Statisztikai módszerek a genetikában

Szakdolgozat

Írta: Németh László

Alkalmazott matematikus MSc

Témavezető:

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

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2016

Tartalomjegyzék

1. Bevezetés **4**
1.1. Motiváció 4
1.2. Orvosi alapfogalmak 5
1.3. Matematikai alapfogalmak 6
1.3.1. Statisztika 6
1.3.2. Gráfelmélet 7

2. Lineáris modellek **10**
2.1. Lineáris regresszió 10
2.1.1. Legkisebb négyzetek módszere 11
2.1.2. Maximum-likelihood módszer 13
2.2. Stepwise algoritmus 15
2.2.1. Hátra fele haladó 15
2.2.2. Előre fele haladó 16
2.3. Szimuláció 16

3. Általánosított lineáris modell **19**
3.1. Exponenciális eloszláscsalád 19
3.2. A modell működési elve 20
3.3. Logisztikus regresszió 20
3.3.1. Szimuláció 21

4. Modellezés strukturális egyenletekkel **23**
4.1. Markov-ekvivalens modellek 23
4.2. Alkalmazhatóság feltételeihez szükséges alapok 24
4.2.1. Injektivitási feltétel 26
4.2.2. Stepwise inverzió 26
4.2.3. A gráf-feltétel szükségessége 30
4.2.4. A gráf-feltétel elégségessége 32
4.2.5. Ciklikus gráf esete 33
4.3. Half-trek kritérium 33
4.4. A gráf-feltétel és a Markov-ekvivalens modellek 36

5. Alkalmazások genetikai kutatásban **38**
5.1. Adatok 38
5.2. Logisztikus regresszió 39

5.3. P-korrekció 42
5.4. Folytonos változók 43
5.5. SEM-modell 44

6. Összefoglaló **47**

1. Bevezetés

1.1. Motiváció

Az utóbbi évtizedekben jelentősen kibővültek az emberi tulajdonságok öröklődésével kapcsolatos ismereteink. A 20. században felfedezték a dezoxiribonukleinsavat, röviden DNS-t, a ribonukleinsavat (RNS), illetve ezeken belül több különböző gént, melyek a legfontosabb élettani tulajdonságokat kódolják. Felderítették a teljes emberi génkészletet 2000-re, megközelítőleg 3 milliárd alkotóelemet, $20000 - 25000$ génen belül. Felfedeztek számos gént, ami összefüggésbe hozható komolyabb betegségekkel. Ezen tudományos felfedezések következtében sokat fejlődött a technológia is, hiszen megszületett az igény, hogy jobban megismerhessük a génekben rejlő információt. Jelenleg léteznek olyan robotok, amelyek egy kevés vérből képesek izolálni a DNS-t és megadni a nukleinsavak pontos sorrendjét, azonban sajnos még komoly költségek mellett. A kutatások eredményeként sok betegséget az orvosok már egyértelműen tudnak egy adott génhez kötni. Azonban a DNS-ben rengeteg információ rejlik, a már megismert betegségek mellett sok más esetben is lehetne kockázati faktorokat, hajlamokat meghatározni. Dolgozatom célja olyan statisztikai módszerek bemutatása, amelyekkel költséghatékonyan, kevés vizsgálat elvégzésével is van lehetőség jelentős előrelépést tenni valamilyen betegségre hajlamosító gének keresésében.

Tanulmányaim során egy orvosi csapattal dolgoztam együtt a Semmelweiss Ignác Orvostudományi Egyetemről. A terhességi cukorbetegség genetikai hátterét próbáltuk felderíteni. A több éves adatgyűjtő munka eredményeképpen körülbelül 800 anya és gyerekeik DNS-e került analizálásra. A vizsgálatban részt vevő alanyok véréből az orvosok megállapították a vércukorszintet, többféle enzim szintjét, valamit rendelkezésünkre állt az édesanyák terhesség előtti testtömeg-indexe és életkora. Minden anya esetében a vizsgáló orvos a vércukorszintből és egyéb látható jelekből megállapította, hogy egészséges, vagy cukorbeteg-e. A kutatásunk célja a rendelkezésünkre álló genetikai adatok és a betegség közötti összefüggés megállapítása valamint a cukorbetegség legfőbb ismérve, a vércukorszint előrejelzése.

Szakdolgozatomban a kutatás során alkalmazott statisztikai módszereket fogom bemutatni. Először a lineáris regresszió elméletét ismertetem, majd annak általánosításait. Közben említést teszek a módszerek működését garantáló becslésekről. A dolgozat második felében egy általánosabb elméletet, a strukturális egyenletek modelljét mutatom be, a működés szükséges és elégséges feltételeivel együtt. Ennek segítségével egy olyan modellt készítek el, amelyben a genetikai változatok, életkor,

testtömeg-index és vércukorszint között felállított struktúrában egyszerre magyarázza meg a tényezők közötti kapcsolatot, önmagába foglalva az esetleges interakciókat is. A dolgozat végén bemutatom a modellek konkrét alkalmazását, kitérve a kutatás során felmerülő egyéb problémákra, mint például a "false discovery". Ezen módszerek és algoritmusok ismerete már jó kiindulási alapot szolgáltathat más típusú genetikai vizsgálatok elvégzéséhez is.

1.2. Orvosi alapfogalmak

A genetikai vizsgálatok hátterében sok orvosi eredmény áll, és hogy ezeket kellően tudjuk modellezni meg kell ismernünk néhány alapfogalmat. A vizsgálatok nagy része a DNS körül zajlik. A DNS négy különböző nukleinsav (adenin - A, guanin - G, citozin - C, timin - T) kombinációjaként áll elő. Kettős spirálba rendeződnek és minden nukleinsavnak van egy őt kiegészítő párja. Így a DNS egyértelműen leírható az egyik szálon található alkotók sorrendjével.

1. ábra. A DNS egy szakasza, kettős spirál szerkezet ábrázolásával és nukleinsav párok feltüntetésével

Az 1. ábra a DNS egy szakaszát mutatja be. Bármely két ember DNS-e 99 százalékban megegyezik, ám vannak bizonyos pontok, ahol tapasztalhatunk eltérést. Single nucleotid polymorphysm-nak, röviden SNP-nek nevezünk egy ilyen eltérést,

ha adott ponton a vizsgált alany DNS-e eltér a populációban gyakrabban előfordulótól.

A DNS-ben a kromoszómák párban állnak és együttesen határozzák meg a különféle betegségekre kialakuló hajlamokat. Ezért nem elég, ha egy kromoszómában egy ponton változást tapasztalunk, meg kell vizsgálnunk, hogy a párján, ugyanezen a ponton is van-e változás. Ha csak egyik pont módosult, akkor heterozigóta variánsnak nevezzük, ha pedig mindkettő, akkor homozigóta variánsnak.

Orvosilag bizonyított, hogy bizonyos génváltozatok kapcsolatban állnak betegségekkel. Ezeken a pontokon vizsgált SNP-k közelebbi képet adhatnak a háttérben meghúzódó hajlamosító tényezőkről. Mivel egy SNP három különböző párosítást tartalmazhat, így a hatása is több típusba sorolható.

- Domináns modellnek nevezzük azt, ha az SNP-ben megjelenik a ritkább változat és a hatás máris észlelhető, függetlenül attól, hogy heterozigóta, vagy homozigóta módon hordozza az egyén.

- Recesszív modellt alkalmazunk, amikor a ritkább változat egyszeres hordozása lényegiben nem tér el attól, mintha a gyakorit hordozná az egyén heterozigóta módon, azaz egyszeresen. Csak akkor észlelhető eltérés, ha a kevésbé gyakori változatból kettő is van.

- Additív modell esetén minden hordozott ritkább változatnak azonos hatása van, így a kétszeres hordozás kétszer akkora hatást fejt ki, mint az egyszeres.

- Multiplikatív modell használatakor a heterozigóta variánsa x, míg a homozigóta variánsé x^2.

További modellek is elképzelhetők, azonban a gyakorlatban ezek fordulnak elő a leggyakrabban, így a génváltozatok legnagyobb része feltételezhetően ezeknek megfelelő módon hat.

1.3. Matematikai alapfogalmak

1.3.1. Statisztika

1.1. Definíció. *Legyen* $(\Omega, \mathbb{A}, P)$ *valószínűségi mező. Egy* $X : \Omega \to \mathbb{R}$ *függvényt valószínűségi változónak nevezünk, ha* $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathbb{A}$ *minden* $x \in \mathbb{R}$ *esetén.*

Átfogalmazva X egy olyan véletlentől függő mennyiség, amelyről meghatározható, hogy milyen valószínűséggel vesz fel x-nél kisebb értéket. Meggondolható, hogy ez a tulajdonság ekvivalens azzal, ha tetszőleges Borel-halmazbeli értéket vizsgálunk.

1.2. Definíció. *Legyenek adottak $X_1, X_2, \ldots, X_n$ független azonos eloszlású valószínűségi változók, továbbá $\omega \in \Omega$ véletlen esemény. Ekkor $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ értékeket véletlen mintának nevezzük.*

1.3. Definíció. *Statisztikai mezőnek nevezzük azt az $(\Omega, \mathbb{A}, \mathbb{P})$ hármast, ahol $(\Omega, \mathbb{A})$ egy mérhető tér, $\mathbb{P}$ pedig az $\mathbb{A}$-n értelmezett valószínűségi mértékek egy családja*

1.4. Definíció. *Statisztikának nevezzük a minta egy függvényét.*

1.5. Definíció. *Egy hipotézisvizsgálati feladat esetén p-értéknek nevezzük azt a statisztikát, amely a nullhipotézis mellett egyenletes eloszlást követ $(0,1)$-en és általában annak a valószínűségét adja meg, hogy a nullhipotézist egy nullhipotézisbeli minta esetén tévesen elutasítjuk. A döntés minőségének ellenőrzésére használatos statisztika. Használata annyiban egyszerűsíti a hipotézisvizsgálati feladatot, hogy a $(0,1)$ intervallum egy α mértékű, tetszőleges részhalmaza egy α szintű próbának felel meg. Optimális hipotézisvizsgálati módszer meghatározásához a $(0,1)$ olyan α hosszúságú részhalmazát kell kiválasztani, amelybe a p érték, ha a minta H_1 szerinti, a lehető legnagyobb valószínűséggel esik.*

1.6. Definíció. *Egy paraméter becslés során kapott érték általában nem egyezik meg pontosan a keresett eloszlás paraméterével. Adott α mellett konfidencia intervallumnak nevezzük a paramétertartomány azon összefüggő részhalmazát, amelyre igaz, hogy a minta alapján a becsült paramétert $1-\alpha$ valószínűséggel tartalmazza.*

1.3.2. Gráfelmélet

A gráfelméleti alapfogalmakat, mint gráf, csúcs, él, irányított gráf, összefüggő gráf ismertnek tekintem. Egy G gráf esetén hagyományosan V (vertex) jelöli a csúcsok halmazát, E (edge) pedig az élek halmazát. Irányított gráf esetén az élek halmazát D-vel szokás jelölni az angol directed szó után.

1.7. Definíció. *Egy n csúcsú G gráfot vegyes gráfnak nevezünk, ha az élhalmazában szerepelnek irányított és irányítatlan élek is. A vegyes gráfot $G = (V, D, B)$ alakban szokás megadni, ahol V a csúcsok, D az irányított élek, B (bidirected) pedig az irányítatlan élek halmazát jelöli.*

1.8. Definíció. *Legyen adott egy n csúcsú $G = (V, D)$ irányított gráf. Indexeljük át a csúcsokat egy tetszőleges, olyan $R : \{1, 2, \dots n\} \to \{1, 2, \dots n\}$ függvénnyel úgy, hogy ha $(i \to j) \in D$, akkor $R(i) \leq R(j)$. A csúcsok R-beli képét növekvően rendezve kapjuk a csúcsok topologikus sorrendjét.*

A definícióból látszik, hogy R leképezés nem feltétlenül egyértelmű, sőt nem is biztos, hogy létezik. Ezért a topologikus sorrend sem létezik minden esetben és a meghatározása sem mindig egyértelmű. A topologikus rendezés szemléletesen a csúcsoknak egy olyan sorrendjét határozza meg, ahol egy csúcsból mindig csak tőle nagyobb indexű csúcs felé mutathat él.

1.9. Definíció. *Ha egy $G = (V, D)$ irányított gráfban létezik az éleknek egy olyan $u_1, u_2, \dots u_k \in D$ sorozata, hogy minden $i = 1, \dots, k$ esetén u_i végpontja megegyezik u_{i+1} kezdőpontjával, valamint u_k végpontja és u_1 kezdőpontja egybe esik, akkor azt mondjuk, hogy G gráf kört tartalmaz.*

Ha egy G irányított gráf nem tartalmaz kört, akkor aciklikusnak nevezzük.

1.10. Állítás. *Egy G irányított gráfban akkor és csak akkor létezik topologikus rendezés, ha a gráf aciklikus.*

Bizonyítás. Tegyük fel, hogy G-be létezik topologikus rendezés. Ha létezik kör, akkor válasszuk ki azokat a pontokat, amelyeken a kör áthalad. Vegyük ebből a legmagasabb indexűt. Innen mindenképpen kisebb indexű csúcs felé mutatna él a kör definíciója miatt, ez viszont ellentmond a topologikus rendezettségnek.

Tegyük fel, hogy az n csúcsú G gráfban nincs kör. Kiválasztva egy tetszőleges csúcsot, irányított éleken előre lépkedve véges lépés alatt eljutunk egy olyan csúcsba, ahonnan már nem fut ki él. Ez a csúcs kapja az n indexet, majd töröljük G-ből. Az így kapott G' gráf továbbra is körmentes marad, és hasonló módon kiválasztható belőle egy legmagasabb, $n - 1$ indexű csúcs. A lépéseket folytatva meghatározunk egy olyan R bijektív leképezést, amelyre teljesül a topologikus rendezés definíciója, így G gráf rendezhető. □

Mivel a bizonyítás során a csúcsok választása nem determinisztikus, innen is látható, hogy a topologikus sorrend nem feltétlenül egyértelmű.

