A statisztikai modellezés filozófiája*

Rappai Gábor
egyetemi docens,
a Pécsi Tudományegyetem
Közgazdaságtudományi
Karának dékánja
E-mail: rappai@ktk.pte.hu

A szerző azt vizsgálja, hogy miért övezi annyi félreértés még a leggondosabban készített statisztikai modellek eredményeit is. Összefoglalja azokat a legfontosabb premisszákat, melyek nélkül a statisztikai modellezés eredményei nem vagy nem korrektül értelmezhetők. Kitér a statisztika tudományterületi besorolására; a sztochasztikus modellezés általános formájára, illetve alapgondolatára; valamint néhány olyan modellezési körülményre, amely a szakavatott statisztikusoknak remélhetőleg triviális, ám a laikusok számára nem kellőképpen tisztázott.

A tanulmány számos olyan momentumot érint (a modellbecslés célfüggvénye, mérési hiba, kiinduló specifikáció, ceteris paribus elv stb.), melyek pontatlan alkalmazása, illetve meg nem értése azt eredményezi, hogy a modellező és a felhasználó között információs rés keletkezik. Összefoglalásként a szerző hangsúlyozza a statisztika önálló módszertudományként való elismerésének, valamint a sztochasztikus modellezés minél szélesebb körben történő oktatásának szükségességét.

TÁRGYSZÓ:
Statisztika.
Sztochasztikus modell.
Tudományfilozófia.

* A szerző sok-sok közvetlen, illetve tágabb értelemben vett kollégának tartozik köszönettel, hiszen jelen tanulmány nem született volna meg az elmúlt két évtized beszélgetései nélkül. Természetszerűleg a dolgozat hibáiért csak engem illet felelősség. A végső szöveg elkészültéért külön köszönettel tartozom közvetlen tanszéki oktatótársaimnak, *Herman Sándornak*, *Kehl Dánielnek*, *Tiszberger Mónikának*, valamint a tanulmány opponensének, *Hunyadi László* professzornak.

A statisztika tudományának elkötelezett és felkészült művelői számára teljesen világos, hogy a szeretett tudományukkal, illetve tevékenységükkel kapcsolatos közvélekedés mennyire torz, nemritkán dehonesztáló. Biztos sok statisztikus feltette már magának a kérdést, miért tapad nagyon sokszor össze a „csúsztatás", az „adatmanipuláció", sajnos egyre gyakrabban a „ferdítés" vagy „hazugság" szó a statisztikával. Megítélésem szerint a szakma nagyon sokáig abba az illúzióba ringatta magát, hogy mindez csak a hozzá nem értő, kontár „álstatisztikusok" negatív hatásának tudható be, mára azonban egyre nyilvánvalóbbá válik, hogy más (is) rejtőzik a jelenség mögött.

Tény, hogy a statisztikustársadalom nagyon sok mindent megtesz annak érdekében, hogy a sztochasztika gondolkodásmódja jobban beépüljön az általános műveltségbe.[1] Az utóbbi évek (évtized) egyik legnagyobb vívmányának tekinthető, hogy komoly előrelépések tapasztalhatók a statisztika etikus művelésének minden adatszolgáltatóra, illetve felhasználóra vonatkozó szabályozásában, illetve ezen tevékenységek minőségbiztosításában.[2] Mindezen törekvések ellenére – úgy érzem – a hétköznapok gondolkodásában nem következett be áttörés, vagyis az „átlagemberek", de ami ennél szomorúbb a közgondolkodók, politikusok, nemritkán gazdasági újságírók sem értik a statisztikus eredmények (adatok, modellszámítások) szemléletét.

A tanulmány címe valószínűleg túlságosan „fellengzős", mindössze arra próbáltam utalni vele, hogy olyan általános érvényű elméleti-módszertani meghatározásokkal kívánok foglalkozni, melyek túlmennek egy-egy konkrét elemzési-előrejelzési feladaton, azaz általános keretfeltételeit képezik a statisztikus munkájának, ezzel egyidejűleg elengedhetetlenül szükségesek lennének ahhoz, hogy a statisztika szempontjából laikus is pontosan értse a legfontosabb megállapításokat. Dolgozatom műfaja – legalábbis szerintem – esszé, ebből következően az átlagos módszertani jellegű tanulmányokhoz képest rengeteg lábjegyzet, idézőjel, „elvarratlan szál" maradt benne, mindezzel leginkább jelezni szeretném, hogy sok kérdésben még további polémiákat tartanék szükségesnek.

Jelen tanulmányban, néhány lényeges elemét tekintve, megpróbálom összefoglalni azokat a – szakma által természetesen teljes mértékben ismert – sarokpontokat,

[1] Elég csak a matematika érettségi vizsga részletes követelményeire gondolni (lásd 40/2002. (V.24) OM rendelet), miszerint: „*A modern tudományelmélet egyik fontos pillére az a gondolkodásmód, amellyel a sztochasztikus jelenségek leírhatók. A társadalomtudományi, a természettudományi és a közgazdasági törvényeink nagy része csak statisztikusan igaz. A mindennapi élet történéseit sem lehet megérteni statisztikai ismeretek nélkül, mivel ott is egyre gyakrabban olyan tömegjelenségekkel kerülünk szembe, amelyek a statisztika eszközeivel kezelhetők. A sztochasztika gondolkodásmódja a XXI. század elejére az emberi gondolkodásnak, döntéseknek és cselekvéseknek olyannyira alapvető része lesz, hogy elsajátítása semmiképpen sem kerülhető meg.*"

[2] A statisztika korrekt műveléséről lásd például „Az európai statisztika gyakorlati kódexét" (*Statisztikai Szemle*. 2007. évi 85. évf. 10–11. sz. 885–896. old.).

melyek megértése nélkül a statisztikusmunka eredményei nehezen értelmezhetők. Nem gondolom azt, hogy a dolgozat minden előzményt nélkülöző, hiszen számos, a témát boncolgató írás látott már napvilágot, akár a *Statisztikai Szemle* hasábjain is, az elmúlt évtizedekben. Feltétlen előzményként tartom számon a tragikusan hirtelenül elhunyt *Mundruczó György* professzor poszthumusz könyvét (*Mundruczó* [1998]), valamint *Szilágyi György* professzornak a statisztikai „olvasni tudásról" szóló cikkét (*Szilágyi* [2000]). Írásomban mindvégig arra törekszem, hogy a profi modellező és a laikus (vagy nem teljesen jártas) felhasználó között gyakran fennálló, elsősorban terminológia nem értésből, illetve az elmélyültség hiányából származtatható kommunikációs rést feltárjam, illetve – lehetőség szerint – csökkentsem. Ennek megfelelően kitérek a *statisztikatudomány bizonytalan besorolásá*ból, a *sztochasztikus modell alapgondolatának, illetve céljának nem pontos ismeret*éből, illetőleg a *modellezés körülményeinek, az eredmények értelmezésének félreértésé*ből eredő, nem feltétlenül teljesíthető elvárásokra. Meggyőződésem, hogy a sztochasztika gondolkodásmódja kevéssé ismert az átlagemberek körében, oktatásának, vagy ami fontosabb megismertetésének megkezdése a középiskolában már késő.[3]

1. A statisztikatudomány legáltalánosabb metodológiai jellemzői

Gyakorta hallható definíció, nemritkán statisztika tárgyú felsőoktatási kurzusok bevezető mondataként is elhangzik, miszerint *„a statisztika a társadalomtudományok matematikája"*. Noha – megítélésem szerint – az előbbi kijelentés számos „sebből vérzik", a mögötte meghúzódó gondolat rendkívül figyelemreméltó.

