TANULMÁNYOK

Moksony Ferenc

A KICSI SZÉP. A DETERMINÁCIÓS EGYÜTTHATÓ ÉRTELMEZÉSE ÉS HASZNÁLATA A SZOCIOLÓGIAI KUTATÁSBAN*

> *„Ezeknek az illeszkedési mutatóknak végzetes vonzerejük van. Bár a hozzáértők rendszerint elismerik, hogy semmit sem jelentenek, magas értékeik mégis büszkeséggel és elégedettséggel töltik el létrehozóikat, bármennyire igyekeznek is titkolni ezeket az érzéseiket*" (Cramer 1987: 253).

Kevés statisztikai mutató örvend akkora népszerűségnek és tiszteletnek a társadalomkutatók körében, mint a determinációs együttható. Az R^2 úgyszólván kötelező tartozéka minden valamirevaló tudományos publikációnak, és sokan szinte megszállottként törekednek a növelésére.[1] Olyan mutató is kevés akad azonban, amelyet gyakrabban használnának fölöslegesen vagy éppen hibásan, és amelyhez több téves értelmezés, megalapozatlan várakozás tapadna. Ennek az írásnak a célja a determinációs együttható értelmezésével és alkalmazásával kapcsolatos néhány probléma áttekintése.

A kutatás célja és a determinációs együttható szerepe

Az R^2 kiszámítása és közlése úgyszólván reflexszerű eljárás a legtöbb kutatónál és eközben rendszerint fel sem merül a kérdés: indokolt-e egyáltalán a mutató használata. A válasz erre a kérdésre alapvetően függ a kutatás céljától. Amennyiben a vizsgálat valamely jelenség *előrejelzésére* irányul, akkor nyilvánvalóan nem mellékes, hogy a magyarázó változó alapján mennyire pontosan tudjuk meghatározni a függő változó értékét; mennyire tudjuk leszorítani a becslési vagy előrejelzési hibát. Ilyenkor valóban indokolt lehet a determinációs együttható figyelembevétele, az R^2 ugyanis többnyire arra utal, hogy a függő változónak a magyarázó változó ismeretében megjósolt értéke kevéssé tér csak el a ténylegestől, vagyis a becslési hiba viszonylag csekély.[2] Hamarosan látni fogjuk azonban, hogy az

* A cikk egy korábbi változatához fűzött értékes megjegyzéseiért köszönettel tartozom Hegedűs Ritának, Lengyel Györgynek és Róna-Tas Ákosnak.

R^2 nagysága nem csupán e hiba mértékétől függ, s ezért ez a mutató csak korlátozottan alkalmas az előrejelzés sikerességének mérésére.

Alapvetően más a helyzet, ha a kutatás célja elméleti *magyarázat ellenőrzése*. Ilyenkor rendszerint tapasztalati következményeket fogalmazunk meg; olyan várakozásokat, amelyek azt fejezik ki, miként kell kinéznie a világnak akkor, ha az általunk kidolgozott magyarázat valóban igaz. Ha például annak az elképzelésnek a helyességét vizsgáljuk, amely szerint a neurózisnak a nők körében tapasztalt nagyobb gyakoriságáért a két nem eltérő társadalmi szerepei, a nőknek a nemek közötti munkamegosztásból eredő nagyobb leterheltsége a felelős, akkor ésszerűnek látszik arra számítani, hogy a férfiak és a nők lelki egészségi állapota városban kevésbé tér el egymástól, mint falun, hiszen a nemi szerepek, a nemek közötti munkamegosztás városban minden bizonnyal kiegyenlítettebb, mint vidéken. Ez a várakozás vagy tapasztalati következmény – és számunkra ez most a fontos – három változó összefüggését, egymásra hatását írja le: nevezetesen azt, hogy a nem hatása a neurózisra függ a település típusától. Márpedig egy változó másikra gyakorolt hatását – e hatás nagyságát és irányát – a standardizálatlan regressziós együttható tükrözi; *a determinációs együttható értéke ebből a szempontból teljesen közömbös*.[3] Egy alacsony R^2 legfeljebb arra utal, hogy a függő változót az általunk vizsgált magyarázó változón kívül még egy sereg más tényező is befolyásolja; ez azonban lényegtelen, hiszen bennünket egy meghatározott oksági kapcsolat érdekel, s nem arra a lehetetlen, egyszersmind fölöttébb kétes értékű feladatra vállalkoztunk, hogy teljes körű leltárt készítsünk valamely jelenség okairól.

