fejénél található. Ez a megközelítés nem teljesen precíz és ezért nem is helyénvaló. Valójában arról van szó, hogy ha a magyarázó változónk értéke megváltozik, azt várjuk, hogy a magyarázott változó értéke is más legyen. Ez önmagában még nem tételez fel ok-okozati viszonyt a két változó között. Előfordulhat például, hogy valamely harmadik, számunkra ismeretlen körülmény megváltozása okoz mindkét oldalon változást. Sőt, az is előfordulhat, hogy a magyarázott változó megváltozása is változást eredményezne a magyarázó változóban. Ez viszont értelmetlenné teszi a megnevezésüket, és egyúttal azt jelenti, hogy a gráfban az él irányítását megfordítva olyan modellhez jutunk, ami ugyanolyan jól írja le az adatainkat. Ezért a gyakorlati alkalmazásban létfontosságú kérdés ezen jelenségek figyelembe vétele, hiszen ha ez elmarad, az alkalmazó könnyen fals eredményhez juthat. Ezzel a kérdéssel a későbbiekben még foglalkozunk, de most rátérünk a modell számításának módszereire.

3. A SEM számítása

Hasonlóan a faktoranalízishez, a SEM-et is a tapasztalati kovarianciamátrix approximációjával számoljuk. Tehát azt szeretnénk, hogy a modellből számolt kovarianciamátrix minél jobban közelítse a minta alapján számolt tapasztalati kovarianciamátrixot. Jelölje a tapasztalati kovarianciamátrixot Σ, és particionáljuk a következő módon:

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

Ekkor az egyes részek a fenti egyenletekből könnyen számolhatók. Jelölje ϕ, ψ, Θ_ε és Θ_δ a ξ, ζ, ε és δ kovarianciamártixát, I az egységmátrixot, és legyen $(I - B)$ invertálható.

$$\Sigma_{yy} = \Lambda_y(I - B)^{-1}(\Gamma\phi\Gamma^T + \psi)(I - B)^{T^{-1}}\Lambda_y^T + \Theta_\varepsilon$$
$$\Sigma_{xx} = \Lambda_x\phi\Lambda_x^T + \Theta_\delta$$
$$\Sigma_{xy} = \Lambda_x\phi(I - B)^{T^{-1}}\Lambda_y^T$$

Ezen egyenletek megoldására explicit képlet a legritkább esetekben adódik, hiszen eleve az kell hozzá, hogy pontosan annyi ismeretlenünk legyen, ahány egyenletünk. Alulhatározott esetben nem is várhatunk megoldást, ekkor más modell felírásával kell próbálkoznunk. Általában a rendszer túlhatározott, azaz egyenletek száma nagyobb. Az egyenletek számának és az ismeretlenek számának különbségét a modell szabadsági fokának nevezzük. Két fő módszer létezik a SEM modellek illesztésére. Az első egy iteratív eljárás, a második pedig a Bayesi megközelítés. Az iteratív eljárások számítják a legkisebb négyzetes illetve a Maximum Likelihood becslést. Azért kell iterációval számolni ezeket a becsléseket, mert a kovarianciamátrixoknak és az együtthatómátrixoknak minden modellben eltérő az alakjuk, így az egyes esetekben a Likelihood függvény és így az ő deriváltja is eltérő. Egy adott modell illesztésénél megtehetjük, hogy kiszámoljuk a Likelihood függvényt majd annak a deriváltjait, és megpróbáljuk kézzel megoldani a kapott egyenleteket. Ez a munka fáradságos és hosszadalmas, az iteráció viszont kevésbé időigényes, és a legtöbb esetben jó megoldást ad. A Bayesi hozzáállás szintén egy idő és számításigényes megközelítés, azonban sok esetben érdemes használni, mert a számítási nehézségek árán a paraméterek becslésén kívül további hasznos eredményeket kaphatunk.

3.1. Iteratív számítás

Számos SEM illesztésre alkalmas program készült, mint a LISREL vagy az AMOS. Ezek a programok iteratív módszerrel számolnak, és bár az alapötlet megegyezik, a konkrét számítási módok eltérőek lehetnek. A módszert röviden összefoglalva, először beállítunk egy kezdőértéket minden ismeretlennek, majd kiszámoljuk a modellből a kovarianciamátrixot. Ennek és a tapasztalati kovarianciamátrixnak valamely függvénye megadja a két mátrix távolságát. Ezt követően a kezdőértékben kiszámoljuk a távolságfüggvény parciális deriváltjait minden egyes ismeretlen szerint, és az értékeiket ez alapján megváltoztatjuk. Az egyik legkézenfekvőbb például, hogy a derivált értékek mínusz egyszeresét hozzáadjuk az ismeretlenjeinkhez. Utána az új pontban ismét deriváltakat számolunk, majd arrébb lépünk a paramétertéren. Ezt addig ismételgetjük, amíg a derivált értékeink valamely alacsony küszöb alá nem kerülnek.

Ez a módszer számos nehézséggel küzd. Egyrészt szükség van valamilyen függvényre, ami két mátrix távolságát méri, aminek a megválasztása nem egyértelmű. Másrészt egyáltalán nem biztos, hogy az optimális megoldást találja meg az eljárás, sőt lehet, hogy nem is konvergál. Ezekre a problémákra kielégítő megoldás még nem áll rendelkezésre. Még az sem tisztázott, hogy milyen feltételek mellett lesz egyértelmű a megoldás. Tudjuk, hogy a faktoranalízisben az együtthatómátrix és a faktorok vektora csak forgatás erejéig meghatározott. Ugyanez igaz a SEM-ben a külső változók mérési modelljére.

A két kovariancia mátrix távolságának, más szóval a modell illeszkedésének jóságára leggyakrabban három mérőszám valamelyikét használjuk. Ezek a Legkisebb Négyzetes eltérés (Ordinary Least Squares vagy OLS), az Általánosított Legkisebb Négyzetes eltérés (Generalized Least Squares vagy GLS) és a Maximum Likelihood mérőszám (ML). Jelölje S a tapasztalati kovarianciamátrixot, Σ a modellből számolt kovarianciamátrixot, melyek mérete $m \times m$.

$$OLS = tr(S - \Sigma)^2$$
$$GLS = \tfrac{1}{2} tr[(S - \Sigma)S^{-1}]^2$$
$$ML = log[det(\Sigma)] - log[det(S)] + trS\Sigma^{-1} - m$$

ahol tr a nyomoperátor, det a determináns és log a természetes alapú logaritmus.

A mérőszámoknak két fő célja van. Segítsék a keresést az algoritmus minden lépésében, és értékeljék a kapott megoldást. Az előzőhöz fontos, hogy gyorsan számolható legyen, hiszen a keresés minden lépésében meg kell határoznunk, és hogy jól reprezentálja a relatív távolságokat a mátrixok között. Az utóbbihoz nem annyira lényegbevágó a számítási bonyolultság, sem pedig a mérőszám viselkedése az opti-

mumtól távol, azonban jó ha tudunk róla valamilyen statisztikai tulajdonságot.

A számítási bonyolultságot tekintve a legjobb az OLS , majd a GLS, a legrosszabb pedig az ML, hiszen a modell kovarianciamátrixának determinánsát és inverzét is meg kell határoznunk a kiszámításához. Ez azért okoz problémát, mert ennek az értéke minden iterációs lépésben más és más. Statisztikai szempontból viszont az ML és a GLS kritériumok bizonyulnak jobbnak. Ugyanis a legjobb illeszkedés pontján számított valamely előbb említett mérőszámot $(N-1)$-el megszorozva a kapott érték megközelítőleg χ^2 eloszlálst követ, ahol N a minta elemszáma. A χ^2 eloszlás szabadsági foka a modell szabadsági foka. Az ML mérőszám ezen tulajdonságához szükségesek bizonyos regularitási feltételek és hogy a minta többdimenziós normális eloszlást kövessen. A GLS-re gyengébb feltételek mellett is igaz marad. Ez előnyös, hiszen arra használhatjuk, hogy a modell illeszkedését statisztikai próbának vessük alá, és különböző mérőszámokat vezessünk be az illeszkedés jóságára. Ami viszont kevésbé vonzóvá teszi az ML mérőszámot, hogy bizonyos esetekben az optimumtól távol rosszul méri a távolságot. Egy példa:

Legyen a becsülni kívánt mátrix

$$S = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

és tegyük fel, hogy az iterációs lépés során a becsült mátrixunk

$$\Sigma_1 = \begin{bmatrix} 11 & 7 \\ 7 & 5 \end{bmatrix}$$

és ebből a

$$\Sigma_2 = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$$

mátrixba lépnénk. Ekkor az OLS mérőszámunk 154-ről 4-re változik, azaz ezt a lépést elfogadná egy olyan iteráció, amely legkisebb négyzetes becslést számol. Ellenben az ML mérőszám értéke 4.513-ról 6.054-re változik, azaz a Σ_1 mátrixból nem lépne a Σ_2-be.

Ez a tulajdonság kellemetlen, hiszen az iterációt elvezetheti a megoldástól, ha még távol vagyunk tőle. Sőt, ha az optimális megoldás olyan, hogy nem illeszkedik túl jól, akkor hiába vagyunk a közelében, akkor is kivezetheti onnan az iterációt. Ez

is egy olyan probléma, amin nem tudunk úrrá lenni. Egy lehetőség, hogy a keresés elején az OLS kritériumot használjuk, ami gyors és nem esik olyan csapdába, mint az ML, majd a megoldás közelében váltunk, és az ML vagy a GLS mérőszámmal finomítjuk a kutatást.

Egy másik gond, hogy a célfüggvények nem biztos, hogy szépek a paramétertér felett. Előfordulhat, hogy lokális minimumhelyei vannak a függvénynek, esetleg bizonyos tartományokon konstans. Ilyen esetben megeshet, hogy az iteráció egy ilyen helyen ér véget, nem optimális megoldást adva. Mivel a célfüggvény tulajdonságaira sem ismert karakterizáció, azt szokás tenni, hogy több, lehetőleg távol eső pontból indítjuk az iterációt, és megvizsgáljuk, melyik milyen eredményt ad. Ha azt tapasztaljuk, hogy sok indításból ugyanazt az eredményt értük el, bízhatunk benne, hogy az lesz az optimális megoldás.

3.2. Bayesi SEM illesztés

Az előzőektől nagyban eltér a Bayesi módszerrel illesztett SEM. Több előnnyel is bír az iteratív számításokkal szemben, azonban ezekért az előnyökért hosszú számításokkal fizetünk. Ennek a módszernek a lényege, hogy az összes változónknak megadunk egy feltételes a priori eloszlást, majd a Bayes-tétel segítségével ebből és a prediktív eloszlálsból kiszámítjuk normalizáló konstans erejéig a feltételes a posteriori eloszlásokat. Ezekből aztán egy MCMC módszerrel közelítjük az együttes a posteriori eloszlását az ismeretleneknek. Így nem csak egy pontbecslést kapunk a paramétertérből (amit most is adhatunk, hiszen a Bayes-becslés az a posteriori eloszlás várható értéke), hanem egy eloszlást, amelynek segítségével árnyaltabb képet kaphatunk a látens változók struktúrájáról.

Ezen felül további előnye a módszernek, hogy azzal, hogy a paramétereknek a piori eloszlást adunk meg, plusz információt van módunk beépíteni a modellbe. Az a priori eloszlások megválasztását indokolhatják korábbi kutatások eredményei, vagy bizonyos tételek, amelyeket az alkalmazáshoz kapcsolódó tudomány biztosít számunkra. Ha nem indokolja semmi egy konkrét eloszlás használatát, akkor érdemes valamilyen nem informatív a priori eloszlást választani. Ezért a Bayesi módszernek az iterációkkal szemben akkor van létjogosultsága, ha van valamilyen fogalmunk a paraméterek vagy a megfigyeléseink eloszlásáról. Ellenkező esetben a módszer használatát az indokolhatja, hogy nem pontbecslést szeretnénk számolni. Ezen felül még egy ok lehet, hogy az előző módszerek nagyobb mintákra működnek jól, viszont a Bayesi hozzáállással illesztett SEM nem annyira érzékeny a mintanagyságra. Ezért ha nem áll rendelkezésre elég nagy minta és az iteratív módszerek nem adnak eredményt, Bayesi illesztéssel érdemes próbálkoznunk.