1.11. Definíció. *Legyen adott G irányított gráf, $i, j \in V$. Ekkor*

- *i és j közti sétának nevezzük a csúcsok és élek olyan $v_i, e_1, v_2, \dots, e_k, v_{k+1}$ sorozatát, ahol a csúcsok és élek felváltva helyezkednek el, $v_1 = i, v_{k+1} = j$, illetve minden e_s él v_s és v_{s+1} csúcsok között fut.*

- *irányított sétának nevezzük azt a sétát, amelyben minden e_s él v_s-ből fut v_{s+1} csúcsba.*
- *útnak nevezzük azt a sétát, amelyben nem szerepel kétszer ugyanaz a csúcs.*
- *irányított útnak nevezzük azt az irányított sétát, amelyben nem szerepel kétszer ugyanaz a csúcs.*

A fenti definíció vegyes gráfokra is ugyanígy fogalmazható meg.

2. Lineáris modellek

Tegyük fel, hogy van két valószínűségi változónk, X és Y. A modellalkotás első és legegyszerűbb megközelítése, hogy a valószínűségi változók között lineáris kapcsolatot feltételezünk, azaz feltesszük, hogy $Y = a \cdot X + c$ egyenlőség közelítően teljesül. Ez nem egy túlzottan szigorú feltétel, hiszen ha egy függvény kellően sima, akkor lokálisan közelíthető lineáris függvényekkel.

Vegyünk egy N elemű mintát, (Y, X) változópárra vonatkozóan. A méréseket vektorba rendezve $\overline{Y} = a \cdot \overline{X} + c \cdot \mathbb{1}$ vektoregyenletet kell megoldanunk, ahol a és c a keresett paraméterek valamint $\mathbb{1}$ az N-dimenziós egységvektort jelöli. Ha $N \geq 3$ akkor az egyenletrendszer túlhatározott, így valószínűleg nem kapunk megoldást, csak közelítésről beszélhetünk.

Tegyük fel, hogy az Y változó magyarázatára több $n \geq 2$ változó áll rendelkezésre, legyenek ezek $X_1, X_2, \ldots, X_n$. Szeretnénk meghatározni azon $\overline{a}, c$ paramétereket, amelyekre a $Y = \overline{a}^T \cdot \overline{X} + c$ lineáris egyenlet a lehető legkisebb hibával teljesül, azaz a legpontosabban közelíti lineáris módon az összefüggést.

2.1. Lineáris regresszió

Célunk a folytonos változók (például vércukorszint, testmagasság) és a genetikai adatok közötti kapcsolat vizsgálata. Tételezzük fel, hogy a variánsok hordozása és a vércukorszint között valamilyen lineáris kapcsolat van. Ekkor lineáris regressziót alkalmazhatunk.

Legyen Y a célváltozó, $X_1, X_2, \ldots, X_n$ pedig a magyarázó változók, melyeket egy vektorváltozóként $\overline{X}$-el jelölünk. A szokványos, független változó megnevezés megtévesztő lehet, ugyanis ezek a változók összefüggethetnek egymással, vagy akár Y-nal is, sőt, ez utóbbi kapcsolatot szeretnénk felderíteni.

Vegyünk egy N elemű mintát. Legyen $\overline{Z}$ az ismeretlen célváltozóból kapott értékek N hosszú vektora, és $\mathbb{S}$ az ismert magyarázó változókból kapott $N \times n$-es mátrix. Keressük azt az n-dimenziós $\overline{a}$ vektort és b számot, amelyre $\overline{Z} = \mathbb{S} \cdot \overline{a} + b \cdot \mathbb{1}$. Ha létezik ilyen $\overline{a}$ vektor, akkor Y lineárisan kifejezhető $\overline{X}$ segítségével. Ekkor $\overline{a}$ vektor koordinátáit az $\overline{X}$ változókhoz tartozó együtthatóknak hívjuk. Részletesebben kiírva

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b.$$

Ilyen vektor azonban a gyakorlatban nem létezik, ezért azt az $\overline{a}$-t keressük, amelyre $\overline{a}^T \overline{X}$ legjobban közelíti Y-t. Így tehát a korábbi egyenlet helyett most a

$\overline{Z} = \mathbb{S} \cdot \overline{a} + b \cdot \mathbb{1} + \overline{\varepsilon}$ megoldását keressük, ahol $\overline{\varepsilon}$ egy N dimenziós a hibavektor. A megoldás során $\overline{\varepsilon}$-t minimalizálni szeretnénk, hogy a legjobban illeszkedő modellt kapjuk. Ekkor:

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b + \varepsilon^*$$

ahol ε^* egy általános hibatag, minden megfigyelés esetén más-más lehet. A minta alapján $\overline{a}$ és b értékeit meg tudjuk becsülni különböző becsléselméleti módszerekkel, amelyekről később lesz szó.

Az egyszerűség kedvéért a becsült együtthatókat továbbra is jelöljük $a_1, \ldots, a_n, b$ módon, hiszen a becsléssel kapjuk meg a legjobban illeszkedő egyenes együtthatóit.

2.1.1. Legkisebb négyzetek módszere

A legkisebb négyzetek módszere egy adathalmazra legjobban illeszkedő egyenes meghatározására alkalmas, tisztán algebrai eszközökkel.

Legyen N darab megfigyelésünk $x_1, x_2, \ldots, x_N$ és $y_1, y_2, \ldots, y_N$ az (X, Y) valószínűségi változópárra vonatkozóan. Az alapvető feladat, meghatározni a pontokhoz legjobban illeszkedő egyenest. Szeretnénk találni olyan a, b értékeket, amelyekre $y_i = ax_i + b$ fennáll minden $i = 1 \ldots N$ esetén. Ekkor a valószínűségi változók között is $Y = aX + b$ összefüggésre következtetünk. Ha X és Y között nincs lineáris kapcsolat elméletileg sem várhatunk pontos eredményt, sőt a gyakorlatban mérési hibával is számolnunk kell.

Ha találnánk olyan egyenest, amely pontosan illeszkedik a méréseinkre, akkor $y_i - ax_i - b = 0$ lenne $i = 1 \ldots N$-re. Egy jól illeszkedő egyenes esetén tehát azt várjuk, hogy $y_i - ax_i - b$ kicsi legyen minden $i = 1 \ldots N$-re. A módszer a négyzetes hibát minimalizálja, tehát olyan együtthatókat keres, hogy $\sum_{i=1}^{N} (y_i - ax_i - b)^2$ minimális legyen. Azaz a nagyon kiugró értékeket jobban bünteti, mint az egyenestől közepesen eltérő többi értéket.

A módszer leírása

Legyen $E(a, b) = \sum_{i=1}^{N} (y_i - ax_i - b)^2$ a kétváltozós hibafüggvény. Mivel a minta elemei adottak, $E(a, b)$ mindkét változójától folytonosan függ, 0 egy alsó korlátja, valamint differenciálható. Ha a, vagy b abszolút értékben tart végtelenhez, akkor az egyenes egyre kevésbé fog illeszkedni, tehát a differenciálással kapott szélsőérték minimum lesz.

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{N} 2(y_i - ax_i - b)(-x_i)$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{N} 2(y_i - ax_i - b)$$

Oldjuk meg a $\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0$ egyenletet. A megoldásként kapott a és b értékek lesznek a legjobban illeszkedő egyenes együtthatói.

$$\begin{aligned} 0 &= \sum_{i=1}^{N}(y_i - ax_i - b)(-x_i) \\ 0 &= \sum_{i=1}^{N}(y_i - ax_i - b) \end{aligned}$$

Átrendezve a következő egyenleteket kapjuk:

$$\begin{aligned} \sum_{i=1}^{N} x_i y_i &= \sum_{i=1}^{N} x_i^2 \cdot a + \sum_{i=1}^{N} x_i \cdot b \\ \sum_{i=1}^{N} y_i &= \sum_{i=1}^{N} x_i \cdot a + N \cdot b \end{aligned}$$

A két egyenlet felírható az alábbi 2×2-es mátrix segítségével

$$\mathbb{A} = \begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix}.$$

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix}.$$

Ha az $\mathbb{A}$ mátrix determinánsa nem nulla, a mátrix inverzével mindkét oldalt balról megszorozva a jobb oldal csupa ismert elemeket tartalmaz, míg a bal oldalon a keresett paraméterek jelennek meg. Ekkor

$$\begin{aligned} det\mathbb{A} &= N \cdot \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} x_i = N \cdot \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2 \quad \geq 0. \end{aligned}$$

Egyenlőség akkor áll fenn, ha minden $i, j = 1 \ldots N$ esetén $x_i = x_j$. Ekkor a determináns 0, így nem tudunk invertálni. A megoldás az x_1 ponton átmenő függőleges

egyenes lenne,ami azt jelenti, hogy X az Y érték magyarázatához nem járul hozzá. Minden más esetben a mátrix invertálható és algebrai módszerekkel megkaphatók azok az a, b értékek, amelyek a legjobban illeszkedő egyenest reprezentálják.

A módszer általánosítható, nem csak egyenest, hanem legjobban illeszkedő polinomot is lehet keresni a segítségével. Bebizonyítható, hogy a megoldhatóság feltétele ekkor is az, hogy ne legyen minden $i, j \in 1 \dots N$ esetén $x_i = x_j$. A legkisebb négyzetek módszerével a [2] foglalkozik részletesebben.

További általánosítás a többdimenziós eset, amikor m darab valószínűségi változóhoz szeretnénk illeszteni l-dimenziós legjobban illeszkedő hipersíkot, ahol $l < m$. Ekkor a mátrix determinánsa több esetben is lehet 0, így az alkalmazás további vizsgálatokat igényel.

2.1.2. Maximum-likelihood módszer

Legyen adott $X_1, X_2, \dots, X_n$ egy F eloszlásból vett független minta értéke. Legyen az F eloszlás egy ismeretlen paramétere θ. Az $x_1, x_2, \dots, x_n$ minta alapján szeretnénk megbecsülni, hogy mi lehet θ értéke. Legjobb becslésnek tartjuk azt a θ_0 értéket, amelyre a lehetséges θ-k közül a legnagyobb valószínűséggel kapjuk az $x_1, x_2, \dots, x_n$ körüli értékeket.

2.1. Definíció. *Likelihood függvénynek nevezzük a θ paraméter azon függvényét, amely megadja annak a valószínűséget, hogy $X_1, X_2, \dots, X_n$valószínűségi változók $x_1, x_2, \dots, x_n$ körüli értéket vesznek fel. A függetlenséget is felhasználva, folytonos esetben az F sűrűségfüggvényét alkalmazva:*

$$\begin{aligned} L(\theta) &= \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots X_n = x_n) \qquad \textit{folytonos esetben:} \\ &= f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta). \end{aligned}$$

A legjobb becslést tehát $L(\theta)$ függvény maximalizálásával kaphatjuk. Amennyiben a likelihood függvény differenciálható, analízisből ismert okokból a maximumhelye vagy az értelmezési tartomány határán található, vagy egy olyan pontban, ahol a függvény deriváltja 0. Nem differenciálható függvény esetén ez a módszer nem alkalmazható.

A logaritmus függvény szigorú monotonitását kihasználva tudjuk, hogy $L(\theta)$ és $l(\theta) := \log L(\theta)$ maximumhelyei ugyanott vannak. Ezért a deriválás megkönnyítése érdekében gyakran $l(\theta)$ függvényt szokás maximalizálni.

A módszer további előnye, hogy több ismeretlen paraméter esetén is használható.

2.2. Definíció. *Legyen $X_1, X_2, \dots X_n$ független azonos eloszlásból származó minta, melynek $\theta_1, \theta_2, \dots \theta_k \in \Omega$ paraméterei ismeretlenek. A sűrűségfüggvényt jelöljük f-el. A likelihood függvény*

$$L(\theta_1, \theta_2, \dots \theta_k) = \prod_{i=1}^{n} f(x_1; \theta_1, \theta_2, \dots \theta_k).$$

Ha az

$$[m_1(x_1, x_2, \dots x_n), m_2(x_1, x_2, \dots x_n), \dots m_k(x_1, x_2, \dots x_n)]$$

vektor maximalizálja a likelihood függvényt, akkor a paramétereket becslő függvény

$$\theta_i^* = m_i(X_1, X_2, \dots X_n)$$

minden $i = 1 \dots k$ esetén.

A paraméterek maximum likelihood (röviden ML) becslése

$$\theta_i = m_i(x_1, x_2, \dots x_n),$$

ahol $x_1, x_2, \dots x_n$ a konkrét megfigyelés.

Az ML becslés definíciója után vizsgáljuk meg a tulajdonságait. Feltételezzük, hogy a mérési hibák független, azonos eloszlásúak (általában normális, de nem feltétlenül), továbbá a likelihood függvény logaritmusa, (röviden log-likelihood függvény) kétszer differenciálható.
A becslés tulajdonságai az alábbiak.

- Ezen feltételek mellett bizonyítható, hogy az ML becslés egyértelmű.
- A becslés konzisztens, tehát ha θ^* a becsült paraméter, akkor

  $$\lim_{n \to \infty} \mathbb{P}(|\theta^* - \theta| > \varepsilon) = 0$$

  minden $\varepsilon > 0$-ra, ahol n a minta elemszáma.
- Aszimptotikusan torzítatlan, azaz $\mathbb{E}(\theta^*) = \theta$, ha a minta elemszáma végtelenbe tart.
- A maximum likelihood becslés aszimptotikusan hatásos, azaz nincs olyan torzítatlan becslés, aminek kisebb lenne a szórása, ha az elemszámmal a végtelenhez tartunk.

- Aszimptotikusan normális eloszlású.

A tulajdonságok alapján a maximum likelihood becslés ideális aszimptotikus értelemben, ezért a minta elemszámát nagyra kell választani, hogy jól használható becslést kapjunk.

2.2. Stepwise algoritmus

Többdimenziós lineáris regresszió esetén nem egy egyenest, hanem egy hipersíkot keresünk. A többdimenziós regresszió elvégzése után az eredmény együtthatói közt gyakran van, amelyik nem szignifikáns. Ez viszont nem feltétlenül jelenti azt, hogy semmit nem tudunk mondani az adatainkról. Előfordul, hogy a magyarázó változók vektorából csak néhány koordináta adatai miatt nem lesz szignifikáns az eredmény. Algoritmikus módszerekkel kiválaszthatók azok a változók, amelyek együttesen ténylegesen hatnak a célváltozóra.

2.2.1. Hátra fele haladó

Első lépésben végezzük el a többdimenziós lineáris regressziót. Ha az eredményünk nem szignifikáns válasszuk ki a modellből azt a változót, amelynek a legkisebb a p-értéke. Ezt vegyük ki a modellből és végezzük el újra a regressziót.

Minden vizsgálat után két lépés közül választhatunk. Vagy kivesszük a modellből a legkevésbé szignifikáns változós, vagy bevesszük a modellbe a korábban már kikerültek közül a legszignifikánsabbat. A modellen belül vett szignifikancia összefügg a többi vizsgált változóval is, így ha újonnan kiveszek egy változót, az nem lesz automatikusan a legszignifikánsabb az addig kivettek közül.