Az R^2 és a „magyarázó erő”

A determinációs együtthatóról gyakran állítják, hogy a regressziós modell – illetve az abban szereplő változók – *magyarázó erejét* fejezi ki. Ez a megfogalmazás kétségkívül jól hangzik (sokak számára éppen ezért igen vonzó), azonban meglehetősen félrevezető, ugyanis összekeveri egymással a statisztikai és a tartalmi magyarázatot. Statisztikai értelemben megmagyarázni valamit annyit jelent, hogy a függő változó teljes szóródásának minél nagyobb hányada esik a magyarázó változó egyes értékei vagy kategóriái közé, és minél kisebb hányada marad ezeken az értékeken vagy kategóriákon belül. Ebben a tisztán statisztikai értelemben az R^2 valóban a „megmagyarázott variancia” nagyságát jelzi; ennek azonban az égvilágon *semmi köze a vizsgált jelenség tartalmi magyarázatához.* Gondoljuk csak meg: ha magyarázó változóként magát a függő változót használnánk, akkor az R^2 garantáltan a lehető legnagyobb, éspedig 1 lenne, vagyis a függő változó teljes szóródását meg tudnánk „magyarázni”. Mégis, aligha mondaná bárki, hogy ezáltal akár csak egyetlen lépéssel is közelebb jutottunk a vizsgált jelenség megértéséhez, tartalmi értelemben vett magyarázatához (lásd Lewis-Beck 1993: 16; King 1986: 677).

Az, hogy a determinációs együttható azonosítása a magyarázó erővel mennyire téves lehet, azt az immáron klasszikusnak mondható tankönyvi példával is érzékeltethetjük. A születések száma egy adott településen elég nagy pontossággal megbecsülhető a házak kéményein fészkelő gólyák száma alapján; ha lefuttatunk

egy regressziót, amelyben a magyarázó változó a gólyák száma, a függő változó pedig a születések száma, akkor az R^2 értéke valószínűleg meglehetősen magas lesz. De következik-e ebből, hogy a gólyák száma magyarázza – tartalmi értelemben – a termékenység szintjét? Nyilvánvalóan nem; statisztikai magyarázó erejét – ami a magas R^2-ben tükröződik – ez a változó kizárólag annak köszönheti, hogy erősen korrelál a születésszám valódi meghatározójával, a település típusával. Falun egyrészt gyakoribb a gólya, mint városban, másrészt itt a termékenység is eleve magasabb.

Vegyük észre, hogy pusztán az *előrejelzés* szempontjából ez a probléma voltaképpen nem probléma: ebben a tekintetben tökéletesen mindegy, hogy a magyarázó változó valóban oka-e a függő változónak, vagy az összefüggés látszólagos csupán (Cook–Campbell 1979: 296–297; Elster 1997: 18). Sőt, mivel a valódi oksági tényezők gyakran nehezebben mérhetők, mint a velük korreláló egyéb változók, tisztán gyakorlati megfontolásból ez utóbbiak alkalmasint még hasznosabbak is lehetnek. Egészen más a helyzet, ha nem előrejelzésről, hanem *magyarázatról* van szó. Ekkor már távolról sem közömbös, mi húzódik meg a nagy R^2 mögött: tényleges oksági hatás vagy hamis kapcsolat. Ennek megfelelően ekkor már nagyon is tudatában kell lenni annak, hogy a determinációs együttható magas értéke egyáltalán nem feltétlenül utal valódi oksági magyarázatra.