Miért is gondolom, hogy az analógia sántít? Egyrészt nem hiszem, hogy a megfogalmazásban indirekt módon megjelenő állítás, mely szerint a matematika egyértelműen természettudomány, bizonyított lenne. Másrészt nem gondolom, hogy a természettudományok és a társadalomtudományok ilyen, módszertudományi szempontból történő szembeállítása releváns kérdés, hiszen gondoljunk bele, hogy a meghatározás következetes továbbvitele az előzőnél – szerintem – sokkal bizarrabb kijelentésekre ragadtathat:

> – a matematikát a társadalomtudományokban nem érdemes használni, vagy

[3] A rémüldözőket megnyugtatandó, nem gondolok az óvodában varianciát számíttatni, de annak felismertetése a gyermekekkel, hogy szüleik délutáni megérkezése egy adott időintervallumban „szóródik", talán nem túlzott elvárás.

– a statisztika alkalmazása a természettudományokban felesleges, esetleg

– „a mikroökonómia a gazdaságtudomány fizikája”.

Ugyanakkor érthető, hogy a statisztikával foglalkozó kutatók, tudósok időről időre megkísérlik elhelyezni tudományukat a diszciplínák rendszerében, és – valószínűleg éppen ez motiválja ennek a dolgozatnak a megírását is – rendre elhatárolási, besorolási problémákba ütköznek. Az említett „definíció” – tömörsége ellenére – számos rendkívül fontos és helyes gondolatot tartalmaz, legalábbis implicit módon. Egy ilyen meghatározás véglegesen lezárná azt a vitát, miszerint a statisztika valóban „csak” a matematika része-e, vagy önálló módszertudomány.[4] Nyilvánvaló, hogy a statisztika nemcsak azért nem tekinthető pusztán a matematika részének, mert jelentős mértékben gyakorlati tevékenységet is értünk a fogalom alatt, de azért sem, mert önálló fogalomrendszerrel, saját tartalmi felépítéssel rendelkezik.

A „definícióban” a következő – igen figyelemreméltó – gondolat rejlik: a természettudományokban túlnyomórészt determinisztikus problémákkal kerülünk szembe (vagy addig alakítjuk a keretrendszerünket, amíg a problémák megfelelően választott axiómarendszerbe ágyazhatók), a társadalomtudományokban a törvényszerűségek csak sztochasztikus jelleggel érvényesülnek, az általános (hibrid, sokszor a valóságban nem is létező) esetekre igazak lehetnek, de biztos nem érvényesek a társadalom (statisztikus szóhasználattal a sokaság) minden egyedére. Amennyiben ez valóban így van, akkor érthető, hogy miért akarjuk a természettudományokat a függvényekkel, a definíció–tétel–bizonyítás „szentháromságával” operáló matematikával leírni, és miért támaszkodunk oly sokszor a társadalomtudományokban (gazdaságtudományok, szociológia, politológia stb.) a sztochasztikus kapcsolatok elemzésére is képes, a reziduummal szinte állandóan operáló statisztikára. Nyilvánvalóan az ilyen megkülönböztetés továbbra is sántít, hiszen határterületek, átfedések, összemosódások mindig lesznek. A tudományfilozófusok küszködnek is a megoldásokkal, gondoljunk például egy olyan osztályozási „döccenőre”, mint az élő és élettelen természettudomány szétválasztása, többek között éppen azért, mert az élőlények – az emberek – sokszínűségének kezelése érdekében a statisztika módszertana is szükségessé válik. (Természetesen ellenpéldát is lehetne hozni, hiszen egy szállításszervezési feladat optimális megoldása a lineáris algebra, illetve az operációkutatás „tisztán matematikai” eszközeivel mitől is ne lenne társadalomtudományi kérdés.)

Az előbbi bekezdés már ráirányítja a figyelmet arra a kérdésre, amelyben a természet versus társadalomtudomány, illetve a matematika versus statisztika dilemmák felvetődnek. Egyrészt kijelenthetjük, hogy egy-egy konkrét kutatás, problémamegol-

[4] A kérdéssel kapcsolatos álláspontunkat egy korábbi tanulmányunkban már részletesebben kifejtettük, lásd *Hunyadi–Rappai* [1999].

dás tudományos besorolása a megválaszolandó kérdés hovatartozásától, és nem az alkalmazott módszertan, vagy főleg gyakorlati tevékenység hovatartozásától függ. Nyilvánvalóan elborzadna a radiológus, ha a tüdőszűrést és az ebből (is) táplálkozó rákkutatást – pusztán csak a röntgensugár használatának okán – a fizika tudományterületére sorolnánk; éppen így megütközést keltene a marketingesben, ha – mivel primer lekérdezésének eredményét számítógéppel dolgozza fel – az informatikához sorolódna a fogyasztói trendek vizsgálata. Másrészt viszont tény, hogy a két nagy tudományterület (természet-, illetve társadalomtudomány) valóban markánsan különbözik egymástól az általában (leggyakrabban) alkalmazott módszertani megfontolásait tekintve:[5]

– *Kísérleti vagy tapasztalati tudomány?* Nyilvánvaló különbség a természet-, illetve társadalomtudományok között, hogy míg az előbbiben nemcsak lehetséges, de helyénvaló, sőt megkövetelt a kísérletekkel történő bizonyítás, addig az utóbbi esetében erre csak rendkívül korlátozottan van mód. Szigorúan statisztikai értelemben véve, miközben a kísérleti tudományok élhetnek a végtelen elemszámú alapsokaság feltételezésével, ráadásul a modellezésükben használt minták elemszáma szinte korlátlanul növelhető,[6] azalatt a tapasztalati tudományok általában egy konkrét empirikus megfigyeléshalmazzal kell, hogy beérjék. (A kérdés sokkal bonyolultabb, ám talán nem jár messze az igazságtól, hogy míg a természettudományokban a független azonos eloszlású minták, és az ezekre alapozott matematikai statisztikai eljárások alkalmazása korrekt, addig a társadalomtudományokban ennél kevésbé „vegytiszta" következtetéses statisztikai eljárásokra van általában szükség.)

– *Dedukció vagy indukció az általánosan használt megközelítés?* Közismert, hogy – nagyon leegyszerűsítve – a dedukció a részlegesnek az egyetemesből való leszármaztatását, az indukció a részlegesből az egyetemesre való következtetést jelenti. Miközben a modern logika felfogása szerint dedukció és indukció semmiképpen sem állnak ellentétben, azért könnyen belátható, hogy a törvények, ok-okozati összefüggések megállapítása tekintetében a két módszer különböző felfogásokat tükröz. Anélkül, hogy komolyabb metodológiai fejtegetésekbe

[5] Az itt következő, pontokba szedett különbségtétel nyilvánvalóan szubjektív, és valószínűleg minden megállapításával szemben hozható ellenpélda. Ugyanakkor azt remélem, hogy a kijelentések „sztochasztikusan" igazak.

[6] Ez az állítás nyilvánvalóan megkérdőjelezhető, hiszen a mintaelemszám növelésének a kísérleti tudományokban is gátat szab a költségvetési korlát (sőt!), ugyanakkor az elvi lehetősége megvan a mintanagyság növelésének.

bonyolódnánk, kijelenthető, hogy a társadalomtudományok alapvető következtetési eszköze az indukció, pontosabban a következetes empirizmus. Ennek értelmében az általános oksági tételek is a tapasztalatból származnak, vagyis – alaptémánk szóhasználatával megfogalmazva – a statisztika megfigyeléseiből következtethetünk a társadalomban vagy gazdaságban meglevő törvényszerűségekre, elfogadva, hogy az így származtatott oksági tételek nem adnak feltétlen bizonyosságot (sztochasztikus jelleggel érvényesek). Mindez azt is jelenti, hogy a természettudományokban (és itt most kifejezetten nem csak a matematikára gondolok) egy egyszer bizonyított tételt nagyon sokáig állandónak tekinthetünk, vagyis mindaddig, amíg a tétel környezetét jelentő axiómarendszer nem változik, a „törvények" is változatlanok. Ezzel szemben a társadalomtudományokban nem feltétlenül kell paradigmaváltás ahhoz, hogy a „törvényeket" folytonosan teszteljük, módosítsuk, pontosítsuk.