Legyen L a likelihood függvény az aktuális modellben szereplő változók mintája alapján.

2.3. Definíció. *Egy modell Akaike információs kritériumának nevezzük az*

$$AIC = 2 \cdot k - ln(L)$$

számot, ahol k a modellben szereplő változók száma.

Két modell összehasonlítható az Akaike kritérium segítségével. Amelyik esetén kisebb az érték, az a modell számít jobbnak. Egy modell likelihhood függvénye egyre nagyobb értékeket venne fel, ahogy több változót vizsgálok, azonban ekkor egyre több lenne a kevésbé releváns változók száma is. Ezt korrigáljuk, hogy a kevesebb változót tartalmazó modelleket preferáljuk.

A stepwise algoritmus két lépése közül mindig azt hajtjuk végre, amelyik jobban csökkenti az AIC értéket. Amennyiben egyik lépés sem csökkentené, akkor az algoritmus leáll és az aktuális modellt tartjuk a legjobbnak, azzal érdemes folytatni a vizsgálatot.

2.2.2. Előre fele haladó

Kezdetben kiindulunk az önmagában legszignifikánsabb változóból. Ez után minden lépésben a már korábban leírt két lépés közül választunk, vagy kivesszük a modellből a legkevésbé szignifikáns változót, vagy bevesszük a modellbe a korábban már kikerültek közül a legszignifikánsabbat. Itt is akkor áll le az algoritmus, ha már nem tudunk javítani az Akaike információs kritériumon.

A két módszer nem biztos, hogy ugyanazt az eredményt adja, sőt, ha véletlenszerűen választott változókból indulunk ki elképzelhető, hogy egy még újabb modell esetén áll le az algoritmus. Azonban a leglényegesebb változókat mindegyik modell tartalmazni fogja és összességében hasznos információkat szolgáltat az összefüggésekről.

Teljes megoldás lenne, ha az összes lehetséges módon kiválasztanánk a változókat és azt a modellt fogadnánk el, amelynek globálisan a legjobb az információ tartalma. Ez azonban exponenciális nagyságrendű lépés lenne, amit a mostani informatikai eszközökkel nem tudunk hatékonyan elvégezni.

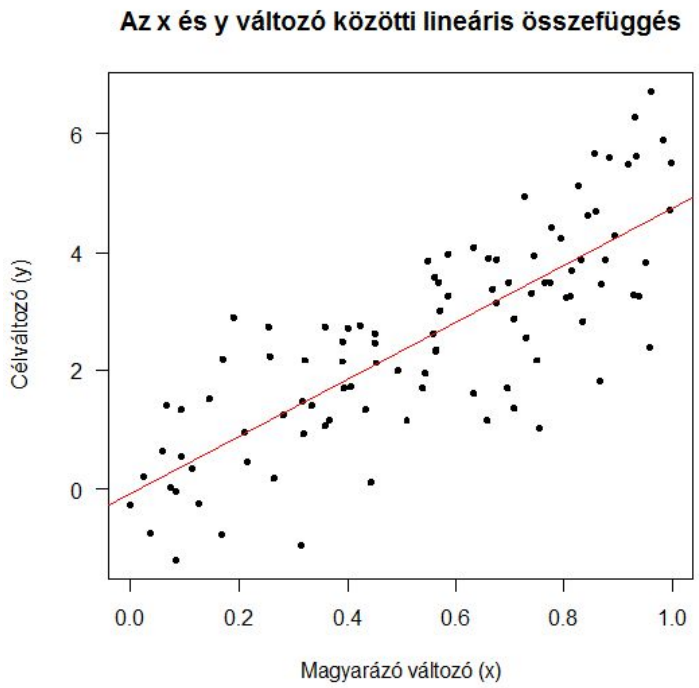
2.3. Szimuláció

A lineáris regresszió működését egy R-ben írt programmal szemléltetem. Legyen X magyarázó változó, egyenletes eloszlású $(0, 1)$-en. Az Y célváltozó ennek lineáris függvénye lesz. Az $Y = 5 \cdot X$ összefüggés álljon fenn. Vegyünk egy 100 elemű mintát X eloszlásából, legyen ez az x vektor. Számítsuk ki az x-hez tartozó Y értékeket. Vegyünk továbbá egy 100 elemű standard normális eloszlású hibavektort. A célváltozónk tapasztalati értékét jelöltjük y-nal és értéke legyen az elméleti érték és a hibavektor összege. Vizsgáljuk meg, hogy a lineáris regresszió módszerével felderíthető-e, hogy a célváltozó elméletileg a magyarázó változó 5-szöröse!

Az alábbi programmal előállítjuk a mintákat és megkeressük a legjobban illeszkedő egyenest:

```
set.seed(521)
darab<-100
```

```
x<-runif(darab,0,1)
eps<-rnorm(darab,0,1)
y<-5*x+eps
LM<-summary(lm(y~x))
plot(x,y, main="Az x és y változó közötti lineáris összefüggés",
 xlab="Magyarázó változó (x)", ylab="Célváltozó (y)", las=1, pch=20)
abline(LM$coefficients[1,1],LM$coefficients[2,1],col="red")
LM$coefficients[1,1]
LM$coefficients[2,1]
LM$coefficients[2,4]
```



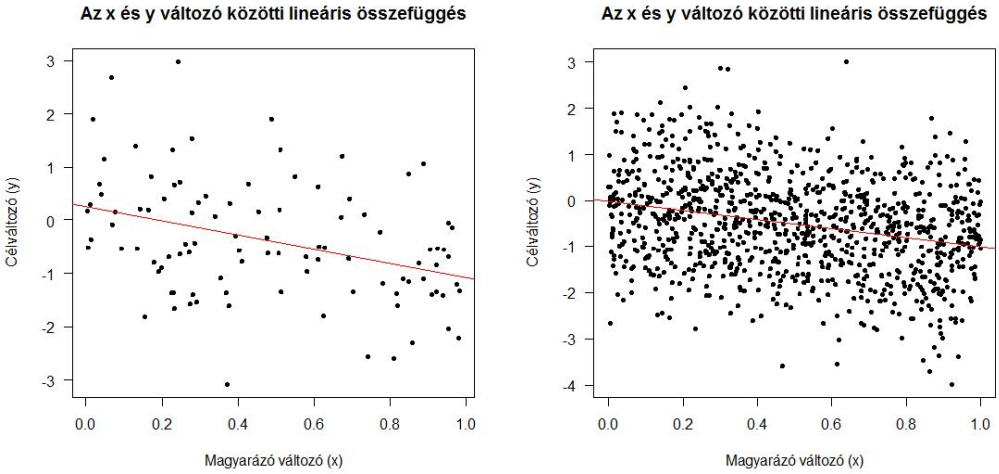
2. ábra. Y változó az X ötszöröse, Y megfigyelésében standard normális eloszlásból származó hibával, valamint a legjobban illeszkedő egyenes

A program alapján az $Y = 4.82 \cdot X - 0.077$ a legjobban illeszkedő egyenes, ami elég közel áll a valósághoz. Továbbá a modell szignifikanciája $p = 1.559 \cdot 10^{-23}$, azaz annak a valószínűség, hogy X és Y független egymástól p. A statisztikában $0,05$-nél kisebb p értékre már elfogadhatónak tekintjük a modellt, így a jelenlegi eredmény nagyon erősen bizonyítja, hogy a két változó között van összefüggés, amit a példa

konstruálásából mi is látunk. A 2 ábrán látható az összefüggés grafikus ábrázolása.

Vizsgáljuk meg, hogy a módszer milyen eredményt ad, ha az elméleti összefüggés $Y = -X$. Ekkor a hiba aránya nagyobb az elméleti értékhez képest, mint a korábbi esetben. A programot lefuttatva $Y = -1.33 \cdot X + 0.248$, ami nem annyira pontos, habár a szignifikancia lényegesen kisebb 0.05-nél: $p = 0.0001668503$, azaz felderítettük, hogy X és Y összefügg, de kevésbé lehetünk biztosak, mint a korábbi esetben.

Ez az eredmény felveti a problémát, hogy mit lehetne tenni, ha a két változó közötti összefüggés olyan kicsi, hogy a hiba miatt nem kapunk pontos, de még csak szignifikáns eredményt sem. A megoldás az, hogy növeljük a mintát, ha lehetséges. Ez a jelenlegi szimulált adatoknál nem okoz problémát, legyen a mintánk 1000 elemű.



(a) 100 mintaelem (b) 1000 mintaelem

3. ábra. X és Y változók egymás ellentettjei, Y megfigyelésében standard normális eloszlásból származó hibával, valamint a legjobban illeszkedő egyenes

A bővebb minta esetén, a 3 ábrán látható módon a lineáris regresszió a $Y = -0.986 \cdot X - 0.0276$ összefüggést találta, $p = 1.719 \cdot 10^{-18}$ szignifikancia mellett, vagyis kijelenthetjük, hogy ekkora mintára a módszer biztosan megfelelő modellt ad. De mi a teendő, ha a modell nem jó, viszont a minta nagy mértékű növelése nem lehetséges? Ilyen esetben a probléma gyakran az, hogy a két valószínűségi változó közötti kapcsolat nem lineáris, csak esetleg lineárissal közelíthető. Ekkor érdemes más, nemlineáris módszert keresni a modellalkotáshoz, aminek elvégzéséhez a [12] ad vázlatos áttekintőt.

3. Általánosított lineáris modell

Ebben a fejezetben a lineáris modell egy, gyakorlatban is sokat használt általánosítását ismerhetjük meg. Ehhez először definiáljuk az exponenciális eloszláscsaládot, amelyre kiterjesztjük a lineáris modellt. A modell ismertetése után az egyik típusával, a logisztikus regresszióval részletesen is foglalkozunk.

3.1. Exponenciális eloszláscsalád

Az exponenciális eloszláscsalád egy eltolás és skálaparaméteres eloszláshalmaz. Legyenek a, b, c, d rögzített, ismert függvények. Ekkor az egy exponenciális eloszláscsaládba tartozó eloszlás sűrűségfüggvényei

$$f(y;\theta) = exp\{a(y)b(\theta) + c(\theta) + d(y)\}$$

formában írhatók fel. A sűrűségfüggvényben θ a paraméterek vektora. Ezek tetszőlegesen megválaszthatók azzal a feltétellel, hogy f az y változó szerint sűrűségfüggvény legyen, azaz nemnegatív, és az integrálja 1.

Egy eloszlást természetes paraméterezésűnek hívunk, ha $a(y) = y$. A továbbiakban természetes paraméterezésű eloszlásokkal foglalkozunk, azaz

$$f(y) = exp\{yb(\theta) + c(\theta) + d(y)\}.$$

Az ismert eloszlások nagy része valamilyen exponenciális eloszláscsaládba tartozik, vannak köztük diszkrét és folytonos eloszlások is. Ezért érdemes megvizsgálni néhány speciális esetet, hogy b, c, d függvények ügyes választásával hogyan vezethetjük vissza a sűrűségfüggvényt a szokványos paraméterekre. Speciálisan vizsgáljuk a Poisson, Normális és Binomiális eloszlások esetén a b, c és d paraméterfüggvényeket.

	$b(\theta)$	$c(\theta)$	$d(y)$	Sűrűségfüggvény
Poisson:	$log(\theta)$	$-\theta$	$-log(y!)$	$\frac{\theta^y \cdot e^{-\theta}}{y!}$
Normális:	$\frac{\mu}{\sigma^2}$	$-\frac{\mu}{\sigma^2} - \frac{1}{2}log(2\pi\sigma^2)$	$-\frac{y^2}{2\sigma^2}$	$\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Binomiális:	$log(\frac{p}{1-p})$	$n \cdot log(1-p)$	$log(\frac{n}{y})$	$(\frac{n}{y}) \cdot p^y \cdot (1-p)^{n-y}$

Az paraméterezés után látható, hogy a fenti eloszlások valóban az exponenciális családba tartoznak. Az általános állítások megfogalmazásához azonban nincs szükség ezekre az átalakításokra.

3.2. A modell működési elve

Az általánosított lineáris modell a lineáris regresszió, exponenciális eloszláscsaládból származó célváltozókra vett általánosítása. Legyenek adottak $X_1, X_2 \ldots, X_n$ magyarázó változók és Y exponenciális családból származó célváltozó. Ekkor az Y változó várható értéke kifejezhető a magyarázó változók segítségével az alábbi módon:

$$E(Y) = \mu = g^{-1}(\sum_{i=1}^{n} X_i \beta_i),$$

a g az úgynevezett linkfüggvény, β_i-k pedig az egyes változókhoz tartozó együtthatók.

A megfigyelésünk az alábbi sűrűségfüggvényből származik

$$f(y) = exp\{y \cdot b(\theta) + c(\theta) + d(y)\}.$$

Ha ismert Y eloszlása, meghatározható g linkfüggvény. A meghatározás sok esetben nem triviális, ezért itt nem részletezzük. A $g(\mu)$ linkfüggvény néhány speciális eloszlás esetén:

Gamma:	$\frac{1}{\mu}$
Poisson:	$log(\mu)$
Normális:	μ
Binomiális:	$log(\frac{\mu}{1-\mu})$

A regresszió illesztése során először a gyakorlati probléma ismerete alapján feltételezzük Y eloszlását. Ez után megadjuk a hozzá tartozó linkfüggvényt. A linkfüggvény segítségével Y értékei helyett $g(Y)$-ra lineáris regressziót alkalmazunk, melynek során maximum-likelihood vagy legkisebb négyzetek módszerének használatával megbecsüljük β_i paramétereket. Ez után visszahelyettesítve $g^{-1}(\sum X_i \beta_i)$ megadja a legjobb modellt, amellyel Y várható értékét megbecsülhetjük X_i-k ismeretében.

3.3. Logisztikus regresszió

Legyen Y célváltozó, és $X_1, X_2, \ldots, X_n$ magyarázó változók. Tegyük fel, hogy Y binomiális eloszlású. Az X_i változók ismeretében szeretnénk megbecsülni Y paraméterét.

