

Gondolatok a járvány elleni védekezés értékeléséről, és a járvány hatásának vizsgálatáról

Ferenci Tamás

Tartalomjegyzék

Egy gondolkodási keret, és a confounding minden átható problémája	2
Egy előkészítő pont: kimenet megválasztása	4
Az elméleti megalapozás: kauzális diagram egy járvány hatására	5
Technikai részletek	6
Az empirikus vizsgálat lehetőségei és nehézségei	9
A confounding problémája	9
Kihagyott változós torzítás	14
A regresszió eszköze	14
A mintavételi ingadozás és a torzítás fogalma	15
A függvényforma kérdése	19
Több változó modellezése	21
A torzítás-variancia dilemma	25
A változószelekció kérdésköre	27
A kényelmetlen tudomány	30
Záró gondolatok	30
Ajánlott olvasmányok	31

“Ne fogjon senki könnyelműen

A húrok pengetésihez!

Nagy munkát vállal az magára,

Ki most kezébe lantot vesz.”

(Petőfi Sándor: A XIX. század költői)

(A dolgozat letölthető PDF és EPUB formátumokban is.)

E pillanatban Európa a koronavírus-járványt lényegileg lezártnak tekinti. Noha a dolog véglegességében azért van még kérdőjel, az alkalom tökéletesen megfelelő arra, hogy feltegyünk két olyan kérdést, melyek legjobban visszatekintve válaszolhatóak meg, egy járvány végén relevánsak – de akkor viszont nagyon is. (Ebben a dolgozatban az „a járvány” kitétel alatt a COVID-19 járványt értem, de az elmondottak lényegében teljesen általánosak, és más esetre is alkalmazhatóak.) Jelesül:

- Hogyan értékeljük egy ország járványügyi intézkedéseit? Megfelelő lépéseket hoztak és a kellő időben? Jó volt a járványügyi rendszer, a tesztelés, a korlátozó intézkedések, a tájékoztatás vagy lehetett volna jobban is eljárni?
- Mi hat arra, hogy a járvány egy adott országban mekkora pusztítást végez? Miért van az, hogy valahol ez nagyobb, máshol viszont kisebb, adott esetben akár jóval kisebb? Mik a közreható tényezők, és melyiknek mekkora a szerepe?

Mostani írásom egyik első állítása az lesz, hogy a két kérdés(csoport) valójában szorosan összefügg, sőt, lényegében ugyanarról a problémakörről szólnak.

A kérdéskör rendkívüli fontosságát az adja, hogy nem csak „tudományos szempontból“ releváns, hanem a hétköznapi beszédben, politikai vitákban, közéleti diskurzusban is lépten-nyomon, és nagy súlyal kerül elő: Magyarországon sikeres volt a járvány elleni védekezés! Magyarországon sikertelen volt a járvány elleni védekezés! A kormány megfelelő intézkedéseket hozott! A kormány megkésett és nem elégsges intézkedéseket hozott! Magyarországon sokan haltak meg a járvány miatt! Magyarországon nem haltak meg sokan a járvány miatt! Magyarországon sokan haltak meg a járvány miatt, de ez nem a kormány hibája, hanem a lakosság egészségi állapotán múlt!

E kérdések megválaszolását számos irányból kísérelhetjük meg: használhatunk járványügyi megfontolásokat, biológiai modellek, epidemiológiai elméleteket és így tovább, én azonban most egyetlen módszerrel fogok foglalkozni: az *empirikus* vizsgálattal, vagyis amikor a begyűjtött tényadatok alapján próbáljuk megválaszolni ezeket a kérdéseket. Gyakori vita tárgya ennek pontos szerepe, ám e kérdésben – mind tudományosan, mind a közvélemény előtt – a legmegbízhatóbb módszerként tűnik fel az empirikus kutatás.

Előre mondjam, hogy az írásomnak nem az a célja, hogy „végeredményt hirdessen“. Én most a módszerekre akarok fókusztálni, helyesekre és hibásakra egyaránt, bemutatva a megfelelő eljásárokat, és – különös hangsúlyjal – a tipikus csapdákat és buktatókat is. Remélem, hogy ezzel hozzá tudok járulni a már most is zajló, és – az előbb említett megbízhatóság miatt – várhatóan a jövőben sem elhalkuló viták színvonalának emeléséhez. Aggodalomra ugyanis lehet okunk e téren: a téma jellegéből adódóan lényegében óhatatlan a politikai szempontok szerinti értelmezése az adatoknak (ami soha nem tesz jó), de hiszek benne, hogy a tudományos szempontok hozzáérhető összefoglalása segít a diskurzus javításában.

A keretes szerkezet kedvéért elmondam az egyik utolsó állításomat is: a kérdést rendkívül nehéz megválaszolni empirikusan, tehát a különböző országok tényadatai alapján következtetve. Bár megpróbálok helyes vizsgálati módszereket bemutatni, és lehetőségeket a nehézségek enyhítésére, a kérdésre nem lehet egyszerű és perdöntő választ adni. Mégis, azt gondolom, hogy ennek ellenére nem haszontalan végiggondolni a problémát, sőt – részint mert segít elkerülni a tipikus csapdákat, buktatókat, téves következtetéseket, részint pedig a probléma jó megértése, átlátása néha még értékesebb is, mint egy egyszerű eredményközlés.

Egy gondolkodási keret, és a confounding mindent átható problémája

Az első állításom tehát az, hogy a két kérdéscsoport lényegében ugyanazon probléma megoldását teszi szükséges. Ennek megértéséhez kezdjük az első kérdéssel. A probléma ennek kapcsán nyilvánvaló: annak eldöntéséhez, hogy „jó“ volt-e a járvány kezelése, muszáj valahogy definiálni, hogy mit értünk „jó“ alatt. Ezt legtermészetesen úgy lehet megtenni, amit angolul counterfactual, magyarul a – kissé furcsán ható – tényellenes szóval szoktak jellemzni: megnézzük, hogy a járvány ideális kezelése esetében mi lett volna a kimenet, és ehhez hasonlítjuk a tényleges kimenetet. (Ez is mutatja, hogy a kérdés nem triviális: lehet olyan helyzet, hogy 10 halott nagyon rossz kezelést jelent és az 1000 jót – ha az ideális esetben az első helyzetben 1 lett volna, a másodikban pedig 999.) Igen ám, de mi az, hogy „ideális kezelés“? Amennyiben van időgépünk, akkor ez semmiféle problémát nem jelent: visszamegyünk időben, próbálkozunk más döntésekkel, és megkeressük, hogy melyik a legjobb. Időgép híján azonban gondban vagyunk. Mihez viszonyítsunk? Az egyetlen épkezél, empirikusan vizsgálható viszonyítási pontot az jelenti, hogy *más országokban* mi a helyzet (feltéve, hogy nem mindenki pontosan ugyanúgy kezelte a járványt), csakhogy ez elég problémás viszonyítási pont: más országok ezernyi, milliónyi egyéb tényezőben is eltér(het)nek tölünk azon túl, hogy máshogyan kezelték a járványt, mint mi – ha találunk is különbséget a végeredményben, a járvány okozta pusztítás méretében, honnan tudjuk, hogy ez mennyiben tudható be az eltérő kezelésnek, és mennyiben az egyéb különbségeknek? Ha valahol kevesebb halott van, az azt jelenti, hogy ott jobban kezelték a járványt? (És így nézzük azt, hogy mit csináltak másképp?) Mi van, ha ugyanolyan jól kezelték a járványt, csak fiatalabb a korfájuk, és ezért lett kevesebb halálozásuk? Mondok még jobbat: mi van, ha rosszabbul kezelték a járványt, csak annyival fiatalabb az ottani lakosság, hogy ez többet javított, mint amennyit a rosszabb kezelésük rontott...?

Itt térhetünk át a második kérdésre, illetve arra, hogy miért ugyanaz lényegében a kettő. A fenti okfejtés ugyanis egy speciális esete egy általános kérdésnek: annak, hogy milyen tényezők befolyásolták, és milyen mértékben a járvány hatását egy adott országban. Azért speciális eset, mert leszűkítettük a kérdést egy tényezőre, a járvány kezelésének jóságára. De nem muszáj ezt a leszűkítést megtennünk, nyugodtan

vizsgálhatjuk általában (sőt, mint majd látni fogjuk, vizsgálnunk is *kell*, a speciális kérdés jó megválaszolásához *is*), hogy milyen tényezők alakították ki, hogy hány halálozás volt, és melyik tényezőnek mekkora a szerepe. És ebből, mintegy mellesleg, majd azt is kiolvashatjuk, hogy mi a helyzet a járvány kezelésének a jósága kapcsán. Ezért mondtam, hogy elég erről a kérdésről beszélni, és ez mindenkorrel megválaszolja.

A probléma, hogy itt *pontosan ugyanabba* a helyzetbe futunk bele, mint az előbb. Amennyiben tudnánk egyetlen és csakis egyetlen tényezőt változtatni, majd megnézni úgy a járvány áldozatainak a számát, akkor meg tudnánk mondani, hogy az a tényező hat-e rá, és ha igen, milyen mértékben. De a probléma ugyanaz: ezt csak időgéppel lehet megtenni, ha vissza tudnánk menni az időben, és pontosan egy dolgot megváltoztatni. (Azért mondok időgépet, hogy látszódjon: az egyetlen dolgon kívül *minden más* változatlanul marad.) Jelen tudásunk szerint e megoldás kivitelezése technikai nehézségekbe ütközik, így visszajutunk ugyanahhoz a gondolathoz: nézzünk e helyett különböző országokat, amik a vizsgált tényezőben eltérnek. Mondjuk, ha az érdekel minket, hogy a 100 ezer lakosra jutó nővérek száma hogyan hat a halálozásra, akkor vegyük országokat, ahol ez a szám – az egyszerűség kedvéért kereken mondva – 2000, és vegyük olyanokat, ahol csak 1000, majd hasonlítsuk össze a járvány-halálozást. De a probléma ugyanaz: mi van, ha a 2000-es, tehát jobb nővér-ellátottságú országokban egyúttal mondjuk az elhízottak aránya is alacsonyabb? Azért hoztam épp ezt a példát, mert ez is a kisebb halálozás irányába hat. Innen kezdve, ha azt is találjuk, hogy a 2000-es országokban kisebb a halálozás, honnan tudhatjuk, hogy az tényleg a több nővér miatt van? Mi van, ha valójában a nővéreknek a világon semmi szerepe, csak ezekben az országokban kevesebb az elhízott, és ez a valódi oka annak, hogy ott kisebb a halálozás?

Ezt a problémát szokták úgy hívni angol szóval, hogy confounding. (Angolul nagyon találó kifejezés: szó szerint „egybemosódást“ jelent, és valóban arról van szó, hogy a különböző tényezők hatása egybemosódik, a több nővér egybemosódik a kevesebb elhízotttal. Magyarra leginkább „zavaró változós hatásként“ szoktak fordítani, csak sajnos ez jóval nyakatekertebben, mint az angol kifejezés.) Ez egy általános probléma, ami akkor jelentkezik, ha adott tényező szerint eltérő csoportokat hasonlítunk össze, de azok a csoportok más tényezőkben is el fognak térní – innen kezdve, ha találunk is különbséget a kimenetben, nem lehet tudni, hogy az mi miatt van: az általunk vizsgált eltérés miatt, a vele együtt járó egyéb eltérés(ek) miatt, vagy ezek valamilyen keveréke miatt. A fenti példából is látható módon leginkább úgy ragadható meg a probléma, hogy van egy – vagy több – háttérben lévő változó, a példában mondjuk a gazdasági fejlettség, ami *egyszerre* függ össze a vizsgált tényezővel (fejlettebb országokban több a nővér) és hat a kimenetre (fejlettebb országokban kevesebb az elhízott, ami szintén csökkenti a halálozást).

A probléma általánosságából adódóan számos más területen fellép, orvostudománytól a közigazdaságig (videó-előadás, írott jegyzet).

Érdemes megjegyezni, hogy *elvileg* van mód a probléma megoldására időgép nélkül is: az, ha tudunk randomizálni. Ez azt jelenti, hogy fogjuk a vizsgálat tárgyat képező alanyokat, és pénzfeldobással két csoportba sorsoljuk őket, az egyik a vizsgált tényező egyik értékét kapja, a másik a másikat, majd ezeket hasonlítjuk egymáshoz. Ez könnyen elképzelhető – és meg is valósítható – egy gyógyszer kipróbálásánál: az „*egyik érték*“, hogy kap gyógyszert, a másik, hogy nem. Ezt valóban kioszthatjuk pénzfeldobással. És ez csakugyan fontos is: ha nem ezt tennénk, hanem mondjuk az orvosokra bízzuk, hogy belátásuk szerint adjanak bizonyos betegeknek gyógyszert, másoknak pedig nem, akkor azonnal jönne ugyanaz a probléma – mi van, ha inkább a súlyosabb állapotúaknak adják a készítményt, vagy pont, hogy az enyhébb eseteknek? A randomizálás megoldja ezt a problémát: ha pénzfeldobással soroltuk őket csoportokba, akkor *biztosan* nem lehet *semmilyen* tulajdonságukban szisztematikus különbség. (Különösen fontos, hogy ez azt is jelenti, hogy azokban a tulajdonságokban sem, amikről nincs információink, sőt, azokban sem, amikről eszünkbe sem jutott, hogy confounding-ot okozhatnak!) Ehhez azonban arra van szükség, hogy mi tudjuk irányítani, meghatározni, hogy ki milyen tényezőnek van kitéve; az ilyen vizsgálatokat szokás kísérletes kutatásnak nevezni.

A mostani esetünkben ez nyilván szóba sem jön: nem lehet pénzfeldobással kisorsolni, hogy melyik országban legyen 1000 nővér és melyikben 2000. Marad az, hogy nézzük a tényadatokat – csak épp ott, szemben a kísérlet példájával, a nővérek száma már összefügghet más jellemzőkkel is. Az ilyen vizsgálatokat szokás megfigyeléses vizsgálatnak nevezni. Az ilyen típusú vizsgálatoknál, általában is, nem csak ennél a példánál, a confounding egy minden átható probléma, olyan értelemben, hogy folyamatosan, minden elemzési lépésnél gondolnunk kell rá, mert Damoklész kardjaként lebeg mindenkorrel a fejünk felett.

Egy előkészítő pont: kimenet megválasztása

A fenti leírásban impliciten feltételeztem, hogy a járvány hatását a halálozásban mérjük. A mostani dolgozat fő kérdése nem ez lesz, de azért pár gondolat erejéig érdemes erről a választásról is beszélni.

Először is, az önmagában választás kérdése, hogy egyáltalán a halálozást használjuk mérőszámként, kimenetként. Ez nem egyértelmű: a „járvány terhe“, a „járvány okozta pusztítás“ egy sokdimenziós fogalom, azaz sok szempont szerint mérhető. Terhe a járványnak természetesen és mindenek előtt, hogy emberek halnak meg miatta, de az is teher, hogy szenvednek tőle (még ha túl is élik), teljesen más szempontból de az is teher, hogy kiesnek a munkából, az is teher, hogy ha maradványtütésekkel gyógyulnak, az is teher, hogy az egészségügyi ellátórendszer kapacitásait igénybe veszik stb. stb. Ezek ráadásul nem is feltétlenül egy irányba mutatnak, elképzelhető, hogy egyik ország egyik dimenzió szerint jobb, de a másikban rosszabb, egy másik ország még épp fordítva.

Mégis, a halálozási szám használata a járvány terhének mérésére elég univerzális, és sok szempontból indokolható. Egyszer a legnagyobb súlyú és legdrámaibb vétület, másrész általában azért jól korrelált a többi szempont is vele (még ha nem is tökéletesen, ahogy az előbbi megjegyzés is mondta), harmadrész az egyik legbiztosabban mérhető mutató (gondoljuk meg mennyivel nehezebb lenne azt megmondani, hogy hány munkaóra esett ki a járvány miatt, de igazából még azt is, hogy hány fertőzött volt – ne feledjük, hogy azok száma rettenetesen függ a tesztelési intenzitástól). Valójában persze a halálozási mutató is elrejt szempontokat, például azt, hogy milyen életkorban történik a halálozás, vagy milyen egészségi állapotú alanyban, egyszóval az elvesztett életévek számának problémáját, de most ezt is utaljuk egy másik vizsgálódás témakörébe.

A második kérdéskör, hogy a halálozást hogyan mérjük. Nagyon kézenfekvőnek látszik a jelentett halálozás használata (hát az épp az, hogy hány halálozás volt, nem?), de sajnos ez sem ilyen egyszerű: a jelentett halálozási számokat is befolyásolja a tesztelési aktivitás (még ha kevésbé is, mint a jelentett fertőzött-számot) és a haláloki besorolásra alkalmazott definíciók. Ezek megnehezíthetik az országok összehasonlítását, de adott esetben akár egy országon belül is változhatnak időben. E kérdések önmagukban is nagyon érdekesek, itt most csak utaltam a problémára, de egy másik írásom részletesen bemutatja a problémát.

Ez a gond vezet el egy másik mutatóhoz: az úgynevezett többlethalálozáshoz. Itt a járvány előtti adatokat használva statisztikai úton készítünk egy előrejelzést a halálozások számára, ezt tekintjük úgy – mivel a járvány hatása nélküli adatokból készült – mint az a halálozás ami akkor lett volna, ha nincs járvány. Kivonva ezt az értéket a tényleges halálozási adatból, kapjuk a többlethalálozást. E mutató hatalmas előnye, hogy biztosan semmennye nem függ a tesztelési aktivitástól és biztosan semmennye nem függ a haláloki besorolástól. Hátránya azonban – azon túl, hogy a leglassabban ismertté váló mutató – egyszerű, hogy függ az előrejelzés jóságától, de még inkább az, hogy egy bruttó jellegű mutató: egybeméri a járvány direkt hatásaival (belehalnak emberek) annak indirekt hatásait is, amik ráadásul egyaránt lehetnek pozitívak és negatívak. Pozitív indirekt hatás, hogy az intézkedések mondjuk az influenza ellen is jót tesznek, de kicsit elengedve a fantáziánkat, az is pozitív indirekt hatás lehet, hogy kevesebb autóbaleset történik. Negatív indirekt hatás, hogy más betegség ellátása nehezedik meg vagy lehetetlenül el, de itt is lehet távlatibb kérdésekre gondolni, például mi van, ha megnő az öngyilkosságok száma a szociális elszigetelődés miatt. A többlethalálozás nem teszi lehetővé ezek biztos elkülönítését. Itt is igaz, hogy e problémakör önmagában több oldalnyi tárgyalást igényelne, ami az említett írásomban megtalálható, itt megint csak a címszavakban összefoglalásra szorítkoztam. Egyetlen megállapítást emelnék ki: Európán belül öt országtól (Bulgária, Románia, Litvánia, Lengyelország és főleg-meddig Szlovákia) eltekintve a két mutató értékei nagyon hasonlóak egymáshoz.