Markov-láncok

Mielőtt az MCMC módszerekre rátérünk, szükségünk van egy kis áttekintésre a Markov-láncokról. Legyen $(X_0, X_1, X_2, ...)$ valószínűségi változók sorozata. A valószínűségi változók lehetséges értékeit állapottérnek nevezzük. Kezdetben véges állapotterű Markov-láncokkal foglalkozunk. A valószínűségi változó sorozatot Markov-láncnak nevezzük, ha az állapottér különböző értékei közötti átmenetvalószínűségek csak a mostani állapottól függnek, azaz

$$P(X_{t+1} = s_j \mid X_0 = s_k, ..., X_t = s_i) = P(X_{t+1} = s_j \mid X_t = s_i)$$

A fenti egyenletet Markov tulajdonságnak nevezzük. Egy Markov-láncot az átmenetvalószínűségek (másnéven átmenetmag) határozzák meg, azaz annak a valószínűsége, hogy a folyamat egy adott állapotból egy másik állapotba kerül egy lépésben. Ezt $P(i, j)$-vel, vagy $P(i \to j)$-vel jelöljük.

$$P(i, j) = P(X_{t+1} = s_j \mid X_t = s_i)$$

Jelölje

$$\pi_j(t) = P(X_t = s_j)$$

annak a valószínűségét, hogy a lánc a t időpontban a j állapotban van, és foglaljuk ezeket a valószínűségeket sorvektorba, legyen ez $\pi(t)$. Úgy indítjuk el a láncot, hogy megadunk egy $\pi(0)$ kezdővektort (ez gyakran olyan, hogy csak egy nem 0 elemet tartalmaz). Ahogy a lánc halad az időben, ez a valószínűség "eloszlik" π koordinátáin. Annak a valószínűségét, hogy a lánc a $(t + 1)$ időpontban az s_j állapotban van, a Chapman-Kolmogorov egyenlet adja meg:

$$\pi_j(t + 1) = P(X_{t+1} = s_j) = \Sigma P(X_{t+1} = s_j \mid X_t = s_k)$$

Ezt az egyenletet kompaktabb formába is írhatjuk. Legyen P az átmenetvalószínűség mátrix, melynek az (i, j)-ik eleme $P(i, j)$. Ezzel a Chapman-Kolmogorov egyenlet

$$\pi(t + 1) = \pi(t)P$$

Ebből a felírásból könnyen látszik, hogy

$$\pi(t + 1) = \pi(t)P = (\pi(t - 1)P)P = \pi(t - 1)P^2 = ... = \pi(0)P^t$$

Ha definiáljuk az n-lépéses átmenetvalószínűséget a következő módon

$$P^n(i, j) = P(X_{t+n} = s_j \mid X_t = s_i)$$

akkor rögtön látszik, hogy P^t (i,j)-ik eleme éppen $P^t(i,j)$.

Azt mondjuk, hogy a Markov-lánc irreducibilis, ha minden (i,j)-re létezik n pozitív egész, hogy $P^n(i,j)$ pozitív, azaz bármely állapotból bármely más állapotba pozitív valószínűséggel el tudunk jutni. Az s_i állapot periódusa k, ha

$$k = lnko(n : P(X_n = s_i \mid X_0 = s_i) > 0)$$

Ha az állapot periódusa 1, az állapotot aperiodikusnak nevezzük. Egy irreducibilis Markov-láncban minden állapot periódusa megegyezik. Ha minden elem periódusa 1, a láncot aperiodikusnak nevezzük.

Egy Markov-lánc stacionárius eloszlásának azt a π^* eloszlást nevezzük, melyre

$$\pi^* = \pi^* P$$

azaz π^* P-nek az 1-hez tartozó baloldali sajátvektora. Ennek a szemléletes jelentése az, hogy ha elindítjuk a Markov-láncot, majd hosszú idő múlva ránézünk, akkor annak a valószínűségét, hogy a láncunk egy adott állapotban van a stacionárius eloszlás adja meg. Ezen eloszlás létezésének feltétele, hogy a lánc irreducibilis és aperiodikus legyen. Az egyértelműséghez a megfordíthatósági feltételnek kell teljesülnie minden (i,j)-re:

$$\pi^*_j P(j,i) = \pi^*_i P(i,j)$$

Vegyük észre, hogy ebből már következik $\pi^* = \pi^* P$, mert $\pi^* P$ j-ik eleme

$$(\pi^* P)_j = \Sigma \pi^*_i P(i,j) = \Sigma \pi^*_j P(j,i) = \pi^*_j \Sigma P(j,i) = \pi^*_j$$

Folytonos állapottérre úgy térhetünk át, ha egy olyan $P(x,y)$ átmenetmagot választunk, amire

$$\int P(x,y)dy = 1$$

A Chapman-Kolmogorov egyenlet folytonos esetben

$$\pi_t(y) = \int \pi_{t-1}(x)P(x,y)dy$$

A stacionárius eloszlás, pedig az a π^* eloszlás, melyre

$$\pi^*(y) = \int \pi^*(x)P(x,y)dy$$

Gibbs módszer

Tegyük fel, hogy adottak X és Y valószínűségi változók, az együttes sűrűségfüggvényük $\pi(x, y)$. Ismerjük mindkét változó feltételes eloszlását a másikra nézve

$$Y \mid X = x \text{ sűrűségfüggvénye } f(y, x)$$
$$X \mid Y = y \text{ sűrűségfüggvénye } g(x, y)$$

és szeretnénk mintát venni az együttes eloszlásból. Ezt a feladatot az előzőekhez hasonlóan egy MCMC eljárással oldjuk meg. Legyen az átmenetmag $P((x, y), A)$, aminek a sűrűségfüggvénye a következő:

$$h((u, v); (x, y)) = f(v, x)g(u, v)$$

Ez azt jelenti, hogy kiindulunk egy (x, y) pontból, először y-ból v-be lépünk ($f(x, y)$ eloszlással), majd pedig x-ből u-ba ($g(u, v)$-ből való mintavételezéssel). Ha f^*, g^* jelöli a marginális eloszlásokat:

$$f(y, x) = frac\pi(x, y)f^*(x) \qquad g(x, y) = frac\pi(x, y)g^*(x)$$

Ezzel pedig ellenőrizhetjük, hogy az így kapott Markov-láncnak valóban $\pi(x, y)$ lesz a stacionárius eloszlása:

$$\int\int h((u, v); (x, y))\pi(x, y)dxdy = \int\int f(v, x)g(u, v)\pi(x, y)dxdy =$$
$$\int f(v, x)g(u, v)[\int \pi(x, y)dy]dx = \int f(v, x)g(u, v)f^*(x)dx = \int g(u, v)\pi(x, v)dx =$$
$$g(u, v)g^*(v) = \pi(u, v)$$

Ha $f(y, x)$ és $g(x, y)$ folytonosak és szigorúan pozitívak minden x, y-ra, akkor ez a lánc irreducibilis, aperiodikus, és a stacioner eloszlása $\pi(x, y)$, aminek így egyértelműnek kell lennie. Ezen felül algoritmust kaptunk arra, hogy hogyan generálhatunk mintát az együttes eloszlásból. Ez a gondolatmenet magasabb dimenzióra is átvihető. Legyen π eloszlás $\mathbf{R}^d$-n ($d > 1$) és tegyük fel, hogy a feltételes eloszlásaival van megadva

$$X = (X_1, ..., X_d)\ \pi$$
$$X_{-i} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_d)$$
$$X_i \mid X_{-i}\ f_i(x_i, x_{-i})$$

Úgy, mint $d = 2$ esetben, $f_i : 1 \leq i \leq d$ itt is meghatározzák π-t, ha minden f_i szigorúan pozitív és folytonos. A Gibbs-sampler algoritmusa tehát:

Legyen $x = (x_1, ..., x_d)$ az együttes eloszlás tartójának egy eleme. $X^0 = x$. Ha már megvan $X^1, ..., X^n$, akkor X^{n+1}-et az alábbi módon kapjuk:

1. Végiglépkedünk i-vel az összes ($\{1, 2, ..., d\}$) koordinátán és minden lépésben az alábbi két pontot hajtjuk végre

2. w-t generálunk $f_i(x_i, x_{-i})$-ből

3. $X_i^{n+1} = w$ és $X_j^{n+1} = X_j^n$ ha $i \neq j$

A fenti algoritmussal kapott minta még nem fog teljesen megfelelni a céljainknak, mivel egyrészt időbe telik, mire a lánc megközelíti a stacionárius eloszlást, másrészt a kapott minta elemei nem lesznek függetlenek. Az első problémát úgy orvosolhatjuk, hogy a lánc első részét egyszerűen elhagyjuk. Ezt beégetésnek (burn-in) nevezik. Nincs egzakt képlet arra, hogy mekkora részét égessük be a Markov-láncnak, de egy alapötlet lehet például a következő: legyen M a legiksebb pozitív egész, melyre $P(X_M = j \mid X_0 = i) > 0$ minden i, j állapotra. Ekkor a beégetés legyen M nagyságrendű, vagy annál egyel nagyobb nagyságrendű.

Afüggetlenség elérése érdekében azt szokás tenni, hogy csak minden l-ik elemet tartjuk meg a láncból (l neve: "lag"). Arra, hoyg mekkora l-et válasszunk, szintén nincs szabály, érdemes a kapott lánc autokorreláció-függvényét megnézni, és azt az l-et választani, ahol már el tudjuk fogadni a függetlenséget.

Bayes módszer a SEM illesztésére

A Bayes becsléseknél a paraméterek (jel.: ϑ) maguk is valószínűségi változók, eloszlásuk a paramétertéren (jel.: Θ) az a priori eloszlás (jel.: Q).

- ϑ értékeit t-vel jelöljük.
- P_t az X minta eloszlása a $\vartheta = t$ feltétel mellett.
- Az $(\mathcal{X}, \vartheta)$ pár együttes eloszlása az $(\mathcal{X} \times \Theta, B_{\mathcal{X}} \times B_\Theta)$ szorzattéren az általánosított szorzatmérték $\mathbf{P}(dx, dt) = P_t(dx)Q(dt)$
- x feltétel nélküli eloszlása $P_Q = \int_\Theta P_t dQ(t)$, ezt prediktí eloszlásnak nevezzük
- Ha $\mathbf{P}$ Lebesgue-abszolút folytonos, akkor a prediktív eloszlás is, és $f_Q(x) = \int_\Theta f_t(x) dQ(t)$
- ϑ feltételes eloszlása a $X = x$ feltétel mellett az a posteriori eloszlás, jelölje ezt $Q^*(. \mid x)$

A Bayes-tétel szerint P-majdnem minden x-re $Q^*(. \mid x)$ abszolút folytonos Q-ra, és $\frac{dQ^*(.|x)}{dQ}(t) = \frac{f_t(x)}{f_Q(x)}$. Valamint, ha Q Lebesgue-abszolút folytonos és $q(t)$ a sűrűségfüggvénye, akkor az a posteriori eloszlás is abszolút folytonos és a sűrűségfüggvénye $q^*(t \mid x) = \frac{f_t(x)q(t)}{f_Q(x)}$. Tehát a Bayes-tétel megadja az a posteriori eloszlást. A paraméter Bayes becslése az a posteriori eloszlás várható értéke.