Mivel a binomiális eloszlás az exponenciális eloszláscsaládba tartozik, alkalmazható az általánosított lineáris regresszió. Binomiális esetben a paraméter éppen a várható érték. Az úgynevezett logit linkfüggvényt alkalmazzuk

$$g(x) = logit(x) = log(\frac{x}{1-x}).$$

Ennek az inverze

$$g^{-1}(x) = \frac{exp(x)}{1+exp(x)}.$$

Ezzel a legjobb közelítés a binomiális eloszlás paraméterére

$$p^* = E(Y) = g^{-1}(a_1X_1 + \cdots + a_nX_n + b) = \frac{exp(a_1X_1 + \cdots + a_nX_n + b)}{1+exp(a_1X_1 + \cdots + a_nX_n + b)}.$$

A feladat tehát megbecsülni $a_1, \ldots a_n, b$ együtthatókat.

A minta ismeretében fel tudjuk írni a likelihood függvényt, ahol Y közelítését p^* adja. A loglikelihood függvény minimalizálásával megkapjuk a legjobb becsléseket a_i együtthatókra.

Elképzelhető, hogy X_i magyarázó változók diszkrét eloszlásúak, $c_1, c_2, \ldots c_k$ értékeket vesznek fel. Ebben az esetben indikátor változók bevezetésével használható a módszer. Ezeket dummy-változóknak nevezzük és D_j-vel jelöljük, ahol D_j 1-et vesz fel, ha $X_i = c_j$, egyébként pedig 0-t, minden $j = 1, 2, \ldots, k$.

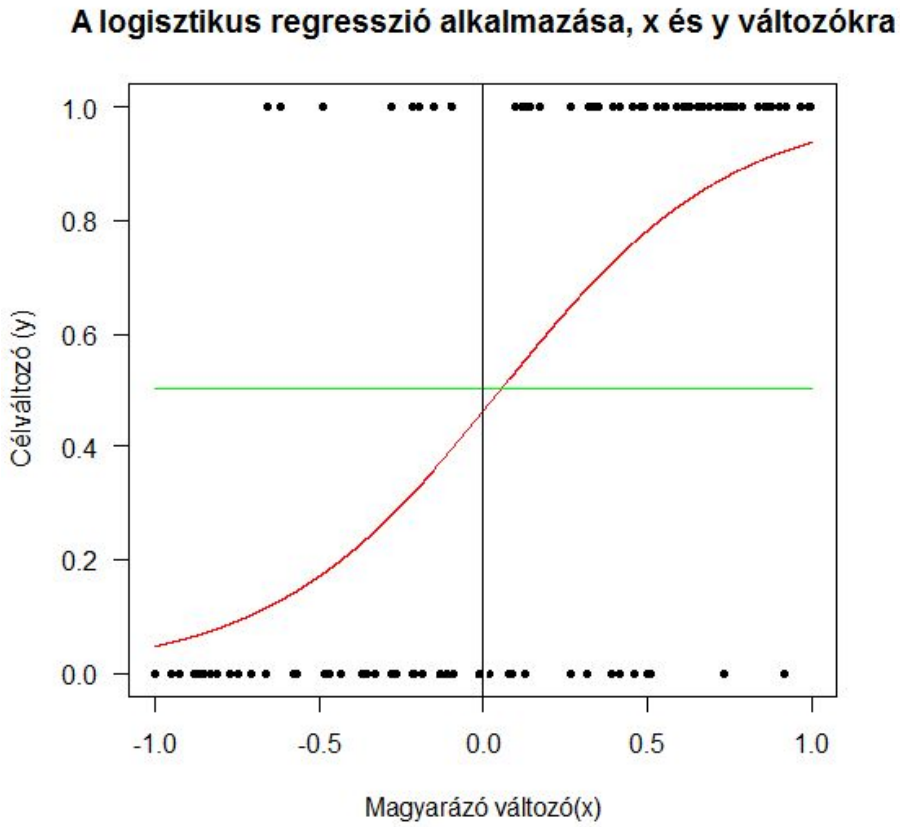
3.3.1. Szimuláció

Legyen Y a függő, X pedig a magyarázó változó. Legyen X egyenletes eloszlású $(-1, 1)$-en. A két változó között kapcsolat, hogy $Y = 0$, ha $X < 0$ és $Y = 1$, ha $X \geq 0$. Az összefüggésbe itt is vegyünk bele egy E-vel jelölt $N(0, 0.5)$ normális eloszlású hibát. Így tehát $Y = I(X + E \geq 0)$.

A logisztikus regressziós modellben feltételezzük, hogy Y binomiális eloszlású. Rögzített X esetén próbáljuk megadni Y paraméterét, vagyis annak a valószínűségét, hogy $Y = 1$. Vegyünk egy x vektort és töltsük fel 100 darab $U(-1, 1)$ eloszlásból származó elemmel. Legyen *eps* a szintén 100 elemű $N(0, 0.5)$ normális hibavektor. Ezekből kiszámíthatóak az y értékek. Az alábbi programmal logisztikus regressziós modellt illesztünk:

```
set.seed(521)
darab<-100
x<-runif(darab,-1,1)
eps<-rnorm(darab,0,0.5)
y<-rep(0,darab)
y[x+eps>0]<-1
LOG<-summary(glm(y~x,family="binomial"))
```

```
darab<-darab*100
xx<-c(-darab:darab)/darab
or<-exp(LOG$coefficients[1,1]+LOG$coefficients[2,1]*xx)
plot(x,y, main="A logisztikus regresszió alkalmazása,
     x és y változókra",xlab="Magyarázó változó(x)",
     ylab="Célváltozó (y)",las=1,pch=20)
lines(xx,or/(or+1), type="l", col="red")
lines(xx,rep(0.5,2*darab+1), col="green")
abline(v=0)
max(xx[or/(or+1)<0.5])
```



4. ábra. A logisztikus regresszió szemléltetése szimulált adatokon

Az 4. ábra tartalmazza y vektort x függvényében, piros vonallal jelölve a modell által becsült valószínűségeket arra vonatkozóan, hogy adott x érték esetén milyen valószínűséggel lesz $y = 1$, illetve az 50%-os valószínűség vonalát.

4. Modellezés strukturális egyenletekkel

A strukturális egyenletekkel való modellezés (Structural equation modeling), röviden SEM egy általános többváltozós módszer, amely speciális esetként tartalmazza többek között a lineáris regresszió típusait, vagy a faktor analízist. Alkalmas olyan változók felderítésére is, amelyeket nem tudunk mérni, viszont kapcsolatban állnak más, mérhető változókkal. Ezeket a nem mérhető változókat látens változóknak szokás hívni. A SEM a meglévő adatokból és egy kovariancia struktúrából olyan együtthatókat ad vissza, melyek a megadott struktúra szerint legjobban jellemzi a változók lineáris kapcsolatát. A módszer abban különbözik a regressziós és faktor analízistől, hogy a változók egymásra hatásának irányát is külön kezeli, valamely esetben függetlenséget is feltételezhet. Ez a módszerben úgy nyilvánul meg, hogy a változókhoz egy gráf struktúrát rendelünk, ami segítségével fogunk majd számolni.

A változók között kétféle kapcsolatot feltételezhetünk, korrelációt, vagy regressziós összefüggést. Első esetet irányítatlan éllel ábrázoljuk a gráfon, míg második esetben a magyarázó változó felől a magyarázottba futó irányított éllel jelöljük az összefüggést. A SEM-modellel vizsgált változók kapcsolatáról készíthető egy $G = (V, D, B)$ gráf, ahol V a csúcsok, D az irányított, B pedig az irányítatlan élek.

4.1. Markov-ekvivalens modellek

A statisztikában problémát okozhat, hogy a két változó esetén nem lehet megállapítani a kovariancia alapján, hogy melyik hat a másikra. Ennek a következménye, hogy többféle gráf is létezik, amelyek esetén az élek típusa más, de az élek által reprezentált kovariancia, vagy regresszió értéke megegyezik. Az ilyen módon különböző esetekben azonban a változók közötti indirekt hatás megváltozik.

4.1. Definíció. *Legyen adott egy kovariancia struktúra, $X_1, X_2, \ldots, X_n$ között, és legyen M a struktúrához kiszámolt modell. Azokat struktúrákat, amelyekben az élek végpontjai megegyeznek, de típusai, vagy iránya más lehet, továbbá az élekhez tartozó kovariancia értékek megegyeznek egy osztályt alkotnak. Az így kapott eltérő modelleket Markov-ekvivalens modelleknek nevezzük.*

Tegyük fel először, hogy a struktúra csak regressziós kapcsolatokat tartalmaz, azaz a gráf csupa irányított élekből áll. Adott M modellhez kereshető Markov-ekvivalens modell az alábbi szabály alapján:

- Legyen $X, Y \in V$ és $(X \to Y) \in D$. Ha minden $(Z \to X) \in D$, $Z \in V$ esetén

$(Z \to Y) \in D$, akkor $X \to Y$ él kicserélhető $Y \to X$ élre. Azaz az X-be vezető csúcsokból vezetnie kell élnek Y-ba is.

1. Megjegyzés. *Egy konkrét modell esetén lehet, hogy csak néhány él fordítható meg a szabály alapján. Azonban a módosított modellek esetén újabb megfordítható élek keletkezhetnek. Egy ekvivalencia osztályba tartoznak tehát azok a modellek, melyek a fenti szabály többszöri egymás utáni alkalmazásával egymásba vihetők.*

Most vegyünk egy olyan M modellt, ahol a feltételezett kovariancia struktúra $G = (V, D, B)$ gráf által meghatározott, azaz tartalmazhat irányítatlan éleket is. Ekkor a cserélési szabályok a következők:

- Legyen $X, Y \in V$ és $(X \to Y) \in D$. Ha minden Z csúcs esetén $Z \to X \in D$ él csak akkor szerepel, ha $Z \to Y \in D$ is, továbbá $(Z, X) \in B$ csak akkor szerepel, ha $(Z, Y) \in B$ is, akkor $(X \to Y) \in D$ él kicserélhető $(X, Y) \in B$ élre.

- Legyen $X, Y \in V$ és $(X \to Y) \in D$. Ha minden olyan $Z \in V$-re, amelyre
  - $(Z, X) \in B$, a G gráf tartalmazza $(Y \to Z) \in D$ élet,
  - továbbá megfordítva, minden olyan $Z \in V$-re, amelyre $(Z, Y) \in B$, a G gráf tartalmazza $(X \to Z) \in D$ élet
  - valamint minden $Z \in V$-re, amelyre $(Z \to X) \in D$, $(Z \to Y) \in D$ él is szerepel a gráfban,

  akkor a $(X \to Y) \in D$ él kicserélhető $(Y \to X) \in D$ élre.

4.2. Alkalmazhatóság feltételeihez szükséges alapok

Ebben az alfejezetben bevezetjük azokat a definíciókat és jelöléseket, amelyekre a továbbiakban szükségünk lesz. Ezek a definíciók a gráfelmélet és lineáris algebra ide vágó területeit érintik. Az alfejezetben említett definíciók [8] és [9]-ben szerepelnek.

Tegyük fel, hogy a (V, B) irányítatlan részgráf egyszerű, azaz nincsenek többszörös és hurok élek. Ezen kívül az irányított (V, D) részgráf legyen körmentes, vagy más néven aciklikus, amiből következően itt sem található hurok él. A csúcsokat jelöljük $V = 1, ..., m$ indexekkel.

4.2. Definíció. *Jelöljük $\mathbb{R}^D$-vel azon valós számokon értelmezett $m \times m$-es mátrixok halmazát, amelyre az (i, j) koordináta 0, ha $i \to j$ él nem szerepel D élhalmazban.*

4.3. Definíció. *Jelöljük $\mathbb{R}^D_{reg}$-gel azon $\Lambda \in \mathbb{R}^D$-beli mátrixok halmazát, amelyre $I-\Lambda$ invertálható.*

2. Megjegyzés. *Mivel (V, D) aciklikus, létezik a csúcsoknak topologikus rendezése. Ha eszerint indexeljük a csúcsokat, akkor $\mathbb{R}^D$ elemei szigorú felső háromszög mátrixok lesznek. Ebből adódóan minden $\Lambda \in \mathbb{R}^D$ esetén $\det(I - \Lambda) = 1$, vagyis invertálható. Azaz $\mathbb{R}^D = \mathbb{R}^D_{reg}$. Emellett $I - \Lambda$ inverze sorba fejthető, amely a szigorú felső háromszög tulajdonságait felhasználva véges sor lesz. Ezért Φ_G polinomiális leképezés lesz.*

4.4. Definíció. *Jelöljük $PD(B)$-vel azon $m \times m$-es pozitív szemidefinit mátrixok halmazát, amelyre az (i, j) koordináta 0, ha i és j között nem fut él B élhalmazban.*

4.5. Definíció. *Egy $G = (V, D, B)$ aciklikus vegyes gráfhoz tartozó lineáris strukturális egyenlet modell egy $N_m(0, \Sigma)$ többváltozós normális eloszláscsalád, melyre*

$$\Sigma = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1},$$

ahol $\Sigma \in \mathbb{R}^D_{reg}$ és $\Omega \in PD(B)$.

Egy i csúcs szüleinek nevezzük azon $j \in V$ pontokat, melyekre $(j \to i) \in D$, és ezek halmazát $pa(i)$-vel jelöljük. Ennek segítségével felírható a lineáris strukturális egyenletrendszer:

$$Y_j = \sum_{i \in pa(i)} \Lambda_{i,j} Y_i + \varepsilon_j.$$

Itt ε egy $N(0, \Omega)$ többdimenziós normális eloszlásból származó véletlen hibavektor. Ekkor $Y_1, \ldots, Y_m$ jól definiált megoldását adja az egyenletrendszernek, valamint $N_m(0, \Sigma)$ többdimenziós normális eloszlást követ.

Egy statisztikai elemzés során Σ lesz az, amit ki tudunk számolni. Ennek segítségével szeretnénk meghatározni Λ, és Ω mátrixokat. A mátrixok elemeiből felírható a strukturális egyenletrendszer, amiből megkaphatjuk a változók közötti összefüggéseket. A mátrixok meghatározása akkor lehetséges egyértelműen, ha

$$\Phi_G : (\Lambda, \Omega) \to (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}$$

leképezés injektív.

4.6. Definíció. *Egy irányított (V, D) gráfot faszerűen konvergensnek nevezünk, ha élei egy fát alkotnak és létezik egy olyan i pontja, hogy minden $j \in V$ pontra igaz, hogy létezik $j \to i$ út a D éleken haladva. Másképpen fordított fenyőnek, vagy befenyőnek is nevezik (V, D) gráfot, illetve fogalmazhatunk úgy, hogy a gráf tartalmaz globális nyelő csúcsot.*

4.2.1. Injektivitási feltétel

Bár még konkrét feltételt nem ismerünk a módszer alkalmazhatóságára, hogyan lehetne mégis egyszerűsíteni G gráf esetén a Φ_G injektivitására vonatkozó feltételt? Erre a kérdésre kapunk választ a [8]-ban.