– *Az általános vagy a kivétel az érdekes?* A harmadik, általam kiemelendőnek vélt, markáns különbség természet- és társadalomtudományok között, hogy az alkalmazott modellek milyen alapcéllal keletkeznek. A természettudományok esetében a modellezés célja – szinte kivétel nélkül – az általános érvénnyel bíró állítás, ha úgy tetszik a törvényszerű megfogalmazása. Abban az estben, ha egy természettudományos modell által megállapított törvényszerűség a későbbi kísérletek, illetve észlelések során alapvetően megkérdőjeleződik, akkor vagy elvetjük a tételt, vagy az egész axiómarendszerünket változtatjuk. A társadalomtudományok esetében ez nem feltétlenül van így. Nyilván érvelhetnénk azzal az előbbi pontban már említett gondolatsorral, miszerint a társadalomtudomány törvényei nem determinisztikusak, ám most egy másik aspektusra kívánom felhívni a figyelmet. A társadalomtudományi (markáns példákat az általam viszonylag jól ismert gazdaságtudományok területéről tudnék mondani) kutatásokban nagyon gyakran a reziduum meghatározása (becslése) az alapcél. (Gondoljunk például a kockázatmodellezésre, ahol például egy árfolyammodell esetében a kockázatot pontosan a várt és a tényleges érték különbsége jelenti, és modelljeink sokszor éppen ezen maradéktag leírására tesznek kísérletet.[7]) Egyrészről társadalomtudományi modellezési specifikumnak gondolom, hogy nem mindig a várható érték becslése a cél, hiszen sokszor a szóródás előrejelzése éppen ilyen fontos. Másrészt szintén a társadalomtudományi modellezés jellemzője, hogy egy-

[7] Lásd például az elmúlt két évtizedben gyakorlatilag alapvetően meghatározóvá váló ARCH-modellcsaládot.

> egy modellbecslést követően, amikor – akár számos – ellenpélda található a modell által leírt jelenségre, nem mindig a specifikáció változtatásával, hanem gyakran új becslési módszer választásával, vagy éppen az ellentmondó paraméter értékének megmagyarázásával reflektálunk a problémára.

Mindez nyilvánvalóvá teszi, hogy a statisztikai modellek alapvetően a véletlen (véletlenszerűség) kezelésében-megítélésében, illetve ennek kívánatos (pontosabban fogalmazva, még elviselhetőnek tartott) nagyságrendjében különböznek a matematikai modellektől. Így amennyiben elfogadjuk, hogy a természettudományok inkább függvényszerű kapcsolatokat leíró matematikai-, míg a társadalomtudományok inkább empirikus következtetéseken alapuló statisztikai modellekkel operálnak, akkor egyfajta „tudomány–módszertan megfeleltetés" talán nem teljesen irrealisztikus.

2. A statisztikai modell általános alakja és célja

Tekintsük ezek után a legáltalánosabb formájú statisztikai modellt! Az alábbiakban nem kívánok különbséget tenni abból a szempontból, hogy modellünk az adatokban rejlő információk tömörítése, illetve valamely nem ismert adat becslése (előrejelzése) céljával készül. A jelen tanulmányban tárgyalandó általános formájú statisztikai modell a következő:

$$y_i = \hat{y}_i + e_i ,$$

ahol a szokásos jelölésekkel: y_i a tényleges (ismert, vagy megismerendő) adat, $\hat{y}_i$ az előbbi adatra vonatkozó modellezett (számított vagy becsült) érték, e_i a modellezés során elkövetett hiba, a tény és a modellezett érték különbsége (a maradéktag vagy reziduum). Az összefüggésben szereplő i index azt jelzi, hogy az y ismérv (változó) valamennyi megfigyelésére vonatkozóan rendelkezünk modellezett értékkel (itt most nem foglalkozunk azzal a kérdéssel, hogy ezek egyenlők, vagy minden esetben különböznek). Annak alapján, hogy az általános modellünkben vizsgált y „eredményváltozó közelítése"[8] során alkalmazunk-e további (értsd a modellezendő

[8] Az idézőjel használatával érzékeltetni kívánom, hogy itt nem a szokásos, regressziós szemléletű eredményváltozóról (tehát a függő változóról, melyet a független- vagy más néven magyarázó-, illetve tényezőváltozók magyaráznak) van szó, hanem az általános modellben vizsgált, valamilyen okból fontos, modellezendő ismérvről. Meglehetősen nehéz egy olyan konzisztens kifejezésrendszert találni a dolgozatban vizsgált, szándékoltan rendkívül általános problémára, mellyel nem ütközünk bele a statisztika módszertanába foglalt vagy legalábbis megszokott szóhasználatba. Annak érdekében, hogy elkerüljem azon „áthallásokat", miszerint a leírás vagy tömörítés kifejezés az alapsokasági (leíró) statisztikai elemzésekben szokásos, ugyanakkor a becslés fogalom a következtetéses statisztikára asszociál, a következőkben a vizsgált általános eljárást *modellezésnek* vagy ennek szinonimájaként *közelítésnek* fogom nevezni.

változón kívül, további) változó(ka)t, az előbbi általános alak tovább komplikálható lenne, de a dolgozat szempontjából most ez sem lényeges.

Mi tehát a modellezés célja, avagy mikor tekinthetünk egy statisztikai modellt jónak? Úgy gondolom, hogy önmagában ennek a kérdésnek könyvtárnyi irodalma van, melynek rövid áttekintésére sem vállalkozom. Jelen tanulmányban mindössze annyit kívánok felvillantani, hogy az előbbi kérdés megválaszolása során alkalmazott „célfüggvény" sem mindig triviális a laikus (statisztikával nem professzionális szinten foglalkozó) felhasználó számára, így önmagában ez is félreértésekre adhat okot. A teljesség igénye nélkül nézzünk néhány elvet, amely a statisztikai modellezés célkeresztjébe kerülhet, és amelyek közötti választás alapvetően meghatározza, de ami lényegesebb, különbözővé teheti a modellspecifikációt, sőt az alkalmazott statisztikai eszköztárat is.

1. Hibaminimalizálás, melynek során az általános modellben szereplő maradéktag minél kisebb értéke, vagy – talán így elfogadhatóbb – a modell minél jobb illeszkedése a cél. Rendkívül nagy számban találunk képzett mutatókat,[9] amelyek abból a szempontból közösek, hogy a tényleges és a modellezett érték különbségét vizsgálják, és a különbség minél kisebb értékét preferálják. Az említett mutatószámok kis mértékű vizsgálata nélkül is könnyen belátható, miként az ilyen elven történő modellezés során gyakran előfordul, hogy a modellünk sokszor (akár mindvégig!) olyan eredményeket szolgáltat, melyek a valóságban (a tényleges értékek, a tapasztalati adatok között) elő sem fordulhatnak, vagyis a modellezés eredményeként kapott közelítő érték elméletileg kizárt.

2. Találatmaximalizálás, vagyis annak megcélzása, hogy minél többször forduljon elő olyan modellezési eredmény, amely megegyezik az adott megfigyeléshez tartozó tényadattal. Könnyen átlátható, hogy egy ilyen célfüggvénnyel készülő modell az előbb említett (egyébként szokásos) minimumra közelítésektől teljesen eltérő eredményre vezet(het). Ugyanakkor vitathatatlan, hogy – noha a megoldás során alkalmazott matematikai-statisztikai eszköztár sokszor nagyon bonyolult is lehet – a laikus számára mindez egyszerűbben követhető. (Képzeljünk el egy fogyasztóiárindex-számítást, amely a vizsgálatba vont árucikkek árváltozásának móduszát használja „árindexként".) Nyilvánvalóan az ilyen elv ellen szól, hogy – főképpen a többváltozós közelítések esetén – nagyon sokszor nem jutunk zárt megoldásra, vagyis a modell alapján történő általánosítás sokkal kevésbé lesz kézenfekvő, ám ez biztos csak a megszokásaink miatt tűnik zavarónak.