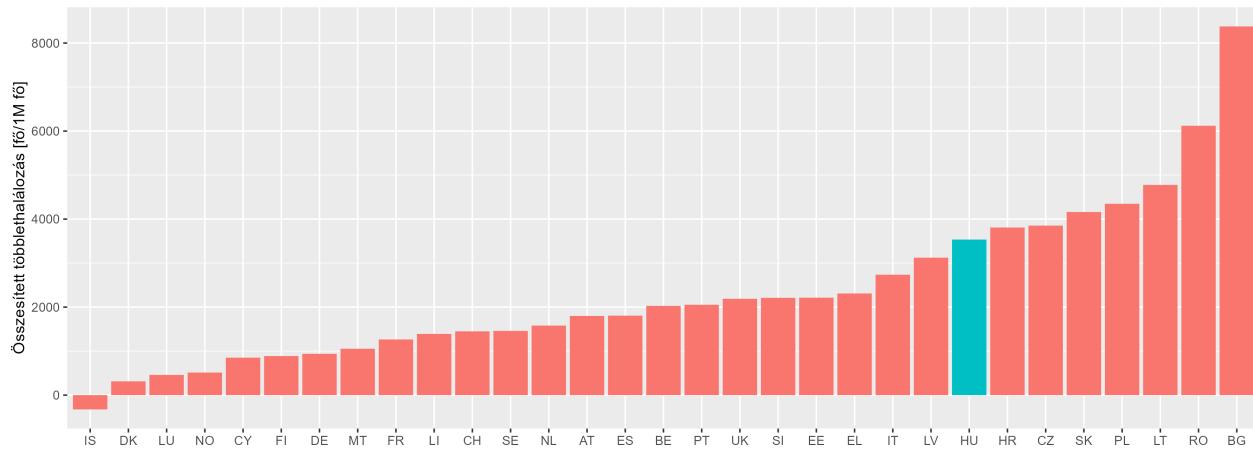
Ez azért is fontos, mert azt mondja, hogy – szerencsére – a két mutató közti választás nem várható, hogy nagy hatást gyakoroljon a végeredményre. Így különösebben sokat nem kell gondolozni azon, hogy melyiket választjuk, én most a többlethalálozást fogom használni a teher végső mutatójaként, azért, mert az empirikus elemzés különböző országok adatait fogja egybevetni, így fontos, hogy országok között robusztusan összehasonlítható mutatót használunk.

Ennek az értékei így néznek ki a 2021. 52. heti állapot szerint (ezt az indikátort nagy késleltetéssel közlik, így sokat visszamegyünk, hogy a lehető legtöbb országra legyen adatunk, és a reprodukálhatóság kedvéért ezt a dátumot most lerögzítjük):

```

RawData <- fread(
  "https://github.com/tamas-ferenci/ExcessMortEUR/raw/main/ExcessMortEUR_data.csv",
  dec = ",") [time=="2021W52" & nuts_level==0 & age=="TOTAL"]
RawData$cumexcessperpop <- RawData$cumexcess/RawData$meanpopulation*1e6
ggplot(RawData[order(cumexcessperpop)], 
       aes(x = factor(geo, levels = geo), y = cumexcessperpop, fill = geo=="HU")) +
  geom_col() + guides(fill = "none") +
  labs(x = "", y = "Összesített többlethalálozás [fő/1M fő]")

```



Az rögtön látszik, hogy Magyarország az *utolsó harmad elején* található; az persze ebből még nem derül ki, hogy ennek mi az oka – pontosan ezt próbáljuk most felderíteni.

Az elméleti megalapozás: kauzális diagram egy járvány hatására

Az ember ezen a ponton késztetést érezhet arra, hogy rögtön el is kezdje nézni az empirikus adatokat: melyik országban mekkora volt a teher (azaz a többlethalálozás)? Összefügg ez a nővérek számával? A gazdasági fejlettséggel? Az intézkedések szigorúságával? Az elhízottak arányával? A korfával?

Mielőtt azonban ebbe belevágunk, némi óvatosságot javasolok. Az ilyen „mindent mindenkel” típusú, elméleti megalapozás nélküli „összevissza” nézegetésekben alapuló adatbázis-masszározások általában nem sok jóra vezetnek. Ennek a későbbiekbén egy sor konkrét okát bemutatom, de most fontosabb egy általános, mondhatni filozofikus gondolat: az ilyen típusú empirikus vizsgálatoknak minden a háttérben lévő elméleti *jó* megértésén kell alapulnia. Nem egyszerűen azért, mert azt alaposan végiggondolva kap jó rálátást az ember a problémára, hanem azért is, mert – pont emiatt – ez teszi lehetővé jobb empirikus modellek készítését is.

Egyelőre tehát – és én azt gondolom, hogy általában is ez a jó hozzállás – ne nyúljunk hozzá semmilyen adathoz. E helyett próbáljuk meg először a fejünkben összeszedni, végiggondolni, és strukturálni a kérdést: milyen tényezők alakítják, hogy hányan halnak meg egy országban a járvány alatt?

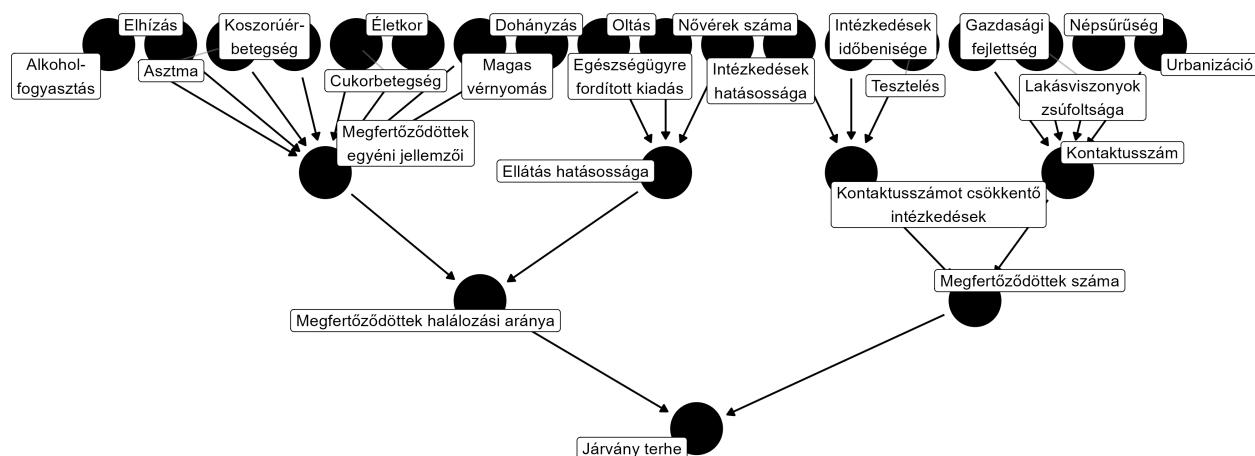
Legfelső szinten függ attól, hogy (A) hányan fertőződnek meg és (B) a megfertőződtek milyen arányban halnak meg. Világos, hogy annál több halálozás lesz, minél több a fertőzött, illetve ha a fertőzöttek minél nagyobb arányban halnak meg. (Vagy, természetesen, mindenki egyszerre.) Legfelső szinten így dekomponálhatjuk a problémát.

Menjünk tovább. Eggyel lejjebb haladva az ‘A’ megint két tényezőtől függ: hogy milyen a kontaktusszám az országban és milyenek az ezt csökkentő intézkedések. Ezek aztán megint tovább bonthatóak. Az előbbi függ a népsűrűségtől, a lakásviszonyok zsúfoltságától, az érintkezésre vonatkozó szociális szokásoktól és hagyományuktól, a különböző generációk együttélési mintázataitól, a szabadidős tevékenységek jellegétől, a városi lakosság arányától, a munkavégzés jellegzetességeitől stb. stb. Számíthat még az ország bekötöttsége a nemzetközi turizmusba, kereskedelemben, a népmozgás intenzitására. Az utóbbi függ az intézkedések szigorúságától, időbeniségétől, betartásuk fegyelmétől, a tesztelési, kontaktuskutatási, karanténozási stratégiától, a járványügy szervezettségtől, hitelességtől stb. stb.

A ‘B’ tényező megint csak két részre bontható: függ az alanyok olyan jellemzőitől, amik befolyásolják a prognózist egyszerűen, másrészt az ellátásuk hatékonyságától. Az előbbi megint rengeteg tényezőre bontható tovább: mindenekelőtt számít az életkor összetétel, tehát a korfa, de ott van az elhízás, a dohányzás prevalenciája, az alkoholfogyasztás mennyisége, a releváns társbetegségek – ez önmagában több tucat – prevalenciája, ami szintén eltérhet országok között. A medikális beavatkozás megint kettéválik: egyik oldalról számít az oltási program (az oltóanyag portfólió összetétele és az átoltottság), másik oldalról a kezelési oldal. A gyógyszerek elérhetősége, felhasználása, orvosok, nővérek, szakdolgozók tudása, de – mindenekelőtt – a túlterhelődésük. Az, hogy az ellátás mennyire zajlik egységes irányelvek mentén, milyen a minőségbiztosítás és a teljesítménymérés (régi vesszőparipám). Stb. stb. stb. És akkor még nem említettem, hogy számíthat az összes felsorolt tényező országon belüli egyenlőtlensége, ami szintén nem biztos, hogy ugyanaz minden országban.

Láthatjuk, hogy a cél az ilyen végiggondolásnál nem is feltétlenül az, hogy minden egyes tényezőt számba vagyunk az utolsó szálig, hanem sokkal inkább, hogy a probléma struktúráját megértsük.

Amit a fentiekben verbálisan igyekeztem körülírni, az lényegében egy egyszerűsített változata annak, amit kauzális diagramnak szoktak nevezni:



A valóságban a kauzális diagramok ennél jóval összetettebbek, itt ez némileg idézőjelben értendő, és inkább csak egyfajta grafikus megjelenítése a fenti leírásnak. Ha csak annyit teszünk, hogy egy ilyet felvázolunk, sokszor már az is segíti a strukturálást és így a jobb megértést.

Technikai részletek

A következő pontban az empirikus számításoknál az azokat megvalósító konkrét kódokat is közölni fogom. Ennek célja az, hogy az ez után érdeklődők azt is lássák, hogy mi történik a háttérben, milyen kódokkal lehet kivitelezni ténylegesen ezeket a vizsgálatokat. Több esetben elő fog fordulni, hogy módszertani részleteket nem írok le, hogy az ez iránt nem érdeklődő olvasót ne terheljem ilyenekkel, de azok a kódok könnyedén kiolvashatóak lesznek.

Az ebből való tanuláson túl az ilyen kódok közlését elvileg is fontosnak tartom, a reprodukálhatóság és a nyílt tudomány jegyében, hogy az esetleges hibáim könnyebben kiderüljenek, és elősegítsem továbbfejlesztési lehetőségek megfogalmazását. Ha valakit ez a része nem érdekel a kérdésnek, bátran ugorja át a szürke hátterű kódokat és e fejezet hátralevő részét.

A számítások R statisztikai programnyelven készültek, 4.1.2-es verziót használva. Felhasználtam a `data.table` (1.14.2-es verzió) és `ggplot2` (3.3.5-es verzió) csomagokat.

A kimeneti adat, azaz a többlethalálozási számok forrása a ‘Többlethalálozási adatok európai összefektetésben’ című anyagom. Ebben minden további technikai kérdés részletesen le van írva, de egy mondatban összefoglalva: Acosta és Irizarry módszerét használtam a többlethalálozási becsléséhez.

A potenciális magyarázó adatok forrása az egységesség és az egyszerűség kedvéért mindenhol az Eurostat volt. A következő adatokat gyűjtöttem le (nem feltétlenül fogom kivétel nélkül mindegyiket használni a későbbiekben, de az adatbázisban rendelkezésre áll, ha valaki szeretne tovább kísérletezni a témaiban); ígyekeztem mindenhol a legkésőbbi, a járvány időpontjához legközelebbi adatokat használni:

Változó	Eurostat azonosító	Év	Mértékegység
Népsűrűség	tps00003	2019	fő/km2
Túlzsúfolt lakások aránya	tessi170	2019	%
Városi lakosság aránya	ilc_lvho01	2019	%
Egy főre jutó bruttó hazai termék	nama_10_pc	2019	Folyó áron, PPS
65 év felettiek aránya	demo_pjan	2019	%
Elhízás prevalenciája	sdg_02_10	2019	%
Dohányzás prevalenciája	hlth_ehis_sk1i	2019	%
Naponta alkoholt fogyasztók aránya	hlth_ehis_al1i	2019	%
Cukorbetegség prevalenciája	hlth_ehis_cd1e	2019	%
Magas vérnyomás prevalenciája	hlth_ehis_cd1e	2019	%
Asztma prevalenciája	hlth_ehis_cd1e	2019	%
Koszorúér-betegség prevalenciája	hlth_ehis_cd1e	2019	%
Egy főre jutó egészségügyi kiadás	hlth_sha11_hf	2019	euró/fő
Ezer főre jutó nővérek száma	hlth_rs_prsns	2017	fő/ezer fő

Elsőként letöltyük ezeket az adatforrást, majd egy táblában egyesítjük az összeset a későbbi felhasználás céljából. A táblát könnyen kezelhető csv formátumban is mentük ki, hogy megkönnyítsük az adatainkat esetlegesen felhasználók dolgát:

```
RawData <- Reduce(function(...) merge(..., by = "geo"), list(
  RawData,
  as.data.table(eurostat::get_eurostat("tps00003"))[
    unit=="PER_KM2"&time=="2019-01-01",.(geo, popdensity = values)],
  as.data.table(eurostat::get_eurostat("tessi170"))[
    sex=="T"&time=="2019-01-01",.(geo, overcrowding = values)],
  as.data.table(eurostat::get_eurostat("ilc_lvho01"))[
    incgrp=="TOTAL"&building=="TOTAL"&deg_urb=="DEG1"&time=="2019-01-01",
    .(geo, urbanization = values)],
  as.data.table(eurostat::get_eurostat("nama_10_pc"))[
    unit=="CP_PPS_EU27_2020_HAB"&na_item=="B1GQ"&time=="2019-01-01",
    .(geo, gdp = values)],
  as.data.table(eurostat::get_eurostat("demo_pjan"))[
    sex=="T"&time=="2019-01-01",
    .(popold = sum(values[age%in%c(paste0("Y", 65:99),
      "Y_OPEN"))])/values[age=="TOTAL"]*100), .(geo)],
  as.data.table(eurostat::get_eurostat("sdg_02_10"))[
    bmi=="BMI_GE30"&time=="2019-01-01", .(geo, obese = values)],
  as.data.table(eurostat::get_eurostat("hlth_ehis_sk1i"))[
    smoking=="NSM"&quant_inc=="TOTAL"&sex=="T"&age=="TOTAL"&time=="2019-01-01",
    .(geo, smoke = 100-values)],
  as.data.table(eurostat::get_eurostat("hlth_ehis_al1i"))[
    frequenc=="DAY"&quant_inc=="TOTAL"&sex=="T"&age=="TOTAL"&time=="2019-01-01",
    .(geo, alcohol = values)],
  dcast(as.data.table(eurostat::get_eurostat("hlth_ehis_cd1e"))[
    unit=="PC"&isced11=="TOTAL"&time=="2019-01-01"&sex=="T"&age=="TOTAL"&
    hlth_pb%in%c("ASTHMA", "HBLPR", "DIAB", "CHRT_ANGPPEC"), .(geo, hlth_pb, values)],
    geo ~ hlth_pb, value.var = "values"),
```

```

as.data.table(eurostat::get_eurostat("hlth_sha11_hf"))[
  unit=="EUR_HAB"&icha11_hf=="TOT_HF"&time=="2019-01-01",
  .(geo, healthexpenditure = values)],
as.data.table(eurostat::get_eurostat("hlth_rs_prsns"))[
  unit=="P_HTHAB"&wstatus=="PRACT"&time=="2017-01-01"&isco08=="OC2221_3221",
  .(geo, nurses = values)])
names(RawData) <- tolower(names(RawData))

write.csv2(RawData, "RawData.csv", row.names = FALSE)

```

A változók sora – mint a fenti leírás is mutatta – végeláthatatlanul bővíthető, ez pusztán egy illusztrációs kiindulópont.

A vizsgált országok körét meghatározza az a tény, hogy mely országokra van információink; végeredményben 22 ország lesz az adatbázisunkban; angol nevükkel ezek a következők: Austria, Belgium, Bulgaria, Cyprus, Czechia, Germany, Denmark, Estonia, Greece, Spain, Croatia, Hungary, Italy, Lithuania, Luxembourg, Latvia, Netherlands, Norway, Poland, Romania, Sweden, Slovenia.