4.7. Lemma. *Tegyük fel, hogy a G vegyes gráfhoz tartozó Φ_G leképezés injektív. Ekkor G egyszerű gráf és minden H részgráfjára Φ_H is injektív.*

Bizonyítás. Legyen $H = (V', D', B')$, a $G = (V, D, B)$ részgráfja, ahol $V' \subseteq V$, $D' \subseteq D$ és $B' \subseteq B$. Vegyünk egy $\Lambda \in \mathbb{R}^{D'}_{reg}$, és $\Omega \in PD(B')$ megfelelő mátrixot Φ_H leképezéshez. Terjesszük ki őket $\Lambda^* \in \mathbb{R}^{D}_{reg}$, és $\Omega^* \in PD(B)$ mátrixokká úgy, hogy a H-n kívüli élekhez tartozó érték mindenütt 0 legyen. Ebből látszik, hogy Φ_H akkor és csak akkor injektív, ha Φ_G is injektív a Λ^*, és Ω^* pontokat tartalmazó részhalmazon.
Ha G nem egyszerű gráf, akkor létezik i, j pont, hogy több él is fut közöttük. Ha G gráf csak ebből a két pontból áll, akkor $\mathbb{R}^{D}_{reg} \times PD(B)$ legalább 4 dimenziós, mivel a két él két szabad paraméter, valamint a pozitív szemidefinit mátrixok főátlójában újabb két érték megválasztható (megkötéssel ugyan, de ez nem csökkenti a dimenziót). Az élekhez tartozó paraméterek a regressziós, vagy kovariancia paraméterei, míg a főátlóban a szórásparaméterek találhatóak. Φ_G képe azonban a $2 \cdot 2$-es mátrixok 3 dimenziós, pozitív szemidefinit kúpja, így a leképezés nem lehet injektív. Ha G több pontból áll, akkor nézhetjük i és j által generált részgráfját és az állítás így is fennáll. □

4.2.2. Stepwise inverzió

Ebben a részfejezetben egy olyan algoritmust ismertetek, amelynek segítségéből tetszőleges G gráf esetén, tetszőleges kovariancia mátrixból kiszámolhatók a modell együtthatói, vagy a számolás során megkapjuk a megoldhatóságot korlátozó tulajdonságot.

Tegyük fel, hogy G aciklikus vegyes gráf, m csúccsal. Szeretnénk egy olyan eljárást, amely ismert $\Sigma = \Phi_G(\Lambda, \Omega)$ esetén megadja a paraméter mátrixokat (ha nem egyértelmű, akkor egy tetszőleges (Λ, Ω) mátrix párt, amire teljesül). Technikailag itt Φ_G függvény inverzét szeretnénk meghatározni adott pontban. Ezt lépésenként fogjuk elvégezni, egyre több csúcsot figyelembe véve.

4.8. Definíció. *Minden $i \leq m - 1$-re $P(i)$ jelentse azon $j \leq i$ csúcsok halmazát, amelyből irányított él fut $i + 1$-be, nevezzük ezeket az $i + 1$-edik csúcs szüleinek $S(i)$*

pedig azokat a $j \leq i$ csúcsokat, melyek irányítatlan éllel kapcsolódnak $i+1$-hez, ezek a csúcs testvérei.

Az algoritmus

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}$$

egyenletben szeretnénk megkapni Λ és Ω értékeit, ha ismerjük Σ-t. Topologikusan rendezzük a csúcsokat, én eszerint írjuk fel az egyenletet, ekkor $\sigma_{1,1} = \omega_{1,1}$ rögtön adódik, továbbá mivel Λ szigorú felső háromszög, az első oszlopa csupa 0. Így a Λ és Ω mátrixok bal felső eleme megkapható. Induktívan tegyük fel, hogy egy $i \geq 1$ esetén a bal felső $i \times i$-s részmátrixot már ismerjük. Nézzük meg, hogy az $i+1$ csúcs bevonásával mit tudunk kiszámítani.

$$(I - \Lambda)_{[i+1],[i+1]} = \begin{pmatrix} \Gamma & -\lambda \\ 0 & 1 \end{pmatrix}$$

$$\Omega_{[i+1],[i+1]} = \begin{pmatrix} \Psi & \omega \\ \omega^T & \omega_{i+1,i+1} \end{pmatrix}$$

Ahol Γ és Ψ a már ismert $i \times i$ méretű részmátrixok. Ebből felírható Σ egyenlete az $i+1 \times i+1$ méretű bal felső részmátrixra:

$$\Sigma_{[i+1],[i+1]} = \begin{pmatrix} \Gamma^{-T}\Psi\Gamma^{-1} & \Gamma^{-T}\Psi\Gamma^{-1}\lambda + \Gamma^{-T}\omega \\ (\Gamma^{-T}\Psi\Gamma^{-1}\lambda + \Gamma^{-T}\omega)^T & \omega_{i+1,i+1} + \lambda^T\Gamma^{-T}\Psi\Gamma^{-1}\lambda + 2\omega^T\Gamma^{-1}\lambda \end{pmatrix}$$

Ebből a felírásból látszik, hogy λ és ω a jobb felső helyen álló vektorban szerepel. Így λ és ω akkor és csak akkor lesz egyértelmű, ha a $\Sigma_{[i],\{i+1\}} = \Gamma^{-T}\Psi\Gamma^{-1}\lambda + \Gamma^{-T}\omega$ egyenletnek egyértelmű megoldása létezik.

A λ vektorban azokhoz a csúcsokhoz tartozó élek szerepelnek nem 0 értékkel, melyekkel az $i+1$-edek csúcs össze van kötve. A topologikus rendezés miatt ezek csupán a szülő csúcsai lehetnek. Az ω vektorban hasonló állítás igaz az $i+1$-edik csúcs testvéreire. Így az egyenlet átírható

$$\Sigma_{[i],\{i+1\}} = \left(\Gamma^{-T}\Psi\Gamma^{-1}_{[i],P(i)}\right)\lambda_{P(i)} + \left(\Gamma^{-1}_{[i],S(i)}\right)^T \omega_{S(i)}$$

alakba. Azokat a sorokat nem tüntettük fel a mátrixokban, amelyek, amúgy sem befolyásolnák az eredményt. Ebből azt kapjuk, hogy egyértelmű megoldásuk akkor lesz, ha

$$\left(\Gamma^{-T}\Psi\Gamma^{-1}_{[i],P(i)} \;\; \left(\Gamma^{-1}_{[i],S(i)}\right)^T\right)$$

mátrix teljes rangú lesz. Ez az érték a vektorok hosszának összege, vagyis $i+1$-edik csúccsal valamilyen módon összekötött csúcsok számának összege, azaz $P(i)+S(i)$ elemszáma.

A feltételt szebb alakra hozhatjuk, ha balról megszorozzuk Γ^T mátrixszal, mivel ez egy felső háromszög mátrix, a diagonálisában csupa 1-essel, így nem fogja megváltoztatni a rangot. Ekkor a vizsgált mátrix:

$$\left(\Psi\Gamma^{-1}_{[i],P(i)} \;\; Id_{[i],S(i)}\right)$$

Mivel a jobb oldali blokk egy egységmátrix része, így a rangfeltétel átfogalmazható úgy, hogy a bal oldali blokk megfelelő $\Psi_{[i]\setminus S(i),[i]}\Gamma^{-1}_{[i],P(i)}$ részének a rangja legyen $P(i)$.

Ha ez teljesül, akkor λ és ω egyértelműen meghatározható, és a segítségükkel $\omega_{i+1,i+1}$ is megkapható, ha Σ jobb alsó elemére felírt összefüggést alkalmazzuk. Mivel Ω pozitív szemidefinit kell, hogy legyen, ezért további szükséges feltétel, hogy a kapott $\omega_{i+1,i+1}$ nemnegatív legyen.

A $\Sigma = (I-\Lambda)^{-T}\Omega(I-\Lambda)^{-1}$ képletben Ω mátrix konjugálva van $(I-\Lambda)^{-1}$-zel. A konjugálás azonban nem változtatja meg a determináns értékét. Így mivel Σ és annak minden főminorja is pozitív szemidefinit, Ω is az kell, hogy legyen. Ez azt eredményezi, hogy ha Σ korreláció mátrix, akkor az algoritmus minden lépésében az adott $\omega_{i,i}$ nemnegatív lesz, és a végső Ω mátrix is szintén pozitív szemidefinit.

Amennyiben a rangfeltétel nem teljesül, úgy szabad paraméterek állnak rendelkezésünkre, ezeket tetszőlegesen választva az algoritmus tovább folytatható, de a végső megoldásunk nem lesz egyértelmű.

4.9. Lemma. *Legyen $G=(V,D,B)$ aciklikus vegyes gráf, topologikusan rendezett csúcsokkal. A Φ_G injektív akkor és csak akkor, ha*

$$rang(\Omega_{[i]\setminus S(i),[i]}(I-\Lambda)^{-1}_{[i],P(i)}) = |P(i)|$$

fennáll minden $i \in 1,...m-1$-re, $\Lambda \in \mathbb{R}^D$-re és $\Omega \in PD(B)$-re.

Bizonyítás. Az algoritmus lépéseit használva az egyértelműséghez az kell, hogy a mátrix teljes rangú legyen. Ezért a rangfeltétel teljesülése esetén a leképezés injektív lesz. Tegyük fel, hogy a rangfeltétel nem teljesül egy $i \leq m-1$ csúcs esetén. Az algoritmussal kapott mátrixok legyenek Λ és Ω. Ha $i=m-1$ akkor $\Phi(\Lambda,\Omega)$ pont legalább egy dimenziós részhalmaz képeként előáll, így Φ_G nem injektív. Az $i<m-1$ esetekben az indukált részmátrixot véve hasonló állítást kapunk Φ_G megszorítására, amiből következik az injektivitás hiánya a korábbi lemma alapján. $\square$

Ha egy adott G gráf, és hozzá tartozó Σ kovariancia mátrix esetén, a rangfeltétel nem teljesül, így Λ és Ω mátrixokat nem lehet egyértelműen kiszámítani, még létezhet olyan másik Σ^* kovariancia mátrix, melyre mégis egyértelmű a feladat megoldása.

1. Példa. *Az alábbi példában azt a 4 csúcsú teljes vegyes gráfot vizsgálom, melyben az irányított élek $1 \to 2; 2 \to 4; 3 \to 4$, a többi él pedig irányítatlan. Válasszuk Σ mátrixot a csupa 1 mátrixnak. A gráf struktúrája miatt Ω és Λ mátrixok a következőféleképpen néznek ki.*

$$\Omega = \begin{pmatrix} d_1 & 0 & \omega_1 & \omega_2 \\ 0 & d_2 & \omega_3 & 0 \\ \omega_1 & \omega_3 & d_3 & 0 \\ \omega_2 & 0 & 0 & d_4 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_2 \\ 0 & 0 & 0 & \lambda_3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

A feladat megoldhatósága Σ mátrixon múlik, ami most legyen:

$$\Sigma = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

A mátrix determinánsa 0 és minden főminor ≥ 0 így pozitív szemidefinit, azaz elképzelhető, hogy ezt kapjuk kovariancia mátrixként. Vizsgáljuk meg, hogy milyen eredményre jutunk a lépésenkénti algoritmus során.

- *1. lépés*

  $d_1 = \sigma_{1,1} = 1$

- *2. lépés*

  $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \lambda_1 \\ \lambda_1 & d_2 \end{pmatrix}$ *amiből egyértelműen látszik, hogy $\lambda_1 = 1$ és $d_2 = 1$.*

- *3. lépés*

  Itt a Λ-ból származó együtthatók mind 0-k, így a megfelelő mellékszámítások elvégzésével:

  $\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \omega_1 \\ \omega_3 \end{pmatrix}$

  amiből $\omega_1 = 0$ és $\omega_3 = 1$ eredmény kapható. Továbbá $d_3 = 1$ a harmadik lépés utolsó egyenletének megoldása.

- *4. lépés*

  Írjuk fel a számításokhoz szükséges segédmátrixokat:

$$\Gamma^{-T}\Psi\Gamma^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \qquad \Gamma^{-T} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

  A vektoregyenlet a következőképpen alakul:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \omega_3 \\ 0 \\ 0 \end{pmatrix}$$

  Itt most $i = 3$ lépésnél járunk, azaz a 4. csúcsot vizsgáljuk. Ennek szomszédja az 1. csúcs, szülei a 2. és 3. csúcs. A szülők száma tehát 2, amiből következik, hogy $rang(\Omega_{[3]\setminus S(3),[3]}(I-\Lambda)^{-1}_{[3],P(3)}) = 2$ egyenlőségnek kellene teljesülnie.

$$\Omega_{[3]\setminus S(3),[3]}(I-\Lambda)^{-1}_{[3],P(3)} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Látható, hogy a sorok összefüggnek, így a rang csak 1, azaz nem teljesül a feltétel. Az algoritmus lépését végrehajtva azt kapjuk, hogy λ_2 szabad paraméter, $\lambda_3 = 1-\lambda_2$, $\omega_2 = 0$, $d_4 = 0$. Azaz a Λ mátrix tényleg nem egyértelmű.

$$\Omega = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_2 \\ 0 & 0 & 0 & 1-\lambda_2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

4.2.3. A gráf-feltétel szükségessége

A [8]-ben tárgyalt fő tétel egy szükséges és elégséges feltételt ad a gráf struktúráját tekintve arra, hogy mikor lehet minden Σ kovariancia mátrix esetén egyértelműen meghatározni Λ és Ω mátrixokat. Ez a feltétel nem azt jelenti, hogy hibás struktúrájú gráfok esetén nem kaphatunk egyértelmű megoldást, csak azt, hogy ilyen esetben valamely Σ esetén nem oldható meg a feladat.

4.10. Tétel. *Egy aciklikus vegyes $G = (V, D, B)$ gráf által meghatározott Φ_G leképezés nem injektív akkor és csak akkor, ha létezik G-nek egy G_A részgráfja, amelynek az irányított része faszerűen konvergens, irányítatlan része pedig összefüggő.*

Gráf-feltételnek nevezzük azt, hogy ha a fenti tétel feltétele nem teljesül. Meggondolható, hogy legfeljebb három pontú gráfokra a gráf-feltétel mindig teljesül, mert ha az irányított rész fa, akkor az irányítatlan nem lehet összefüggő.