[9] MSE, MAPE stb., ezek meghatározása itt most lényegtelen.

3. Hírérték-maximalizálás, vagyis azon közelítések preferálása, amelyek a meglepőbb (váratlanabb) eredmények esetén jól viselkednek, azokkal szemben, melyek csak a várakozásoknak megfelelő értékeket „találják el". (Valami hasonlóval kísérleteznek a tőzsdei árfolyamok előrebecslésénél is, hiszen közismert tény, hogy jelentős hozamra csak úgy lehet szert tenni, ha a befektető a piaci szereplők nagy részével ellentétes irányba mozog.) A szokásos statisztikai modellezési szóhasználat szerint ez esetben azon modelleket preferáljuk, melyek az outlierekre jól illeszkednek, még akkor is, ha a várható érték közelében sokszor (remélhetőleg nem túl nagyot) tévednek.

Nyilván nem kell magyarázni, hogy a különböző modellezési optimumfeltételek teljesen különböző modellekhez vezet(het)nek. Nem is ezt szánom a lényeges megállapításnak, hanem azt, hogy a statisztikai modellezésben járatlan felhasználó fejében nem feltétlenül a hagyományos (általában legkisebb négyzetek elvén vagy valamilyen ehhez hasonlóan „közismert" megfontoláson nyugvó) célérték keresési elv a kívánatos, így az általunk optimálisnak ítélt modell neki nem feltétlenül „tetszik".

3. A modell eredményeinek korrekt értelmezéséhez elengedhetetlenül fontos meggondolások

A statisztikai modellezés eredményeképpen megjelenő információk – meggyőződésem szerint – az esetek túlnyomó hányadában korrektek, a statisztikusok a szakma szabályainak megfelelően, legjobb tudásuk szerint használják modelljeiket elemzési, előrejelzési vagy szimulációs célra. Miért van mégis ilyen sok kritika a statisztika eredményeit illetően, miért érzi gyakran az átlagember, hogy a statisztika torz, illetve pontatlan következtetésekre jut? Biztosan oka ennek az is, hogy a – valószínűsíthetően a közérthetőség kedvéért félreértelmezett – statisztikai nyelv (terminológia) rengeteg olyan kifejezést használ, melyek a köznyelvben mást (nem teljesen ugyanazt) jelentenek, illetve melyeknek korrektül definiált jelentéstartalmáról a laikusnak nincs, vagy nem megfelelő az információja.[10]

[10] Nyilván más tudományterületek is küszködnek hasonló problémákkal. Gondoljunk például a pénzügyekre, ahol a „kockázat" fogalma a várható értéktől való eltérést (szóródást) jelenti, vagyis semleges fogalom; ugyanakkor a – pénzügy szempontjából – laikus egyértelműen negatív tartalmat rendel a szóhoz. Lehet, hogy mégis igaza van azon tudományoknak (tudományterületeknek), melyek a szakértők számára megengedik, hogy – egyszerűen – egy másik (idegen) nyelven kommunikáljanak, ilyen például a latint használó orvos- és egészségtudomány vagy az angolról le nem mondó informatika? A magam részéről egyébként az előbbi költői kérdésre egyértelműen *nem* a válaszom: szerintem egyetlen tudomány sem idegenedhet el az eredményeit használó közegtől.

A továbbiakban tekintsünk néhány olyan modellezési körülményt, melyek félreértéséből – legalábbis megítélésem szerint – a statisztikai eredmények félreértelmezése következhet.[11]

Miből táplálkozik a statisztikai modell?

Amint ezt a korábbikakban tisztáztuk, a statisztikai modellek alapja az empirikus (tapasztalati, megfigyelt) adatbázis. Ahogy ezt minden bevezető statisztika tankönyv tartalmazza, az egyedek azonosítása ismérvek alapján történik, a statisztikai modellezés feladata az egyedek összességét jelentő sokaság tömör leírása, illetve a különböző ismérvek közötti összefüggések feltárása. (Talán elsőre meglepőnek tűnhet, hogy a – bevett szóhasználattal – leíró, illetve következtetéses statisztika módszertanában egyaránt modellezést emlegetek, de a korábbiakból ez viszonylag logikusan következik.) Nos, az empirikus adatbázist minden (laikus) felhasználó mindig meglevőnek, tökéletesnek, ideálisnak, de legalábbis a vizsgálat szempontjából optimálisnak tételezi fel. Ugyanakkor a megfigyelt (megfigyelhető) valóság, a tapasztalati adatok jelentik az első komoly korlátot, melyet a statisztikus modellezőnek le kell győznie.

Az első problémát az jelenti, hogy a társadalomtudományokban[12] az elméletek által kutatott jelenségek nem mindig azonosíthatók be egyértelműen megfigyelhető ismérvekkel (változókkal). Gondoljunk egy triviális példára, a termelési függvényre! Valószínűleg valamennyi közgazdász számára egyértelmű, hogy $Q = f\left(K, L\right)$ egy termelési függvény, ahol Q a kibocsátás, K a holtmunka-ráfordítás és L az élőmunka-ráfordítás. De mivel mérjük az előbbi kategóriákat? Egyáltalán mérhető egy ilyen összetett fogalom, mint például a holtmunka-ráfordítás? Nyilvánvalóan nem kívánok a termelési függvény könyvtárnyi irodalmába belekontárkodni, de könnyen belátható, hogy az output (kibocsátás) változó éppúgy lehet árbevétel, mint hozzáadott érték, mérhetjük akár naturáliában, akár pénzben és így tovább. A nehézség persze nem ez, hanem az, hogy a látszólag azonos specifikációjú (például Cobb–Douglas-típusú), de különböző jelentéstartalmú változókat tartalmazó termelési függvények azonos(nak tűnő) paramétereit (például a határtermelékenységet vagy a helyettesítési rugalmasságot) elkezdjük összehasonlítani. Természetesen mondhatjuk, hogy ez inkorrekt, ám komoly tudományos publikációkban is találkozhatunk ilyen jellegű „tévedésekkel".

Az előző példában ráadásul olyan ismérvek jelentek meg, melyek jól (könnyen) mérhetők. De nem egy statisztikai modell támaszkodik – kényszerűségből –

[11] Reményeim szerint jelen tanulmányt a statisztika szakértői, illetve – a statisztika tudományát tekintve – laikusok is olvasni fogják. A következőkben, illusztratív céllal, néhány standard statisztikai módszert is nevesítek. Előre elnézést kérek a szakértőktől a túlzott egyszerűsítésért, ugyanakkor biztatom a laikusokat, hogy ezen módszerek alapos ismerete nélkül is kövessék a gondolatmenetet.