A begyűjtött adataink:

```
knitr::kable(RawData[, c(1, 20, 21:35)], digits = 1)
```

		nuts_	geo-	cum-	ex-	pop-	overc-	urba-		al-		healt-					
		vel	name	geo-	cess-	per-	den-	niza-	gdp	pop-	obe-	co-	hex-				
				ge-	pop-	pop-	sity	row-	gdp	old	se	smok-	pen-				
AT	0	Aust-	ria	1799.1	107.6	15.1	31.0	39519	18.8	17.1	26.2	5.7	4.3	3.2	6.0	21.8	4671.6
BE	0	Bel-	gi-	2027.9	377.3	5.7	29.5	36919	18.9	16.3	19.4	9.7	5.8	1.5	5.8	17.4	4418.1
BG	0	Bul-	ga-	8377.7	63.4	41.1	44.8	16665	21.3	13.6	36.2	10.2	2.2	7.0	6.9	29.7	625.6
CY	0	Cyprus		851.3	95.7	2.2	51.8	28803	16.1	15.2	25.5	4.0	4.0	2.1	7.0	18.9	1771.2
CZ	0	Cz-	echia	3851.9	138.2	15.4	30.0	29155	19.6	19.8	26.4	7.8	4.6	2.8	8.8	26.3	1644.1
DE	0	Ger-	many	938.3	235.2	7.8	36.3	37860	21.5	19.0	28.3	7.5	8.0	4.4	8.7	26.2	4855.3
DK	0	Den-	mark	312.3	138.5	10.0	37.6	39916	19.6	16.5	20.0	9.6	7.2	1.4	5.3	18.9	5355.1
EE	0	Es-	to-	2213.6	30.5	13.9	61.0	25789	19.8	21.8	24.8	1.3	4.1	4.7	6.0	23.3	1426.0
EL	0	Gree-	ce	2309.8	82.4	28.7	36.9	20651	22.0	16.7	28.6	5.9	3.3	2.9	8.0	19.6	1340.8
ES	0	Spa-	in	1805.8	93.8	5.9	49.6	28382	19.4	16.0	22.1	13.0	4.1	0.7	7.5	19.3	2411.7
HR	0	Cro-	atia	3808.6	72.8	38.5	29.6	20768	20.6	23.0	25.7	10.2	4.8	8.9	12.1	37.3	930.6
HU	0	Hun-	gary	3534.6	107.1	20.3	32.8	22800	19.3	24.5	27.2	6.3	5.0	3.6	8.9	31.5	949.4
IT	0	Italy		2736.0	201.5	28.3	35.3	30189	22.9	11.7	22.4	12.1	4.6	2.1	6.5	20.4	2599.2

		nuts_geo	legeo- vel	name	cum- ex- cess- per- pop	pop- den- sity	overc- row- ding	urba- niza- tion	gdp	pop- old	obe- se	smok- hol	al- ma	asth- ma	chrt_anti- ab	healt- hblpr	hex- pendi- ture
LT	0	Lit- hua- nia		4776.0	44.6	22.9	43.2	26219.19.8	18.9	23.7	0.8	2.8	6.0	5.3	29.9	1223.8	
LU	0	Lu- xem- bourg		458.4	239.8	7.1	19.6	79634.84.4	16.5	18.2	8.9	6.0	1.7	4.6	15.5	5502.1	
LV	0	Lat- via		3123.4	30.2	42.2	43.8	2169720.3	23.0	26.8	1.2	3.8	5.8	5.7	31.7	1045.6	
NL	0	Net- her- lands		1579.9	507.3	4.8	56.2	40140.19.2	14.7	21.1	8.3	6.4	2.6	5.8	16.1	4748.7	
NO	0	Nor- way		512.8	17.3	6.1	28.9	45442.87.2	14.1	18.1	1.4	7.9	1.4	4.5	15.1	7126.7	
PL	0	Pol- land		4347.8	123.6	37.6	35.0	22740.87.7	19.0	22.6	1.6	4.1	7.5	8.1	26.5	906.1	
RO	0	Rom- ania		6120.8	82.7	45.8	28.8	21674.88.5	10.9	27.3	2.9	1.5	1.5	5.0	15.7	661.3	
SE	0	Swe- den		1458.6	25.2	15.6	40.3	37143.89.9	15.3	12.6	1.8	7.5	1.3	6.3	18.2	5041.8	
SI	0	Slo- ve- nia		2210.7	103.7	11.6	19.5	27659.19.8	19.9	23.2	6.6	4.8	3.2	7.8	25.4	1975.2	

Érzékelhetőek a hatalmas különbségek: a többlethalálozás az egymillió lakosonként 500 alattitől (Dánia) a 8000 felettiig (Bulgária) terjednek, és több magyarázó jellegű változóban is vannak drámai eltérések, nagyságrendi különbségek vannak a népsűrűségben, az alkoholfogyasztásban, de még az egészségügyre fordított kiadásokban is.

Az empirikus vizsgálat lehetőségei és nehézségei

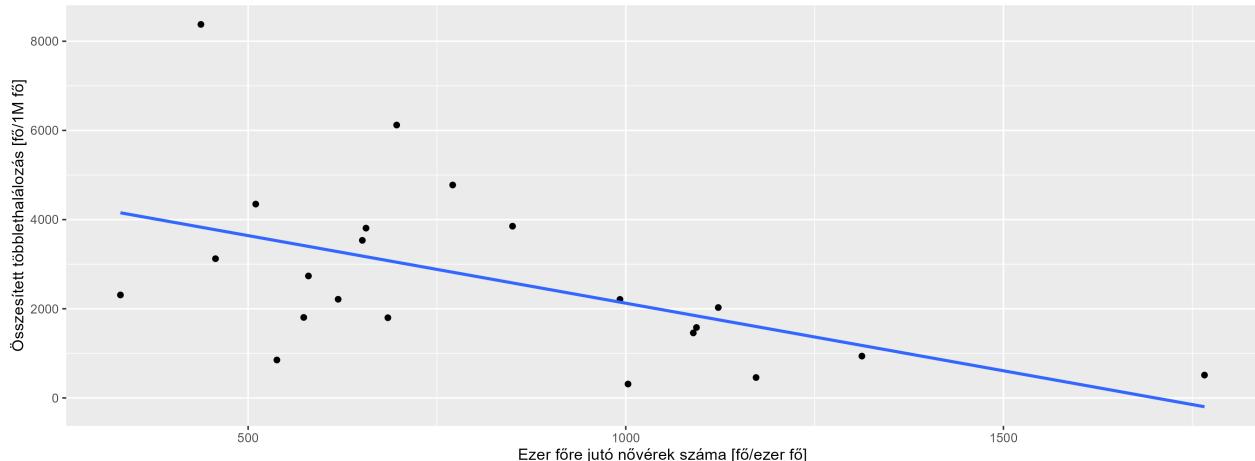
Miután nagyon alaposan végiggondoltuk a háttérben lévő elméleti modellt, elkezdhetünk foglalkozni azzal a kérdésből, hogy ebből mit és hogyan tudunk empirikus adatok alapján megbecsülni (illetve mit nem).

A confounding problémája

Először, hogy benyomást kapjunk az elemzési lehetőségekről, és azok problémáiról, kezdjünk néhány egyszerű vizsgálattal.

Sokan mondják, hogy a kimenet összefügg a nővérek számával: ahol több van, ott kedvezőbben alakult a járvány-halálozás. Nézzük is meg a kérdést empirikusan! Ezt látjuk az európai országok körében:

```
ggplot(RawData, aes(x = nurses, y = cumexcessperpop)) + geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  labs(x = "Ezer főre jutó nővérek száma [fő/ezer fő]",
       y = "Összesített többlethalálozás [fő/1M fő]")
```

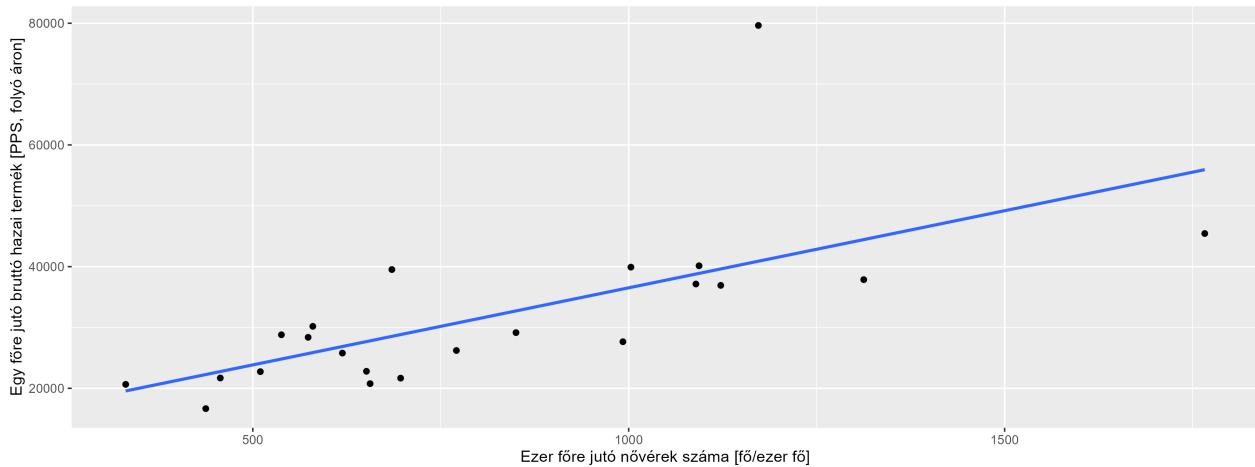


Első ránézésre teljes mértékben megerősítjük a felvetést: a több ápoló valóban alacsonyabb halálozás jár együtt. A behúzott vonal a pontokra legjobban illeszkedő egyenes; később ennek majd nagyobb jelentősége lesz, egyelőre fogjuk fel úgy, hogy azért van ott, hogy vezesse a szemet.

Ha figyelmen kívül hagyjuk a felvezetésben mondottakat, azaz megfeleldkezünk a confounding lehetőségéről, akkor akár le is vonhatnánk a következtetést, hogy a nővérek száma *csökkenti* a halálozást. (Figyeljünk a szóhasználatra! A „csökkenti” már egy *kauzális* szó, azt mondja, hogy okozati összefüggés van, szemben a „jár együtt” megfogalmazással.) Csakhogy eszünkbe jut a confounding problémája: mi van, ha a nővérek száma összefügg valamilyen más változóval, ami a *valódi* oka a kedvezőbb adatoknak...?

Például eszünkbe jut, hogy a jobb gazdasági állapotú országokban több az ezer főre jutó nővér. (Megengedhetik maguknak? Más az ottani felfogás a nővérek szerepére vonatkozóan? – ez most, ilyen szempontból, mindegy is.) Ellenőrizzük is ezt gyorsan le, a gazdasági fejlettség mérésére a GDP-t használva:

```
ggplot(RawData, aes(x = nurses, y = gdp)) + geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  labs(x = "Ezer főre jutó nővérek száma [fő/ezer fő]",
       y = "Egy főre jutó bruttó hazai termék [PPS, folyó áron]")
```



Ez bejött! De akkor viszont... lehet, hogy baj van? Lehet, hogy az ápolók száma nem is fontos, csak a fejlettségé...? Nézzük meg:

```
ggplot(RawData, aes(x = gdp, y = cumexcessperpop)) + geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  labs(x = "Egy főre jutó bruttó hazai termék [PPS, folyó áron]",
       y = "Összesített többlethalálozás [fő/1M fő]")
```

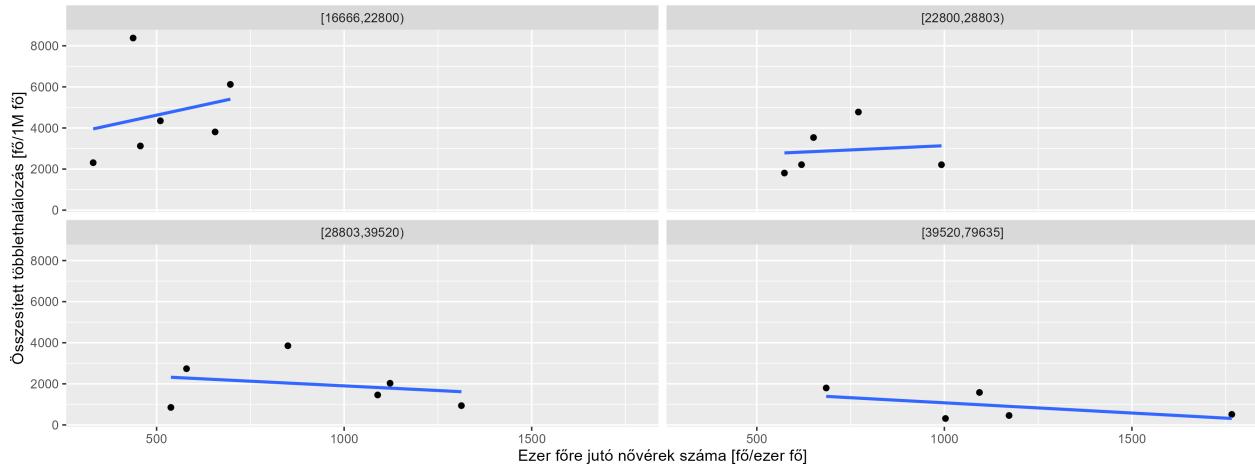


És akkor most tényleg bajban vagyunk.

Persze vigyázat: nem arról van szó, hogy bebizonyítottuk, hogy a nővérek számának nincs hatása – ettől még éppenséggel lehet! Csak azt bizonyítottuk be, hogy erre vonatkozólag az első ábra, bármennyire is szuggerálná ezt a konklúziót, valójában nem bizonyító erejű.

Hogy lehet ennek a kérdésnek utánajárni? Próbálkozzunk egy trükkös ábrázolással. Ismét a nővérek száma és a kimenet közti kapcsolatot ábrázoljuk, ugyanúgy mint a legelső ábrán, de úgy, hogy megbontjuk a GDP értékei szerint:

```
ggplot(RawData, aes(x = nurses, y = cumexcessperpop)) + geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  facet_wrap(~Hmisc::cut2(gdp, g = 4)) +
  labs(x = "Ezer főre jutó nővérek száma [fő/ezer fő]",
       y = "Összesített többlethalálozás [fő/1M fő]")
```



Bár az eleve sem túl nagy mintanagyság további sztosztódása miatt ez kicsit nehezebben értelmezhető, az azért így is látszik, hogy valami nagyon érdekes történt: míg összességében nézve egyértelmű és negatív kapcsolat volt (több nővér – kevesebb halálozás), addig most ez a kapcsolat teljesen megszűnt! Néhol egy kicsit negatív kapcsolat van, néhol egy kicsit pozitív, ilyen mintaméretek mellett egyiknek sincs jelentősége, nem látunk érdemi kapcsolatot sehol sem.

Ahhoz, hogy megértsük, hogy itt mi történik, gondoljuk végig, hogy az első ábra esetében mit jelent a confounding problémája. Mi ott a probléma? Az, hogy ha megyünk jobbra (több ápoló), akkor *együttal* a gazdasági fejlettség is javul, nem csak az ápolók száma nő. Azaz, amit látunk (csökkenő halálozás), valójában nem pusztán a több ápoló, hanem a több ápoló és magasabb GDP *együttet* hatása (és nem tudjuk, hogy

milyen arányban). A probléma tehát, hogy egy változót akartunk vizsgálni (ápolók száma), de ahogy az változik az ábrán, vele együtt valami más is odébbmászik.

...ez a mostani ábrázolás azonban pont ezt oldja meg! Hiszen az egyes kis részábrákban a GDP adott értékű (közelítőleg), azaz, amikor a kis részábrákat nézzük, akkor a fenti hatást kikapcsoltuk! Ilyen értelemben ez az ábrázolás az ápolók számának valódi hatását igyekszik megragadni, tisztítva a fejlettség miatti confounding-tól. (Ezt a módszert szokás egyébként rétegzésnek nevezni.)

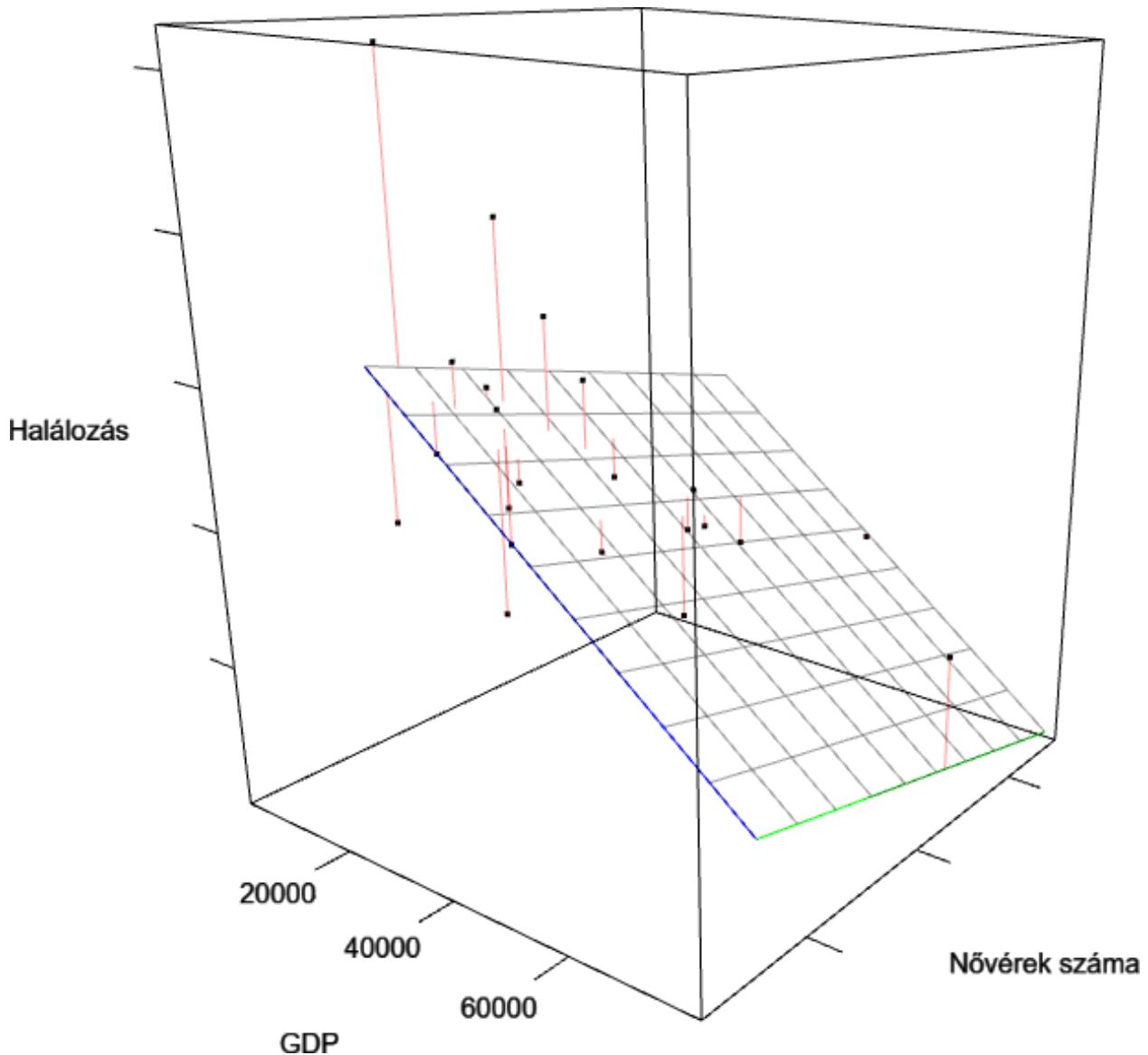
Ez itt egy nagyon fontos általános gondolat: hogy egy változó hatását úgy kapjuk meg, ha *csak* a kérdéses tényezőt változtatjuk. És most értünk körbe, hiszen honnan indultunk? Abból, hogy az a probléma, hogy nem tudunk „egyetlen és csakis egyetlen tényezőt” változtatni – ezt látjuk a fenti ábrákban megjelenői. Csakhogy most már, ezzel az utolsó ábrával, az e probléma elleni küzdelem eszköze is kezd a kezünkbe kerülni.