A tétel bizonyítását a következő lemmák adják.

4.11. Lemma. *Legyen $G = (V, D, B)$ egy $m+1$ csúcsú aciklikus vegyes gráf, aminek a csúcsait topologikusan rendeztük. Ha az irányított rész faszerűen konvergál az $m+1$-edik csúcshoz, az irányítatlan rész pedig egy feszítő fa, akkor léteznek olyan $\Lambda \in \mathbb{R}^D$ és $\Omega \in PD(B)$ mátrixok, hogy*

$$\Omega_{[m]\setminus S(m),[m]}(\mathbb{I} - \Lambda)^{-1}_{[m],P(m)}$$

magtere nem üres.

Ha egy $G = (V, D, B)$ gráf ilyen, akkor a hozzá tartozó Φ_G nem injektív. Ha egy részgráfra teljesülnek ezek a feltételek, akkor a részgráf és gráf közti összefüggést leíró lemma miatt szintén nem lehet injektív. Az állítás bizonyításához a továbbiakban leírt két lemmát használja fel a [8] cikk.

Legyen $L(\Lambda) \subseteq \mathbb{R}^D$ az $(I - \Lambda)^{-1}_{[m],P(m)}$ mátrix oszlopai által generált lineáris altér.

4.12. Lemma. *Ha (V, D) egy $m + 1$ csúcsú, faszerűen konvergáló irányított gráf, amelynek nyelője az $m + 1$-es csúcs, akkor a $\Lambda \in \mathbb{R}^D$ mátrixokhoz tartozó $L(\Lambda)$ lineáris alterek uniója tartalmazza az $(\mathbb{R}\setminus\{0\})^m$ halmazt.*

$R(m)$-mel definiáljuk azokat a csúcsokat, amelyek az m-edik csúcsnak nem testvérei, azaz $R(m) = [m]\setminus S(m)$.

4.13. Lemma. *Legyen (V, B) egy $m + 1$ csúcsú fa. Ekkor létezik olyan $\Omega \in PD(B)$ mátrix, amelyre $\Omega_{R(m),[m]}$ magtere tartalmazza az egységvektort.*

Vegyük a legszűkebb olyan G' részgráfot vesszük, amelyre teljesülnek a tétel feltételei. Ezek teljesítik a fenti két lemma feltételét is. Választható egy olyan Ω, mely az egységvektort a 0-ba viszi. Mivel $(\mathbb{R}\setminus\{0\})^m$ halmaz része az irányított rész mátrixának oszlopai által kifeszített lineáris térnek, választható egy olyan Λ, melyre

$$\Omega_{R(m),[m]}(Id - \Lambda)^{-1}_{[m],P(m)}$$

mátrix az egységvektort a 0-ba viszi, azaz a magtere nem üres. Ekkor pedig $\Phi_{G'}$ nem lesz injektív. Mivel egy részgráfra $\Phi_{G'}$ nem injektív, így az eredeti G gráf esetén sem lesz Φ_G injektív.

4.2.4. A gráf-feltétel elégségessége

Most belátjuk, hogy ha $G = (V, D, B)$ gráfhoz tartozó Φ_G leképezés injektív, akkor teljesülnie kell a gráf-feltételnek. Itt tehát feltesszük, hogy Φ_G nem injektív és belátjuk, hogy egy részgráfjára teljesülnie kell a gráf-feltételnek.

4.14. Állítás. *Legyen $G = (V, D, B)$ egy $m+1$ csúcsú aciklikus vegyes gráf, melynek csúcsait topologikusan rendeztük. Legyen $\Lambda \in \mathbb{R}^D$, $\Omega \in PD(B)$ valamint $\alpha \in \mathbb{R}^{|P(m)|}$ nem nulla vektor úgy, hogy*

$$\Omega_{R(m),[m]}(Id - \Lambda)^{-1}_{[m],P(m)}\alpha = 0.$$

Tegyük fel, hogy a gráf irányított része nem tartalmaz fordított fenyőt. Legyen A azon $v_i \in V$ csúcsok halmaza, melyekből irányított úton elérhető az v_{m+1}, továbbá legyen még $v_{m+1} \in A$. Ekkor $A \subsetneq V$ és Φ_{G_A} nem injektív.

4.15. Állítás. *Legyen $G = (V, D, B)$ egy $m+1$ csúcsú aciklikus vegyes gráf, melynek csúcsait topologikusan rendeztük. Legyen $\Lambda \in \mathbb{R}^D$, $\Omega \in PD(B)$ valamint $\alpha \in \mathbb{R}^{|P(m)|}$ nem nulla vektor úgy, hogy*

$$\Omega_{R(m),[m]}(Id - \Lambda)^{-1}_{[m],P(m)}\alpha = 0$$

Tegyük fel, hogy a gráf irányítatlan része nem összefüggő. Legyen A azon csúcsok halmaza, melyek összefüggnek $m+1$-gyel, ebbe beleértjük magát az $m+1$-edik csúcsot is. Ekkor $A \subsetneq V$ és Φ_{G_A} nem injektív.

Az állítások szerint ha találunk egy olyan α vektort, amelyre igaz, hogy benne van $\Omega_{R(m),[m]}(Id - \Lambda)^{-1}_{[m],P(m)}$ magterében, akkor találunk egy olyan részgráfot is, amelyre Φ nem lesz injektív, sőt meg is adható, hogy hogyan.

4.16. Állítás. *Ha $G = (V, D, B)$ gráf nem tartalmaz irányított befenyőt, valamint az irányítatlan része nem összefüggő, viszont Φ_G nem injektív, akkor létezik egy olyan A részgráfja, melyre Φ_A nem injektív.*

Mivel egy véges gráfnak csak véges számú részgráfja lehet, a tétel alapján létezik egy olyan részgráf, amelyre már igaz, hogy irányított befenyőt feszít ki, valamint irányítatlan értelemben összefüggő. Tehát ha nem injektív Φ_G, akkor valamely részgráfja biztosan nem teljesíti a gráf-feltételt.

4.2.5. Ciklikus gráf esete

Egy $G = (V, D, B)$ gráf elképzelhető, hogy tartalmaz irányított kört. Ebben az esetben a korábbi tétel nem alkalmazható, viszont megfogalmazható egy egyszerűbb állítás:

4.17. Tétel. *Ha egy G vegyes gráf tartalmaz kört, akkor Φ_G nem injektív.*

4.18. Lemma. *Legyen $G = (V, D, B)$ gráf egy $n \leq 3$ csúcsú irányított kör. Legyen $\Lambda \in \mathbb{R}^D_{reg}$ és $\Omega \in PD(B)$. Ekkor $\Sigma = \Phi(\Lambda, \Omega)$ mátrixhoz pontosan még egy olyan $(\Lambda^*, \Omega^*) \neq (\Lambda, \Omega)$ mátrix pár, amelyre $\Sigma = \Phi(\Lambda^*, \Omega^*)$ is teljesül.*

A lemma következménye, hogy Φ_G nem injektív. Így ha egy H gráf nem aciklikus, akkor részgráfként tartalmaz egy H' kört, melyre $\Phi_{H'}$ nem injektív, és így egy korábbi lemma miatt Φ_H sem lehet az.

4.3. Half-trek kritérium

Az előző alfejezetben olyan feltételt sikerült találni, amely egy G gráfról eldönti, hogy minden kovariancia mátrix esetén alkalmazható-e a SEM modell, léteznek-e a feltételezett együtthatók. Azonban a gyakorlatban, nem túl speciális kovariancia mátrixok esetén egy általánosabb gráfosztályon is megoldható a probléma. Ezzel a témával a [9] foglalkozik részletesen.

Legyen $\Theta := \mathbb{R}^D_{reg} \times PD(B)$ az összes lehetséges regressziós és korrelációs struktúrát leíró mátrixok halmaza. Legyen G egy n csúcsú gráf, $\Phi_G : \Theta \to \mathbb{R}^{n \times n}$ pedig az előző fejezetben már bevezetett leképezés, amely a lehetséges kovariancia mátrixok halmazába képez.

Amennyiben Φ_G nem injektív, de azok a kivételes (Λ, Ω) párok, amelyeknek több ősképük is lenne, egy nullmértékű halmazt alkotnak, a feladat majdnem mindenütt megoldható. Mivel Φ_G racionális leképezés, egy ilyen kivételes halmaz egy algebrai részhalmaz lehet.

4.19. Definíció. *Legyen S a Θ halmazon értelmezett $\mathbb{R}$-be képező polinomfüggvények halmaza, továbbá legyen tetszőleges $T \subset S$. Ekkor*

$$V(T) = \{\theta \in \Theta | f(\theta) = 0, \ ha \ f \in T\}$$

halmazt Θ egy algebrai részhalmazának nevezzük.

4.20. Definíció. *A G vegyes gráfot általánosan azonosíthatónak nevezzük, ha Φ_G injektív $\Theta \backslash V$ halmazon, ahol $V \subset \Theta$ egy algebrai részhalmaz.*

4.21. Definíció. *Egy vegyes G gráf racionálisan azonosítható, ha létezik egy algebrai $V \subset \Theta$ részhalmaz és egy racionális Ψ leképezés, hogy $\Psi \circ \Phi_G(\Lambda, \Omega) = (\Lambda, \Omega)$ minden $(\Lambda, \Omega) \in \Theta \backslash V$ esetén.*

Az általánosabb gráfosztály tehát az általánosan azonosítható és a racionálisan azonosítható gráfok. Ezek vizsgálatához további gráfelméleti fogalmak következnek.

4.22. Definíció. *Legyen $G = (V, D, B)$ egy vegyes gráf. Legyen $v_i, v_j \in V$ olyan, hogy létezik v_i és v_j között $v_i = v_0, v_1, v_2, \ldots, v_{k-1}, v_k = v_j$ séta, mely irányított vagy irányítatlan élet is tartalmazhat, valamint a következők egyike áll fenn:*

- *Létezik egy olyan $0 \leq s \leq k$ index, hogy minden $l < s$ esetén v_l, v_{l+1} között $v_{l+1} \to v_l$ irányított él szerepel, továbbá minden $l > s$ esetén v_{l-1}, v_l között $v_{l-1} \to v_l$ él fut.*
- *Létezik egy olyan $0 \leq s$, $s + 1 \leq k$ pár, hogy v_s, v_{s+1} között irányítatlan él fut, valamint minden $l < s$ esetén v_l, v_{l+1} között $v_{l+1} \to v_l$ irányított él szerepel, továbbá minden $l > s + 1$ esetén v_{l-1}, v_l között $v_{l-1} \to v_l$ él fut.*

Az ilyen i, j pontpár közt futó sétát trek-nek nevezzük.

Néhány példa trek-re:

$$① \longleftarrow ② \longrightarrow ③ \longrightarrow ④$$

$$① \longleftarrow ② \longleftrightarrow ③ \longrightarrow ④$$

$$① \longleftarrow ② \longleftarrow ③ \longleftarrow ④$$

Néhány példa olyan sétára, amely nem trek:

$$① \longleftarrow ② \longrightarrow ③ \longleftarrow ④$$

$$① \longleftrightarrow ② \longrightarrow ③ \longleftrightarrow ④$$

Szemléletesen ezt úgy lehet megfogalmaz, hogy ha találunk i és j között egy sétát úgy, hogy ha az irányítatlan élet kétirányúnak tekintve semelyik pontba nem fut be egyszerre két él, akkor trek-ről beszélünk.

4.23. Definíció. *Half-treknek nevezzük azt a trek-et, amennyiben a definícióban szereplő $s = 1$ vagy v_i-ből rögtön irányítatlan élen indulunk.*

Példák half-trek-re:

$$① \longleftarrow ② \longrightarrow ③ \longrightarrow ④$$

$$① \longleftrightarrow ② \longrightarrow ③ \longrightarrow ④$$

3. Megjegyzés. *Half-trek úgy keletkezhet, ha i-nek egyik szülőjéből, vagy testvéréből irányított úton elérhető j. Értelemszerűen a definíció úgy is áll, ha j-re teljesül ez a feltétel, hiszen akkor a pontokat felcserélve visszakapjuk az eredeti definíciót.*

Egy adott π trek esetén az irányítatlan éltől (amennyiben nem tartalmaz ilyet, akkor a definícióbeli s-edik csúcstól) balra eső csúcsokat $L(\pi)$ a jobbra eső csúcsokat pedig $R(\pi)$-vel jelöljük.

Egy alternatív definíció a half-trek tulajdonságra, hogy olyan π trek, melyben $L(\pi) = 1$.

4.24. Definíció. *Legyen $X = x_1, x_2, \ldots x_i$ és $Y = y_1, y_2, \ldots, y_i$ csúcshalmazok. Trek rendszernek nevezzük azon $\pi_1, \pi_2, \ldots, \pi_i$ trek-ek halmazát, melyekre igaz, hogy minden $j = 1, 2, \ldots, i$ esetén π_j x_j és y_j közötti trek.*

Amennyiben a halmazok közt futó trek-ek half trekek, akkor egy half trek rendszerről beszélünk.

4.25. Definíció. *Egy trek rendszerre azt mondjuk, hogy nincs oldalsó metszőpontja, ha minden π_i, π_j esetén $L(\pi_i) \cap L(\pi_j) = 0$ és $R(\pi_i) \cap R(\pi_j) = 0$.*

4.26. Definíció. *Egy $Y \subset V$ csúcshalmaz teljesíti a half-trek kritériumot, a $v \in V$ csúcsra nézve, ha*

- $|Y| = |pa(v)|$,
- *Y nem tartalmazza v-t és v testvéreit sem,*
- *létezik egy olyan half-trek rendszer Y és $pa(v)$ között, amelyben nincs oldalsó metszőpont.*

Jelöljük $htr(v)$-vel azon csúcsok halmazát, melyből létezik half trek v-be. A trek kritérium felhasználásával a következő tételek teljesülnek.

4.27. Tétel. *Legyen $G = (V, D, B)$ egy vegyes gráf, $(Y_v : v \in V)$ pedig a csúcsokon értelmezett részhalmazok egy családja. Ha*

- *minden $v \in V$ csúcsra Y_v teljesíti a half trek kritériumot,*
- *létezik a csúcsokon egy olyan teljes rendezés, hogy $v < z$ akkor és csak akkor, ha $v \in Y_z \cap htr(z)$*

akkor G racionálisan azonosítható.