[12] Lehet, hogy a természettudományokban is így van, de azt a területet nem ismerem olyan mélységben, hogy ezt merjem állítani.

proxykra. Az ilyen helyettesítő változók esetében a korábbi probléma eszkalálódik. Tessék néhány gyakran elhangzó gazdaságpolitikai fogalomra, például a „*nemzeti vagyonra*" vagy a „*nemzetgazdaság kibocsátására*", esetleg a „*gazdasági fejlődésre*" gondolni! Milyen éles viták tudnak kialakulni politikusok között abban a tekintetben, hogy az „ország fejlődésének az üteme" elérte-e az „EU fejlődési ütemét", vagy sem. Könnyű préda a statisztikus, amikor a fejére olvassák a különböző publikációkban megjelenő eltérő „tényszámokat" (például hiába védekezik azzal, hogy a GDP volumenváltozásának nagysága – definíció szerint – nem esik egybe a folyó áron mért GNP változási ütemével).[13]

Szintén az empirikus adatbázis összeállítása során keletkező (egyébiránt a hivatalos közlések során, de még az alapozó tankönyvekben is gyakran elkerült) probléma a mérési hiba kérdése.[14] Nem kívánok részletesen foglalkozni a felvételi hibák fajtáival; arra sem térek ki, hogy egyes hibák (definíciós hiba, nemválaszolási hiba) gyakorta kiküszöbölhetők lennének, míg a mintavételi hiba egyértelműen a részleges felvétel velejárója; sőt szintén csak említem, hogy az adatok feldolgozása és közlése során is gyakran keletkeznek hibák; mindössze egy dolgot hangsúlyozok: az empirikus adatbázisban szereplő adatok, az esetek jelentős részében pontatlanok, hibát tartalmaznak. A statisztikai modellek tehát kényszerűségből ilyen hibás adatokból építkeznek, és ezek alapján tesznek, nemritkán megfellebbezhetetlennek tűnő megállapításokat. Természetesen ott, ahol erre lehetőség nyílik, mindent meg kell tenni azért, hogy a mérési hibát kiküszöböljük, ám ha ez nem sikerül, vagy túl nagy befektetést igényelne, akkor sokkal korrektebb ezt közölni, mint a pontosság látszatát kelteni. Az előbbi kijelentés triviális, olyannyira, hogy a már említett „gyakorlati kódex" 12. alapelve követelményként határozza meg a mintavételi és nem mintavételi hibák mérését és rendszeres dokumentálását. Ugyanakkor a tényleges problémát szerintem az okozza, hogy mást gondol kézenfekvőnek, ezáltal közlendőnek a szakértő és mást a laikus. Nem hiszem, hogy sokan lennének a gyakorló politikusok vagy szakpublicisták között, akik a központi költségvetés tavalyi –861,7 milliárd forintos egyenlegét[15] korrektül tudnák értelmezni, vagyis értenék például az adatközlési hiba jelentőségét, miszerint a vonatkozó adat azért tizedmilliárd forint pontosságú, mert „50 millió forint ide vagy oda, még belefér". Gondoljunk arra, hogy milyen mértékű és mélységű politikai vitákat generál a fogyasztói árindexnek az „ígérthez képest" 0,1 százalékpontos eltérése, miközben mindez egyenlő az adatközlési hiba nagyságrendjével.[16] Mindennek tükrében érthető, hogy miért szorgalmaznám a felhasználók alapvető statisztikai műveltségének gyarapítását.

[13] És akkor még kedvencemet, a „negatív GDP-növekedési ütemet", nem is említettem…

[14] A témával az alapozó tankönyvekben szokásosnál bővebben foglalkozik *Hunyadi–Vita* ([2008] 30–35. old.).

[15] Lásd „A KSH jelenti 2009/4", 37. old.

[16] Mivel az adat tized százalékpont pontossággal szerepel (például 6,2 százalék), ez azt jelenti, hogy ez az „utolsó" még pontos számjegy (vagyis értéke a 6,15–6,25 százalékos felülről nyitott intervallumban van), azaz lehet, hogy a politikai vita okafogyott.

Az utolsó, az adatbázis sajátos jellegéből adódó, ugyanakkor a statisztikai modellezésben nem járatos felhasználók előtt nem kellő mélységben ismert probléma, az ún. *outlierek* (kiugró értékek) kezelése.[17] Az adatbázisban szereplő kiugró értékek a modellezés szempontjából egyértelműen káros következményekkel járnak: egyrészről – természetesen az optimalizálási elvek függvényében különböző mértékben – „elmozdítják" a modellt az ideális (vagy a többségre jobban jellemző) pályáról; másrészt – részben az előzőekből is következően – rontják a modell illeszkedését (növelik a standard hibát). Mit tehetünk a probléma kiküszöbölésére? A legáltalánosabban alkalmazott megoldás valamilyen robusztusabb eljárás használata, tipikusan ilyenek a helyzeti középértékek a számított átlagok helyett, vagy bizonyos jól megválasztott csonkolt (censored) mintás eljárások stb. Lényeges megjegyezni, hogy az előbb említett módszerek kellően megalapozottak, statisztikai szempontból tulajdonságaik ismertek, az értő felhasználó számára alkalmazásuk kézenfekvő.[18] Ennek ellenére, legalább két ok is nevesíthető, melyek miatt az outlierek kiszűrése, a csonkolt eljárások a laikusokban ellenérzéseket váltanak ki:

– nyilván nem egyszerű megérteni, hogy arra a kérdésre: mit tekintünk kiugró értéknek, többé-kevésbé egzakt válasz adható; a laikus mindig gyanakodni fog, hogy ami nem tetszik a modellezőnek (például mert rontja a modell illeszkedését, vagy nem illenek az eredmények a prekoncepcióba) azt kiugró értéknek tekinti és kihagyja az adatbázisból;

– sokszor éppen az első ránézésre outliernek tűnő adatok hordozzák a lényeges információkat, gondoljunk a trendforduló előrejelzésére a pénzügyi idősorokban vagy a strukturális törések hatásának első megjelenésére stb.

Összegezve az empirikus adatbázisról leírtakat, annyit mindenképpen meg kell állapítani, hogy a bevezető statisztika kurzusok (tankönyvek) méltatlanul keveset foglalkoznak az adatbázis összeállításának nehézségeivel. Az oktatásban is könnyen elintézzük a problémát azzal, hogy létezik a gyakorlati és az elméleti statisztika, az adatokat az előbbi szolgáltatja, ez a modellező számára adottság. Nyilvánvalóan sokat segítene a statisztikai modellek elfogadottsága szempontjából, ha a felhasználók jobban ismernék azt a „vajúdást", amit minden modellező átél, amikor nem talál megfelelő (elégségesen nagy, konzisztens, azonos módszertannal végzett gyűjtésből

[17] Többször említettem már, hogy minden egyes (jelen tanulmányban egy-két bekezdésben tárgyalt) probléma önmagában is szakkönyvek tucatjait tölti meg; hangsúlyozottan így van ez az outlierek kérdéskörével.

[18] A biometriában (általában az egészségügyi alkalmazásokban) sok évtizede bevett gyakorlat a MoM (a mediánhoz viszonyított egyedi érték, multiple of median) használata, nyilvánvalóan abból a felismerésből, hogy a laboratóriumi diagnosztikában néhány kiugró érték az átlagot jelentősen képes befolyásolni, ugyanakkor a medián „érzéketlen" a fals adatokra.

származó) adatbázist, és nem tudja eldönteni, hogy korlátozottan hasznosítható empirikus modellt építsen, vagy megmaradjon az elméleti fejtegetések szintjén.

Mi van a modellen „túl”?

A második részletesebben tárgyalt körülmény a modellen kívüli világ, vagy, ahogy a problémát gyakran, de nem teljesen korrektül interpretálják: mi a véletlen szerepe a statisztikai modellekben. Úgy gondolom, nem lehet elégszer hangsúlyozni, hogy az alapmodellben szereplő e_i, vagyis a tényadatok és a modellezett értékek különbsége a statisztikai modellek szükségszerű velejárója, megjelenése nem a modellező hozzá nem értéséből vagy tévedéséből fakad.[19] Ennek ellenére érdemes az előbbiekben említett maradékot (reziduumot) keletkezés szempontjából, két nagy típusra osztani:

– „tiszta” véletlenre, illetve
– a modellből kimaradó tényezők hatására.

Anélkül, hogy komolyabb valószínűségelméleti fejtegetésekbe kívánnék bonyolódni, tételezzük fel, hogy mindkét előbb említett esetben a reziduális változó (vagyis a tényleges, elkövetett hiba elméleti értéke, amit a továbbiakban ε-nal jelölünk) 0 várható értékű, konstans szórású. Felmerülhet, hogy ez esetben miért kell megkülönböztetnünk a két reziduumtípust. Megítélésem szerint azért, mert a laikus felhasználó jelentősen különböző megítéléssel viseltetik a „tiszta” véletlen, illetve a modellből kimaradó tényezők tekintetében.