Egy pillanatra még érdemes elidőzni az ábrázolásoknál. Ez keményebb dió lesz az eddigieknel, de kiegészítő pontként nem felesleges: egy utolsó vizualizációt mutatok, ami az előbbieknél jobb térlátást igényel, de megéri, ha az ember becsukja a szemét, és igyekszik meglátni maga előtt a helyzetet, mert segíti, hogy mélyebben megérte a confounding mibenlétét. Készítünk egy három dimenziós ábrát! Itt természetesen nem egyenest, hanem egy síkot illesztünk a pontjainkra (a függőleges piros vonalak csak a síktól vett eltérést mutatják):

```
fit <- lm(cumexcessperpop ~ gdp + nurses, data = RawData)

gdpgrid <- seq(min(RawData$gdp), max(RawData$gdp), length = 10)
nursegrid <- seq(min(RawData$nurses), max(RawData$nurses), length = 10)
predgrid <- predict(fit, expand.grid(gdp = gdpgrid, nurses = nursegrid))

rgl::plot3d(RawData$gdp, RawData$nurses, RawData$cumexcessperpop,
            xlab = "GDP", ylab = "Nővérek száma", zlab = "Halálozás")
# xlab = "Egy főre jutó bruttó hazai termék [PPS, folyó áron]",
# ylab = "Ezer főre jutó nővérek száma [fő/ezer fő]",
# zlab = "Összesített többlethalálozás [fő/1M fő]"
rgl::view3d(userMatrix = matrix(c(0.74, -0.18, 0.44, 0, 0.67, 0.23, -0.70, 0, -0.02, 0.95,
                                 0.30, 0, 0, 0, 0, 1), nc = 4))
rgl::segments3d(rbind(RawData$gdp, RawData$gdp),
                rbind(RawData$nurses, RawData$nurses),
                rbind(RawData$cumexcessperpop, predict(fit)),
                alpha = 0.4, col = "red")
rgl::surface3d(gdpgrid, nursegrid, predgrid, alpha = 0.4, front = "lines")
rgl::lines3d(c(gdpgrid[1], gdpgrid[10]), rep(nursegrid[1], 2),
             c(predgrid[1], predgrid[10]), col = "blue")
rgl::lines3d(rep(gdpgrid[10], 2), c(nursegrid[1], nursegrid[10]),
             c(predgrid[10], predgrid[100]), col = "green")
```



Amit látunk, hogy a GDP–halálozás vetületben ferde a sík (nézzük a kék élet) – ez fejezi ki azt, hogy a GDP növekedtével csökken a halálozás. Viszont a nővér–halálozás vetületben (zöld él) szinte vízszintes – a GDP-t is figyelembe véve már nincs hatása a nővérek számának! Ez egy fontos megállapítás, vagy jobban mondva újabb elmondása a korábbi megállapításnak: *ha* lerögzítjük a GDP-t, akkor a nővérek változatára már nem számít. Ha rögzítjük a GDP értékét, azaz kiválasztunk egy pontot a kék él mentén, majd *adott, rögzített GDP mellett* elkezdjük változtatni a nővérek számát, azaz a kiválasztott ponttól elindulunk a sík mentén a nővér él növekvő irányába, akkor szinte vízszintesen haladunk – nem változik a halálozás.

Ez egyúttal azt is jelenti, hogy a sík valami nagyon fontos dolgot tud: annak ellenére, hogy a valódi adatokban (lévén, hogy nem kísérletről van szó) együttesen változik a két tényező, mi *mégis*, pusztán *matematikai úton* megmondjuk, hogy mi lenne, ha csak az egyik változna. Úgy tudunk erre válaszolni, hogy pusztán megfigyeléses adataink vannak! Erre a gondolatra mindenki visszatér.

Az ábra azt is szemlélteti, hogy miért lép fel a confounding. Hiszen, kérdezhetné a fentiek alapján valaki, akkor miért láttuk a legelső ábránkon azt, hogy az ápolók száma és a halálozás között van összefüggés? Egy egyértelműen negatív meredekségű egyenest tudtunk behúzni. Most meg azt mondjuk, hogy „vízszintes“ abban az irányban a sík, és „nem számít“ az ápolók száma...?! Akkor most vízszintes, vagy negatív meredekségű? Számít a nővérek száma vagy sem? Az ellentmondás csak látszólagos. Nézzük meg jobban az ábrát! A

pontok nem akárhogy helyezkednek el a síkon: alapvetően a jobb alsó sarok – bal felső sarok átló mentén szóródnak. (Ez fejezi ki azt, hogy a GDP és a nővérek száma *egymással is összefügg!*) Azaz amikor mi csak a nővéreket vizsgáltuk, akkor nem *pusztán* balról jobbra haladtunk a nővér tengely mentén, hanem *egyúttal* jöttünk fentről is le! És itt van a magyarázat: ezért láttuk úgy, hogy van összefüggés, és negatív a kapcsolat, hiszen ez utóbbi, a fentről lefelé jövés okozta a csökkenő halálzást.

Tudom, hogy minden elég agyzibbasztó, és valószínűleg kell rá aludni egyet (többet), ha valaki először látja, de e kép megértése sokat segít a confounding átlátásában és a megoldás megtalálásában is.

Kihagyott változós torzítás

Valamit nagyon fontos megérteni részleteiben is a confounding problémája kapcsán: ha kihagyunk egy lényeges változót, tehát olyat, aminek *valójában* van hatása a kimenetre, akkor annak a hatását a bentmaradt változók veszik át. (Attól a gyakorlatban nem túl izgalmas esettől eltekintve, ha egyik vizsgált változóval sem függ össze a kihagyott.) Ezt szokták kihagyott változó okozta torzításnak nevezni.

Ha meggondoljuk, akkor ez a fenti, két magyarázó változót tartalmazó példán is tökéletesen elmondható: ha kihagyjuk a GDP-t, akkor a nővér változó valójában már nem *csak* a nővérek hatását fogja mutatni, hanem a nővérek és a GDP együttes hatását. A nővér változó, legalábbis részben, *átvette* a GDP szerepét is! Általánosan megfogalmazva: a vizsgált változóinkra kimutatott hatások – sajnos – a nem vizsgált változók hatásait (is) tartalmazzák, ha nem vizsgáltunk olyat, ami valójában lényeges.

Bár tényleg csak átfogalmazásról van szó, azért fontos külön is említeni, mert egy gyakori tévedés, ha egy nem vizsgált változóról automatikusan kijelentjük, hogy nincs hatása. Ez nem csak hogy nem igaz, de a helyzet rosszabb: ezt a hatást, ha van neki, sajnos a vizsgált változókra szétesztva fogjuk kimutatni...! Ha a társadalmi távolságtartásra vonatkozó szokásokra nincs adatunk (hogy mondjak egy csakugyan nehezen mérhető változót), akkor nem mondhatjuk, hogy annak nincs hatása, de ez még hagyján, az igazi baj, hogy a hatását, ha van neki, a többi változóban fogjuk elszámolni, mert azok átveszik a kimaradt változó hatását. Ha például a déli országokban szorosabbak a szociális kontaktusok mint az északiakban, és a déliek és az északiak között GDP-ben is van eltérés, akkor a szociális távolság hatását vidáman ki fogjuk mutatni GDP címszó alatt!

A kihagyott változó okozta problémák legegyszerűbb kezelési megoldása elég kézenfekvő: ne hagyunk ki lényeges változót... A dolog azonban nem ilyen egyszerű, két okból sem. Az egyik, kézenfekvő probléma, hogy mi van, ha nincs információnk a lényeges változóról? Akár azért, mert nem tudtuk begyűjteni az információt, akár azért, mert nem is gondoltunk rá, hogy be kellene gyűjteni! (Ugye ez a kísérletes vizsgálatok nagy előnye.) Megjegyzem, jelen esetben az információ begyűjtése sem feltétlenül nyilvánvaló; lehetnek változók amiről nincs megfelelő nemzetközi adatgyűjtés, vagy – mint az előző bekezdés példája is mutatja – az is előfordulhat, hogy nem is könnyű mérni, számszerűsíteni a tulajdonságot. Ilyenkor szoktak néha úgynévezett proxy változót használni, azaz olyat, ami – remélhetőleg szoros – kapcsolatban van a vizsgálni kívánt változóval, de szemben azzal, mérhető. Egyébként a GDP nagyon sok változónak lehet a proxy-ja (a probléma inkább pont az, hogy sok minden keveredik benne). A másik probléma, hogy még ha ismerjük is a lényeges változókat, baj származhat abból, ha nagyon sok van belőlük (a minta nagyságához képest). Ezt a kérdéskört fogjuk hamarosan körbejárni.

A regresszió eszköze

A háromdimenziós ábrával már nagyon közel kerültünk a confounding problémájának egy lehetséges kezeléséhez. (A „megoldásához” szó használata talán túlzás lenne, hiszen megfigyeléses adatból igazán bombabiztosan soha nem tudunk okozati hatásra következtetni. De az ezzel kapcsolatos problémákat enyhíthetjük.) Ez lesz az egyenes – és később a sík – behúzásának igazi értelme: nem a görbeillesztés a lényeg, hanem, hogy feltételezzük egy modellt, és annak paramétereit becsüljük meg a begyűjtött tényadatok alapján. Az egyenes és a sík épp egy ilyen modell megjelenése; ezt a modellt hívjuk lineáris regresszióknak. Ebben a modellben annak a hatása, ha az ezer lakosra jutó nővérek száma egygyel nagyobb miközben a GDP értéke rögzített, nem függ sem attól, hogy milyen értéken rögzített a GDP, sem attól, hogy honnan indulva növeljük meg egygyel a nővérek számát.

Az egész kulcsa a „miközben a GDP értéke rögzített“ kitétel: a confounding problémája *épp* az volt, hogy a nővérek számának növelésével a GDP is változik közben, de most mit látunk? Hogy a modell ezen paramétere azt jelenti, hogy mennyi a hatás *akkor* ha a GDP nem változik! Ilyen értelemben meg tudtuk azt tenni, hogy bár az adataink megfigyelések voltak, confounding-gal terhelve, mi mégis, *pusztán matematikai úton* kiszedtük belőle a confounding-tól tisztított értéket! Úgy szokták szép szóval mondani: kimatattuk a nővérek hatását úgy, hogy *kontrolláltunk* a GDP-re.

Mindezek eredményét mutatja a következő táblázat, a ‘Becsült hatás’ oszlop adja meg a fenti értelemben vett hatást:

```
knitr::kable(data.frame(`Becsült hatás` = signif(coef(fit), 3),
                        `95% CI` = apply(signif(confint(fit)), 3, 1, paste,
                                         collapse = " -- "),
                        p = Hmisc::format.pval(summary(fit)$coefficients[,4], digits = 3,
                                         eps = 0.001),
                        check.names = FALSE, row.names = c("Tengelymetszet",
                                                         names(coef(fit))[-1])))
```

	Becsült hatás	95% CI	p
Tengelymetszet	5990.0000	3990 – 7990	<0.001
gdp	-0.0746	-0.146 – -0.00277	0.043
nurses	-1.1400	-3.91 – 1.62	0.398

Észrevehető, hogy ez a módszer lényegében a rétegzés továbbfejlesztése: ott is arra törekedtünk, hogy a nővérek számának hatását úgy mutassuk ki, hogy a GDP nem változik vele együtt, azt állandóan tartjuk, de most ezt ügyesebben tesszük meg (pl. nem kell kategóriára osztani a GDP-t, amik ha túl szélesek, akkor nagyon különböző dolgokat mosnak egybe, ha túl szűkek, akkor kevés pont jut egy kategóriába, közvetlenül számszerű választ kapunk stb.).

Természetesen annak, hogy a modell által becsült paraméterek tényleg jó választ adjanak, bizonyos feltételeknek meg kell felelniük. Hát persze: nem fordulhat elő például, hogy van egy *harmadik* változó, amiről még mindig megfeledkezünk! Hiszen honnan tudjuk, hogy csak a nővérek és a GDP számít? Mi van, ha valamit így is kihagyunk? Ami, ha hat a halálozásra, megint pont a confounding problémáját hozza be! Itt jutunk vissza ahhoz a téTELmondathoz, amit már az elején is láttunk: megfigyeléses adatoknál mindenkor fejünk felett lebeg, hogy mi van, ha egy confounding-ot okozó változóról megfeledkezünk. Pontosan ez tükröződik itt is vissza: ha egy lényeges változót kihagyunk, akkor baj lesz.

Valójában nem ez az egyetlen feltevés, mi van például, ha a hatás nem lineáris? Ezek a kérdések mindenkorral jól vizsgálhatóak a regressziós keretrendszerben; a linearitás kérdését hamarosan külön is meg fogjuk nézni.

A mintavételi ingadozás és a torzítás fogalma

Az ilyen modellek becslésére rendelkezésre állnak jól bevált módszerek. Ezek milliónyi kérdése közül csak egyet emelnék most itt ki: az így kapott számok bizonytalanságának ügyét. A probléma az, hogy ezek becsült értékek, melyekben van attól függő ingadozás, hogy éppen milyen adatbázisból (mintából) becsültük, pusztán a véletlen szeszélye folytán: ha pontosan ugyanabbnál a valódi összefüggésből pontosan ugyanúgy véletlen mintát veszünk, akkor sem mindenkor ugyanazt kapjuk. A kapott becslés mintáról-mintára ingadozik, amiből persze egyúttal az is következik, hogy nem kaphatjuk meg mindenkor a valódi értéket. Nagyon fontos tehát, hogy most nem arról beszélek, hogy a mintát bármilyen értelemben is „hibásan“ vettük: a legtökéletesebb véletlen mintavétel esetén is lesz a becsült értékben ingadozás, pont úgy, ahogy a kihúzott lottószámok átlaga sem pont 45,5 minden héten. Pedig az összes számnak (1-től 90-ig) ennyi az átlaga, a mintavétel – remélhetőleg – itt aztán tökéletesen véletlen, és mégis, a mintának, tehát a kihúzott 5 számnak az átlaga néha kicsit kisebb mint 45,5, néha kicsit nagyobb! Ezt hívjuk mintavételi ingadozásnak.

Nézzük meg ezt egy konkrét példán! Sajnos az adatbázisunk erre a cérla nem használható, hiszen az csak egyetlen minta, és mi magunk sem tudhatjuk, hogy mi a háttérben lévő valóság. Ezért egy más eszközökhöz

nyúlunk: *szimulálni* fogunk, azaz fogunk egy valódi összefüggést, abból veszünk egy véletlen mintát, és az alapján becsüljük a regressziót. Ilyen módon egyszerűt mi is tudjuk mi a valóság, másrészről a mintavételt akárhányszor meg tudjuk számítógépen ismételni, és megvizsgálhatjuk a kapott eredményeket. Nézzünk is egy egyszerű példát, ahol a változó hatása – a fentiekben bemutatott értelemben – 2; a fekete pontok jelentik az újabb és újabb szimulációkat, a kék görbék az egyes szimulált esetekből becsült regressziót, a piros a valódi összefüggés:

```
n <- 20
nSim <- 20
SimData <- data.table(sim = rep(1:nSim, each = n), x = runif(n*nSim))
SimData$y <- 1 + 2*SimData$x + rnorm(n*nSim, 0, 0.5)
fits <- lapply(1:nSim, function(i) lm(y ~ x, data = SimData[sim==i]))
for(i in 1:nSim)
  print(ggplot(SimData[sim<=i], aes(x = x, y = y, alpha = 1 - 0.5*(sim < i))) +
    geom_point(color = "blue") +
    coord_cartesian(xlim = c(-0.1, 1.1), ylim = c(0, 4)) +
    geom_abline(intercept = coef(fits[[i]])["(Intercept)"],
                slope = coef(fits[[i]])["x"], color = "blue") +
    labs(title = paste0("Valódi hatás: 2, mintából becsült hatás: ",
                        round(coef(fits[[i]])["x"], 1), " (", i, ". szimuláció)")) +
    guides(alpha = "none") +
    if(i>1) lapply(1:(i-1), function(j)
      geom_abline(intercept = coef(fits[[j]])["(Intercept)"],
                  slope = coef(fits[[j]])["x"], color = "blue", alpha = 0.2)) +
    geom_segment(aes(x = x, xend = xend, y = y, yend = yend),
                 data.frame(x = -0.2, xend = 1.2, y = 1 + 2*(-0.2),
                             yend = 1 + 2*(1.2)), inherit.aes = FALSE, color = "red",
                 size = 1.2))
```

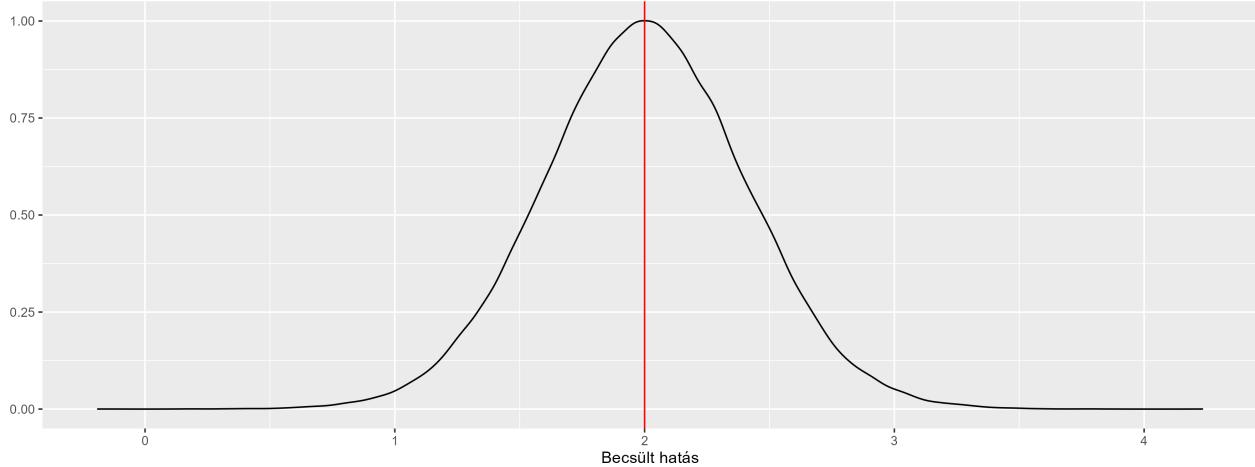
Ezekből az ábrákból egy doleg látható és egy doleg sejthető. Ami látható, az a mintavételi ingadozás: ahogy megbeszéltek, hiába adott, állandó, rögzített értékű a valódi összefüggés (piros vonal), hiába teljesen véletlen a mintavétel, a mintából becsült hatás ingadozik mintáról-mintára, pusztán a mintavétel szeszélye miatt tehát. Ez elkerülhetetlen. De az ábrák sugallanak is valamit: hogy van ugyan ingadozás, de azt remélhetjük, hogy ennek kellemes tulajdonságai lesznek! Az ábrákból ugyanis eléggé az látszik, mintha a becsült érték ingadozna ugyan, de átlagosan jó lenne, mert a jó érték körül ingadozik. Vajon igaz ez?