4.28. Definíció. *Amennyiben egy G gráfhoz tartozó Φ_G leképezés esetén minden képpont végtelen sok (Λ, Ω) mátrixpár esetén megkapható, akkor azt mondjuk, hogy G-ben általánosan végtelen jut egyre.*

4.29. Tétel. *Legyen $G = (V, D, B)$ egy vegyes gráf. Ha minden $(Y_v : v \in V)$ csúcsokon értelmezett részhalmazok családjában*

- *létezik olyan $v \in V$ csúcs, amelyre Y_v nem teljesíti a half trek kritériumot v-re,*
- *vagy létezik olyan Y_v, Y_w pár, melyre $w \in Y_v$ és $v \in Y_w$,*

akkor G-ben általánosan végtelen jut egyre.

Összefoglalva [9] ad egy olyan bővebb gráfosztályt, amelyre majdnem mindig alkalmazható a SEM modell, illetve megad egy másikat, amelyről bizonyítja, hogy semmilyen körülmények között nem kaphatunk egyértelmű megoldást. A két osztály nem fedi le az összes gráfot, így létezhetnek ennél pontosabb kritériumok is.

4.4. A gráf-feltétel és a Markov-ekvivalens modellek

A fejezet elején definiáltuk a Markov-ekvivalens módszereket. Egy ekvivalencia osztály meghatározásához egy konkrét G gráfból indultunk ki és kicserélési szabályok használatával új gráfokat kapunk. A G által meghatározott rendszer statisztikailag egyértelműen meghatározható. Vajon az ekvivalencia osztály minden elemére igaz, hogy egyértelműen megadható hozzá Λ és Ω?

Ennek megválaszolásához a ellenőrizzük a gráf-feltételt a módosított gráfon. Legyen $G = (V, D, B)$ egyszerű aciklikus gráf, melyre teljesül a gráf-feltétel.

Első esetben G' legyen az a gráf, amelyben egy feltételnek megfelelő $X \to Y$ élet megfordítunk. Ekkor definíció szerint X testvérei Y gyerekei, Y testvérei egyben X gyerekei, továbbá X szülei egyben Y szülei is. Tegyük fel, hogy nem teljesül G'-re a gráf-feltétel. Az irányítatlan élek G'-ben összefüggő részgráfot kell, hogy alkossanak, ám mivel a csere során az irányítatlan élek nem változtak, már G-ben is összefüggőnek kellett lennie. G' tehát akkor nem teljesíti a feltételt, ha létezik benne globális nyelő úgy, hogy közben G-ről tudjuk, hogy benne nem volt. Mivel az irányítatlan rész összefüggő, X-nek és Y-nak is van testvére, ami a cserélési szabály alapján azt jeleni, hogy X-nek is és Y-nak is van gyereke. Y és X testvérei között nem lehet közös, mert ha S közös testvér lenne, egyben közös gyerek is lenne, azaz (X, S) között irányított és irányítatlan él is futna, így a gráf nem lenne egyszerű. Legyen X egyik gyereke K, Y egyik gyereke pedig L. Tegyük fel, hogy létezik G'-ben

fordított fenyő. Mivel G-ben még nem volt, ez csak úgy lehet, ha éppen $Y \to X$ éllel keletkezett fordított fenyő. Ekkor az Y biztosan nem a fenyő csúcsa. Ha X a fenyő csúcsa, akkor vegyük hozzá $X \to K$ élet is a fához. Így mindenképpen keletkezett egy kör. A kör át kell, hogy menjen $Y \to X$ élen, különben már G-ben is lett volna kör. Ekkor azonban $Y \to X$ él eltávolításával egy újabb fordított fenyőt kapunk, melynek csúcsa Y, és csupa olyan éleket használ, amelyek már G-ben is megvoltak, ami ellentmondás. Ha nem X a csúcs, akkor $Y \to X$ csúcs helyett $Y \to L$ élet bevéve kapnánk egy másik fordított fenyőt, ami szintén csak G-beli éleket használ, ez szintén ellentmondás. Vagyis ezzel az élcserével továbbra is megfelelő gráfot kapunk.

A második típusú csere, amikor $X \to Y$ helyett $(X, Y) \in B$ élet veszem be. Mivel az irányított élek csak csökkennek, ha volt fordított fenyő, akkor annak már G-ben is meg kellett lennie. Tehát a feltétel azért romolhat el, mert az irányítatlan rész összefüggő lett. A cserélési feltétel alapján azonban X és Y testvérei közösek. Így az újonnan megjelenő irányítatlan él G egyazon komponensén belül van, így ha G' nem összefüggő, akkor több olyan komponensből áll, melyek már G-ben is külön komponensek voltak. Ekkor azonban már G megsértette volna a gráf-feltételt, ami ellentmondás.

Tehát ha G-re teljesül a feltétel, akkor egy belőle szabályos élcserével kapott gráfra is, amennyiben az új gráf aciklikus. Ha kört tartalmazó gráfot kapunk, az alkalmazhatósági tétel feltételei nem teljesülnek, azonban ebben az esetben tudjuk, hogy nem egyértelmű Λ és Ω mátrix.

Ebből következik, hogy az azonos ekvivalencia osztályba tartozó aciklikus gráfok egyszerre teljesítik a gráf-feltételt, vagy egyszerre nem teljesítik azt.

5. Alkalmazások genetikai kutatásban

A dolgozatban leírt módszereket a Semmelweis Ignác Orvostudományi Egyetem, II. számú belgyógyászati klinikáján egy kutató csapatban, genetikai kutatás során alkalmaztam. A csapat célja a terhességi cukorbetegség, más néven a Gestational Diabates Mellitus kialakulásának hátterének megismerése volt, elsősorban genetikai és élettani adatok szerinti elemzése volt. Ez a betegség terhes nők esetén alakulhat ki a terhesség során, az anyán a cukorbetegség tünetei figyelhetőek meg. Megfelelő kezelés esetén a szülés után problémamentesen elmúlik, azonban bizonyított, hogy köze lehet a 2-es típusú cukorbetegséghez, így ezek a nők az életük folyamán később ismét megbetegedhetnek.

Magyarországon már bevett gyakorlat a terhes nők szűrése, úgynevezett terheléses vizsgálattal. Megmérik az anya vércukorszintjét reggel, éhgyomorra, majd tömény cukrozott vizet itatnak meg vele. Két órával később ismét megmérik a vércukorszintet. Ha a két érték közül valamelyik elér egy küszöbértéket, akkor az anyát betegnek nyilvánítják, ellenkező esetben egészséges. Külföldi kórházakban a terhelés után egy órával mért értéket is figyelembe szokták venni. Speciális esetekben a vizsgáló orvos más tényezők alapján is betegnek, vagy veszélyeztetettnek nyilváníthatja az anyát. A vérvizsgálat után így három különböző adatot kapunk az anyáról, a terhelés előtti (orvosi szakszóval élve éhomi) vércukorszintet, a 120 perces terhelt értéket és a klasszifikációt, hogy beteg, vagy egészséges-e.

5.1. Adatok

A cukorbetegség más típusaival korábban már nagyon sok vizsgálatot végeztek. Meghatároztak sok olyan SNP-t, amelyek összefüggésbe hozhatóak a megbetegedéssel. Mivel a terhességi cukorbetegség is kapcsolatban áll más cukorbetegségekkel, feltételeztük, hogy a kialakulást elősegítő SNP-k között is lehet átfedés. Ezért orvosi szempontok alapján kiválasztottunk 100 SNP-t, melyeket a vizsgált alanyoknál figyelembe vettünk. A későbbi analízis során kiderült, hogy ezek közül 23 polimorfizmus annyira ritkán változik meg, hogy a vizsgált alanyok között mindenkiben azonos volt, így csak 77 változó maradt, amiket figyelemmel kísértünk. Minden SNP esetén két konkrét nukleinsav állhat az adott pozícióban, egyszeres, vagy kétszeres formában. Az emberiségben gyakrabban előforduló változat kétszeres hordozását 0-val kódoltam, egy gyakori és egy ritkább allél esetén 1, két ritka allél esetén pedig 2-es szám került a táblázatba. Ez egy additív modellt feltételező adatbázis.

2. Példa. *Egy speciális SNP helyen Adenin (A) vagy Citozin (C) fordul elő. A gyakoribb változat az A. Ekkor ha egy alany AA-t hordoz,* 0 *ha AC-t, akkor* 1*, CC esetén pedig* 2 *besorolást kap.*

Több betegség a hordozás esetén már kifejti hatását, azaz domináns módon hat, ezért további hasznos megközelítés, ha csak azt vizsgáljuk, hogy az alany hordoz-e a ritkább változatból. Ez egy újabb adatsort eredményez.

Végül elképzelhető, hogy az egyszeres hordozásnak még nincs semmilyen hatása, viszont a kétszeres megbetegedéshez vezet, vagyis recesszív módon hat. Ekkor elegendő azokat az alanyokat megkülönböztetnünk a többitől, akik kétszeresen hordozzák a ritkább variánst.

A multiplikatív modellt nem vizsgáltuk, mivel olyan kicsi értékeket vártunk, hogy nem lett volna lényeges eltérés valamely másik modelltől.

Ezeket az adathalmazokat közvetlenül előállíthatjuk az eredeti $0, 1, 2$ értékű adatsorból, Dummy változók segítségével, és ezek ügyes használatával a modellek is felírhatók. Egy X SNP-ből készítsük el X_1 és X_2 változót, ahol $X_1 = \mathbb{I}(X = 1)$ és $X_2 = \mathbb{I}(X = 2)$. Ekkor a különböző modellek az alábbi módon kaphatók:

- Domináns: $X_1 + X_2$
- Recesszív: X_2
- Additív: $X_1 + 2 \cdot X_2$

5.2. Logisztikus regresszió

A vizsgálat első szempontja az volt, hogy határozzuk meg melyek azok a génvariánsok, amelyek hatással vannak a terhességi cukorbetegség kialakulására. A betegség egy 0, 1 kódolású logikai változó, jelöljük Y-nal. A génvariánsok mindig három értéket vehettek fel. Attól függően, hogy hányszorosan hordozzák az adott ponton ritkább változatot $0, 1$ vagy 2 értéket rendeltem hozzá. A vizsgálat során az éppen aktuálisan használt SNP változóját jelöljük X-szel. A betegség vagy kialakul, vagy nem, így feltételezhető, hogy binomiális eloszlást követ az azonos genetikával rendelkező emberek között. Ezért a logisztikus regresszió használata mellett döntöttem.

Az adatokat a *be* nevű tömbbe olvastam be, és k-val jelöltem a génvariánsok számát, ami jelen esetben 77. Az egyes SNP-k és a betegség vektorát az alábbi algoritmussal hasonlítottam össze:

```
  #Génvariánsok száma, lényeges adatok kigyűjtése, inicializálás
y<-be$Dg.Kód.WHO.ANYA
g<-matrix(0,k,2)
rownames(g)<-colnames(be)[1:k]
colnames(g)<-c("koeff","p érték")
#SNP-k egyenkénti vizsgálata
for(j in 1:k){
 x<-be[,j]
 M <- glm(y ~ x, family=binomial(logit))
 S<-summary(M)
 w<-S$coef
 g[j,]<-c(w[2,1],w[2,4])}
 #szignifikáns adatok
round(g[g[,2]<0.05,],6)
#szignifikánsak Odds Ratio-ja
exp(g[g[,2]<0.05,][,1])
```

Eredményül 8 szignifikáns variánst kaptam, amik befolyásolják a terhességi cukorbetegség kialakulásának valószínűségét.

A fenti analízis az additív modell esetét vizsgálta. Azonban mivel a génvariánsok konkrét viselkedése még nem ismert, nem tudjuk meghatározni, hogy melyik modell a helyes. A vizsgálatot elvégeztem domináns és recesszív modell esetén is. Az analízis itt is a logisztikus regresszió módszerével, a már ismertetett algoritmussal történt.

Ezek után el kellett dönteni, hogy az adott SNP viselkedését a fenti modellek közül melyik írja le a legjobban. Ehhez dummy változókká alakítottam az SNP-ket az adatbázisban. Így meg tudtam vizsgálni minden variáns esetén, hogy az első és a második hordozott változat milyen hatást fej ki. Amennyiben csak az elsőnek volt 0-tól lényegesen eltérő hatása, akkor domináns modellt feltételeztem. Amennyiben a másodiknak is körülbelül annyi volt, mint az elsőnek, akkor additív, míg ha csak a másodiknak volt hatása, akkor recesszív tulajdonságúnak tekintettem. Volt olyan eset, amikor a második variáns hatása ellentétes volt, ekkor az SNP egyik ismert modellbe sem fért volna bele, így ha valóban szignifikánsnak találtuk volna, akkor orvosi szempontból is további vizsgálatot igényelt volna. Azonban ilyen változat nem volt, így a felmerülő eltérések a minta egyediségéből származtak. Így minden SNP-ről sikerült eldönteni, hogy melyik típusú modell adja rá a legjobb becslést.

A következő problémát az jelenti, hogy genetikai vizsgálatok alkalmával nem csak az SNP-k ismertek. A vizsgált személyekről vérvétel során vércukorszintet, különböző enzimek szintjét lehet megállapítani, valamint az életkor, testsúly, és egyéb testi értékek is rendelkezésre állnak. Egy ilyen sokténye zős betegségnél, mint a cukorbetegség, nem tehetjük meg, hogy ezeket figyelmen kívül hagyjuk. A mi esetünkben a vércukorszint rendelkezésünkre állt, azonban mivel közvetlenül ebből lett klasszifikálva a beteg csoport, úgy döntöttünk, hogy nem mint okozó tényező, hanem mint maga a célváltozó tekintünk rá és egy későbbi vizsgáltban foglalkozunk vele. Ezen kívül az életkor (jelöljük A-val a változót) és a testtömegindex (jelöljük B-vel) mutatott összefüggést a betegséggel, így ezeket kellett figyelembe vennünk.

Az analízishez ezeket is vektorba rendeztem és többváltozós logisztikus regressziót alkalmaztam. Így a $logit(Y) = c_0 + c_1 \cdot A + c_2 \cdot B + c_3 \cdot X$ egyenlet együtthatóit kellett megbecsülni. Ebből is elsősorban c_3 volt fontos számunkra. Orvosi értelemben így megkaptam azt az értéket, ami a génvariáns hajlamosító kockázatát jellemzi, abban az esetben, hogyha mind az életkort, mind a testtömegindexet szem előtt tartjuk.