A „tiszta” véletlen (példaként szokás említeni a háborút, elemi csapást, vis maior szituációkat stb.) definíció szerint nem előre jelezhető kategória. Ráadásul – szigorúan statisztikai szempontból nézve – az említett $E(\varepsilon)=0$, $Var(\varepsilon)=\sigma^2$ paraméterpár úgy keletkezik, hogy a maradék a megfigyelések nagy részénél 0, de néha akár jelentős mértékben is eltérhet attól. Mindez az empirikus adatbázis esetében úgy jelenik meg, mint néhány „váratlanul” kiugró érték, melynek előidéző okára a modellezőnek nem is kell gondolni. (Az ilyen eseményekre szokták használni azt a jól hangzó kitételt, miszerint *nem lehetetlen, de a bekövetkezésének valószínűsége 0*.) Az elmúlt közel negyedszázados modellezői gyakorlatom, amelynek viszonylag jelentős része piaci szereplők felkérésére végzett modellépítés volt, azt mutatja, hogy a laikus felhasználók az ilyen típusú véletlennel szemben megértőbbek, elnézőbbek, valahogy elhiszik, hogy az említett események nem prognosztizálhatók. Ez a megértő szemlélet természetesen több veszélynek is forrása lehet: egyrészt a modellezőt arra sarkall-

[19] A gondolatmenetet nem megszakítva utalok egy előbbi, a félreérthető szóhasználatra utaló, lábjegyzetre: a „standard hiba” sok laikus számára nyilvánvalóan valamilyen kiküszöbölhető és kiküszöbölendő jelenségnek tűnik, ami egyébként a statisztikus munkája során elkövetett tévedésből fakad. A valóságban a „standard hiba” a statisztikus munkájának szükséges velejárója.

ja, hogy úton-útfélen vis maiort emlegessen, másrészt arra csábítja, hogy alkalmasan választott dummy változókkal a modell illeszkedését mesterségesen javítsa.

Egészen más a felhasználók hozzáállása a modellből kihagyott tényezők megítélésének tekintetében. Ez esetben az alkalmazók mindig azt kérik számon a modellezőtől, hogy miért nem épített be a modellbe általuk kézenfekvőnek tartott megfontolásokat, miért nem vettek észre fontos körülményeket. Gyakran merül fel ez a kérdés az idősori modelleknél, ahol a laikusok nem is értik a megfontolásbeli különbséget például egy ARMA- és egy VARMAX-modell között. De talán egyszerűbb, ha arra gondolunk, hogy mennyiszer kérik számon egy havi bontású empirikus idősorra vonatkozóan a periodikus ingadozás figyelembe vételét a modellben olyankor, amikor kimutathatóan nincs szezonalitás az adatokban.[20] Nehogy lekicsinyeljük ennek a problémának a nagyságrendjét! Talán elég csak emlékeztetnem arra, hogy az elmúlt egy-két évtizedben a „mainstream" (ún. főáramlathoz tartozó) közgazdaságtan mellett megjelent a „behavioral economics", ami – természetesen nagyon leegyszerűsítve – nem más, mint a korábbi, alapvetően a racionális várakozások hipotézisén alapuló modellek maradéktagjában meglevő további, nem véletlen mintázatok keresése, és ezek viselkedéstani magyarázata.

Azt gondolom tehát, hogy szakítanunk kell azzal a korábbi sommás megállapítással, mely szerint a modell a valóság leegyszerűsített változata, ami a lényeges elemeket kiemeli, a lényegtelenektől pedig eltekint, ugyanis a lényeges–lényegtelen megkülönböztetés rendkívül szubjektív, a modellező felelősségét nagyban felerősíti. Úgy érzem, hogy nagy szükség lenne arra, hogy a modellspecifikáció során többet hangsúlyozzuk az elemzési módszer megválasztásának problémáit, hívjuk fel a laikus felhasználók figyelmét arra, hogy sokszor az elmélet által megalkotott specifikáción túl is vannak olyan kérdések, melyek a modellek készítésében jelentős szerepet játszanak. Ismét csak a transzparencia szükségességére hívom fel a modellezők figyelmét, valószínűleg el lehet kerülni, vagy a megalapozott vita kategóriájába lehet terelni a modellspecifikáció körül kirobbanó polémiákat, ha előre közöljük azokat a kereteket (beleértve a kihagyott elemeket is), melyeken belül a modellünk (és ebből adódóan a modellezés eredménye) alkalmazható.

Hol érvényesek a modellek?

Rendkívül fontos, ugyanakkor nagyon sok félreértésre okot adó kérdés, hogy mikor érvényesek egy statisztikai modell megállapításai. Természetesen most nem kifejezetten a modell időbeli érvényességét vizsgáljuk,[21] hanem a modell értelmezési

[20] Tipikus példája a jelenségnek az újságírók folyamatos „erőlködése", mellyel például a havi fogyasztói árindex vagy a költségvetési deficit-adatokba szezonalitást próbálnak „belelátni".

[21] Önmagában az is rendkívül érdekes kérdés, hogy egy sztochasztikus kapcsolat vagy oksági modell milyen időtartamon keresztül verifikálható.

tartományát. A laikus felhasználók, és közöttük is leginkább az átlagosnál jobb matematikai előképzettséggel rendelkező megrendelők számára nehezen érthető, hogy egy függvényszerű kapcsolat feltételezésére épített modell megállapításai nem érvényesek az adott függvényformula analízisben tanult teljes értelmezési tartományán.

A talán kissé elméleti, és éppen ezért nem feltétlenül érthető kijelentést világítsuk meg egy példával. Az egyik legegyszerűbb statisztikai modell a kétváltozós lineáris regresszió modellje

$$y_i = \hat{y}_i + \varepsilon_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i ,$$

ahol – a már bevezetett szimbólumokon túl – jelölje x_i a modell független (magyarázó-) változóját, valamint $\hat{\beta}_0$, $\hat{\beta}_1$ a modellezés során becsült paramétereket. A már korábban körülírt analízisbeli ismeretekkel rendelkező felhasználók nagyon nehezen értik meg, hogy sokszor annak ellenére az előbbi, ún. konstans tagot (tengelymetszetet) tartalmazó regressziós modellt alkalmazzuk, hogy ab ovo tudjuk, a magyarázóváltozó zérus értékéhez az eredményváltozó zérus értéke tartozik.[22] Nyilvánvalóan érthető a laikus felhasználó gyanakvása, hiszen a modell paramétereinek (konkrétan a β_0 paraméternek) értelmezése – elvben – úgy szól: amennyiben a magyarázóváltozó értéke 0, úgy modellünk szerint az eredményváltozó $\hat{\beta}_0$ értéket (a tárgyalt probléma esetén tehát nem feltétlenül 0 értéket) vesz fel. Nehezen fogadják el a laikusok azt az érvelést az előző kritikus gondolatmenettel szemben, hogy az empirikus adatbázison (a szokásos jelöléseket alkalmazva n megfigyeléspáron) alapuló modellezés során pontosan n konkrét x_i esetén ismerjük a konkrét (tényleges) y_i értéket, és mégsem érezzük ennyire igazságtalannak, hogy ezeken a pontokon nem feltétlenül megy át az egyenesünk.[23]

Az előbbi konkrét példában felmerülő problémánál sokkal gyakrabban kerül előtérbe az a kérdés, amikor a modell magyarázóváltozójának az értelmezési tartománya nem is tartalmazza a 0-t. A paraméterértelmezést könnyen elintézzük, hiszen a tipikus tankönyvi válasz valahogy így szól: β_0 *közgazdaságilag nem értelmezhető*. De mit kezdjünk azokkal a nem ritka felvetésekkel, mely szerint inkorrekt olyan modellt specifikálni, amely modell elméleti (matematikai) tulajdonságai nem felelnek meg a

[22] Nem kívánom mélyebben kifejteni, hogy ennek milyen – egyébiránt könnyen bizonyítható – oka van, itt elég megjegyeznünk, hogy mindezt a jobb illeszkedés kedvéért tesszük. (A statisztikában járatos kollégák számára az is közismert, hogy a konstans tagot nem tartalmazó regresszió esetén még az a „csúnyaság” is megeshet, hogy a modell magyarázó erejét mérő determinációs együttható negatív lesz.)