A SEM modell illesztésére ezeket az eredményeket úgy használjuk, hogy a modell paramétereire, mint valószínűségi változókra tekintünk. Célunk meghatározni az együttes a posteriori eloszlásukat. Ehhez meg kell adnunk a minta feltételes eloszlását a $\vartheta = t$ feltétel mellett, valamint a paraméterek a priori eloszlását. A minta eloszlásának leggyakrabban többdimenziós normális eloszlást választanak. A paraméterek a priori eloszlásának megválasztása mindig az adott modelltől függ. Sokszor felteszik, hogy a paraméterek függetlenek egymástól, így megadhatóak külön-külön az a priori eloszlások, és nem kell együttesen megadni az eloszlásukat. Ez könnyít a számolásokon, azonban nem minden esetben tudjuk megtenni. A hibák függetlenek a többi paramétertől, így például az ő eloszlásuk megadható a többitől külön.

Ezek után a Bayes tétel segítségével meghatározzuk az a posteriori eloszlást. A számítás megkönnyítése érdekében konjugált eloszláspárokat választanak a legtöbb alkalmazásnál. Ennek az eloszlásnak a várható értéke lesz a Bayes becslés. Ezt kiszámolni azonban nagyon bonyolult, hiszen egy többdimenziós integrálást igényel, ezért egy MCMC eljárással határozzuk meg az értékét. A Gibbs módszer éppen arra használható, hogy egy többdimenziós eloszlásból mintavételezzünk. Az algoritmusnak szüksége van a paraméterek feltételes eloszlásaira. Ezeket az előbb meghatározott együttes a posteriori eloszlásból kiszámolhatjuk. Így futtathatjuk a Gibbs módszert, ami mintát ad nekünk a paraméterek együttes a posteriori eloszlásából, és az ebből számított tapasztalati várható érték lesz a Bayes becslés.

Ennél azonban többet kaptunk ennek az eljárásnak az alkalmazásával, hiszen nem csak egy pontbecslésünk van, hanem egy mintánk is, ami alapján ezt számoltuk. Így képünk van arról, hogy milyen a paraméterek együttes eloszlása, és ezzel árnyaltabb képet kapunk a modellünkről. Például intervallumbecslést írhatunk fel az egyes paraméterekre, vagy megtalálhatjuk az extrém egyedeket az eloszlás farkainál. Ezek azok az többlet eredmények, ami miatt érdemes lehet ezt az összetettebb és időigényesebb módszert alkalmazni a SEM modell illesztésére.

4. Illeszkedési indexek

4.1. A klasszikus illeszkedési mutató

Szükségünk van valamilyen mutatóra, ami meghatározza a modell illeszkedésének jóságát. Erre a célra számos érték áll rendelkezésre, amelyek mind valamilyen módon támpontot adnak arról, hogy mennyire elfogadható a modellünk.

A klasszikus illeszkedési mutatóról már korábban szó esett az iteratív illesztések során. Ez nem más, mint az ML mérőszám megszorozva a mintaelemszámmal. Erről tudjuk, hogy az optimális illeszkedési pontban χ^2 eloszlást követ, melynek szabadsági foka a modell szabadsági foka. Ezt használhatjuk arra, hogy próbának vessük alá a modell illeszkedését. A nullhipotézisünk az, hogy a modell jól illeszkedik, és ezt akkor fogadjuk el, ha a fenti módon számolt statisztika a megfelelő szabadsági fokú χ^2 eloszlás elfogadási tartományába esik. Ez azonban nem informatív eredmény. Ha elutasítjuk a nullhipotézist, akkor valóban állíthatjuk, hogy a modell rossz. Az viszont, hogy elfogadjuk még nem jelenti, hogy a modell valóban jól illeszkedik, csak annyit, hogy nem tudtuk elutasítani a modellünket. Egy további probléma a klasszikus illeszkedési mutatóval, hogy függ a mintanagyságtól. Ha nem elég nagy a mintánk, előfordulhat, hogy olyan modelleket is elfogadunk, amik szemlátomást rosszul illeszkednek. Másrészt ha a mintánk nagyon nagy, kimondottan jól illeszkedő modelleket is elutasítunk. Ezért ez a mutató inkább csak tájékoztató jellegű és nem lehet messzemenő következtetéseket levonni ennek segítségével a modell illeszkedéséről.

Fontos, hogy alternatív modellek illeszkedését is meg tudjuk vizsgálni. Attól, hogy egy adott modellt elfogadtunk, nem záhatjuk ki, hogy létezik egy másik modell, ami az előzőnél lényegesen jobban írja le az adatainkat. Ekkor megtehetjük, hogy mindkét modellre kiszámoljuk az illeszkedési mutatókat. Előfordulhat, hogy a kettő közül az egyiket elutasítjuk, a másikat pedig elfogadjuk. Ez esetben gyakran azt mondják, hogy a két alternatív modell közötti választás egyértelmű, hiszen az a jobb, amit el tudtunk fogadni. Ez azonban helytelen, hiszen lehet, hogy az első modell illeszkedési mutatója épphogy az elfogadási tartományba esett, a másodiké pedig épphogy kicsúszott belőle. Ez esetben hiba volna azt mondani, hogy az első jobb, mint a második, hiszen valójában alig van különbség a kettő között.

Bizonyos esetekben azonban lehetőségünk nyílik két modell egyenes összehasonlítására. Ezt hierarchikus vagy más szóval egymásba ágyazott modellek esetén tudjuk megtenni. Akkor beszélünk ilyen modellekről, ha az egyik modell megkapható a másikból úgy, hogy ez utóbbiban egyes paraméterek értékeit rögzítjük. Ekkor a két

χ^2 statisztika különbsége szintén χ^2 eloszlású, a szabadsági foka pedig a két szabadsági fok különbsége. Ha ez a statisztika a megfelelő eloszlás elutasítási tartományába esik, akkor azt mondhatjuk, hogy a kisebb szabadsági fokú modell szignifikánsan jobban illeszkedik, mint a nagyobb szabadsági fokú.

Ezt lehet például arra használni, hogy megvizsgáljuk, hogy egy adott regressziós együttható fontos része-e a modellünknek. Mivel nekünk a célunk az, hogy az adatainkat minél jobban, ugyanakkor minél egyszerűbb modellel írjuk le, érdemes lehet megvizsgálni, hogy az egyes regressziós együtthatók elhagyásával lényegesen romlik-e a modell illeszkedése. Tehát azt tesszük, hogy egy együtthatót kiválasztunk és nullára állítjuk az értékét. Így egy újabb modellt kapunk, ami az eredetibe be van ágyazva. A két modellt összehasonlítjuk a fenti módon, és annak alapján dönthetünk az adott regressziós együttható szignifikanciájáról. Ha az derült ki, hogy ezen paraméter elhagyása nem ront lényegesen a modell illeszkedésén, akkor meggondolandó, hogy egyszerűen elhagyjuk a továbbiakban. Ezt követően az újonnan kapott modellben tovább vizsgálhatjuk a paraméterek szignifikanciáját ugyanúgy, mint előbb.

Ezt érdemes több ágon is elvégezni, azaz több beágyazott sorozatot készíteni egy alapmodellből a különböző paraméterek elhagyásával. Az egyes ágakon belül ezután ki tudjuk választani azokat a modelleket, amelyek a legjobban illeszkednek. Arra nincs módunk, hogy a különböző ágakon levő modelleket közvetlenül összehasonlítsuk.

4.2. Javított illeszkedési mutatók

Ahogy arról a korábbiakban szó esett, a klasszikus illeszkedési mutató nem ad elég árnyalt képet a modell illeszkedéséről. Például nagy szerepe van a mintanagyságnak is: kis minta esetén hajlamosak vagyunk elfogani a modelleinket, nagy minta esetén hajlamosak vagyunk elvetni őket. Ezen hiányosságok áthidalása érdekében számos másik illeszkedési indexet használnak, amelyek közül a legfontosabbakat az alábbiakban sorra vesszük.

Az első a Karl Joreskog féle χ^2/df mutató. A klasszikus illeszkedési mutató értékét elosztva a χ^2 eloszlásunk szabadsági fokával egy olyan értéket kapunk, ami támpontot nyújt abban, hogy mennyire jó a modell illeszkedése a szabad paraméterek számához viszonyítva. Ha ennek a mutatónak az értéke jóval 1 alá esik, az azt mutatja, hogy az illeszkedés "túl jó", vagyis lehet, hogy csak szerencsénk volt, és más minta esetén a modell nem állná meg jól a helyét. Ellenben ha a mutató értéke túl nagy, az azt jelenti, hogy az illeszkedés nem jó, és új modellt illestésével kell próbálkozunk, amelyben több út van, vagy más utak vannak. Az, hogy pon-

tosan mekkora értékeknél vonjuk le ezt a következtetést, nincs explicit megszabva. Összegezve, nagyjából 1/2 és 2 közötti értékeket szeretnénk látni ezen a mutatón, és minél közelebb van az egyhez, annál jobb. Túl kicsi érték esetén érdemes csökkenteni a paraméterek számát, túl nagy érték esetén érdemes újabb paramétereket belevenni a modellbe.

A következő mutató Bentler & Bonett normált illeszkedési mutatója (Normed Fit Index, NFI). Azt javasolták, hogy egy modell illeszkedésének a jóságát mérjük egy olyan skálán, ami a tökéletes illeszkedéstől egy ún. "null modell" illeszkedéséig fut. A null modell egy önkényes, erősen megszorított modell, ami minden korrelációt nullával becsül. Ez egy alapszintet jellemezne, amit minden használható modellnek át kell lépnie. A mutató azt mondja meg, hogy a vizsgált modellünk hova esik a null modell és a tökéletes illeszkedés közötti skálán. Formálisan a következőképpen definiálható:

$$NFI = \frac{\chi_0^2 - \chi_k^2}{\chi_0^2}$$

ahol a k index jelöli a vizsgált modellt, a 0 index pedig a null modellt. Ez a mutató akkor jelez jó illeszkedést, ha az értéke közel van az egyhez.

Az előzőnek egy kicsit módosított változata a James, Mulaik & Brett féle javított illeszkedési index (Parsimonious Fit Index, PFI). Ez a mutató figyelembe veszi azt is, hogy mekkora szabadsági fokot áldoztunk fel annak érdekében, hogy az adott illeszkedéshez eljussunk. Ennek oka az, hogy egy olyan modellt tekintünk igazán jónak, ami jól illeszkedik ugyan, de aránylag egyszerű, azaz kevés paraméter van benne és így nagyobb a szabadsági foka. Tehát ha rengeteg él van az útdiagramban és így jó illeszkedést kapunk nem olyan értékes, mintha kicsit gyengébb illeszkedést érünk el, de jóval kevesebb regressziós paraméterrel. Ezen mutató alakja az alábbi:

$$PFI = \frac{df_k}{df_0} NFI$$

A PFI index értékei az előzőhöz hasonlóan nulla és egy közöttiek, és minél magasabb az érték annál jobb.

Az utolsó, előzőekhez hasonló index Akaike információs kritériuma (Akaike's Information Criterion, AIC). Ez szemléletét tekintve hasonló az előzőhöz, mivel szintén figyelembe veszi, hogy a modell mennyi paraméterrel tudja elérni az adott illeszkedést. Egy adott modellre az információs kritérium

$$AIC = -\frac{\chi^2}{2} - q$$

Ez AIC mutató értéke mindíg negatív, és minél közelebb van a mullához, annál több információt ad az adott modell. Ezért több modell összehasonlításánál azt választjuk, amelynek a maximális ez a mutatója.

Az előzőeken felül szokás még olyan mutatókat is használni, amik a becsült és a megfigyelt kovarianciamátrixok eltérését mérik. A legegyszerűbb ilyan index egyszerűen a két mátrix átlagos négyzetes különbségéből vont négyzetgyök. Ez az RMR, teljes nevén Root-Mean-square Residual. Az átlagos eltérést mutatja a két mátrix elemei között. Legjobban akkor interpretálható, ha a mátrixok elemei nagyjából azonos skálán mozognak, például ha korrelációmátrixok.