Futassuk most ne 20-szor a szimulációt, hanem 10 ezerszer, jegyezzük fel minden egyes futtatásnál a becsült hatást, majd nézzük meg ezek eloszlását:

```

simbeta <- as.data.table(replicate(1e5, {
  x <- runif(20)
  lm.fit(cbind(1, x), 1 + 2*x + rnorm(20, 0, 0.5))$coefficients[2]
}))
ggplot(simbeta, aes(x = V1)) + geom_density() +
  geom_vline(xintercept = 2, color = "red") + labs(x = "Becsült hatás", y = "")

```



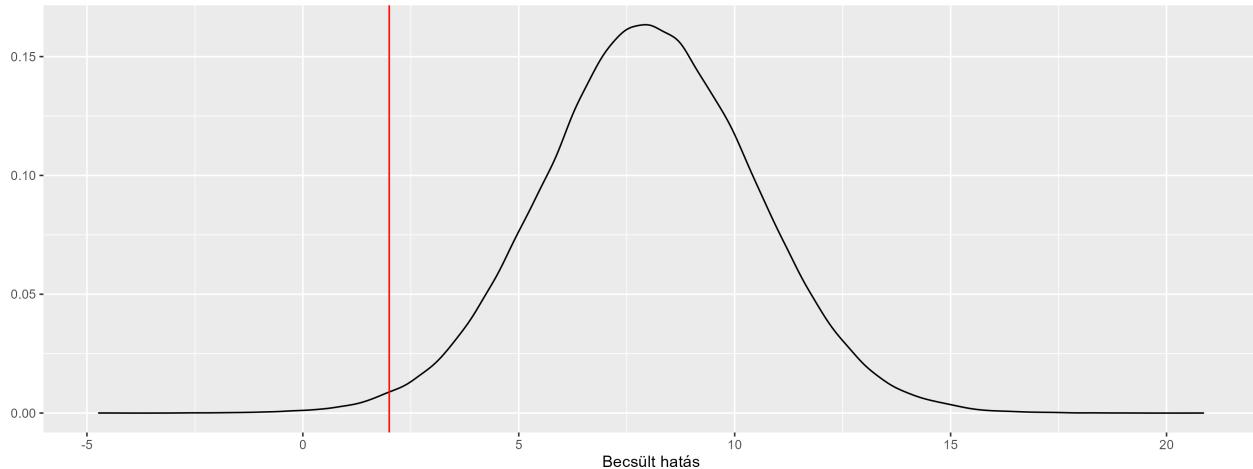
A piros jelzi a valódi értéket, a fekete pedig a mintából kapott becslések eloszlását. Csakugyan beigazolódott a sejtésünk, és valami megnyugtató dolgot látunk: ingadozik ugyan a becslés (amint láttuk, ez elkerülhetetlen), de átlagosan jó az értéke! A statisztikában ezt a tulajdonságot hívjuk torzítatlanságnak. Ha nem teljesül, tehát a becsléseknek az átlaga sem egyezik a valódi értékkel, akkor azt mondjuk, hogy torzított a becslés (és a torzítás az átlag és a valódi érték eltérése). A fentiből úgy tűnik tehát (nem akarom azt mondani, hogy „bizonyítottuk“, hiszen ez csak egy szimuláció), hogy a változó hatásának értékére kapott becslés torzítatlan. Megfelelő matematikai eszközökkel bizonyítani is lehetne, hogy valóban torzítatlan az így kapott becslés.

De álljunk meg egy pillanatra: korábban kihagyott változó okozta *torzítást* emlegettünk! Elég gyanús egybeesés, hogy itt is pont a torzítás szó került elő, és elárulhatom, hogy a dolog nem véletlen. Készítsünk most egy újabb szimulációt: a valóságban most a kimenetre két – összefüggő – változó hat, csak épp az egyiket kihagyjuk a regresszióból. Például a GDP és a nővérek száma hat a halálozásra, de mi mégis csak a nővérek hatását vizsgáljuk. Mi történik ilyenkor a regresszióban bent maradt változó kimutatott hatásával, példánkban a nővérekre kimutatott hatással? A piros vonal továbbra is a valódi (általunk a szimulációban beállított) értéket jelöli, a fekete görbe pedig ugyanúgy a mintákból kapott becslések eloszlása:

```

simbeta <- as.data.table(replicate(1e5, {
  x1 <- runif(20)
  x2 <- 1 + 2*x1 + rnorm(20)
  lm.fit(cbind(1, x1), 1 + 2*x1 + 3*x2 + rnorm(20, 0, 0.5))$coefficients[2]
}))
ggplot(simbeta, aes(x = V1)) + geom_density() +
  geom_vline(xintercept = 2, color = "red") + labs(x = "Becsült hatás", y = "")

```



Baj van! A becslések továbbra is ingadoznak, van mintavételi ingadozás (ez eddig nem meglepő, láttuk, hogy ez elkerülhetetlen) – csak épp most már *egy rossz érték körül*. Átlagosan sem jó az értékük, szép szóval élve: torzított a becslésünk. Hát ezért hívtuk ezt „kihagyott változó torzításnak“...!

Valaki megkérdezheti, hogy ez hogyan lehet, hiszen az előbb még azt mondta, hogy „matematikai eszközökkel bebizonyítható“, hogy a kapott becslések torzítatlanok. A válasz az, hogy igen, bebizonyítható, ha bizonyos feltételek teljesülnek. Rögtön el is árulhatom tehát, hogy a confounding, a lényeges változó kihagyása épp e feltételek egyikét fogja megsérteni. Ezért nem fog ebben az esetben működni ez a bizonyítás, és ezért lesz ilyenkor torzított a kapott eredményt.

A fentiek a mintavételi ingadozást úgy fogalmazták meg, hogy ha tudom, hogy a valódi érték mondjuk 10, akkor a mintából becsült lehet 9 vagy épp 11 is. A dolognak azonban van egy ezzel egyenértékű, de a mi szempontunkból fontosabb megfogalmazása: ha a mintában 20-at kaptam, attól még a valódi lehet 19 vagy 21 – egyenértékű, hiszen az előbbi megfogalmazás azt mondja, hogy a 19 vagy a 21 is beingadozhat a 20-ra, pusztán a véletlen szeszélye folytán. Azért mondta, hogy a mi szempontunkból ez utóbbi megfogalmazás a fontosabb, mert mi az utóbbi helyzetben vagyunk ténylegesen, a gyakorlatban (azaz ha épp nem szimulációs vizsgálatot csinálunk): ekkor ugyanis mi sem tudhatjuk, hogy mi a valóság, csak azt látjuk, hogy 20-at kaptunk a mintából becslésként. Ez a megfogalmazás tehát egy nagyon fontos üzenettel bír számunkra: vigyázat, ettől még nem biztos, hogy 20 a valódi érték! És most nem a confoundingról vagy hasonló gikszerről beszélek, ha minden ilyet elkerültünk, és minden tökéletesen csináltunk, ez *akkor is* igaz lesz, pusztán a véletlen ingadozás miatt. Nem biztos, hogy 20 a valódi érték, ebben bizonytalanság van, lehet éppenséggel 19 vagy 21 is. De lehet 18 vagy 22 is? 15 vagy 25 is? Lehet 0 is...? (Ez utóbbi pláne fontos, mert az azt jelentené, hogy a változónak igazából nincs is hatása!)

A kérdésre biztos válasz nincs, de valószínűségi választ adhatunk. Ezt teszi meg a harmadik oszlopban feltüntetett úgynevezett konfidenciaintervallum (CI). Ez tartalmazza azokat az értékeket, amikre igaz, hogy ha az lenne a valódi, akkor könnyen beingadozhatnának abba, amit ténylegesen kaptunk is: a -0,142 és a -0,003 közti valódi értékekből könnyen kaphatnánk, pusztán a véletlen ingadozás miatt, -0,072-t. (A fejlécben feltüntetett 95% szabályozza azt, hogy mit értünk „könnyen“ alatt.) Úgy is szokták mondani, hogy ilyen valódi értékek esetén a tényleges érték attól való eltérése betudható a véletlen ingadozásnak, szép szóval: nem *szignifikáns* az eltérés. Ha a konfidenciaintervallum tartalmazza a nullát (ami ugye, ahogy az előbb is mondta, azt jelenti, hogy a kérdéses változónak valójában nincs hatása az eredményre!), az magyarra lefordítva azt jelenti, hogy nincs okunk feltételezni, hogy a változónak van hatása, mert ha nem lenne, akkor is kényelmesen kijöhetett volna az, ami ki is jött. Ilyenkor mondjuk azt, hogy a változó hatása nem szignifikáns.

A regresszió kérdésköre ezen túlmenően is egy hatalmas, de érdekes, és nagyon sok területen – így a biostatisztikában kiemelten – fontos téma. Most szinte csak utalásszerűen tudtam megemlíteni pár fontos kérdését, de a részletek előadásaimban és jegyzeteim között elérhetőek (néhány adott témaiba vágót a szövegben is belinkelek a megfelelő helyen).

A függvényforma kérdése

A fentiekben hangsúlyosan szerepelt a linearitás, mint a modellünk alapfeltevése. Érdemes erről egy picit bővebben is beszélni, azért is, mert egyúttal jó általános illusztráció az ilyen modellfeltevések szerepére, ellenőrzésére és feloldására is.

Először is kezdjük azzal, hogy egyáltalán miért tételezzük fel, hogy a valóság lineárisan működik? Mi van, ha a hatás nem lineáris, például eleinte nagyon számítanak a plusz nővérek, de később már egyre kevésbé? Mi van, ha a nővérek hatása függ a másik magyarázó változótól, a gazdasági fejlettségtől, például alacsony fejlettségnél jobban számít plusz egy nővér, de magasnál már kevésbé? (Ez utóbbi esetben automatikusan igaz lesz az is, hogy a gazdasági fejlettség halálozásra gyakorolt hatása is függ a nővérek számától; ilyenkor szokták azt mondani, hogy interakcióban van a két változó.)

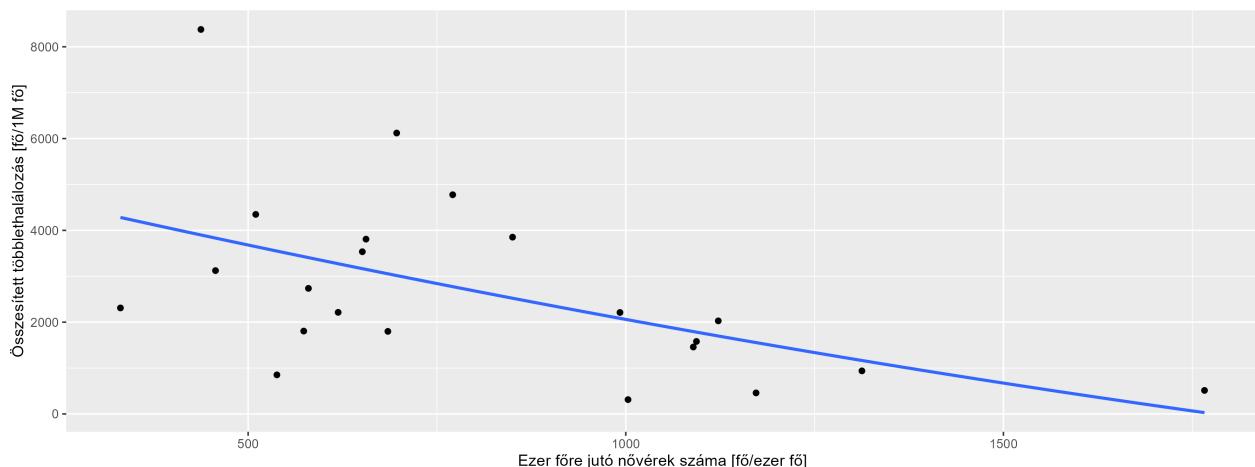
Kezdjük ott, hogy igen, csakugyan, a valóság működése általában pont hogy nem lineáris. Mégis, jó okaink vannak ennek ellenére is a linearitás használatára, legalábbis első közelítésként. Az egyik, hogy a lineáris modellek kényelmesek: a kapott eredmények nagyon jól interpretálhatóak: egy változó hatása egyetlen szám, azzal a nagyon egyszerű értelmezéssel, hogy +1 egység növekedés minden mást változatlanul tartva hogyan hat a kimenetre. Az, hogy egyetlen számot kell becsülni, ráadásul statisztikailag is nagyon előnyös, kis mintánkon is ez működik a legjobban. Mindemellett a lineáris modellek jól használhatóak extrapolációra, azaz, ha a rendelkezésre álló adatok tartományán kívül eső területről kell nyilatkoznunk, akkor egyszerűen meghosszabbíthatjuk az egyenest minden matematikai nehézség nélkül.

Persze a kényelem nem sokat ér, ha a valóság nem így működik! De itt jön a második előnyös vonás: az, hogy valóságban nem lineáris összefüggések is elég jól közelíthetők lineárissal, ha kellően szűk tartományon dolgozunk. Viszont ez úgy képzelhető el, hogy ha veszünk is egy akármilyen hepe-hupás függvényt, ha kellően jól ránagyítunk, akkor egy kis tartományban elég jól közelíthető lesz egyenessel. (Ez az ilyen, kissé ráolvasás jellegű érvelésen túl azért ez matematikailag is alátámasztható.)

Összességében véve arról van szó, hogy ha nincs kellő adatunk a mintából, akkor ezt az információhiányt valamilyen feltevéssel kell kipótolnunk, és erre a fenti okok miatt nagyon csábító a linearitás. Természetesen vannak alternatívái, de ezek között választani csak akkor fogunk tudni, ha van kellő információink. Azt, hogy milyen függvényt használunk, szokták a függvényforma megválasztásának nevezni.

Az egyik ilyen alternatíva a függvényformára, ha a magyarázó változó mellett annak négyzetet is felhasználjuk; ezzel egy egyenes helyett egy parabolát illesztünk:

```
ggplot(RawData, aes(x = nurses, y = cumexcessperpop)) + geom_point() +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE) +
  labs(x = "Ezer főre jutó nővérek száma [fő/ezer fő]",
       y = "Összesített többlethalozás [fő/1M fő]")
```



Ha például kilaposodó hatást feltételezünk, akkor ennek leírására ez alkalmas lehet. Ez ráadásul az érvényességi tartományok kérdésére is jó példa: a parabola ugyan kilaposodik, de egy ponton túl vissza is fordul, tehát

fontos kérdés, hogy mely tartományban kell használnunk a modellt. De ugyanez a lineárisra is igaz: ahogy az előző megjegyzés is mutatta, lehet, hogy egy tartományban még jó közelítés, de máshol már nem (ezért kell az extrapolációval óvatosnak lenni!).

A fenti függvény a potenciálisan jobb illeszkedésért cserében már nem ad olyan kézenfekvő és egyszerű értelmezést mint a lineáris modell (de továbbra is könnyen extrapolálható). Még egy dolog fontos: ez már két paraméter becslését igényli, így statisztikailag nehezebb dió, illusztrálva azt a korábbi megjegyzést, hogy ahoz, hogy ilyen kérdéseket, tehát a nemlineáritás ügyét értelmesen vizsgálni tudjuk, nagyobb mintára lesz szükség.

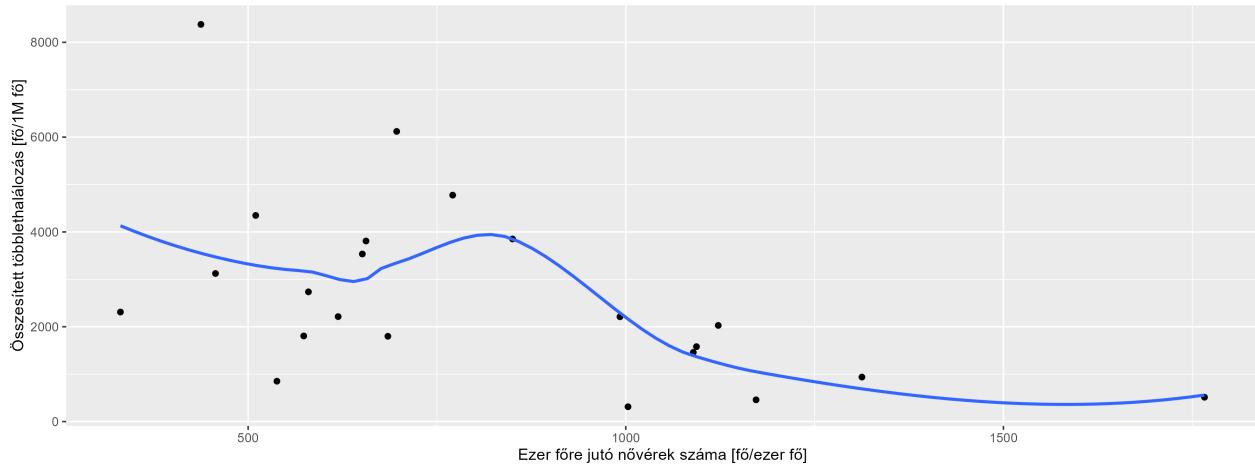
A fenti ábra nagyon egyértelműen mutatja, hogy túl nagy kilaposodó hatás ebben a halálozás–nővér összefüggésben nincsen: a görbe szemmel alig láthatóan tér el az egyenestől. Ennek vizsgálatára statisztikai tesztet is lehetne konstruálni. Ez egy példa a *modelldiagnosztikának* nevezett nagyon fontos lépére: ennek során vizsgáljuk, hogy a modellünk feltevései vajon teljesülnek-e abban a konkrét esetben, amit elemzünk.