Még egy dolgot érdemes megemlíteni ezen a ponton. A determinációs együtthatót gyakran használják a változók *relatív* – egymáshoz viszonyított – magyarázó erejének megállapítására. Ez az alkalmazás rendszerint – bár nem szükségszerűen – a lépésenkénti regresszióhoz kötődik; olyan eljáráshoz, ami – ha lehet – még kétesebb értékű, mint az R^2 nyakló nélküli növelése. A lépésenkénti regresszió általában annak alapján állít fel fontossági sorrendet az egyes magyarázó változók között, hogy milyen mértékben járulnak hozzá a determinációs együttható növeléséhez. Ezzel nem is volna különösebb baj, ha a magyarázó változók függetlenek lennének egymástól; ekkor ugyanis minden változóhoz egyértelműen hozzá lehetne rendelni azt az R^2-növekményt vagy „magyarázó erőt", ami kizárólag neki tulajdonítható. A gyakorlatban azonban a magyarázó változók rendszerint többé-kevésbé erősen *korrelálnak egymással*. Ebben az esetben a „magyarázó erőt" már nem lehet egyértelműen hozzárendelni az egyes változókhoz; túl azon a mértéken, ami minden változót a „saját jogán" megillet, van egy olyan rész is, ami közös, ami egyiknek sem kizárólagos „tulajdona". Az, hogy ezt a közös „magyarázó erőt" melyik változó kapja meg, *a változók bevonásának sorrendjétől függ*: az a változó, amely elsőként kerül be a modellbe, saját részén kívül „magával viszi" ezt a közös részt is, és így aránytalanul fontosnak, jelentősnek látszik; annak a változónak pedig, amelyet másodikként vonunk csak be, a közös részből már semmi sem marad, és így a ténylegesnél kevésbé fontosnak tűnik. Korreláló magyarázó változók esetén tehát az R^2-növekmény mértéke nem használható annak megítélésére, melyik változó fontosabb, melyiknek nagyobb a „magyarázó ereje", ez ugyanis teljes egészében attól függ, milyen sorrendben vonjuk be őket az elemzésbe. (Minderről bővebben lásd Lewis-Beck 1978; Pedhazur 1982: 167–171; Kennedy 1992: 63–64.)

Hogy mennyire hibás következtetésekhez vezethet, ha az R^2-növekmény alapján foglalunk állást egy változó súlyáról, szerepéről, azt a gólyákkal és a születések számával kapcsolatos iménti példával is érzékeltethetjük. Tegyük föl, hogy a

termékenység szintjét két, egymással korreláló változóval: a gólyák számával és a település típusával próbáljuk megmagyarázni, és arra vagyunk kíváncsiak, e két tényező közül melyik a fontosabb. Tegyük föl továbbá, hogy valamilyen oknál fogva – mondjuk, apró mérési hiba vagy más ehhez hasonló jelentéktelen dolog miatt – a gólyák száma hajszálnyival erősebben korrelál a termékenységgel, mint a másik magyarázó változó, a település típusa. Ebben a helyzetben valószínűleg a gólyák száma kerül be elsőként a modellbe – hiszen a beválasztás szempontja az első szakaszban általában a függő változóval való egyszerű korreláció mértéke –, magával vive annak a közös „magyarázó erőnek" a teljes egészét is, amely pedig részben a másik változót, a településtípust illetné meg. Ez utóbbi változónak így aztán már semmi sem marad a közös „magyarázó erőből", és ennek megfelelően kevésbé fontosnak, kisebb „magyarázó erejűnek" látszik. Történik mindez annak ellenére, hogy oksági szempontból nyilvánvalóan épp a településtípus a fontos, és a gólyák száma a lényegtelen. Ha tehát pusztán az R^2-növekmény alapján döntünk, akkor kihagyjuk a valódi oksági tényezőt, és bevonjuk azt a változót, amelynek a hatása látszólagos csupán.

Az R^2 és az „illeszkedés szorossága"

Másik gyakori nézet szerint a determinációs együttható a regressziós modell *illeszkedését* méri; azt, hogy a regresszió segítségével a függő változó értékére adott becslések mennyire esnek közel a tényleges értékekhez; vagy – képszerűbben fogalmazva – hogy az adatpontok mennyire „simulnak rá" a regressziós egyenesre. Láttuk, hogy bár olyan vizsgálatokban, amelyek elméleti magyarázat ellenőrzésére irányulnak, ennek a dolognak nincs túl nagy jelentősége, azokban a kutatásokban, amelyeknek célja az előrejelzés, nem lényegtelen a becslések pontossága. Ilyen esetben tehát valóban szükség lehet az illeszkedés valamiféle mutatójára, kérdés azonban, az R^2-e a legalkalmasabb erre a feladatra.