Egy másik lehetőség, hogy a négyzetes eltérések összegét vizsgáljuk, ennek a neve Goodness-of-Fit Index, röviden GFI. Ez informatívabb akkor, ha a mátrixok elemei nem azonos skálán mozognak, vagy eltérő nagyságrendűek. Egy verziója ennek, amikor az ML illesztést használjuk, és az $S\Sigma^{-1}$ és az egységmátrix közötti négyzetes eltérések összegét számoljuk, ahol S a megfigyelt, Σ pedig a modell által becsült kovarianciamátrix. Minél hasonlóbb a két mátrix, ez a mutató annál kisebb értéket vesz fel. Ennek a javított verziója az $AGFI$, amit az előző értéket a szabadsági fokok arányával felszorozva kapunk, hasonlóan ahhoz, ahogy az NFI-ből a PFI-hez jutunk.

A fenti mutatók mindegyike más és más szempontból vizsgálja a modellek illeszkedését. Ezért nehéz dönteni, hogy melyeket használjuk, és melyek alapján válasszuk ki azt a modellt a sok közül, amit végül használni szeretnénk. Idealizált esetnek tűnik az, hogy a modellek közül majd lesz egy olyan, ami az összes többinél jobb minden egyes illeszkedési index szerint. Azonban tudnunk kell, mit akarunk elérni és mindig az adott alkalmazásnál dől el, hogy mely mutatókat részesítjük előnyben. Ezen felül szem előtt kell tartanunk azt is, hogy valószínűtlen, hogy a legtöbb mutató nagyon jó értéket adna. Ennek az oka az, hogy a célunk, hogy egy bonyolult rendszerre egy aránylag egyszerű, mégis használható modellt építsünk fel. Nem valószínű, hogy egy ilyen helyzetben rá fogunk találni minden egyes magyarázó viszonyra a jelenségen belül, már csak azért sem, mert bizonyára rengeteg a modellen kívülről érkező hatás is jelen van a leírni kívánt rendszerben.

EAz illeszkedési indexek jó támpontot adnak arra nézve, hogy modellünk elfogadható-e, és jól lehet őket alkalmazni arra, hogy különböző modelleket összehasonlítsunk. Másfelől pusztán egy modell indexeit megvizsgálva nem állíthatjuk egyértelműen, hogy az jó vagy rossz, puszán egy hozzávetőleges képet kaphatunk arról, hogy hogyan teljesít az adott mintával.

5. Útdiagramok a SEM-ben

Korábban már szó esett az útdiagarmokról, mint hasznos, jól áttekinthető eszközökről a modell felírása során. Azonban ennél sokkal hasznosabb feladatot is elláthatnak, mert számos kérdésre könnyen választ kaphatunk az útdiagramok megvizsgálásával.

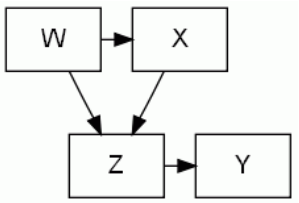
Több probléma is felmerülhet egy jelenség modellezése során. El kell tudjuk dönteni, hogy melyiket válasszuk a sok szóbajövő modell közül, amelyek mind más magyarázatot adnak a leírni kívánt jelenségre. Erre egy lehetőség a korábban tárgyalt illeszkedési mutatók vizsgálata. De mit tegyünk akkor, ha van több modell is, amelyeknek mind ugyanolyanok az illeszkedési indexei? Ezen modellek száma nagy lehet, és fontos, hogy megtaláljuk az összeset annak érdekében, hogy ki tudjuk választani a számunkra optimálisat.

Ha vannak az előző értelemben ekvivalens modelleink, van-e valamilyen karakterizációjuk, vagy közös vonásaik? Például vannak-e azonos együtthatók, vagy mindkettőben jelen levő korrelált hibák? Ha ehhez hasonló közös vonásokat tudunk felfedezni, akkor ha nem is tudunk választani egyet a modellek közül, legalább valamilyen információnak a birtokába jutunk.

5.1. A problémák felírása

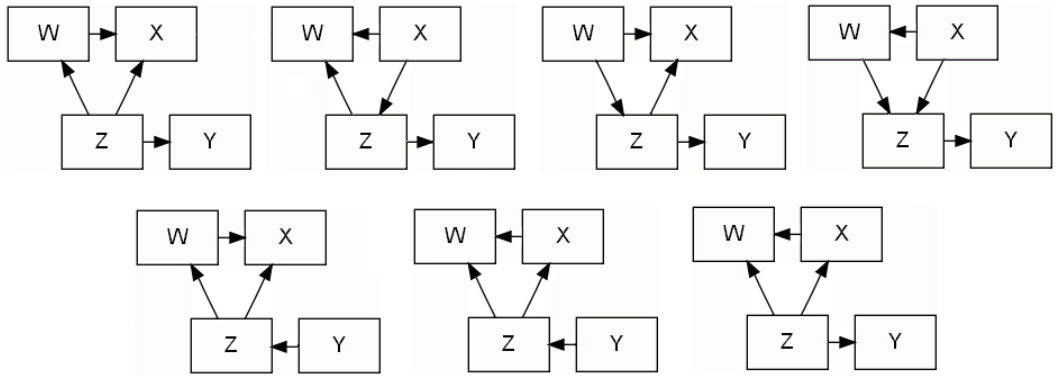
Egy útdiagramban A csúcsból B csúcsba akkor fut irányított és, ha nem-nulla együttható van a B-re felírt regresszióban A-nál. A két csúcs között akkor fut két fejű él, ha az ő hibatagjaik korreláltak. A továbbiakban az olyan útdiagramokat, melyekben nincs irányított él, irányított gráfnak nevezünk. Ha adott egy SEM, jelölje őt M, akkor $\Sigma(M)$-el jelöljük az általa implikált kovarianciamátrixot, és $G(M)$-el az útdiagramját.

Legyen M egy SEM, amelynek az útdiagramja az alábbi ábrán látható, és tegyük fel, hogy az illeszkedési mutatói mind jók.



Annak ellenére, hogy minden illeszkedési index jó, nem lehetünk biztosak ab-

ban, hogy a modellünk valóban kielégítő magyarázatot ad a leírni kívánt adatokra. Előfordulhat ugyanis, hogy még számos olyan modell létezik, amely ugyanilyen jó illeszkedést mutat az adatokon. A mi esetünkben az alábbi ábrán látható bármelyik SEM tetszőleges adathalmazra ugyanolyan jól illeszkedik, mint a fenti (azaz megegyeznek az illeszkedési mutatók).



Ahogy a fenti példából látható, előfordulhat, hogy sok egyformán illeszkedő modell létezik, ami egy magyarázó elemzés elkészítésekor rengeteg gondot okozhat. Ha az elemző nem ismeri az övével azonosan illeszkedő modelleket, az könnyen rossz következtetésekhez vezethet a leírandó jelenséggel kapcsolatban. Ezért létfontosságú, hogy rendelkezésre álljon egy módszer, amivel megtalálható az összes ilyen modell.

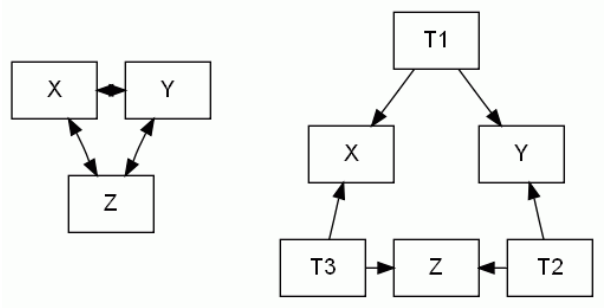
Definíció

Legyen O mért változók egy csoportja G_1 és G_2 diagramokban. Azt mondjuk, hogy G_1 és G_2 koveriancia ekvivalensek O felett, ha bármely M SEM-hez, melyre $G(M) = G_1$ létezik egy M' SEM, melyre $G(M') = G_2$, és $\Sigma(M')$ O-nak megfelelő részmátrixa megegyezik $\Sigma(M)$ O-nak megfelelő részmátrixával, és fordítva.

Egyszerűbb szavakkal a fenti definívió azt mondja, hogy bármely O geletti kovarianciamátrix, amit G_1 parametrizálása generál, generálható G_2 parametrizálásával és fordítva. Ha G_1 és G_2 minden változója O-ban van, akkor röviden azt mondjuk, hogy kovariancia ekvivalensek. Ha két kovariancia ekvivalens modell ugyanolyan megfelelő a háttérismereteink alapján és ugyanakkora a szabadsági fokuk, akkor nem tudunk különbséget tenni köztük, hiszen az adatok sem segítenek a megkülönböztetésükben, mivel minden mutatójuk megegyezik. Ezért nagyon fontos, hogy az összes ilyen modellt.

Modellek ilyen osztályát a továbbiakban kovariancia-ekvivalens osztálynak nevezünk. Ha a diagramokban nincs kétfejű nyíl vagy irányított kör, akkor egyszerű kovariancia-ekvivalens osztálynak hívjuk a modellek ilyen csoportját.

Annak eldöntése, hogy mely modellek tartoznak egy osztályba koránt sem egyszerű. Erre az alábbi ábra szolgál egy egyszerű példával.

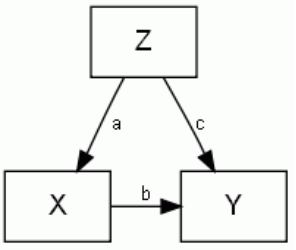


$$\Sigma = \begin{bmatrix} 1 & 0,99 & 0,99 \\ 0,99 & 1 & 0,99 \\ 0,99 & 0,99 & 1 \end{bmatrix}$$

Első ránézésre úgy tűnhet, hogy a két modell kovariancia ekvivalens $\{X, Y, Z\}$ felett, de nem ez a helyzet. Mert létezik egy SEM amelynek a diagramja az ábra bal oldalán levő diagram, és Σ a kovarianciamátrixa, de nincs olyan SEM, melynek a diagramja a jobb oldali, a kovarianciamátrixa pedig Σ $\{X, Y, Z\}$ felett.

Ha sikerült meghatároznunk az ekvivalens modelleket, akkor már bíztatóbb a helyzet. Igaz, hogy előfordulhat, hogy sok ekvivalens modellünk van, melyek közül nem tudunk választani, de előfordulhat, hogy közös vonásokkal rendelkeznek ezek a modellek, például van olyan él, amely mindig azonos irányítással szerepel. Ez mindenképpen informatív eredmény, hiszen ekkor nagyobb magabiztossággal állíthatjuk, hogy a szóban forgó két változó közti hatás egyirányú.

A következő felmerülő probléma a regressziós és struktúrális együtthatókkal kapcsolatos. Ismert, hogy ahhoz hogy két változó regressziójában a regressziós együtthatót interpretálhassuk úgy, mint egyik változónak a másikra gyakorolt hatása, nem szabad léteznie egy harmadik "zavaró" változónak, ami mindkét előzőre hatással van.



Tehát a fenti ábrán ha X és Y közötti b regressziós együtthatót szeretnénk X Y-ra gyakorolt hatásának nevezni, nem létezhet egy Z változó, mely mindkettőre hatással van. Ez könnyen ellenőrizhető:

$$cov(X,Y) = bD^2(X) + acD^2(Z)$$
$$\frac{cov(X,Y)}{D^2(X)} = b + \frac{acD^2(Z)}{D^2(X)}$$

Így ha csak Y és X között írjuk fel a regressziót, akkor X együtthatója csak akkor lesz torzítatlan, ha a vagy c valamelyike nulla. Sőt, a torzítás tetszőleges előjelű és nagyságrandű lehet.