Ezzel pedig már kezd látszani egy lehetséges munkamódunk: próbálkozzunk különböző lehetséges függvényformákkal, és válasszuk ki, hogy melyik a legjobb a mi konkrét mintánkra! Ez első hallásra nagyon csábítóan hangzik, csak egy gond van: a függvényformák próbáltatása, pláne ha végiggondolatlanul történik, a túlilleszkedés nevű jelenséghöz fog vezetni, ami rövidesen sokat fogunk beszélni. Annyi megelőlegezhető, hogy csak néhány, nem túl nagy számú, előre eldöntött függvényforma kipróbálásának van értelme, annak is inkább csak akkor, ha van kellően nagy mintánk.

Ha interakciót is feltételezni akarunk, akkor hatványozottan jelentkezik ez a probléma, hiszen abból aztán nagyon sok potenciális van: bármelyik változó lehet interakcióban bármelyikkal. A gyakorlatban az mondható, hogy – ha csak nincs extrém nagy méretű mintánk – legfeljebb nagyon kis számú, előzetesen, tehát nem az adatok által sugallt módon, hanem tárgyterületi ismeret alapján feltételezett interakció modellbe rakásának van értelme.

Zárásként megjegyzem, hogy az összes fent tárgyalt modell az úgynevet paraméteres modellek körébe tartozik. Ezek lényege, hogy a függvény formáját (egyenes, parabola stb.) mi határozzuk meg, előre, azt úgymond rákényszerítjük az adatokra. Az adatokból magukból pusztán néhány számot (az egyenes meredekségét, a parabola két együtthatóját stb.) becsüljük, de nem a függvény alakját. Épp innen jön a „paraméteres“ elnevezés: ekkor csak egy vagy több számszerű paramétert becsülünk az adatokból. Ezek a modellek általában jól becsülhetők statisztikailag, kisebb mintán is, valamilyen szintű tárgyterületi interpretációt kis szerencsével lehet adni a paramétereiknek, és lehetővé teszik az extrapolációt. Csak épp ott van az a hátrányuk, hogy a függvényformát nem az adatok mondták meg, hanem mi – de mi van, ha rosszat választunk? Elvégire egy hullámszerűen fel-le ingadozó pontfelőre is rá lehet húzni egy egyenest... (csak sok értelme nem lesz). A modelldiagnosztika enyhít ezen a problémán, hiszen kimutathatóvá teszi, hogy valamit rosszul csináltunk, és segíthet egy jobb megoldás megtalálásában is, azaz valamilyen értelemben mégis csak alakítja az adatok alapján a függvényformát, de mint volt róla szó, ennek is megvannak a hátrányai. Létezik azonban egy radikálisan más megközelítés, a nemparaméteres modellek: ezeknél egyáltalán nem kell semmilyen függvényformát feltételeznünk! Ami persze jó hír, mert ha nem kell függvényformát feltételeznünk, akkor nem fenyeget, hogy rosszat feltételezünk... Meglepő lehet, hogy ilyen létezik, és lehet regressziót csinálni így is, de a probléma megoldható. Lényegében arról van szó, hogy egyszerűen követjük az adatokat:

```
ggplot(RawData, aes(x = nurses, y = cumexcessperpop)) + geom_point() +
  geom_smooth(method = stats::loess, formula = y ~ x, se = FALSE) +
  labs(x = "Ezer főre jutó nővérek száma [fő/ezer fő]",
       y = "Összesített többlethalálozás [fő/1M fő]")
```



Ebben a kontextusban ezt szokás simításnak is nevezni, ennek is van irodalma. E modellekknél tehát a rossz függvényforma miatt nem kell aggódunk, ami hatalmas fegyvertény, de cserében rosszabbul becsülhetőek, olyan értelemben, hogy nagyobb mintát igényelnek, paraméter híján nincs egyszerű szám, amihez jó esetben még tárgyterületi értelem is tartozik (maximum kirajzolhatjuk a görbét), és extrapolálni sem lehet, vagy csak trükkökkel. A továbbiakban ilyen modellekkel nem foglalkozunk most.

Több változó modellezése

Látszólag tehát meg is vagyunk: kimutattuk mind a nővérek számának, mind a GDP-nek a hatását, immár korrekt módon – legalábbis, ha az említett feltételek teljesülnek, például más változó nem játszik szerepet, ezek hatása tényleg lineáris. Kezdjük az első megjegyzéssel: ennél a kettőnél ugyanis jóval több – potenciálisan – a kimenetet befolyásoló változónk van, és ezt már szépen ábrázolni nem fogjuk tudni. De sebaj, ha elszakadunk az ábrázolástól, és attérünk arra az értelmezésre, hogy egy lineáris kapcsolatot feltételezzük, és annak a paramétereit (amiknek mind „adott paraméter egységnyi növelésének a hatása a többöt változatlanul tartva“ értelmük van) szeretnénk megbecsülni, akkor ez nem probléma. Márpedig ez az eljárás könnyedén kiterjeszthető tetszőleges számú változóra! (A korábban említett lehetőségek a nemlineáris kapcsolatok vizsgálatára szintén átvihetők többváltozós esetre, de ezzel most nem fogunk foglalkozni, részint mert nem jelent újdonságot – a korábban látott módszerek alkalmazhatóak minden egyes változóra – részint mert mindenjárt látni fogjuk, hogy ennél most nagyobb problémáink lesznek...)

Úgyhogy ezt használva bepakoljuk az összes változót egy nagy modellbe, és abból kiolvassuk az eredményt, azt is, hogy mi az ami befolyásolta a halálozást (és mennyire), és azt is, hogy mi az, ami nem:

```
fit2 <- lm(cumexcessperpop ~ popdensity + overcrowding + urbanization + gdp + popold +
            obese + smoke + alcohol + asthma + chrt_angpec + diag + hblpr +
            healthexpenditure + nurses, data = RawData)

knitr::kable(data.frame(`Becsült hatás` = signif(coef(fit2), 3),
                       `95% CI` = apply(signif(confint(fit2), 3), 1, paste,
                                         collapse = " -- "),
                       p = Hmisc::format.pval(summary(fit2)$coefficients[,4], digits = 3,
                                         eps = 0.001),
                       check.names = FALSE, row.names = c("Tengelymetszet",
                                                         names(coef(fit2))[-1])))
```

	Becsült hatás	95% CI	p
Tengelymetszet	3730.0000	-11800 – 19300	0.587
popdensity	1.5000	-5.85 – 8.85	0.645
overcrowding	56.0000	-46.9 – 159	0.239
urbanization	14.9000	-83.8 – 114	0.732

	Becsült hatás	95% CI	p
gdp	-0.0213	-0.139 – 0.0967	0.683
popold	-102.0000	-640 – 435	0.666
obese	-195.0000	-595 – 206	0.288
smoke	24.0000	-176 – 224	0.784
alcohol	55.6000	-250 – 361	0.680
asthma	-496.0000	-1840 – 852	0.413
chrt_angpec	-27.6000	-809 – 754	0.936
diab	-105.0000	-921 – 711	0.770
hblpr	167.0000	-209 – 543	0.329
healthexpenditure	-0.2830	-1.57 – 1.01	0.621
nurses	2.9800	-1.55 – 7.51	0.164

Akkor most végeztünk? Sajnos a helyzet nem ilyen egyszerű.

A multikollinearitás problémája Az első probléma abból fakad, hogy ezek a magyarázó változók – ahogy azt a GDP és a nővér-szám példáján láttuk is – egymással is összefüggnek. Ez probléma? Ha empirikusan szeretnénk meghatározni az egyes változók szerepét, akkor igen: ahogy arról volt szó, a regresszió azt próbálja megbecsülni, hogy mi a helyzet akkor, ha az egyik változó értéke rögzített, és csak a másik változik. Igen ám, de ha a magyarázó változók összefüggnek egymással, az épp azt jelenti, hogy ha az egyiket lerögzítjük, az elég jól meghatározza a másikat is, azaz kicsi lesz abban a szóródás. Márpedig kicsi szóródásból nehéz lesz megbecsülni a hatást: mikor mondanánk meg szívesebben empirikus alapon, hogy +1 fok hőmérséklet hogyan hat a gázfogyasztásra, ha ezer darab -10 és +30 fok közötti napról van információink, vagy ha ezer darab 20 és 21 fok közötti napról? Ezt hívjuk a multikollinearitás problémájának. Mint ebből is látható, ez a jelenség egyetlen változóra vonatkoztatva is ugyanúgy értelmezhető: amiben kicsi a szóródás, annak nehezebb megbecsülni a hatását. Extrém esetben, ha egy változónak ugyanaz az értéke az egész mintában, akkor annak a hatását nem tudjuk megbecsülni – hiszen egyáltalán nem lesz benne szóródás. (Például ezen adatbázis alapján mondjuk meg, hogy az északi féltekén történő elhelyezkedés hogyan hat a járvány terhére – érthető, hogy ezt nem fogjuk tudni megmondani!) De nem kell ilyen erőltetett példát hozni: ha mindenki pontosan ugyanúgy hozott egy intézkedést, akkor ennek hatása empirikusan nem vizsgálható. A regresszió tehát mindig csak valamelyen szóródásból tud becsülni, ezt sokszor fontos végiggondolni, hogy milyen szóródás áll rendelkezésre egy adott paraméter becsléséhez.

Térjünk vissza a két, összefüggő változó kérdésére. Hangsúlyozom, hogy ilyen esetben a két változó hatása *együtt* továbbra is jól megbecsülhető, csak *külön-külön* nem, magyarázni: nehéz *elkülöníteni* a hatásukat. De ez talán intuitíve is érezhető: nehéz megmondani, hogy *külön* mi a hatása a demenciának és a 60 év feletti életkornak, hiszen kevés alanyunk lesz akinek 60 év alatt demenciája van, így amikor a nem demens helyett demens alanyokat néziünk, akkor *szinte* automatikusan 60 év felettesek is lesznek.

Ez nem egy „kijavítható“ probléma abban az értelemben, hogy nem az általunk megalkotott modell, hanem a felhasznált adatok jellemzője – az meg olyan, amilyen, az nem egy általunk befolyásolható kérdés, hogy a GDP és a nővérek száma összefügg.

Minél több egymással összefüggő változót rakunk bele a modellbe, annál jobban fogja az egyes változókat a többi magyarázni, így annál rosszabb lesz ilyen szempontból a helyzet. Ami biztosan segít, az a nagyobb mintanagyság: ettől ugyan továbbra is nehezebben lesznek szétválaszthatók a hatások, azaz bizonytalannabbak lesznek a becslések ahoz képest, mintha a változók nem függenék össze, de a nagyobb mintanagyság ezt ellensúlyozza, azáltal, hogy *általában* csökkenti a becslések bizonytalanságát.

Így már is jobban érthető, hogy miért lett inszignifikáns az összes változónk!

Valójában azonban ennél rosszabb a helyzet. Ha valami véletlen folytán lett volna szignifikáns változó, akkor sem biztos, hogy túlságosan hihető lenne az eredmény a jelen esetben. Ennek okát fogjuk megnézni a következő pontban.

A túlilleszkedés problémája A másik gond, ami akkor jelentkezik, ha a mintamérethez képest túl sok a változónk, a túlilleszkedés. Amikor mi egy regressziós modellt megbecslünk, azt mindig egy minta alapján végezzük, és az a célunk, hogy a mintát a lehető leg pontosabban leírjuk. (Ha megnéztük volna a becslés konkrét módszerét, akkor láttuk volna, hogy ez szó szerint is igaz: a legnepszerűbb becslési eljárás épp azt adja vissza becsült értékként, ami mellett a modellünk a legjobban leírja a mintát). Igen ám, de közben persze valójában nem magát a mintát akarjuk jól leírni (ha ennyi lenne a célunk, akkor nincs is szükség regresszióra, ott a minta, az teljesen pontosan leírja a mintát, és kész), hanem következtetni a minta alapján a háttérben lévő valóságra, statisztikus nyelven úgy szokták mondani: a sokaságra. Ehhez természetesen fel kell használnunk a mintabeli információt, és ha olyan modellt alkalmazunk, ami nem tud elég információt felhasználni – nevezük ezt alulilleszkedésnek – akkor természetesen nem várható, hogy a sokaságról helyesen tudunk nyilatkozni. Ez eddig elég logikus, de itt jön a csavar, mert az előbbiből úgy tűnhet, hogy minél több információt felhasználni képes, azaz minél komplexebb modellt vetünk be, annál jobb lesz a helyzet. Csakhogy meglepő módon ennek egy ponton túl ennek az ellenkezője lesz az igaz: egyre komplexebb modellekkel használva nem egyszerűen nem javul a sokaság leírása, hanem elkezd kifejezetten romlani! Ezt a jelenséget hívjuk túlilleszkedésnek.

A probléma úgy fogalmazható meg, hogy a túl komplex, azaz túl sok információt felhasználni képes modell egy ponton túl már az adatbázisban lévő véletlen zajokat is képes lesz leírni, márpédig ha képes lesz rá, akkor meg is fogja tenni, hiszen a becslési módszernek ugyebár az a célja, hogy a mintát minél jobban leírja – azt pedig a véletlen zajok leírása is javítja. Csakhogy a sokaság leírását ez nem hogy nem javítja, de kimondottan lerontja, hiszen abban nincsenek benne ezek a véletlen zajok.

Ezt szemléletesen mutatja az alábbi animáció, ahol egy magyarázó változon van, a piros görbe mutatja a valódi összefüggését az eredményváltozóval. Ismét a szimuláció eszközével élünk: ebből az összefüggésből generáljuk a 20 darab fekete pontot úgy, hogy azok véletlenszerűen szóródjanak a valódi görbe körül, majd a pontokra egyre komplexebb és komplexebb, a mintából egyre több információt felhasználni képes regressziós modellekkel illesztünk; ezeket a kék görbék mutatják:

```
set.seed(1)
SimData <- data.frame(x = runif(20, 0, 10))
SimData$y <- 1 + 2*sin(SimData$x) + 0.05*SimData$x^2 + rnorm(20, 0, 1)
for(p in 0:18)
  print(ggplot(SimData, aes(x = x, y = y)) + geom_point() +
    geom_function(fun = ~ 1 + 2*sin(.x) + 0.05*x^2, color = "red") +
    geom_smooth(method = "lm", formula = if(p==0) y ~ 1 else y ~ poly(x, p),
                se = FALSE, n = 500) +
    labs(title = if(p<=3) paste0("Alulilleszkedés (p = ", p, ")") else
         if(p<8) paste0("Nagyjából jó illeszkedés (p = ", p, ")") else
         paste0("Túlilleszkedés (p = ", p, ")")) +
    coord_cartesian(ylim = c(-0.5, 8)))
```

(Érdemes megnézni, hogy több modellre is a „nagyjából jó“ kifejezést használtam: bár a sokaság ismeretében meg lehetne mondani, hogy közülük konkrétan melyik a jó, de az ilyen kis mintanagyság nem teszi lehetővé, hogy ezeket a minta alapján elkülönítsük.)

Azt hiszem a fenti illusztráció nem sok kommentárt igényel a tekintetben, hogy mit jelent a túlilleszkedés, az, hogy a túl komplex modell már a zajokat is megragadja, így bár a mintát egyre jobban leírja, a sokaságot meg egyre kevésbé.

A jelenséget úgy is meg szokták fogalmazni, hogy az általánosítóképesség romlik: ez a kifejezés jelenti azt, hogy a modellnek a mintában látott konkrét dolgokból általánosítania kell a sokaság általános viselkedésére. Ha azonban már az adott konkrét minta zajait is elkezdi leírni, akkor épp ez sérül.

A túlilleszkedés abban jelenik (jelenne) meg, hogy bár a modell az *adott, konkrét* mintát jól leírja, ha vennénk egy másik, új mintát, amit a modell korábban nem látott (értsd: amit a megbecsüléséhez nem használtunk fel), akkor azon már rosszul teljesítene. El lehet képzelni, hogy ha kisorsoltunk volna, szintén teljesen véletlenszerűen, *másik* 20 pontot a piros görbe körül, akkor a nagyjából jól illeszkedő modellek azokat is szépen leírták volna, de a túlilleszkedőek teljesen rosszul – hiába is írták le a fenti 20 pontot jobban, és hiába is ugyanazon görbe körül sorsoltuk. Ez persze ilyen formában csak gondolatkísérlet, hiszen a valóságban csak egyetlen mintánk van, de ezen a nyomon el lehet indulni úgy, hogy nagyon hasznos eredményekhez jussunk, de erről picit később.

Bár a fenti animáció nem a magyarázó változók számának változtatásával oldotta meg a modell komplexitásának változtatását (hiszen az fixen 1 volt), de a helyzet az, hogy a több magyarázó változó is pontosan ugyanezt jelenti! A több magyarázó változó ugyanis több megbecsülendő paramétert jelent (ahogy a fenti táblázat is mutatja, mindegyiknek van egy saját hatása), a több szabad paraméter pedig komplexebb modellt jelent. Ha túl sok magyarázó változót pakolunk bele egy regressziós modellbe, akkor azt fogjuk *hinni*, hogy gyönyörű szép az illeszkedése (és ezt fogják a különböző számszerű jellemzők is, látszólag teljesen objektíven, mutatni), de valójában az történik, amit a fenti animációt látunk.