[23] Ha nagyon „szőrszálhasogatók” akarunk lenni, természetesen kiküszöbölhető lenne a felvetés úgy is, ha egy $n+1$ -edik megfigyeléspárként hozzávennénk az adatbázishoz a $y_{n+1} = x_{n+1} = 0$ „virtuális” megfigyelést, és így végeznénk el a modellbecslést. Könnyen belátható – gondoljunk a normálegyenletekre –, hogy mindez elégségesen sok megfigyelés esetén, minimális módosulásokat okozna a paraméterbecslésben. (A torzulás mértéke az $\frac{n}{n+1}$ hányadostól függ, vagyis kisminták esetén azért óvakodjunk a torzítás elbagatellizálásától.)

jelenségre vonatkozó elméleteknek? Kicsit praktikusabban megfogalmazva a kérdést: mennyit adhatunk fel (vagy feladhatunk-e bármit is) az elméletek generálta modellspecifikációból a jobb illeszkedés vagy – nagyon sokszor – a paraméterbecslés kivitelezhetősége érdekében? Erre a problémára bizony nem könnyű korrekt, minden körülmények között alkalmazható választ adni.[24]

Azt hiszem a megoldás (amennyiben egyáltalán ilyen létezik) ismét csak abban rejlik, hogy igyekeznünk kell körültekintően lehatárolni a problémát, egyértelművé tenni, hogy a statisztikai modellezésben az analitikus függvények vagy a sztochasztikus kapcsolatok „végállapotai" (függetlenség, illetve determinisztikus kapcsolat) csak elméleti eshetőségek, a gyakorlatban nem fordulhatnak elő. Nyilván könnyű lenne ezt a kijelentést ellenpéldákkal cáfolni (például, ha egy termék ára p és x darabot akarunk venni belőle, akkor $c = px$ kiadásunk keletkezik). Ám hangsúlyozni szeretném a statisztikai modell esetén a determinisztikus összefüggés csak elméleti eset, amely azért állhat elő, mert a figyelembe veendő egynémely körülményektől eltekintünk (azaz nem vesszük számba, hogy a termék csomagolásáért díjat kell fizetnünk, a vásárolt mennyiség nagyságától árkedvezményt kaphatunk, vagy akár alkudhatunk stb.) Magam is érzem, hogy az előbbiekben egy meglehetősen „erős" állítást fogalmaztam meg, amely sok vitára adhat okot, hiszen nem kevesebbet mondtam, minthogy statisztikára csak bizonytalan körülmények között van szükség, de így gondolom. Világos, hogy itt arról van szó, el kell különítenünk azt az esetet, amikor a tömegjelenség elvben teljes körűen és pontosan megismerhető, attól az esettől, amikor valóban nem állandó (vagyis sztochasztikus) törvények irányítják az eseményeket.[25] Ez nehéz, de szép feladat.

Hogyan interpretáljuk a modellezés eredményeit?

Az adatbázis, a modellből kimaradó hatások, valamint az értelmezési tartomány okozta félreértések után röviden térjünk ki a modellezési eredmények értelmezése során keletkező disszonanciákra is. Természetesen nem kívánom végigvenni a statisztikatudomány valamennyi eljárását, sőt még csak nem is azzal kívánok foglalkozni, hogy egyes bonyolultabb, vagy többször félre értett módszer esetén hogyan hangzik a korrekt értelmezés. Ezzel szemben a továbbiakban három olyan kérdésre kívánok összpontosítani, melyek – tapasztalatom szerint – a legtöbbször okoznak

[24] Vegyük észre, hogy – az adatbázis elégséges volta, illetve a modellből kihagyott hatások elemzése után – immár harmadszor járunk ugyanazon az „ingoványos talajon". A kérdésünk egyszerű: modellezni vagy nem modellezni, hiszen a modell úgysem tökéletes?

[25] A mintavételben járatos kollégákhoz szólva: kicsit hasonlónak érzem mindezt a FAE-minta kiválasztásához (véges elemszámú alapsokaság esetén visszatevéssel, végtelen elemszámú alapsokaság esetén akár visszatevés nélkül is). Igazodási pontnak kiválók a kitűnő matematikai tulajdonságokkal rendelkező eredmények, de megéri-e az információveszteség.

problémát a „kiművelt” statisztikus és a „laikus” felhasználó között. A nyilvánvalóan szubjektív válogatás eredményeképpen röviden áttekintem

– az eredmények verifikálásából,
– az intervallumbecslésből adódó, illetve
– a ceteris paribus elv alkalmazásából

származtatható „kommunikációs réseket”.

Úgy gondolom nem túlzás, ha azt állítom, hogy még a gyakorlott modellező számára is fejtörést okoznak az olyan eseteket, amikor a modell szolgáltatta legvalószínűbb előrejelzés olyan érték, amely elméletileg kizárt. Gondoljunk az egyik legegyszerűbb modellként felfogható számtaniátlag-számításra! Bizonyos – nem túl bonyolult – feltételek fennállásától eltekintve rendszeresen fordul elő, hogy természetes számok átlagolásával racionális számot kapunk, miközben ilyen elméletileg sem lehet az empirikus adatok között.[26] Számos további – az előzőnél lényegesen bonyolultabb – modell esetében találkozhatunk ugyanezzel a jelenséggel, tehát például a dichotom eredményváltozóra vonatkozó becslésünk folytonos érték, vagy a véletlen bolyongás modelljéből következő prognózis, és még sorolhatnánk. Hasonlóan a verifikáció fázisában vehetünk észre analitikus trendmodellekkel kapcsolatos anomáliákat, például amikor a lineáris trend végtelenben vett határértéke az elképzelhető értékkészleten kívül esik, vagy a polinomfüggvény túljut az utolsó trendfordulón, és a továbbiakban már monoton. Miből adódik itt a félreértés? Az átlagfelhasználó mégiscsak a korábban már tárgyalt matematikai gondolkodásmódon nőtt fel, ezért azt hiszi, hogy ha talál egy ellenpéldát, akkor az egészet „ki lehet dobni”. Ezen a szemléleten megint csak sok munkával lehet(ne) változtatni.