Másfelől, ha Y-t is belevesszük a regresszióba, akkor

$$cov(X, Y \mid Z) = cov(X,Y) - \frac{cov(X,Z)cov(Z,Y)}{D^2(Z)} =$$
$$bD^2(X) + acD^2(Z) - aD^2(Z)(ab+c) = b(D^2(X) - a^2D^2(Z))$$
$$D^2(X \mid Z) = D^2(X) - \frac{cov(X,Y)^2}{D^2(Z)} = D^2(X) - a^2D^2(Z)$$

Ehhez felhasználtuk, hogy

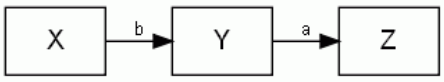
$$cov(X,Z) = aD^2(Z) \qquad cov(Z,Y) = (ab+c)D^2(Z)$$

Az előzőekből kapjuk, hogy

$$\frac{cov(X,Z|Z)}{D^2(X|Z)} = b$$

így az X együtthatójára adott becslésünk torzítatlan lesz, amennyiben "Z"-t is figyelembe vesszük.

Az imént tehát láttuk, hogy ha létezik egy olyan harmadik változó amit nem ismerünk, és hatással van mindkét változóra amelyek között regressziót írunk fel, akkor a harmadik ún. "zavaró" változó kihagyása torzítást eredményez a regressziós együttható becslésében. Másfelől előfordulhat az is, hogy egy nem "zavaró" változó bevezetése is hasonló torzítást eredményez. Erre egy egyszerű példa a következő:



Itt ha csak X és Y változókra írjuk fel a regressziót, akkor

$$cov(X,Y) = bD^2(X)$$

ezért a regressziós együttható becslése torzítatlan. Azonban ha már Z-t is belevesszük, akkor

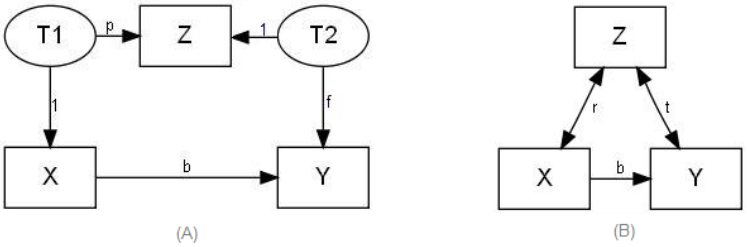
$$\frac{cov(X,Z|Z)}{D^2(X|Z)} = b\frac{D^2(Z) - a^2D^2(Y)}{D^2(Z) - b^2a^2D^2(X)}$$

mivel

$$cov(Y, Z) = aD^2(Y)$$
$$cov(X, Z) = abD^2(X)$$

Tehát ha mindhárom változó jelen van, torzított becslést kapunk b-re, amely ugyan előjelben jó, de abszolút értékben nem. Vegyük észre, hogy a becslésünk akkor és csak akkor nulla, ha $b = 0$.

Nézzünk egy újabb példát!



Az (A) ábrán levő SEM hibaváltozóit jelölje ε_X, ε_Y és ε_Z, a (B) ábrán levő SEM hibaváltozóit pedig ε'_X, ε'_Y és ε'_Z. Az (A) ábrán két látens zavaró változó látható, $T1$ és $T2$, amelyek korrelálatlanok. Bármely SEM, aminek ez az útdiagramja, átírható egy másikká, melynek a diagramja (B) a következő választásokkal:

$$r = pD^2(T1)$$
$$t = fD^2(T2)$$
$$D^2(\varepsilon'_X) = D^2(\varepsilon_X) + D^2(T1)\ D^2(\varepsilon'_Y) = D^2(\varepsilon_Y) + f^2D^2(T2)$$
$$D^2(\varepsilon'_Z) = D^2(\varepsilon_Z) + p^2D^2(T1) + D^2(T2)$$

Ez fordítva nem lehetséges, azaz nem lehet minden modellt, amelyben korrelált hibák vannak $(X \leftrightarrow Y)$, egy látens változós modellre cserélni, amelyben a hibák korrelálatlanok úgy, hogy bevezetünk egy rejtett T változót, amely X-nek és Y-nak őse $(X \leftarrow T \rightarrow Y)$.

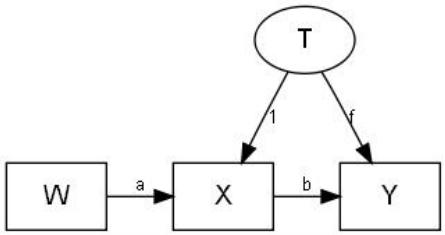
Visszatérve a (B) útdiagramhoz, láthatjuk, hogy az előzőhöh hasonlóan ha csak X és Y változók között írjuk fel a fegressziót, akkor b-re torzítatlan becslést kapunk, ám, ha Z-t is bevesszük a regresszióba, akkor

$$\frac{cov(X,Z|Z)}{D^2(X|Z)} = \frac{cov(X,Y)D^2(Z) - cov(X,Z)cov(Y,Z)}{D^2(X)D^2(Z) - cov(X,Z)^2} = \frac{bD^2(X)D^2(Z) - r(rb+t)}{D^2(X)D^2(Z) - r^2} = b - \frac{rt}{D^2(X)D^2(Z) - r^2}$$

azaz nem lesz torzítatlan a b-re kapott becslés, hacsak nem $r = 0$ vagy $t = 0$. Sőt, az is előfordulhat, hogy más előjelű lesz a becslésünk, mint a becsülni kívánt érték. Ezen felül, ha $b = 0$ X és Y regressziójában, azaz valójában nincs magyarázó viszony a két változó között, az együttható nem nullává válik, ha Z-t is bevezetjük.

Általánosan elfogadott a SEM alkalmazása során, hogy jobb bevezetni változókat, mint kihagyni. Ennek oka az, hogy azzal, hogy bekerült egy új változó, annak a hatását is kezelni tudjuk és így kiküszöbölhetjük a torzítást. De mint az előző példákból láthattuk, ez a hozzáállás nem megfelelő, mert előfordulhat, hogy éppen ezzel egy torzítatlan becslés torzítottá válik.

Végül meg kell jegyeznünk azt is, hogy bizonyos esetekben a keresett együtthatót semmilyen regresszióval sem tudjuk torzítatlanul becsülni.



Ebben a SEM-ben

$$cov(X,Y) = bD^2(X) + fD^2(T)$$

tehát az X együtthatója az $X \to Y$ regresszióban nem torzítatlan, és W bevezetése sem segít a helyzeten, mert

$$\frac{cov(X,Z|W)}{D^2(X|W)} = b + \frac{fD^2(T)}{D^2(X)-a^2D^2(W)}$$

Ennek ellenére létezik torzítatlan becslés, mégpedig

$$\frac{cov(Y,W)}{cov(X,W)} = \frac{abD^2(W)}{aD^2(W)} = b$$

Összegezve tehát a következő kérdések adódnak:

1. Egy adott SEM esetén melyek a vele kovariancia ekvivalens modellek? Van-e ezeknek valamilyen közös jellemzőjük?

2. Ha Y-t W változókkal illesztjük, $x \in W$, mely SEM-ekben lesz X regressziós együtthatója torzítatlan becslése a struktúrális együtthatónak, melyet az $X \to Y$ és reprezentál?

3. Ha Y-t W változókkal illesztjük, $x \in W$, mely SEM-ekben lesz X regressziós együtthatója 0, ha az $X \to Y$ élnek megfelelő struktúrális együttható 0?

4. Adott egy sem $X \to Y$ éllel, az él súlyát jelölje b. Van-e a megfigyelt változóknak egy olyan W halmaza, melyre $x \in W$ és ha Y-t W változókkal illesztjük, akkor X együtthatója b-nek torzítatlan becslése lesz?

5. Adott SEM-ben egy b együtthatóra létezik-e egy $h(S)$ függvény, amely b-nek torzítatlan becslését adja? (Ahol S a tapasztalati kovarianciamátrix)

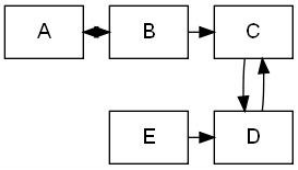
Ezen kérdésekre az utolsó kivételével választ tudunk adni.

5.2. d-szeparáció

A fenti problémák megválaszolásához a "d-szeparáció" fogalmának bevezetésére van szükségünk. Ez az útdiagramok csúcshalmazain értelmezett tulajdonság. A könnyebb áttekinthetőség kedvéért egy példát fogunk felírni. vegyünk egy M SEM-et öt változóval, melyeket A, B, C, D, E jelöl, a hibáikat pedig $\varepsilon_A, \varepsilon_B, \varepsilon_C, \varepsilon_D, \varepsilon_E$. A modell a következő:

$$\begin{aligned} A &= \varepsilon_A \\ B &= \varepsilon_B \\ C &= \beta B + \delta D + \varepsilon_C \\ D &= \gamma C + \eta E + \varepsilon_D \\ E &= \varepsilon_E \end{aligned}$$

A hibák páronként korrelálatlanok, kivéve A és B hibatagjait, köztük a korrelációs együttható α. Ezek alapján $G(M)$ a következő:



Ahhoz, hogy definiálni tudjuk a d-szeparációt, előbb tisztáznunk kell a terminológiákat, amiket az útdiagramoknál használunk. Két féle él húzódhat X és Y csúcsok között, irányított illetve kétfejű. Mindkét esetben azt mondjuk, hogy X és Y az él *végpontjai*, és X és Y *szomszédosak*. Egy $X \to Y$ irányított él esetén azt mondjuk, hogy X az él *farka*, Y az él *feje*, X az Y *szülője*, Y az X *gyermeke*.

Egy U *irányítatlan út* X és Y között élek egy olyan $\{E_1, E_2, ..., E_m\}$ sorozata, melyre E_1 egyik végpontja X, E_m egyik végpontja Y és a sorozatban minden szomszédos E_i, E_{i+1} élre E_i egyik végpontja megegyezik E_{i+1} egyik végpontjával. A példánkban $A \leftrightarrow B \to C \leftarrow D$ egy irányítatlan út.

Egy P *irányított út* X és Y között irányított élek egy olyan $\{E_1, E_2, ..., E_m\}$ sorozata, melyre E_1 farka X, E_m feje Y és a sorozatban minden szomszédos E_i, E_{i+1} élre E_i feje megegyezik E_{i+1} farkával. A példánkban $B \to C \to D$ egy irányított út.

Egy csúcs *előfordul* egy útban, ha létezik egy él az útban, amelynek a csúcs valamelyik végpontja. Egy út *aciklikus*, ha minden csúcs legfeljebb egyszer fordul elő benne. A példánkban $C \to D \to C$ nem aciklikus.

Egy X csúcs *őse* egy Y csúcsnak és Y *leszármazottja* X-nek, ha létezik P irányított út X-ből Y-ba, vagy $X = Y$.

Egy X csúcs ütköző egy U úton akkor és csak akkor, ha U tartalmaz egy részutat az alábbiak közül: $Y \leftrightarrow X \leftrightarrow Z$, $Y \to X \leftrightarrow Z$, $Y \to X \leftarrow Z$ vagy $Y \leftrightarrow X \leftarrow Z$, különben X *nem-ütköző* U-n. A példánkban C ütköző a $B \to C \leftarrow D$ úton, de nem-ütköző a $B \to C \to D$ úton.

Azt mondjuk, hogy X *őse* egy Z csúcshalmaznak, ha Z valamely elemének őse.

Definíció

X,Y és Z diszjunkt csúcshalmazokra X *d-kapcsolódik* Y-hoz Z-t feltéve akkor és csak akkor, ha létezik egy aciklikus (irányítatlan) U út valamely $x \in X$ és $y \in Y$ között úgy, hogy minden ütköző U-n Z őse, és minden nem-ütköző U-n nincs benne Z-ben.