Ugyanígy túlilleszkedéshez vezet bármi más is, ami a becsülendő paraméterek számát növeli, hiszen minden ilyen komplexebbé, plasztikusabbá, több információt felhasználni képessé teszi a modellt. Tehát például a több paramétert használó függvényforma is! Hiszen a parabola is két paramétert igényel, míg a lineáris csak egyet. Ha valaki harmadfokú függvényt akarna illeszteni, az – bár még jobban tudná követni az adatokat – természetesen még érzékenyebb lenne a túlilleszkedésre is, hiszen az már három paramétert igényel. Most már elárulhatom így utólag, hogy igazából pont ez van az animáción: a pontokra egyre magasabb fokú függvényeket illesztünk. Összességében tehát nem az az igazán fontos, hogy azért van két paraméterünk, mert két változót használunk lineárisan, vagy azért mert egyet, de azt másodfokúként modellbe helyezve, hanem a mintából becsülendő szabad paraméterek száma. Szokták néha a modell komplexitása helyett azt a kifejezést is használni, hogy a modell kapacitása, ami ugyanúgy a fentieket fejezi ki: mennyi információt tud magába fogadni, leírni a modell, jelen esetben, hogy mennyi szabad paramétere van, amit mintából becsülünk meg.

A torzítás-variancia dilemma

Tegyünk egy rövid, de tanulságos kitérőt ezen a ponton. Ennek kiindulópontja legyen egy első látásra talán meglepő állítás: a fenti túlilleszkedéses példában, bár az utolsó görbék már nagyon-nagyon csúnyán néznek ki, de – bármilyen meghökkentő is – *torzítás* nem lép fel! (Torzítás abban az értelemben, ahogy korábban pontosan is definiáltuk.) Tehát ingadoznak ugyan a becsült értékek, de átlagosan jók.

De akkor mégis mi történt, hogy ilyen csúnya lett az eredmény? Az, hogy bár *átlagosan* jók vagyunk, de elköpesztően nagy lett a mintavételi ingadozás! Így egy konkrét eredmény sajnos nagyon messze is tud lenni a valóságtól.

Egy rövid megjegyzés: a mintavételi ingadozás háttérében igazából egyszerűen a multikollinearitás van, semmi egyéb varázslat nem történik. Hipotetikusan, ha nem multikollinearitás változókból (tehát olyanokból, amik egyáltalán nem függnek össze a többi változóval) raknánk be újabbakat és újabbakat, akkor nem nőne a mintavételi ingadozás.

Sok változó hozzáadása jót tesz torzítás szempontjából (kevésbé valószínű, hogy kihagyunk lényeges változót), hasonlóképp a sok paraméterű függvényforma is jót tesz torzítás szempontjából (kevésbé valószínű, hogy a valóságot nem tudja leírni a függvényformánk), viszont a jelek szerint minden rosszat tesz a mintavételi ingadozás szempontjából. Fordítva megfogalmazva, ha csökkentjük a változók számát, ha egyszerűsítjük a függvényformát, akkor javítjuk a mintavételi ingadozást, de torzítást okozhatunk.

És az izgalmas az, hogy ez nem véletlenül volt így, csak pont ebben az esetben, hanem egy *általában is* fennálló jelenség: e két tulajdonság csak egymás rovására javítható. A mintanagyság növelése az egyetlen, ami a torzítás befolyásolása nélkül csökkenti a mintavételi ingadozást, de rögzített mintanagyság mellett csak abban dönthetünk, hogy melyik ujjunkba harapjunk: ha javítjuk az egyik szempontot, akkor *automatikusan* rontjuk a másikat. E szomorú, de másrészt rendkívül érdekes jelenség neve: torzítás-variancia dilemma (angolosabban torzítás-variancia trade-off, azaz átváltás, utalva arra, hogy egyiket csak lecserélhetjük a másikra). A variancia szó itt azt jelenti, hogy mekkora a mintavételi ingadozás mértéke, mérve azzal, hogy a kapott becslések mennyire ingadoznak a saját átlaguk körül (ami vagy egybeesik a valódi értékkel vagy nem – ugye ez a torzítatlanság kérdése).

Nézzük meg mindezeket egy szimulált példán! Az előbbi példát folytatjuk, ahol egyre magasabb fokú függvényeket illesztünk. Csak épp most nem egyszer tesszük meg, hanem nagyon sokszor, és így ki tudjuk számolni a torzítást és a varianciát is, minden komplexitás mellett. Ezt kapjuk:

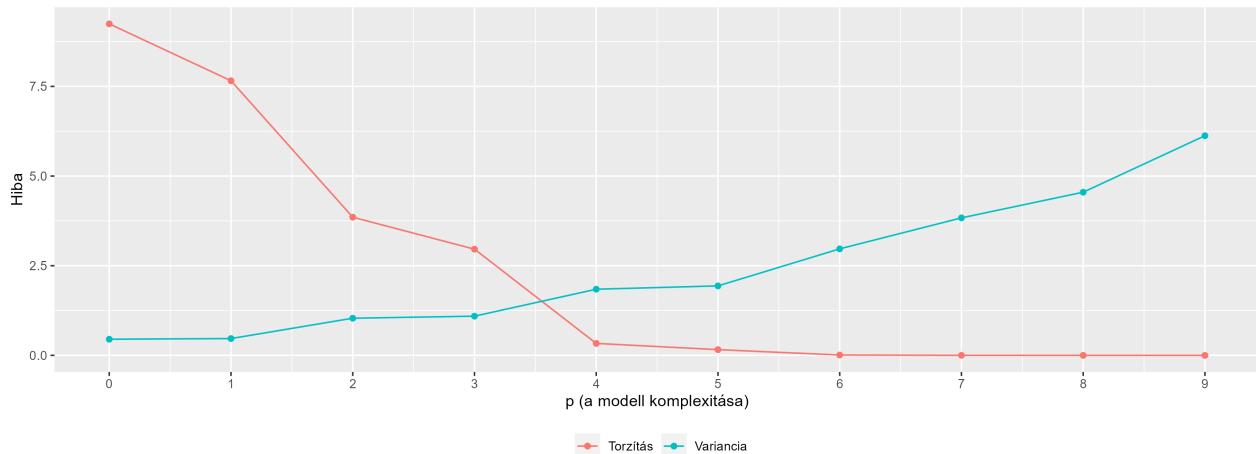
```
if(!file.exists("BiasVarianceSimRes.rds")) {  
  set.seed(1)  
  cl <- parallel::makeCluster(parallel::detectCores()-1)  
  
  x <- runif(20, 0, 10)  
  parallel::clusterExport(cl, "x")  
  
  SimRes <- parallel::parSapply(cl, 0:9, function(p) replicate(1e6, {  
    y <- 1 + 2*sin(x) + 0.05*x^2 + rnorm(20, 0, 3)  
    if(p==0) mean(y) else lm.fit(cbind(1, poly(x, p, raw = TRUE)),  
                                 y)$coefficients%*%c(1, poly(5, p, raw = TRUE))  
  }))  
  
  parallel::stopCluster(cl)  
  
  SimRes <- data.table(p = 0:9,  
                        `Torzítás` = apply(SimRes, 2,  
                                         function(x) (mean(x)-(1 + 2*sin(5) + 0.05*5^2))^2),  
                        `Variancia` = apply(SimRes, 2, function(x) mean((x-mean(x))^2)))
```

```

  saveRDS(SimRes, "BiasVarianceSimRes.rds")
} else SimRes <- readRDS("BiasVarianceSimRes.rds")

ggplot(melt(SimRes, id.vars = "p"),
       aes(x = p, y = value, group = variable, color = variable)) + geom_point() +
  geom_line() + scale_x_continuous(breaks = 0:9) +
  labs(x = "p (a modell komplexitása)", y = "Hiba", color = "") +
  theme(legend.position = "bottom")

```



Szépen látszik, hogy valóban trade-off áll fenn: ahogy az egyik hiba javul, a másik romlik. Ennél ráadásul nemileg több is leolvasható: a dolog a a komplexitáson múlik, a nagyobb komplexitás a torzítást csökkenti, de a varianciát növeli, kisebb komplexitás mellett a variancia kisebb, de a torzítás nagyobb.

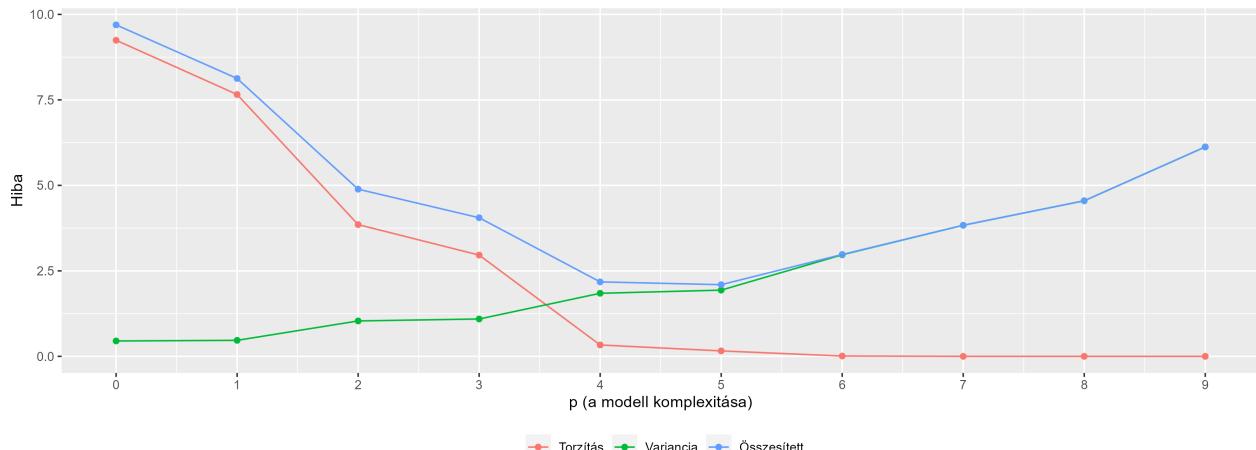
Egy lépéssel tovább is mehetünk: definiáljuk *általában* azt, hogy mit értünk hiba alatt! Tehát ne egy-egy szempontról beszélünk, hanem összességében, a teljes hibáról. Egy elég kézenfekvő választás: hiba az, hogy a becsült érték milyen messze van a valódítól. Az érdekes az, hogy ez visszavezethető a fenti két tényezőre, ami végülis azért annyira nem meglepő, ha arra gondolunk, hogy a variancia az, hogy a becsült milyen messze van az átlagától, a torzítás meg az, hogy ez az átlag milyen messze van a valodi értéktől. A pontos számítás ennyire azért nem nyilvánvaló, nem lehet egyszerűen összeadni a két tényezőt, de aránylag könnyen elvégezhető, csak a „milyen messze van“ mérésére kell elfogadnunk egy – kézenfekvő, és máshol is alkalmazott – konvenciót. Ezt használva a következőt kapjuk:

```

SimRes$`Összesített` <- SimRes$`Torzítás` + SimRes$`Variancia`

ggplot(melt(SimRes, id.vars = "p"),
       aes(x = p, y = value, group = variable, color = variable)) + geom_point() +
  geom_line() + scale_x_continuous(breaks = 0:9) +
  labs(x = "p (a modell komplexitása)", y = "Hiba", color = "") +
  theme(legend.position = "bottom")

```



Így már határozottan érdekes a dolog! Úgy tűnik, hogy az egyik hibaforrás úgy romlik miközben a másik javul, hogy a kettő között van egy optimum! Ebben a példában olyan 4-5 körüli p -nél. Természetesen itt sem nulla a hiba, valamennyi része a hibának megszüntethetetlen, de ez a legjobb választás, amivel élhetünk.

De hogy találjuk meg ezt az optimumot? Ugyanis vigyázat: a fenti ábra egy *szimuláció*, amit úgy kaptunk, hogy *ismertük* a sokaságot! Ezzel csak szemléltetni lehetett a jelenséget, de a valódi helyzetben csak egyetlen mintánk van, nem tudjuk mi a sokaság, így mi magunk sem tudhatjuk, hogy egy adott modell hol van a fenti skálán. (A dolog ott jelent meg, hogy amikor a fenti ábrákhöz számoltam a hibák értékeit, akkor ahhoz a számításhoz *felhasználtam*, hogy mi a *valódi* összefüggés – amit persze a valóságban nem fogunk ismerni.) El kell kezdenünk tehát boncolgatni kérdést, hogy ilyen körülmények között, a valódi helyzetben hogyan tudjuk megválasztani helyesen a modell-komplexitást.

A változószelekció kérdésköre

Általánosságban véve a torzítás a nehezebben kézben tartható hibaforrás, és a variancia írható le jobban: bár ez is függ a multikollinearitástól (azaz a változók egymás közötti összefüggéseitől), a szóródásoktól, de azért összességében jól jellemzi egyszerűen a felhasznált magyarázó változók számára. (Pontosabban szólva, ha nem csak egyszerű függvényformákat használnánk: a becsült paraméterek száma.) Azt is láttuk, hogy a torzítás-variancia dilemmában egyedül a mintanagyság növelése jelent univerzális javítást. E két tényt összerakva nem meglepő, hogy a gondolkodás merre indult el: mivel a varianciát jól jellemzi a magyarázó változók száma, ezért azt, mint jobban leírható hibaforrást limitáljuk – azaz korlátozzuk a magyarázó változók számát. De az előbbiek miatt nem akárhogy: a mintanagyság függvényében! Ha nagyobb a mintánk több változót (becsült paramétert) is merhetünk használni, ha kisebb, akkor sajnos csak kevesebb vizsgálható értelmesen. Az irodalom a mintanagyság és a modellezhető változók számának összefüggéséről nagyon sok eredményre jutott, egészen számszerű formában is; egy gyakori mondás például, hogy olyan regresszióknál, mint amit most mi is használunk, legyen a felhasznált változók száma legfeljebb a mintaméret osztva 15-tel. Avagy, fordítva megfogalmazva, legyen legalább 15-ször annyi megfigyelésünk a mintában, mint ahány változót a modellbe akarunk rakni! A pontos szám sok mindenről függ, a multikollinearitás fokától, attól, hogy a valóságban mennyire szoros a kapcsolat a kimenet és a magyarázó változók között, a magyarázó változók eloszlásától, így a helyzet akár ennél rosszabb is lehet, de nagyjából irányiszámunk most tökéletesen megteszeli ez a 15.

A fenti szabály azért fontos, mert azonnal ordítóvá teszi, hogy milyen drámai helyzetben vagyunk. Nekünk a legjobb esetben is 25-30 európai országunk van, és mivel ezzel nem tudunk mit kezdeni, így ez lényegében azt jelenti, hogy 2 változót tudunk mindenkorral vizsgálni!

Ez egy egészen drasztikus probléma – emlékezzünk vissza a felrajzolt diagramra a megvizsgálandó tényezőkről! (És persze még az sem volt teljes, milliónyi további ötletünk lehet!) És mi 2 változót tudunk vizsgálni? Hogy fogunk így egyáltalán *bármit* mondani?! Azonnal elkezd járni az ember agya azon, hogy mit tehetünk a megoldás érdekében. Az első ötlet nagyon kézenfekvő: valahogy válogassuk ki a változók egy részét, és csak azokat rakjuk be a modellbe. Ez nagyon csábító lehetőség, hiszen így csökkentjük a köztük lévő, lehetséges

összefüggéseket, így javítjuk a multikollinearitást, és egyidejűleg a túlilleszkedés ellen is védekezünk. Ez nem egyszerűen „csábító“, de egy ilyen helyzetben szinte megkerülhetetlennek is tűnik: valahogy be kell a „mintanagyság osztva 15-tel“ küszöb alá pofozni a magyarázó változók számát.

A gond az, hogy ez az egész terület egy hatalmas aknamező, ahol nagyon könnyű hibás megoldást választani (és sok közülük sajnos még a szakirodalomban is újra meg újra felbukkan).

A téTELmondat nagyon egyszerű: csak olyan módszert használhatunk a változók kiválogatására, ami nem használja fel a kimenet értékeit.

Tehát néhány jó módszer:

- Pusztán a magyarázó változók struktúráját, azaz egymás közti kapcsolataikat vizsgálva elhagyni olyanokat, amik redundánsak (azaz a bennük lévő információk más magyarázó változókban is járászt megvannak).
- Pusztán a magyarázó változók struktúráját, azaz egymás közti kapcsolataikat vizsgálva új magyarázó változókat képezni összevonással (erre vannak statisztikai módszerek).
- Nem statisztikai, hanem szakmai alapon képezve összevonásokat. Egy tipikus példa, hogy az „alan cukorbeteg“, az „alanynak magas vérnyomása van“, az „alanynak perifériás érbetegsége van“ változók helyett berakunk egy darab az „alanynak krónikus betegsége van“ változót. Igen, jobb lenne megmondani, hogy *külön-külön* mi a hatása a cukorbetegségnek, a magas vérnyomásnak és a perifériás érbetegségnek, de ha egyszerűen nincs elég nagy mintánk, akkor el kell fogadni, hogy erre nem leszünk képesek.

A végére hagytam a legjobb jó módszert: szakmai alapon, tárgyterületi ismereteket használva megpróbálni szűrni a felhasznált változók körét. Igen, ez nem empirikus – miközben pont ez lenne a célunk! – de néha nem tudunk jobbat tenni. (A „néha nem tudunk jobbat tenni“ gondolatra még visszatérök később.)