Az általános vélekedéssel ellentétben a determinációs együttható csak korlátozottan használható a regressziós modell illeszkedésének mérésére. E mutató értéke ugyanis nem csupán attól függ, mennyire szorosan tömörülnek az adatpontok a regressziós egyenes körül – vagyis mennyire kicsi a becslési hiba –, hanem attól is, mekkora a magyarázó változó szórása. *Ugyanolyan illeszkedés nagyobb R^2-et eredményez, ha a magyarázó változó értékei szélesebb sávban szóródnak.* A szórásnak ez a hatása világosan kitűnik az alábbi egyenlőségből, amelyben $\hat{Y}$ a függő változó becsült értéke, $\overline{Y}$, illetve $\overline{X}$ a függő, illetve a magyarázó változó átlaga, b_1 pedig a magyarázó változó hatását kifejező standardizálatlan regressziós együttható:[4]

$$\sum(\hat{Y}-\overline{Y})^2 = b_1^2 * \sum(X-\overline{X})^2 \qquad (1)$$

Látható, hogy a regressziónak tulajdonítható eltérésnégyzet-összeg – ami a bal oldalon szerepel, s ami nem más, mint a determinációs együttható számlálója – függ a magyarázó változó szóródásától, ami a jobb oldalon áll. Feltéve, hogy b_1 értéke nem módosul, minél szélesebb sávban szóródnak az X értékek, annál nagyobb a regressziónak tulajdonítható eltérésnégyzet-összeg, és így – amennyiben a reziduális eltérésnégyzet-összeg állandó – annál nagyobb az R^2 értéke is.

Túl a tisztán matematikai bizonyításon, érdemes ezt a kérdést a kutatási gyakorlat oldaláról is szemügyre venni. A társadalomtudományokban viszonylag ritkán adódik alkalom kísérletezésre, a magyarázó változó aktív befolyásolására; rendszerint kénytelenek vagyunk beérni a passzív megfigyeléssel, a változó tőlünk függetlenül kialakult értékeinek puszta feljegyzésével. A mintavétel révén olykor mégis lehetőségünk van arra, hogy a magyarázó változó eloszlását módosítsuk. Ezt tesszük például akkor, amikor szándékosan olyan eseteket vonunk be az elemzésbe, amelyek a magyarázó változó *szélső pontjait* képviselik, vagy amikor egy dichotóm magyarázó változó kategóriáiból *azonos számú* esetet választunk ki. Mindezek a mintavételi „trükkök" növelik a magyarázó változó szórását[5], ezen keresztül pedig a determinációs együttható értékét.

A mintavételnek ezt a hatását jól szemléltetik Blalock (1964: 114–124), Ezekiel és Fox (1970: 18. fejezet), valamint Weisberg (1985: 74–76) munkái, amelyekben a szerzők mesterségesen módosítják a magyarázó változó szórását, majd megvizsgálják, miként befolyásolja ez a különböző statisztikai mutatók értékét. Ez a fajta szimuláció vagy módszertani kísérlet azért is tanulságos, mert rávilágít arra, hogy miközben az R^2 értéke számottevően ingadozik aszerint, hogy széles sávban szóródnak a magyarázó változó értékei, addig a reziduumok szórása – a regressziós becslés standard hibája – nagyjából állandó marad. Ez utóbbi mutató tehát nem függ szisztematikusan a magyarázó változó szórásától[6], és így a determinációs együtthatónál alkalmasabbnak tűnik a regressziós modell illeszkedésének, a becslési hiba nagyságának a mérésére.[7] A reziduális szórás további előnye, hogy az illeszkedés „jóságát" a függő változó természetes mértékegységében fejezi ki – ellentétben az R^2 -tel, ami dimenzió nélküli mutató, és ezért általában nehezebben kapcsolható közvetlenül a vizsgált jelenséghez (Achen 1982: 61–64).