Definíció

X,Y és Z diszjunkt csúcshalmazokra X *d-szeparált* Y-hoz Z-t feltéve akkor és csak akkor, ha X nem d-kapcsolódik Y-hoz feltéve Z-t.

A fenti példánkban a $C \to D \to E$ út d-kapcsolja C-t és E-t feltéve $\varnothing$-t, valamint A-t, B-t vagy $\{A, B\}$-t. $E \to D \leftarrow C$ d-kapcsolja E-t és C-t feltéve D-t, $\{D, A\}$-t, $\{D, B\}$-t, $\{D, A, B\}$-t. A példánkban levő összes d-szeparáció reláció:

- A és C feltéve B, $\{B, D\}$, $\{B, E\}$, $\{B, D, E\}$
- A és D feltéve B, $\{B, C\}$, $\{B, E\}$, $\{B, C, E\}$
- A és E feltéve $\varnothing$, B, $\{B, C\}$, $\{B, D\}$, $\{B, C, D\}$, $\{C, D\}$
- B és E feltéve $\varnothing$, $\{C, D\}$

Tétel

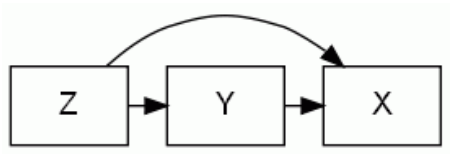
Ha M egy SEM, és $G(M)$-ben X és Y d-szeparáltak feltéve Z, akkor $cov(X, Y \mid Z) = 0$ $\Sigma(M)$-ben.

Tétel

Ha X és Y nem d-szeparáltak G-ben, akkor létezik egy M SEM, melyre $G(M) = G$

és $cov(X, Y \mid Z) \neq 0$ $\Sigma(M)$-ben.

Az első tétel azt mondja ki, hogy a d-szeparáció egy G útdiagramban elégséges feltétel arra, hogy minden G diagramú SEM-ben az X, Y-nak a feltételes kovarianciája Z-re 0. A második tétel szerint a d-szeparávió szükséges feltétele annak, hogy 0 legyen a feltételes kovariancia. Ez persze nem mondja azt, hogy nincs olyan M SEM, amelyben nulla a feltételes kovariancia olyan változók között, amelyek nem d-szepaáltak. Például vegyük az alábbi diagramot.



A kapott modellünk legyen a következő:

$$X = 0,3Y + 0,6Z + \varepsilon_X$$
$$Y = -2Z + \varepsilon_Y$$
$$Z = \varepsilon_Z$$

Ez esetben $cov(X, Y) = 0$, pedig X és Y nem d-szeparáltak $\varnothing$-ra. Ez azért van mégis így, mert a regressziós együtthatók pont kiejtik a szórásnégyzeteket. De az előző tétel értelmében létezik olyan M SEM, melynek a fenti a diagramja és nem lesz 0 X és Y kovarianciája. Ezen felül megmutatták (Spirtes et. al, 1993), hogy azok a paraméterek, melyek nulla kovarianciát eredményeznek olyan változók között, melyek nem d-szeparáltak, Lebesgue 0-mértékűek a paramétertéren.

5.3. A d-szeparáció alkalmazása

Az előbb tárgyalt d-szeparáció hasznos eszköznek bizonyul a korábban felvetett kérdések megválaszolásához. Az első felmerülő probléma a kovariancia ekvivalens modellek felismerése, illetve megkeresése volt. Azaz, ha valamely M SEM-hez létezik egy M' SEM, amelynek más a diagramja, de ugyanakkora szabadságfokú és az illeszkedési mutatói is megegyeznek, akkor a második modell ugyanolyan jó, mint az első és ezért nem tudunk választani a kettő közül. Az ilyen modellek voltak kovariancia ekvivalens modellek, és célunk az volt, hogy egy adott M SEM-re megtaláljuk az összes vele kovariancia ekvivalens modellt.

Először nézzük azokat a modelleket, melyekben nincs irányított kör, illetve korrelált hiba.

Definíció

G_1 és G_2 *d-szeparáció ekvivalensek*, ha minden X, Y, Z csúcshalmazra X akkor és csak akkor d-szeparált Y-tól feltéve Z-t G_1-ben, ha G_2-ben is.

Tétel

Legyenek G_1 és G_2 irányított gráfok. G_1 és G_2 akkor és csak akkor kovariancia ekvivalensek, ha d-szeparáció ekvivalensek.

Ezzel választ adtunk a fenti kérdésre. Ha van olyan élfelcserélési eljárás, mely egy M SEM-ből egy M' SEM-be vezet, akkor ez a két modell kovariancia ekvivalens. Azonban ezt körülményes ellenőrizni, de ebben a segítségünkre lesz a következő tétel. Azt modjuk, hogy X *unshielded collider* G aciklikus gráfban akkor és csak akkor, ha vannak G-ben $A \to X \leftarrow B$ élek és A nem szomszédos B-vel.

Tétel

Két aciklikus irányított gráf akkor és csak akkor d-szeparáció ekvivalensek, ha megegyeznek a csúcsaik, a csúcsok szomszédsági viszonyai és az unshielded colliderei.

Ebből azonnal következik hogy két kovariancia ekvivalens aciklikus irányított gráfú SEM-nek egyanakkora a szabadsági foka. Valamint ezzel egy egyszerű módszert kaptunk a kovariancia ekvivalens modellek megtalálálára. Ahhoz, hogy irányított kört és korrelált hibát is tartalmazó modellekre is ilyen eredményt kapjunk, szükségünk van még egy definícióra. Legyen $O \subseteq V(G_1), V(G_2)$. Azt mondjuk, hogy G_1 és G_2 d-szeparáció ekvivalensek O felett, ha minden diszjunkt X, Y, Z O-beli halmazra, X d-szeparált Y-tól feltéve Z-t G_1-ben akkor és csak akkor, ha ugyanez igaz G_2-ben is.

Tétel

Ha G_1 és G_2 kovariancia ekvivalensek O felett, akkor d-szeparáció ekvivalensek O felett.

Ennek a megfordítása nem igaz, ezt láttuk az előző bekezdés második példájában. Azzal, hogy megtaláltuk a kovariancia ekvivalens modelleket, egyúttal választ kaptunk arra is, hogy egy adott esetben milyen közös vonásai vannak az azonos ekvivalencia osztályban levő modelleknek.

Áttérhetünk a regressziós kérdések megválaszolására. Adott egy SEM G dia-

grammal. Jelöljük $G\backslash\{X \to Y\}$-el azt a diagramot, amelyet úgy kapunk G-ből, hogy elhagyjuk az $X \to Y$ élt. Az első kérdésünk az volt, hogy ha egy SEM-ben Y-t W változókkal illeszjük, $X \in W$, mely esetekben lesz a regressziós együttható torzítatlan becslése a struktúrális paraméternek? Azt tudjuk mondani, hogy ha W-ben nincs Y-nak leszármazottja és X d-szeparált Y-tól W-re $G\backslash\{X \to Y\}$-ben, akkor a becslés torzítatlan lesz. Ha ez nem teljesül, akkor a legtöbb esetben rossz lesz a becslés. Azaz, ha például van $X \leftrightarrow Y$ él, vagy ha X az Y-nak leszármazottja, akkor szinte biztos, hogy torzított becslést kapunk. Az utolsó előtti kérdésre is választ adhatunk ez alapján, mivel ha létezik W változóhalmaz, amelyben nincs leszármazottja Y-nak, és X d-szeparált Y-tól feltéve $G\backslash\{X \to Y\}$-t, akkor ha W-vel illesztjük Y-t, a paraméter becslése torzítatlan lesz.

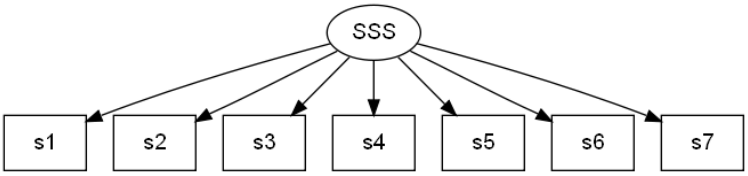
A következő kérdésünk, hogy ha az előzőhöz hasonlóan Y-t W változókkal illeszjük, $X \in W$, mikor lesz X regressziós együtthatója 0, ha az $X \to Y$ élnek megfelelő struktúrális együttható 0? Ha X és Y d-szeparáltak $W\backslash X$-re nézve, akkor nullát kapunk a becslés során. Ha ez nem teljesül, akkor majdnem minden nem-nulla lesz a paraméter becslése, még ha "valójában" nincs is él a két változó között.

A fenti állítások segítségével egy sokkal tisztább képet kaphat az alkalmazó a modellezési eljárásáról. Amikor modellt írunk valamilyen rendszer leírása érdekében, szinte soha nem lehetünk biztosak abban, hogy sikerült-e megtalálni a legjobban illeszkedőt, vagy a kívánalmainknak legmegfelelőbbet. Éppen ezért fontos minél jobban tájékozódni arról, hogy az aktuális modellünk miben teljesít jól és miben rosszul, hogy sikerülhessen egy minél jobb tulajdonsággal bíró modellhez eljutni.

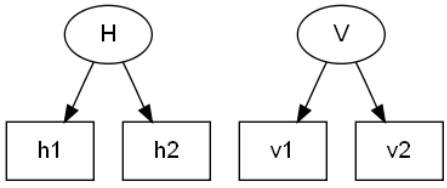
6. Egy példa SEM illesztésre

Az utolsó fejezetben egy példán mutatom be a SEM modell használatát. A példában a 2003-ban lefolytatott "ADE-2003" kutatás[1] keretében felvett országos reprezentatív mintával dolgoztam.

Tegyük fel, hogy a magyar népesség alkoholfogyasztási szokásait szeretnénk modellezni egy bizonyos szempontból. A feltevésünk az, hogy egy olyan embernek, aki kalandvágyó, mások az alkohol fogyasztási szokásai, mint egy otthon ülő típusnak. Azt, hogy valaki mennyire kalandvágyó egy ún. szenzoros élménykeresési skálán mérhetjük meg, amelynek egy rövidített változata megtalálható az adatbázisban. Ez a rövid változat 7 kérdéses, így kezdő lépéseként van ez a hét mért változónk a modellben, melyeket egy közös látens változó magyaráz.



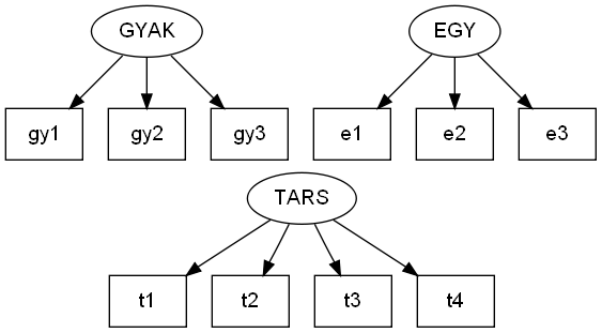
Tegyük fel, hogy ezen felül szeretnénk más magyarázó változókat is. Például azt szeretnénk megvizsgálni, hogy ha valaki helyteleníti, vagy éppen veszélyesnek tartja az alkoholfogyasztást, akkor ez hogyan hat az ő ivási szokásaira. Jogosnak tűnik az a feltételezés, hogy ezek a tényezők hatással vannak arra, hogy valaki mennyit iszik, ezért ezeket a változókat is beletesszük a modellbe. Mindkét attitűdre vonatkozóan két kérdés található az adatbázisban, így bevezetünk további két látens változót, amelyek ezeken a mért változókon ülnek.