Az így kapott c_3 értéknek exponenciálisát véve Odds Ratio-t kaptam:

$$OR = e^{c_3}$$

Az Odds Ratio-ból kiszámítható a betegség kialakulásának valószínűsége, ha egy alanyról csak annyit tudunk, hogy milyen formában hordozza az adott génvariánst. Az additív, domináns és recesszív modellek esetén ez az érték különböző, valamint az SNP-ről is más-más ismeretek szükségesek. Domináns modell esetén például elég, ha tudjuk, hogy az alany hordozza a variánst, és megadható a megbetegedés valószínűsége, míg additív modell esetén magasabb kockázati valószínűség társul ahhoz, aki kétszeresen hordoz, mint ahhoz, aki csak egyszeresen.

Tegyük fel, hogy c_3 együtthatót domináns modell alapján számoltuk, ekkor hordozó ember esetén a megbetegedés valószínűsége:

$$p = \frac{OR}{OR + 1}$$

Ezeket a valószínűségeket egyenként számoltuk ki, azonban az emberre egyszerre több is hatással van. Ha az SNP-k függetlenek, akkor a hozzájuk tartozó Odds Ratio-k összeszorzódnak és együttesen határozzák meg a fenti képlet alapján az alany megbetegedésének valószínűségét. A tapasztalatok azonban azt mutatják, hogy ezek a hatások jelentősen befolyásolják egymást is, ezért a konkrét értékek meghatározásához a későbbiekben többdimenziós vizsgálatra lesz szükség.

5.3. P-korrekció

A statisztikában ritkán ugyan, de előfordulnak véletlen eredmények, amelyek hamis következtetésre vezetnek. Például ha 100 különböző próbát végzek el egyszerre, ahol mindegyik esetben 5% esély van arra, hogy hibásan vetem el a nullhipotézist, és mindegyikben olyan eredményt kapok, hogy el kell vetnem, akkor várhatóan lesz 5 eset, amikor hibásan fogadtam el az ellenhipotézist. Ez a mi esetünkben azt jelenti, hogy felmerülhet olyan SNP, amelyet mi szignifikánsnak deklaráltunk, de valójában csak a véletlen műve, hogy épp olyan embereket vizsgáltunk, akik esetén felmerülhet ez a következtetés. Az ilyen eseteket false discovery-nek nevezzük.

Az együtthatók kiszámítása után a következő feladatunk a false discovery esélyének lecsökkentése volt. Mivel a statisztikákat $0,05$ szinten szokták elvégezni, mert ez még az elfogadható hibahatár, első ötletként gondolhatunk arra, hogy az egész vizsgálat hibahatárát vigyük $0,05$ alá, azaz ez annak a valószínűsége legyen, hogy a vizsgálat során lesz egyáltalán egy olyan változónk, ami hamis pozitív. Mivel 77 génváltozatot analizáltunk, ha egy SNP szignifikanciája $p < \frac{0,05}{77} = 0,00065$, akkor biztosan megfelel a követelményeknek, és elfogadhatjuk, mint jelentősen befolyásoló tényezőt. A módszer kidolgozása Bonferroni nevéhez fűződik.

Sajnos azonban hiába állt rendelkezésünkre nagy adathalmaz, csupán egy SNP p-értéke csökkent a megfelelő küszöbszint alá. Ennek az oka, hogy a Bonferroni módszer nagyon szigorú. Ahhoz, hogy az egész vizsgálat szignifikanciáját $0,05$ alá vigyük, nem kell minden egyes p-értéket ennyire alacsonyra venni. Egy másik, megengedőbb módszert dolgozott ki Benjamini és Hochberg.

A módszer elve az, hogy sorba rendezhetjük a változókat szignifikancia szerint növekvő sorrendben. Ekkor a sor elején álló változó valóban csak $0,00065$ alatt tekinthető ténylegesen szignifikánsnak, azonban a második változóra már tekinthetünk úgy, mint a megmaradt 76 közül a legszignifikánsabb, így $p < \frac{0,05}{76} = 0,00066$ alatt elfogadhatjuk. A további változóknál ez a küszöb index egyre növekszik. Végül megnézzük, hogy melyik volt a legnagyobb szignifikanciájú elfogadott változó és az összes tőle balra lévőt elfogadottnak tekintjük.

Ez a módszer sokat nem javít ugyan, viszont ha több lényeges SNP-t felderítünk, amelyek önmagukban a $0,00065$-ös határt nem érik el, de a legnagyobb mégis elfogadásra kerül, akkor a többi eredményünket sem kell elvetni.

5.4. Folytonos változók

A genetikai változók mellett rendelkezésünkre álltak más élettani adatok is, például a terhes nők vércukorszintje cukorterhelés előtt és után. Mivel maga a diagnózis 90%-ban ezeken az értékeken alapul, érdemes megvizsgálni, hogy hogyan hatnak a génváltozatok külön-külön a terhelés előtti és a 120 perces cukorértékekre. Ehhez a lineáris regresszió módszere alkalmazható. Itt is célszerű figyelembe venni, hogy az életkor és a testtömegindex az elsődleges befolyásolója a vércukorszintnek, így csak az általuk megmagyarázatlan rész összefüggését kell vizsgálni a génvariánsokkal. Az alábbi programmal számoltam ki az egyes SNP-khez tartozó regressziós együtthatókat:

```
#BMI-vel és életkorral való korrigáláshoz a segédszámítások
korr<-lm(X120.~kor+BMI,be)$res
ages<-lm(X120.~kor,be)$res

g<-matrix(0,k,4)
rownames(g)<-colnames(be)[1:k]
colnames(g)<-c("koeff","p érték","korrigalt","power")

#Az üres sorok ne kerüljek számításba,
 csak ahol mindenütt lényeges adat áll
f<-(!is.na(be$BMI)&!is.na(be$kor)&!is.na(be$X120.))
f1<-(!is.na(be$BMI)&!is.na(be$kor)
    &!is.na(be$X120.)&!is.na(be$Dg.Kód.WHO.ANYA))
y<-rep(NA,n=dim(be)[1])
darab<-sum(f)
y[f]<-korr
z<-rep(NA,n=dim(be)[1])
z[f1]<-y[f1]
y<-z

for(j in 1:k){
x<-be[,j]
M <- lm(y ~ x)
S<-summary(M)
```

```
w<-S$coef
pow<-pwr.t.test(d=w[2,1],n=darab,sig.level=0.10,
     type="one.sample",alternative="two.sided")
 g[j,]<-c(w[2,1],w[2,4],p.adjust(w[2,4],"BH",77),pow$power)
}
index<-g[,2]<0.05
g[index,]
```

A programom már tartalmazza a korábban ismertetett Benjamini-Hochberg féle p korrekciós eljárást is, valamint az SNP-knek a testtömeg-indexen és életkoron felüli hatását vizsgálja csak.

Habár ez a módszer sok szempontot fegyelembe vesz, mégis felmerülhetnek hiányosságai. Léteznek olyan változatok, amelyek hatnak a betegségre és az elhízásra is, így közvetve a BMI-n keresztül is kifejtik hatásukat. Az ilyen esetekben a kapott eredmény kisebb lehet, mint a valódi hatás. Ahhoz, hogy még pontosabb, minden összefüggést pontosan felderítő modellt kapjunk nem elég az SNP-ket külön vizsgálni, hanem összetettebb módszert kell alkalmazni.

5.5. SEM-modell

Ahhoz, hogy a strukturális egyenletek modelljét használhassuk először ki kell választani a fontos SNP-ket, majd el kell készítenünk a felhasznált változók közötti összefüggőségi gráfot. A génvariánsok kiválasztásánál segítségünkre lehet a korábbi vizsgálat, amelyből a legszignifikánsabb és abszolút értékben legnagyobb hatást kifejtő, de még szignifikáns változókat érdemes használni. Ezek a változók a 0 perces és a terhelt vércukorszintre, illetve a BMI-re is hathatnak. Itt is a korábbi eredményekre támaszkodhatunk, hogy melyik összefüggést érdemes feltételeznünk. Továbbá nyilvánvalóan az életkor és a testtömeg-index hatnak a cukorértékekre, valamint a terheletlentől jelentősen függ a 120 perces érték. Az így kialakított gráf fogja alkotni a modell alapját. A gráf struktúrája elég egyszerű, így látszik, hogy kielégíti a gráf-feltételt, azaz alkalmazhatjuk a SEM-modellt.

Az adatbázisom alapján 8 lényeges SNP-t találtam, amelyeket bevettem a modellbe. Az adatok torzulását elkerülendően csak olyan egyének adatait vizsgáltam, ahol nem hiányzott az anya betegségét vagy egészségességét meghatározó adat. A felépített modell nem volt szignifikáns, ezért a korábbi fejezetben leírt módszerrel elkezdtem kitörölni a gráf csúcsait, míg végül az alábbi eredményre jutottam:

```
f<-(!is.na(be$Dg.Kód.WHO.ANYA))
ter<-be$X120.[f]
alap<-be$X0.[f]
kor<-be$kor[f]
BMI<-be$BMI[f]
AI<-be$AI[f]
AN<-be$AN[f]
AZ<-be$AZ[f]
BL<-be$BL[f]
F<-be$F[f]
K<-be$K[f]
V<-be$V[f]
Z<-be$Z[f]
adatok<-matrix(c(AI,AN,AZ,BL,F,K,V,Z,BMI,kor,alap,ter),820,12)
colnames(adatok)<-c("AI","AN","AZ","BL","F",
     "K","V","Z","BMI","kor","alap","ter")
#Lépésenkénti elhagyással a legjobbnak talált modell
model<-'
BMI~K
BMI~AI
ter~alap
ter~BMI
ter~kor
ter~AI
ter~BL
ter~F
alap~BMI
alap~kor
alap~AI
alap~F
ter~1
'
MTX<-cov(adatok, use="pairwise.complete.obs")
illesztett<-sem(model,data=adatok, sample.cov=MTX, fixed.x=FALSE)
summary(illesztett)
```

Ez a modell 4 fontos génvariáns, életkor, testtömeg-index, valamint egyszeri terhelés nélküli vércukorérték ismeretében jó közelítést ad a terhelt érték nagyságára. A segítségével időben ki lehet szűrni azokat az édesanyákat, akik hajlamosabbak a betegségre, korábban megvizsgálni őket, és ha szükséges, korábban elkezdeni a kezelésüket, ami lényeges segítség az orvosok számára, hogy megelőzzék a komolyabb problémákat, szövődményeket. Továbbá a modell bizonyítja, hogy szükség van a genetikai vizsgálatokra is, hiszen csak ennek a 4 variánsnak akkora hatása lehet, mintha valaki 20 évvel később vállalna gyereket.

6. Összefoglaló

A dolgozatomban ismertettem a génkutatásban leggyakrabban használható módszereket. A lineáris regresszióval folytonos változókkal való kapcsolatot kerestem. Additív modellt használva találtam 6 olyan változót, melyek lényeges összefüggést mutatnak az éhomi vércukorszinttel és 5 olyat, melyek a 120 perces terhelt értékkel állnak kapcsolatban. Domináns modellel 3, illetve 5 változót sikerült felderíteni, amelyek egy része a korábbitól eltérő volt.

Logisztikus regressziót használtam a diszkrét értékű változók magyarázására. Ezzel a módszerrel 8 illetve 9 olyan SNP-t sikerült észlelni, amelyek nagy mértékben beleszólnak a betegség kialakulásába. Ezek közül 2 olyan volt, hogy a Benjamini-Hochberg féle korrigált p-érték mellett is megállta a helyét. Sajnos a mintaelemszám alacsony volt, ezért több ennyire szignifikáns változót nem sikerült találni, azonban az eredmény jó irányvonalat adott a későbbi kutatásokhoz.

Az általánosabb modell kidolgozásához strukturális egyenleteket használtam. A módszer sikerességéről tanúskodik, hogy 4 olyan SNP-t sikerült beépíteni, melyek mind különböző génen helyezkednek el, és ezek a gének az orvosok szerint nagyban befolyásolhatják a vércukorszinttel kapcsolatos megbetegedéseket.

A konkrét modellek mellett több olyan felmerülő problémáról is szót ejtettem, amelyekkel részletesebben is érdekes lehet foglalkozni. Ezekkel a problémákkal a kutatás során találkoztam is és sikeresen alkalmaztam a dolgozatban leírt megoldásokat.

Összefoglalva, a dolgozatban leírtam egy eredményes genetikai kutatás menetét és matematikai hátterét, mintát és módszereket nyújtva egy későbbi hasonló típusú analízishez.

Ábrák jegyzéke

1. A DNS egy szakasza, kettős spirál szerkezet ábrázolásával és nukleinsav párok feltüntetésével 5
2. Y változó az X ötszöröse, Y megfigyelésében standard normális eloszlásból származó hibával, valamint a legjobban illeszkedő egyenes . . . 17
3. X és Y változók egymás ellentettjei, Y megfigyelésében standard normális eloszlásból származó hibával, valamint a legjobban illeszkedő egyenes 18
4. A logisztikus regresszió szemléltetése szimulált adatokon 22

Hivatkozások

[1] William Feller, *An introduction of Probability Theory and its Applications, VOLUME II.*, 1965.

[2] Steven J. Miller, *The Method of Least Squares*, Mathematics Department Brown University, 2006

[3] John Aldrich, *R.A. Fisher and the Making of Maximum Likelihood 1912-1922*, Statistical Science, 1997, Vol. 12, No. 3, 162-176

[4] David A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press, 2009.

[5] Debasis Kundu, G. Murali, *Model selection in linear regression*, Computational Statistics & Data Analysis, 1996, Vol. 22, 461-469

[6] J. A. Nelder, R. W. M. Wedderburn, *Generalized linear models*, Journal of the Royal Statistical Society. Series A (General), 1972, Vol. 135, No. 3 , pp. 370-384

[7] J. J. Hox, T. M. Bechger, *An Introduction to Structural Equation Modeling*, Family Science Review, 2006, Vol. 11, 354-373

[8] Mathias Drton, Rina Foygel, Seth Sullivant *Global identifiability of linear structural equation models*, The Annals of Statistics, 2011, Vol. 39, No.2, 865-886

[9] Rina Foygel, Jan Draisma, Mathias Drton *Half-trek criterion for generic identifiablity of linear structual equation models*, The Annals of Statistics, 2012, Vol. 40, No. 3, 1682-1713

[10] Therese D. Pigott, *A Review of Methods for Missing Data*, Educational Research and Evaluation, 2001, Vol. 7, No. 4, pp. 353-383

[11] Donald B. Rubin, *Inference and missing data*, Biometrika, 1976, Vol. 63, Issue 3, pp. 581-592

[12] Gordon K. Smyth, *Nonlinear regression*, Encyclopedia of Environmetrics, 2002, Volume 3, pp 1405-1411