Azt hiszem minden gyakorló statisztikaoktató megerősít abban, hogy az egyetemi hallgatók is attól kezdve nem értik a következtetéses statisztikai módszereket, amikor elmondjuk, hogy a becslés segítségével nem teljesen pontos, ráadásul bizonytalan eredményekhez jutunk. Közismert, hogy az intervallumbecslés során a hibahatárt a becslés *pontosságának* mérésére használjuk, vagyis annál pontatlanabb egy becslés, minél szélesebb intervallum a végeredmény, ugyanakkor köztudott, hogy minél szűkebb az intervallum, annál kisebb a becslés megbízhatósága. Úgy tapasztaltam, hogy a felhasználók nagy része a *megbízhatóságot* (amit egyébiránt valószínűségi szintként, vagyis 0 és 100 százalék közötti értékként adunk meg) nem folytonos, hanem

[26] Érdekes módon a statisztikai szempontból teljesen laikus szülők is megértik, hogy gyermekük annak ellenére hozott haza 4,5-ös bizonyítványt, hogy egyetlen tárgyból sem kapott 4,5-ös jegyet; ugyanakkor akár gazdasági végzettségű politikusoknak vagy újságíróknak is nehézséget okoz néha annak megértése, hogy „kerek kulcsokat” tartalmazó adórendszerben, hogyan keletkezhet például 33,2 százalékos átlagos adóterhelés…

kategóriás változóként értelmezi. Ráadásul a kategóriákba rendezés sajátságos:[27] létezik a *teljes bizonyosság* (100 százalékos megbízhatóság), a *nagy valószínűség* (90–95 százalék felett – felhasználója válogatja), a *valószínűbb* (50 százalék felett), valamint a *nem valószínű* (50 százalék alatt). Könnyen látható, hogy az eredmények értelmezésében a problémát az jelenti, hogy míg a megbízhatóságot a felhasználó leegyszerűsítve kezeli, a pontosság megítélése keretében érzékeny az apró változásokra is. Tekintsünk egyetlen, unalomig ismert példát! A standard normális eloszlásra visszavezethető intervallumbecslések esetén a 95 százalékos megbízhatósági szint 95,5 százalékosra (tehát fél százalékpontos, avagy 0,526 százalékos) növelése a hibahatárt mintegy 2 százalékkal (1,0204-szeresére) növeli. Megkérdezhetjük, ha egyébként a „kétszigma-szabály" nem lenne olyan „szép", kérné ezt akárcsak egyetlen megbízó is?

A statisztikai modellezés, pontosabban a modellek alkalmazásának (értelmezés, előrebecslés) egyik, ha nem a legfontosabb alapelve, az ún. *ceteris paribus* elv. Ismeretes, hogy a ceteris paribus elv alkalmazása annyit jelent, hogy a komplex, egymással szoros kölcsönhatásban levő összetett jelenségek modellezése során, egy-egy kiválasztott változó hatását úgy vizsgáljuk, hogy az összes többi hatótényezőt konstansnak tekintjük. Nem biztos, hogy ez minden statisztikával foglalkozó laikus és nem laikus felhasználó előtt világos, de ezt tesszük a standardizálásnál, így járunk el a klasszikus indexszámításnál, felhasználjuk az alapelvet a regressziós modell eredményeinek értelmezésénél stb. Talán nem túlzás azt állítani, hogy a ceteris paribus elv alkalmazása a statisztikában annyira triviális, hogy nagyon sokszor magunk is azt gondoljuk, hogy nem is lehetne másképp.

Miért gondolom azt, hogy a ceteris paribus elv sokszor nem is explicit megfogalmazás melletti alkalmazása számos problémának a forrása? Mert úgy vélem, hogy a „vegytiszta" értelmezésekben a lehatárolás megtehető, sőt az eredmények korrekt interpretálása kifejezetten elvárható, ugyanakkor a laikus felhasználó még ha hallja/olvassa is, hogy „minden mást változatlannak feltételezve" átsiklik ennek a jelentésén, ugyanis a gyakorlatban ez a feltevés sokszor elméletileg kizárt. Tekintsük ismét a már említett Cobb–Douglas típusú termelési függvényt, ahol $\hat{Q} = \hat{\alpha} K^{\hat{\beta}} L^{1-\hat{\beta}}$. Hogyan is szól a becsült $\hat{\beta}$ együttható értelmezése? *„Amennyiben 1 százalékkal növeljük a holtmunka-ráfordítást, a kibocsátás* $\hat{\beta}$ *százalékkal változik, feltételezve, hogy az élőmunka-ráfordítás változatlan."* A laikus ezt meghallgatja, de egy pillanatig sem gondolja, hogy a XXI. században reális lenne annak feltételezése, hogy ve-

[27] A következő kategorizálást semmiképpen sem tudnám tényadatokkal alátámasztani, ugyanakkor – azt hiszem, a statisztikus kollégák megerősítenek – nagyjából mindannyian azt érezzük, hogy ilyen osztályok léteznek. A gondolatmenet szempontjából egyébiránt a kategorizálás tényének van jelentősége és nem az osztópontoknak.

szünk egy új gépet, de nem kell hozzá munkás, aki működteti.[28] A statisztikában járatos azonnal tudja, hogy a multikollinearitás jelenségével állunk szemben, a modellező rögtön bonyolult megoldásokat javasol… A probléma ugyanakkor szerintem nem ez, hanem a „közvélekedés", amely a modellezés során egy kérdésre adott válaszunkat elméletileg tartja hibásnak, és ezért a teljes modellezésről is lemondana. De erről már sok szót ejtettünk.

Bizonyára számos kérdést érdemes lenne az előbbieknél részletesebben tárgyalni, sőt azt gondolom, sok olyan modellezés közben felmerülő problémáról tudunk, melyet a dolgozatban nem is említettem. Arra próbáltam fókuszálni, hogy megemlítsek olyan nehézségeket, melyek a statisztika tananyagokban standardként tárgyalt modellezési lépések között kevésbé mélyen tárgyaltak.

4. Összegzés – Mit tegyünk?

Nyilvánvalóan minden statisztikus célja, hogy a tömegjelenségek leírása érdekében jól hasznosítható, korrekt modelleket készítsen. A statisztikatanárok emellett még azt is szeretnék, hogy a diákjaik megértsék a sztochasztikus gondolkodásmód alapjait, későbbi munkájukban helyesen alkalmazzák a tanultakat.

Ugyanakkor azt gondolom, és őszintén remélem, hogy vélekedésemmel nem állok egyedül, mely szerint a statisztikus oktatóknak-kutatóknak talán ennél „emelkedettebb" céljaik is lehetnek:

> – egyrészt helyére illeszteni a statisztikai modellezést a módszertudományok rendszerében, azaz túllépni végre azon a vélekedésen, hogy „a statisztika a gazdaságtörténet segédtudománya", és elindulni az úton, amely a sztochasztika gondolkodásmódját – önálló tudományként – a matematika szintjére juttatja (a ma divatos szóhasználattal eldönteni végre, hogy „art vagy science");
>
> – másrészt mintegy társadalmi felelősségtudattól vezérelve, népművelő feladatokat ellátni, melynek eredményeképpen a laikus tanuljon, de legalább kérdezzen; a statisztikus korrektül, a figyelmet a korlátokra is felhívva modellezzen, és türelmesen magyarázzon.

Szívből remélem, hogy tanulmányommal kis mértékben hozzájárultam ezen célok megvalósulásához.

[28] Szándékosan vulgarizálom a problémát, nyilván elképzelhető (főként naturália helyett értéken mért változók esetén) olyan eset, amikor az említett „karikatúra" hibás.

Irodalom

HUNYADI L. – RAPPAI G. [1999]: Gondolatok a statisztikáról. *Statisztikai Szemle*. 77. évf. 1. sz. 5–15. old.

HUNYADI L. – VITA L. [2008]: *Statisztika I–II*. Aula Kiadó. Budapest.

MUNDRUCZÓ GY. [1998]: *Útmutatás a statisztikai modellezés gyakorlatához*. KSH Könyvtár és Dokumentációs Szolgálat. Budapest.

SZILÁGYI GY. [2000]: Érteni a számok nyelvén. *Statisztikai Szemle*. 78. évf. 1. sz. 5–12. old.

Summary

The study examines why so many misunderstandings surround even the results of the most carefully prepared statistical models. The author summarizes the most important premises without which the results of statistical modelling not, or not fairly can be interpreted. The paper goes into detail about the area of science classification of statistics; the general form and fundamental idea of stochastic modelling; and some modelling circumstances, which are hopefully trivial for skilful statisticians, but not enough clear for laymen.

The study affects a number of events (target function of the model estimate, measurement error, initial specification, ceteris paribus principle, etc.), the inaccurate application or misunderstanding of which results in an informational gap between modellers and users. In conclusion, the author emphasizes the need for acknowledging statistics as an independent method science and for teaching stochastic modelling as broadly as possible.