Nézzünk most néhány rossz módszert:

- Megnézni, hogy melyik potenciális magyarázó függ össze *önmagában* (kétváltozósan) az eredményváltozóval, és csak azokat belerakni a többváltozós modellbe. Ez a módszer teljesen hibás. Az még csak hagyján, hogy az előzetes szűrés is egy statisztikai teszttel történik, ami nem tökéletes (a valóságban az eredményváltozóval nem összefüggőre is mondhatja, hogy összefüggő, vagy, ami a mi mostani szempontunkból még nagyobb probléma, pont fordítva), de az igazi gond, hogy nem veszi figyelembe a többváltozós struktúrát: mi van, ha egy változó csak az *után* válik lényegessé, ha más változók már bent vannak a modellben?
- Megcsinálni a többváltozós modellt, majd elhagyni az inszignifikáns változókat. Ez a módszer szintén hibás. A helyzet ugyanaz pepitában: egyszerű az, hogy „inszignifikáns“, nem ugyanaz, mint hogy biztosan nulla a hatása: az inszignifikanciát is egy statisztikai teszttel ítéljük meg. Ha ez téved, márpedig ez előfordulhat, és lényeges változót hagyunk el, akkor épp a kihagyott változó okozta torzítást fogjuk előhozni! Épp emiatt a dolog ráadásul filozófiaileg is érthetetlen, hiszen attól még, mert egy változó inszignifikáns, nagyon is segíthet a többi paraméterét helyes értékre beállítani.
- Pláne horrorisztikusak azok a módszerek, amelyek oda-vissza veszik be és dobálják ki a változókat, keresve a valamilyen mutató szerinti legjobbat („stepwise regresszió“). Ez lényegében felturbózza az előbbi módszerek hibáit: szinte garantáltan túlilleszkedett modellre vezet, aminek praktikusan minden létező paramétere torzított/hibás lesz.
- Amire kevesebben gondolnak, pedig ugyanaz a helyzet, amikor a kutató az előbbi „kézzel“ hozza létre: össze-vissza keresgél, kidobva és bevonva változókat, próbálkozva különböző modellekkel, hogy mi a legjobb.

Ha valaki nem hiszi el a fentieket nekem (nagyon jól teszi!), akkor statisztikai programmyelven maga is leszsimulálhatja, és saját kezűleg kipróbálhatja; ilyen módon, tehát szimulációval a fenti jellegű állítások könnyen leellenőrizhetőek.

Érdemes kiemelni és mélyebben végiggondolni azt a megjegyzést, hogy az utóbbi módszerek túlilleszkedéshez vezetnek. Ez egyfelől végülis nem olyan meglepő: e módszerek mind a konkrét adatbázishoz „csiszolják“ az eredményt, innen nézve érthető, hogy ahhoz vezethetnek, hogy a modell túl jól fog illeszkedni ahhoz. Mindegyik ilyen változó kihagyás vagy hozzávétel a komplexitást növeli, hiszen megtehetünk volna, hogy nem hagyjuk el

vagy vesszük hozzá a változót, vagy más változóval tesszük azt, így már a változó modellben szerepelésének a ténye is egy adatokon alapuló, konkrét mintához adaptálódó döntés. Azaz a végső modellünkben *benne lesz* az is, hogy milyen döntést hoztunk, a valóságban ez is növelni fogja a komplexitását, noha pusztán a modellt nézve ez nem fog látszódni. De másrészt mégis csak valami nagyon paradox helyzetre jutottunk, hiszen miért is kezdtünk egyáltalán bele ebben az egész változó válogatás dologba? Azért, mert hallottunk róla, hogy a túl sok változó túlilleszkedéshez vezethet, és tenni akartunk ez ellen valamit. Magyarán: nekiálltunk lépni a túlilleszkedés ellen, majd kiderül, hogy amit teszünk, az pont a túlilleszkedést rontja!

Ez egy nagyon nagy csapdája és nehézsége ennek az egész témakörnek: azt szoktuk mondani, hogy a regressziós modellépítés egy „iteratív folyamat” (azaz a modellt ellenőrizni kell, ha nem jó, akkor visszamenni, módosítani, újra ellenőrizni, és így tovább), ami persze igaz is, csak közben a *túl sok* iteráció ugyanúgy hiba forrása lehet! A dolognak még mélyebb megértését nyerhetjük, ha a torzítás-variancia dilemmára gondolunk: a modellépítés iterációi lényegében a modell komplexitását növelik (és most a valós komplexitásról beszélek, amiben a modellhez vezető út is benne van!), ami nem feltétlenül baj, sőt, nem csak, hogy nem baj, de egy pontig szükséges is: az sem jó, ha túl kicsi a modell komplexitása. Ebben igazi feladat megtalálni az egyensúlyt.

Még egy, de nagyon fontos kommentárt fűznék a fentiekhez. Gyakran hallani hivatkozást arra, hogy a modelleknek „egyszerűeknek”, „takarékosnak” kell lenniük (néha erre mondják azt, hogy a parszimónia elve). Ami persze igaz is, ha a túlilleszkedés szempontjára gondolunk. De he ezzel indokolják az inszignifikáns változók elhagyását, vagy a stepwise módszerek alkalmazását, az a parszimónia elvének totális félreértséiről tanúskodik: a parszimónia nem egyszerűen az, hogy hány darab változó van benne a végső modellben! A takarékkosság fogalmába ugyanúgy beletartozik az is, hogy hogyan jutottunk el ahoz a modellhez. Ha a végső modellben csak három magyarázó változó van, de háromszáz lépésen keresztül barkácsoltuk, mire kijött, az a legkevésbé sem „takarékos” – csak ez nem látszik a végeredményből! A probléma *pont* az, hogy a közbenső barkácsoló lépések mind-mind a modell (valódi) komplexitását növelik, teszik azt nem „egyszerűvé” és nem „takarékkossá”, csak épp ezzel sehol nem számolunk el, ha a végső modellt úgy prezentáljuk, mintha előből azt becsültük volna meg. Azaz a parszimónia elve teljesen rendben van, csak épp annak az állításnak, hogy e módszerek ezt segítik elő, épp az ellenkezője az igaz: ezek megsértik ezt az elvet, csak ezt megsértést *elduggják* az útban, amíg eljutunk a végső modellig – ami persze még annál is rosszabb, mintha legalább látnánk, hogy mi a valódi helyzet.

Kitérőként megjegyzem, hogy igazából ugyanez a helyzet a függvényforma megválasztásával is, nem csak a változószelekcióval: az, ha valaki a modellbe rak egyöt paraméteres függvényformát, nem tér el lényegesen attól, mintha egy egyparamétereset használna, csak épp előtte öt különböző lehetségeset próbált végig, és pusztán az adatok alapján választotta ki, hogy melyik a legjobb. Pedig az utolsó esetben könnyen lehet, hogy arra fog hivatkozni, hogy milyen szép takarékos a modell, a parszimónia elvének megfelelően...!

A fentiek abba az irányba mutatnak, hogy lehetőleg egyetlen modellt találunk ki, előre, azt becsüljük meg az adatokon, és bármi is jön ki, ne módosítsuk. (Különben túl fogunk illeszkedni.) Ahogy az iterációra vonatkozó megjegyzésemben is utaltam rá, ez azért túl radikális álláspont: ugyanúgy, ahogy egy *kicsit* a modell komplexitását is növeltejük túlilleszkedés nélkül, mondjuk néhány változó hozzáadásával, sőt, az még jót is tehet, *kicsit* azért pozogathatjuk is. (Amint láttuk, ez a kettő igazából ugyanaz: mindenkitő a modell valódi komplexitását növeli.) Újfent utalnék a torzítás-variancia dilemmára, illetve az ott látott grafikonra: a pozogtatás jobbra tolja a modellt, ami nem feltétlenül baj, sőt, direkt jó is, ha még a grafikon bal szélén vagyunk, de nagyon óvatosnak kell lenni, hogy ne füssünk túl az optimum pontján.

Pár próbálkozást tehát tehetünk változó kihagyására vagy hozzávételére, összehasonlíthatjuk az eredeti modellünkkel ezeket, és kiválaszthatjuk, hogy melyik a legjobb. Két dolog fontos, hogy *néhány* próbálkozást tegyük, ne rengeteget, és hogy ezek *prespecifikáltak* legyenek! (Tehát még az adatokkal való bármilyen munka előtt, előre döntsük el, szakmai megfontolások, tárgyterületi ismeretek alapján hogy melyik lesz az a néhány modellünk, amik közül adat-alapon majd választunk.) A túlilleszkedés szempontjából ugyanis az a legrosszabb, ha adatok által sugallt felvetéseket vizsgálunk meg. Ha pedig rengeteg lehetőséget nézünk végig, akkor hiába is prespecifikáltuk őket, ugyanúgy jönni fog a túlilleszkedés problémája.

A kényelmetlen tudomány

A fenti okfejtés összességében véve egy nagyon nyugtalanító képet sugall: úgy tűnik, hogy ha kevés adatunk van, akkor egyszerűen nem tudunk mit tenni. Nem akarok zsákbamacskát árulni, ez valamelyen értelemben tényleg így van. Olyannyira, hogy a helyzetet a neves statisztikus John Wilder Tukey találóan úgy hívta: ez az „uncomfortable science“, a kényelmetlen tudomány.

A talán leghíresebb példa erre a Titius–Bode-szabály. Ez azt állítja, hogy a Naprendszerben a sorrendben n -edik bolygó távolsága a naptól $0,4 + 0,3 \cdot 2^n$ (egy csillagászati egységnek nevezett mértékegységen mérve). Ezt a 18. században vetették fel, és az akkor ismert bolygókra prímán működött. De itt vajon tényleg valamelyen matematikai összefüggés van? Ez borzasztó fontos, mert ha igen, akkor valamilyen csillagászati, mechanikai okot kell keresni amögött, hogy ez így alakult. Vagy egyszerűen csak véletlen egybeesésről van szó? Azaz: lehet szó túlilleszkedésről? Hogyne, simán, elvégre ki tudja, hogy Titius és Bode vajon hány formulát próbált ki, mire ez működött...! De akkor mit tegyünk, hogyan ellenőrizzük ezt le? És itt jön a kényelmetlen tudomány: aligha tudunk venni még egy bolygót mintának, hogy kipróbáljuk azon is működik-e a szabály...!

De mennyire „kényelmetlen tudomány“ helyzet a mi mostani kérdésünk?

Egyrészt megpróbálhatjuk növelni a mintanagyságot. Lejjebb tudunk menni megyei szintre? Esetleg járásira? Ez csábító lehetőség, de két nehézség lesz. Az egyik az adatokhoz kapcsolódik: van adatunk arra, hogy *megyei* szinten hány nővér jut egy betegre? Egyáltalán, értelmezhető ez a mutató megyei szinten? (Pláne járásin!) A másik gond, hogy sok változó van, amiben az országban *belüli* adatokban nem lesz nagy szóródás (például a bevezetett intézkedések akár teljesen egységesek is lehetnek), ami nehezíti a becslésüket, ahogy a multikollinearitásnál láttuk is.

Kísérletezhetünk azzal is, hogy idődimenziót is adunk a vizsgálatnak, tehát nem egyben vizsgáljuk az egész járványt, hanem időszakonként. Ez egyébként önmagában, a mostani problémától teljesen függetlenül is egy fontos felvetés. Hiszen egészen idáig „összeöntöttük“ a járvány adatait, és bár valóban van sok minden, ami teljesen állandó, vagy lényegében állandó (mondjuk az ország korfája), sok minden nem: oltást bevezettek, időben felfutott az oltottak aránya, intézkedéseket meghoztak, vagy épp kivezettek stb. Ez nagyon is eltérhet országok között, és nagyon is lehet, hogy jelentősége van a végeredmény alakításában. Szép szóval élve a *dinamika* is fontos lehet, így egy finomabb vizsgálat – sajnos azonban, legyünk őszinték, komoly módszertani kihívások és adatszerzésre vonatkozó nehézségek árán – megpróbálkozhat ezt is figyelembe venni.

Tehetünk okosan a változók számának csökkentése érdekében is. Már az ottani pont végén is említettem ennek egy lehetőségét, a néhány, prespecifikált modell összehasonlítását. A korszerű statisztika további eszközökkel is szolgál: vannak módszerek a túlilleszkedés fokának megbecslésére, hogy lássuk, bajban vagyunk-e (validáció, például bootstrap), vannak módszerek amelyek úgy becsülnek regressziót, hogy igyekeznek tenni a túlilleszkedés ellen (regularizáció, vagy más szóval penalizáció), vannak más még jobban eltérő elven felépülő becslések (pl. a LASSO), sőt, elárulom, hogy ha ésszel csináljuk, akkor akár még a stepwise szelekciós módszereknek is lehet szerepük (nagy mintán, megfelelő paraméterekkel, elszámolva az előzetes szelekció tényével).

Záró gondolatok

Talán ezek az utolsó gondolatok érzékelgették legjobban, hogy írásom célja nem egy „lezárt válasz“ közvetítése volt. Sőt, egyáltalán semmilyen választ nem adtam a kérdésre, és szándékosan: én most a módszertanra szerettem volna fókusznál, ezen belül is különösen arra, hogy felhívjam a figyelmet a veszélyes (mert csábító) csapdákra.

A tárgyalásunk lezárásához érve talán érdemes egy lépést tenni hátrafelé. Valójában ugyanis minden amiről beszélünk, csak a szakpolitikai intézkedések értékelésének egyik lehetséges útját jelentik, amit eredményszempontú értékelésnek szoktak hívni. Van egy másik lehetséges megközelítés, az eljárás helyességén alapuló értékelés. (Egészségügyi területen erre hozható egy analógia: a Donabedian-modell, a jól ismert „struktúrafolyamat–eredmény“ hívószavaival.) Mint a fentiekből is kiderült, az eredmény-alapú értékelésnek komoly nehézségei vannak jelen esetben, és a kapott válasz múlhat az alkalmazott feltevéseken. Azonban látni kell, hogy ettől a jellegzetességtől az eljárás-helyesség értékelése sem mentes, legfeljebb ott ez máshol jelenik

meg: az egy értékválasztás folyománya lesz, hogy milyen procedúrát tekintünk helyesnek, olyat, ahol a döntéseket transzparens módon, szakmai integritással rendelkező szervetek véleményének figyelembenével, nyilvánosan hozzáérhető adatokra alapozva hozzák meg, vagy olyat, ahol az adatok titokban tartása mellett politikai vezetők belátásán, nem transzparenlesen dolgozó szervezetek véleményén, a választópolgárok rövid távú kívánságainak való megfelelésen alapulnak a szakpolitikai döntések. A kettő közötti preferencia jellemzően nem pusztán egy-egy vezető személyén múlik, hanem történelmi és társadalmi hatások alapján jön létre, adott országban és adott időpontban.

A területtel foglalkozó kutatók attól tartok nem lesznek egyszerű helyzetben hazánkban. Ezt sajnos nem csak statisztikai okokból mondjam: a kérdéskör átpolitizáltsága véleményem szerint rendkívül romboló a valódi tudományos kutatásokra nézve. Félreérzés ne essék, a „politika“ szót nem egy szükségképp negatív töltetű értelemben használom, sőt, ez pont egy olyan téma, ahol fontos és hasznos is a politikával való kölcsönhatás, hiszen az ilyen vizsgálatokból levonhatóak következtetések, amelyek az (egészség)politikát is vezethetik a magyar ellátórendszer, a prevenciórendszer, a népegészségügyi programok javításának irányában. A gond ott kezdődik, amikor megjelenik a prekonceptiozus szemlélet, amikor minden intézkedésnek vagy teljesen, alapjában és menthetetlenül rossznak, vagy teljesen, tökéletesen és aggálytalanul jónak *kell* lennie. (Pláne, ha vannak olyan szereplők, akik rá is tudják ezt kényszeríteni másokra, például mert elhallgattathatnak nekik nem tetsző véleményeket...) A probléma az, hogy az ilyen kutatások sajnos ennek nagyon kitettek, hiszen nehéz elkerülni, hogy a magyarázó változókat az ember összekösse a kormány tevékenységével: mi az, amire nincs érdemi ráhatása a releváns időhorizonton (pl. a korfa), mi az, amire van, de azért nem túl direkt és azonnali módon (pl. elhízottak aránya), és mi az, amire direkt és azonnali módon van (pl. tesztelési stratégia). minden oldali szereplőnek meg kellene értenie, hogy az ország érdekét csak az szolgálja, ha e kérdéseket olyan légkörben lehet megbeszélni, ami nem a pillanatnyi politikai haszon kinyerésének lehetőségét nézi az eredményekből.

A fentiekből minden bizonnal kiderült, hogy miért gondolom, hogy nehéz ezt a lantot kézbe venni. Ezzel azonban senkit nem elriasztani szeretnék, épp ellenkezőleg, remélem azt is meg tudtam mutatni, hogy a nehézségei egyúttal a terület szépségét is jelentik, valamint, hogy ha valaki kellő óvatossággal, alapossággal és persze elszántsággal vág bele, akkor az egész ország számára értékes eredményekre lehet jutni ebben a témaban. Remélem írásom, ha csak gondolatébresztés erejéig is, de segítséget jelent ebben.

Ajánlott olvasmányok

A következő könyvek érdekesek és tanulságosak lehetnek szerintem a téma iránt mélyebben érdeklődőknek:

- Frank E. Harrell. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, 2015. Link. Az alapmű. Hangsúlyozni kell azonban, hogy ez nem első bevezetés a regresszióba, azoknak szól, akik a regresszió alapjait már ismerik. Számukra azonban fantasztikusan hasznos, szemet felnyitó, a bevezető irodalmakban sokszor nem tárgyalt – de a gyakorlatban nagyon fontos – kérdéseket bemutató könyv, ami megismerteti a modellezés stratégiáját, nem pusztán a regresszió technikai használatát. A könyvhöz egy nagyon jól használható R csomag, az **rms** tartozik, illetve számos kiegészítés elérhető a szerző honlapján.
- Judea Pearl. Causality – Models, reasoning and inference. Cambridge University Press, 2009. Link. Az egyik legismertebb könyv, ami általában tárgyalja az okozatiság kérdését, a filozofikus kérdéseket is érintve. Kitér az olyan kapcsolódó kérdésekre, mint a confounding, és hangsúlyosan alkalmazza a kauzális diagramok eszközét.
- Trevor Hastie, Jerome Friedman, Robert Tibshirani. The Elements of Statistical Learning – Data Mining, Inference, and Prediction. Springer, 2009. Link. Az egyik legalaposabb könyv ami a statisztikai nézőpontból tárgyalja a tipikusan kevésbé statisztika, inkább „adatbányászat“, „gépi tanulás“ címekkel által besorolt témaikat.
- Ewout W. Steyerberg. Clinical Prediction Models – A Practical Approach to Development, Validation, and Updating. Springer, 2009. Link. Mint címe is mutatja, ez a könyv elsősorban a klinikai predikciós modellekkel foglalkozik, de sok hasznos tanulság is leszűrhető belőle általában a regressziós modellezésre nézve.

A szerző klinikai biostatisztikus, orvosbiológiai mérnök. A fent leírtak teljes egészében a magánvéleményét képviselik.