Ezek után megválasztjuk azokat a mért adatokat, és magyarázandó látens változókat, amelyek az alkoholfogyasztási szokásokat mutatják. Legyen például egy olyan változó, ami az alkohol fogyasztásának gyakoriságát méri. Erre vonatkozóan 3

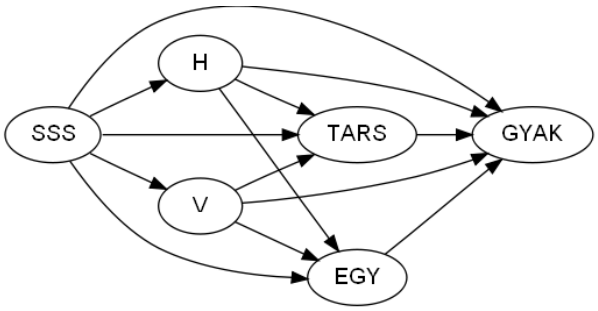
[1]Demográfiai folyamatok társadalmi beágyazottsága program
Finanszírozó: A Nemzeti Kutatási és Fejlesztési Program
Elekes Zsuzsanna, Paksi Borbála

mért adatot választottunk az adatbázisból. Továbbá tegyük fel, hogy minket elemzőként az is foglalkoztat, hogy a kérdezett jellemzően egyedül iszik, vagy inkább olyankor, amikor társaságban vagy szórakozóhelyen van. Az előbbire vonatkozóan 3, az utóbbira vonatkozóan 4 mért megfigyelésünk van, tehát az korábbiakhoz hasonló módon felírjuk a mérési modellünk utolsó részét.



A mérési modellünk ezzel készen van, most a strukturális modellen a sor. Ehhez azt kell eldöntenünk, hogy egy adott változóval melyeket akarjuk magyarázni a többi közül. Más szóval, ha az adott változó értéke megváltozna, mely másik változókban várunk változást? Az élménykeresési változóból ezért kezdetben minden másikba vezetünk élt. Egy kalandvágyó ember valószínűleg más alkoholfogyasztási szokásokkal rendelkezik, mint egy otthon ülő típus, és valószínűleg más a véleményük az ivás veszélyességéről és helytelenítéséről is. Ezt persze nem tudjuk előre, de kezdetben igyekszünk az összes elképzelhető magyarázó viszonyt bevenni a modellbe, és majd a végén kiderül, melyek lényegesek.

Hasonló gondolatmenet alapján a helytelenítés és veszélyészlelés változókból minden alkoholfogyasztási változóba vezetünk élt. Végül, ha valaki szeret mondjuk társaságban inni, akkor valószínűleg gyakrabban is fogyaszt alkoholt, valamint ugyanez érvényes a magányos ivásra, így az alkohol fogyasztásának gyakoriságát mérő változóba élt húzunk mindkettőből. Így elkészült a strukturális modellünk.



Most, hogy felírtuk a modellt, ezt illeszthetjük, majd megvizsgáljuk, hogy a különböző élek elhagyásával hogyan változik a modell illeszkedése. Ezek a modellek

egymásba ágyazottak, így módunk van összehasonlítani őket. Célunk, hogy megtaláljuk azt a legegyszerűbb modellt, ami még mindig jól magyarázza az adatokat. A mintánk 2400 elemű, ezért egyúttal módunkban áll validálni is a kapott eredményt. Ugyanis a modell finomítását a mintánk alapján végezzük, így előfordulhat, hogy ami a mi mintánk szerint nagyon jól illeszkedik, más mintára nem állja meg olyan jól a helyét. Ezért szokásos technika, hogy ha a minta elég nagy, kettévágjuk és az egyik felét használjuk arra, hogy finomítsuk a modellt, majd ha ezzel készen vagyunk, ezt illesztjük a másik felére. Így ha jó modellt kaptunk, a másik felén is ugyanolyan jól kell teljesítsen, mint az előzőn. Az összes illesztést az R statisztikai program segítségével végeztem el.

	χ^2	GFI	RMR	NFI	PFI	
Második Modell	1746,7	0,8688	0,0844	0,8429	0,7225	
Melyik él marad ki	χ^2	GFI	RMR	NFI	PFI	χ^2 Eltérés
SSS -> H	*	*	*	*	*	
SSS -> V	1817,6	0,8661	0,1030	0,8365	0,7090	71,6
SSS -> EGY	1746,6	0,8688	0,0843	0,8429	0,7144	0,6
SSS -> TARS	1760,8	0,8641	0,0784	0,8416	0,7133	14,8
SSS -> GYAK	1746,1	0,8689	0,0842	0,8429	0,7145	0,1
H -> EGY	1808,1	0,8638	0,0893	0,8374	0,7097	52,1
H -> TARS	1815,0	0,8629	0,0880	0,8367	0,7092	69,0
H -> GYAK	*	*	*	*	*	
V -> EGY	1762,9	0,8677	0,0866	0,8414	0,7132	16,9
V -> TARS	1758,2	0,8676	0,0849	0,8418	0,7136	12,2
V -> GYAK	1746,3	0,8688	0,0842	0,8304	0,7038	0,3
EGY -> GYAK	1777,8	0,8569	0,0709	0,8401	0,7121	31,8
TARS -> GYAK	1791,1	0,8545	0,0706	0,8389	0,7111	45,1

A fenti táblázatban láthatóak az eredmények. A legfelső sor tartalmazza az előbb felírt modell illeszkedési mutatóit. A χ^2 statisztika értéke hatalmas, de ezen nem is szabad meglepődnünk, hiszen a mintaelemszámunk is igen nagy. Ezért informatívabb lehet a többi mutató. Az RMR mutató a becsült és a tapasztalati kovarianciamátrix elemeinek átlagos eltérését mutatja. Mivel jelen esetben standardizált változókkal dolgoztam, ezért az értékek a [0,1] intervallumba esnek, ezért az eltérés elég jelentős. A GFI index az előzőhöz hasonló, csak itt a becsült és a tapasztalati értékek aránya látható. Ez a mutató sem túl meggyőző. Az NFI érték az egyedüli, ami bizalomra ad okot, hiszen ez az illeszkedés minőségét egy [0,1] skálán mutatja, ahol 0 a legrosszabb, 1 pedig a legjobb érték. Ez az index győz meg minket arról, hogy egy aránylag jó modellt sikerült felírnunk.

A táblázat második része a modell egyszerűsítéséhez szükséges vizsgálatok eredményeit tartalmazza. Minden strukturális élt elhagyunk, és megvizsgáljuk, hogyan változik a modell illeszkedése. Ha nem romlik jelentősen, akkor az adott élt ki is lehet hagyni a modellből. Előfordult, hogy egyes élek elhagyásával kapott modell illesztésekor nem talált megoldást az iteráció több indítóértékre sem. Az ezeknek az éleknek megfelelő sorok nem tartalmaznak adatot. Ez például az $SSS \rightarrow H$ él esetében nem is meglepő, mivel eredetileg csak az SSS változónk volt exogén, de így a H is azzá vált. Viszont ez utóbbi csak két mért adaton ül, ami nagyon kevés, és ilyen esetekben gyakori, hogy nem talál megoldást az algoritmus. Ilyen megvilágításból már szerencsésnek mondható, hogy az $SSS \rightarrow V$ él elhagyásával nem lépett fel ilyen gond.

Ebből a részből az utolsó oszlop igazán lényeges. Itt látható az aktuális modell és az első modell χ^2 értékeinek különbsége, ami szintén χ^2 eloszlású, egy szabadsági fokkal. Ránézésre látszik, hogy három élt is el lehet hagyni a modellből, az $SSS \rightarrow EGY$, az $SSS \rightarrow GYAK$ és a $V \rightarrow GYAK$ éleket. A többi él elhagyásával viszont drasztikusan romlik a modell illeszkedése. Ez nem feltétlenül jelenti azt, hogy mindhárom élt egyszerre elhagyva is hasonló eredményt látunk, de jelen esetben szerencsénk van, ahogy azt a második táblázat is mutatja.

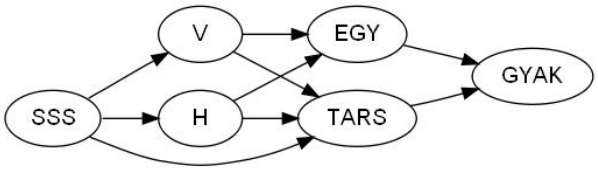
	χ^2	GFI	RMR	NFI	PFI	
Első modell	1746	0,8688	0,0842	0,8429	0,7077	
Melyik él marad ki	χ^2	**GFI**	**RMR**	**NFI**	**PFI**	χ^2 **Eltérés**
V -> GYAK és SSS -> GYAK és SSS -> EGY	1746,7	0,8688	0,0844	0,8492	0,7225	0,7
V -> GYAK és SSS -> GYAK és SSS -> EGY és H -> GYAK	1769,4	0,8675	0,0847	0,8408	0,7287	23,4
Hipotézis: EGY -> GYAK = TARS -> GYAK	1763,8	0,8684	0,0849	0,8413	0,7251	17,8

Azt látjuk, hogy a három él együttes elhagyásával gyakorlatilag egyáltalán nem változik meg a modell illeszkedése. Érdemes viszont megnézni a PFI mutatót, ami az NFI index egy korrigált változata, ami figyelembe veszi a modell szabadsági fokát. Ez a mutató jelentősen megnőtt, hiszen a modellünk ugyanazt tudja, mint a korábbi, viszont már hárommal kevesebb paramétert használ.

Az előzőekben még volt egy él ($H \rightarrow GYAK$), amelyiket nem tudtuk tesztelni. Most, hogy már kevesebb paramétert kell a modellnek becsülnie, előfordulhat, hogy ha elhagyjuk ezt is, már kapunk eredményt. Valóban ez a helyzet, és bár a klasszikus χ^2 statisztikát használva nem hagynánk ki ezt az élt, az előbb említett PFI mutató javult, azaz ha figyelembe vesszük azt is, hogy kevesebb paramétert használunk, megéri elhagyni ezt az élt is.

Ezen felül használhatjuk a SEM-et egyfajta hipotézis tesztelésre is. Például, vajon ugyanakkora hatással vannak e a EGY illetve $TARS$ változók a $GYAK$ változóra? Az például azért merülhet fel, mert a két előző azt méri, hogy inkább társaságban szeret alkoholt fogyasztani a kérdezett, vagy egyedül. Így az előbbi felvetés azt jelenti, hogy mindegy, hogy melyiket preferálja, ezek ugyanúgy befolyásolják a fogyasztás mértékét. Ha ezt elvetjük, akkor a két regressziós együttható különböző, ami azt jelenti, hogy ha például a $TARS \rightarrow GYAK$ él súlya nagyobb, akkor az aki inkább társaságban szeret inni, többet is fogyaszt. Ezt a korábbiakhoz hasonló módon tesztelhetjük: illesztünk egy modellt úgy, hogy a két él súlyát egyformának vesszük, és megnézzük, hogy lényegesen rosszabbul illeszkedik-e így a modell. Az eredmény a táblázat utolsó sorában láthatjuk. Érdekes, hogy a hipotézist elvetjük, hiszen a χ^2 statisztikánk 17, 1-el romlott, viszont érdekes módon a PFI mutató nőtt, azaz azzal, ha így felszabadítanánk még egy paramétert javulna a modell viszonylagos illeszkedése.

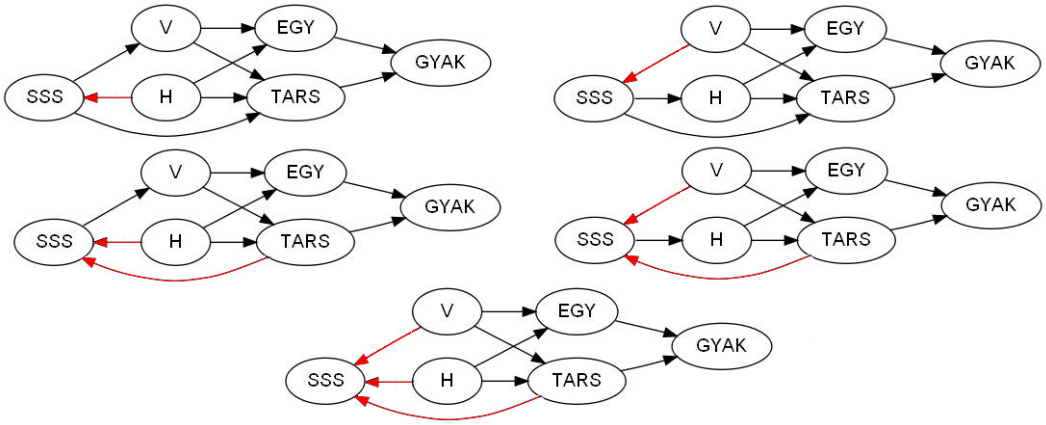
A többi él elhagyásával a modell illeszkedése jelentősen romlik, azaz eljutottunk a legegyszerűbb modellünkhöz, aminek strukturális része az alábbi ábrán látható.



A következő lépésben megvizsgáljuk hogyan teljesít a modell azon a részmintán, amit a modellünk finomítására nem használtunk. Azt várjuk, hogy ha mindent jól csináltunk, az illeszkedési mutatók ugyan rosszabbak lesznek, de nagyságrendileg nem változnak. Az adatok ezt alátámasztják, csak a χ^2 statisztika eltérése nagy, de az összes többi mutató kevéssé változott.

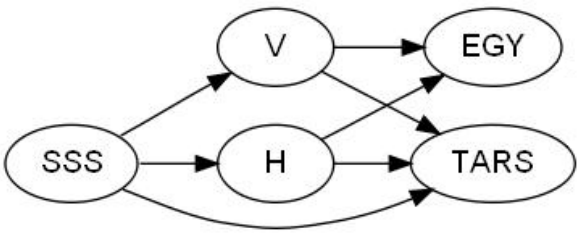
	χ^2	GFI	RMR	NFI	PFI
Eredeti modell	1769,4	0,8675	0,0847	0,8408	0,7282
Kontroll modell	2067,3	0,8528	0,0910	0,8186	0,7079

Azonban ezzel még nem teljes az elemzésünk, mivel tudjuk, hogy lehetnek ekvivalens modellek a mienkkel, és ezen modellek ismerete és kizárása egy elemzés során létfontosságú. Az ekvivalens modellek a következő ábrán vannak felsorolva. Ezeket a modelleket az elemzést végző kutatónak kell kizárnia az adott tudomány tételei és ismeretei segítségével, mivel az adatok önmagukban nem segítenek a modellek megkülönböztetésében.



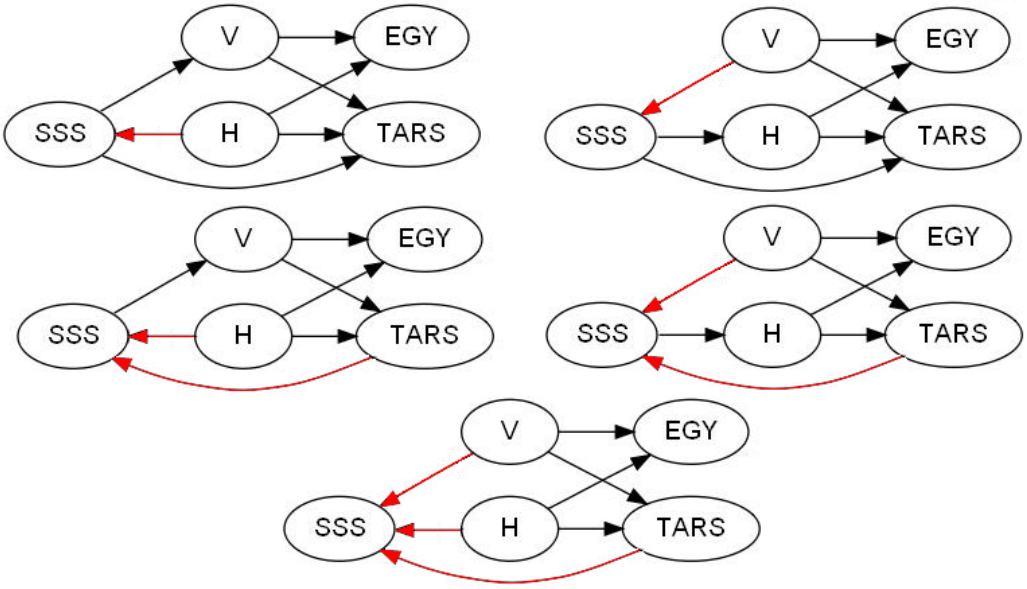
Még egy fontos észrevétel a kapott modellel kapcsolatban az, hogy az alkohol fogyasztásának gyakoriságát csak az ivási szokások magyarázzák, mi pedig elsősorban a másik három változóval szerettük volna magyarázni. Így meggondolandó, hogy a fogyasztási gyakoriságot egyszerűen elhagyjuk, hiszen úgy néz ki a vizsgálódásaink szempontjából irreleváns, a másik két alkoholfogyasztási változó elég a jelenség modellezésére. Azaz ugyanazt a procedúrát végig csináljuk mint előbb, csak a kiinduló modellünkből kihagyjuk a $GYAK$ csúcsot és minden hozzá tartozó mért adatot és élt.

Az előző utat követve eljutunk a korábbi optimális modellünkhöz a $GYAK$ csúcs nélkül. Az alábbi ábrán látható a kapott modell. Az illeszkedési mutatói is hasonlóak, mint a korábbinak.



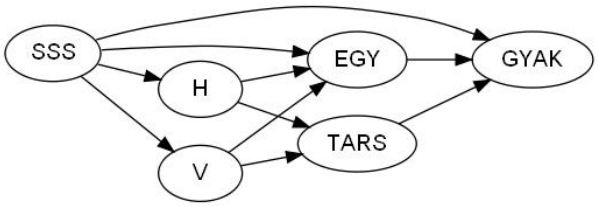
Így a modellező választás elé kerül, hiszen közvetlenül nem hasonlíthatja össze a két modellt. A modellek ugyanakkor nagyon hasonlóak, sőt lényegében ugyanolyanok, de más adathalmazra vannak illesztve, ezért az elemzőnek kell döntenie arról, hogy

	χ^2	df	GFI	RMR	NFI
GYAK nélküli modell	1277,4	129	0,8893	0,0843	0,8239
Kontroll	1623,2	129	0,8717	0,1016	0,7748



melyiket akarja használni. Az első modell mutatói kicsivel jobbak, és a kontroll mintán is jobban teljesít, mint a második. A második viszont egyszerűbb, és az elsőben nem ad fontos információt az alkoholfogyasztási gyakoriság változók használata.

A modell validálásának egy további módja lehet, hogy a két részmintánk szerepét felcseréljük. Ami eddig a kontroll minta volt, az alapján fogjuk a modellt finomítani, amin pedig előzőleg a finomítást végeztük, az fog szolgálni az ellenőrzésre. Ha a modellünk jó, akkor az eredetiből kiindulva ebben az esetben is ugyanazt az optimális modellt kell kapjuk. Azonban azt tapasztaljuk, hogy a szerepek felcserélésével a legjobb modell eltér az előzőtől. Az alábbi ábra mutatja az új optimális modellünket, az alatta levő táblázat pedig az illeszkedési mutatókat.



Az illeszkedési indexek azt mutatják, hogy a most kontrollnak használt első részmintára a modell sokkal jobban illeszkedik a másodiknál még arra a modellre is, amit a második rész alapján finomítottunk. Az első részmintára valamilyen ok-

	χ^2	GFI	RMR	NFI	PFI
Alapmodell	1981,6	0,8539	0,0824	0,8244	0,6909
	χ^2	GFI	RMR	NFI	PFI
Optimális M.	1985,1	0,8549	0,0839	0,8241	0,7024
Kontroll	1725,0	0,8681	0,0788	0,8440	0,7194

ból lényegesen jobban illeszkedik a modellünk minden esetben. Más szóval a modellünk jobban magyaráz a minta első felén. Ez úgy fordulhatott elő, hogy amikor az eredeti mintánkat két részre bontottuk, azt teljesen véletlenszerűen tettük, és semmilyen szempontra nem figyeltünk oda. Így lehet, hogy az országos mintánkat, ami heterogén, két homogénebb részmintára bontottuk. Ha megvizsgálnánk, hogy a két rész miben tér el egymástól lényegesen (például nem, kor, lakóhely...), akkor megtudhatnánk, hogy milyen további tényezők befolyásolják a modellezni kívánt jelenséget, egyúttal magyarázatot kaphatnánk arra, hogy miért illeszkedik jobban a minta első része. Ezeket aztán beépíthetnénk a modellbe új változóként, vagy akár tudatosan is tagolhatnánk a mintánkat a kapott tényezők alapján, és minden részmintára külön-külön elvégezhetnénk a korábbi elemzést. Így eljuthatunk egy jobban illeszkedő modellhez, vagy modellek egy halmazához, ami mutatná, hogy hogyan változnak az alkoholfogyasztási szokások a korábban meghatározott tényezők szerint.

6.1. Összefoglalás

Ahogy az a fenti példán látszik, kevés előre lefektetett szabály áll rendelkezésre a SEM modell használatára, és sok akadály merülhet fel egy alkalmazás során. Azonban ha az alkalmazó tudja mit szeretne elérni, ki tudja választani a céljának leginkább megfelelő modellt. Ezen felül azt is mutatja az előbbi példa, hogy ez a modellezési technika inkább a leírni kívánt adatokkal kapcsolatos elképzelések tesztelésére és validálására használható, a meghúzódó ok okozati viszonyok feltárására kevésbé. Ennek az oka az, hogy előre meg kell adjuk, hogy mely változók magyarázóak, és hogy mit magyaráznak. Ha valamilyen valójában jelen levő magyarázó viszonyt kihagyunk azt a modell nem jelzi nekünk, pusztán annyit látunk, hogy a modell rosszul illeszkedik. Ezért amikor egy SEM-et akarunk használni, előzőleg alaposan meg kell ismerkednünk a leírni kívánt jelenséggel és feltérképezni azokat az esetleges magyarázó viszonyokat, amelyek meghúzódhatnak a háttérben.

Irodalomjegyzék

[1] John C. Loehlin (1987) *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis.* Lawrence Erlbaum Associates, Inc.

[2] James H. Steiger (2001) *The Relationship Between Software Developement, Theory, and Education in Structural Equation Modeling.* Journal of the American Statistical Association, Vol. 96, No. 453

[3] Jesus Palomo, David B. Dunson & Ken Bollen (2007) *Bayesian Structural Equation Modelling.* Handbook of Computing and Statistics with Applications, Vol. 1

[4] Peter Sprites, Thomas Richardson, Chris Meek, Richard Sheines, Clark Glymour (1998) *Using Path Diagrams as a Structural Equation Modeling Tool.* Sociological Methods and Research

[5] J.J. Hox, T.M. Bechger (1998) *An Introduction to Structural Equation Modeling.* Family Science Review

[6] Zhiyong Zhang, Ellen L. Hamaker, John L. Nesselroade (2008) *Comparisons of Four Methods for Estimating a Dynamic Factor Model.* Structural Equation Modeling: A Multidisciplinary Journal, 15:3, 377-402

[7] Paul Barrett (2007) *Adjudging Model Fit.* Persomality and Individual Differencies, 42, 815-824

[8] John Fox (2006) *Structural Equation Modeling with the sem Package in R.* Structural Equation Modeling, 13(3), 465-486

[9] B. Walsh (2004) *Markov Chain Monte Carlo and Gibbs Sampling.* Lecture Notes for EEB 581