Eddig arról beszéltünk, hogy amennyiben a mintavétel folyamán képesek vagyunk mesterségesen növelni a magyarázó változó szórását, akkor a determinációs együttható szinte tetszőlegesen változtatható; épp ezért ilyenkor rendkívül körültekintőnek kell lenni e mutató értelmezésekor. Indokolt lehet azonban az óvatosság fordított esetben is. Gyakori jelenség, hogy a magyarázó változó szórása éppenséggel túl alacsony, és nincs lehetőség a növelésére. Ez a helyzet akkor, ha a magyarázó változó *ritka előfordulású* eseményre vonatkozik, például arra, hogy a vizsgált személy követett-e el fiatal korában öngyilkossági kísérletet vagy súlyosabb bűncselekményt. Az ilyen személyek a teljes mintának vélhetőleg viszonylag csekély hányadát képezik csupán, vagyis – technikailag kifejezve – a magyarázó változó eloszlása meglehetősen ferde: az esetek zöme az egyik kategóriában összpontosul, és a másik kategóriába csak kevés megfigyelés tartozik. Ennek következtében a magyarázó változó szórása viszonylag kicsi lesz, hiszen egy dichotóm változó varianciája egyenlő a két kategória relatív gyakoriságának a szorzatával. Minél eltérőbbek a relatív gyakoriságok – minél

ferdébb a változó eloszlása –, annál kisebb a szorzat értéke, azaz annál csekélyebb a szórás. Ritka események hatásának vizsgálatakor tehát a determinációs együttható értéke különösen csalóka lehet: a hatás – amit a standardizálatlan regressziós együtthatóval vagy annak megfelelő más mutatóval mérhetünk – nagy lehet annak ellenére, hogy az R^2 viszonylag alacsony (erről bővebben lásd Glenn–Shelton 1983).

Az R^2 és a „tökéletes modell”

Gyakran találkozhatunk azzal a nézettel, miszerint a determinációs együttható *a regressziós modell „tökéletességét” vagy „teljességét”* jelzi. Minél magasabb az R^2 értéke, annál jobb – úgymond – a modell; annál hívebben tükrözi a tényleges összefüggéseket. Valóban, sok kutató egyfajta minőségtanúsító pecsétként kezeli a determinációs együtthatót; olyan védjegyként, amely önmagában szavatolja az elvégzett munka értékét, a felállított modell helyességét. Ez a felfogás azonban alapvetően téves, az a törekvés pedig, ami ebből a felfogásból fakad, és ami az R^2 mindenáron való növelésére irányul, teljesen értelmetlen. Először is, tökéletes modell nincs; nem azért, mert a tökéletesség elérhetetlen, hanem azért, mert a modell definíció szerint a valóság leegyszerűsített és így szükségképpen pontatlan képe (King 1991: 1048). Olyan kép, amely bizonyos részeket tudatosan kiemel, felnagyít, másokat viszont szándékosan árnyékban hagy. Minden modell meghatározott elméleten nyugszik és ennek az elméletnek a hangsúlyait tükrözi. És minden modell csak egy másik, a sajátunkéval versengő elmélet talajáról bírálható; nem pedig annak alapján, hogy az R^2 értéke túlságosan alacsony. *Amikor a regressziós egyenletet újabb változókkal bővítjük, a cél nem a determinációs együttható növelése; nem valamiféle teljes vagy végső modell elérése, hanem a különféle alternatív magyarázatok kiküszöbölése* (Achen 1982: 52). Az, hogy valamely modell jó vagy rossz, elméleti érveléssel dönthető csak el; az R^2-nek ebbe nincsen beleszólása. Baj is volna, ha lenne; ha gépies számításokkal lehetne pótolni a tartalmi gondolkodást.

Azt a tényt, hogy a regressziós modell „jósága” mennyire nem a determinációs együttható értékén múlik, egy példával érzékeltethetjük. Tegyük föl, hogy olyan képzési program hatékonyságát vizsgáljuk, amelynek célja a munkanélküliek elhelyezkedésének az elősegítése. Tegyük föl továbbá, hogy a részvétel a programban *önkéntes*: azok az állástalanok, akiket érdekel a dolog, igénybe veszik a felkínált lehetőséget, a többiek pedig kimaradnak belőle. A két csoportot összehasonlítva megállapítjuk, hogy azok, akik részt vettek a képzésben, átlagosan rövidebb idő alatt találtak újra munkát, mint azok, akik nem vettek részt. Tudjuk persze, hogy épp az önkéntesség miatt ez az eredmény nem bizonyítja a képzés hatékonyságát: elképzelhető, hogy azok, akik a részvétel mellett döntöttek, eleve jobban törekedtek az újbóli elhelyezkedésre, s így a program nélkül is könnyebben találtak volna állást. Az is lehetséges, hogy a résztvevők fiatalabbak és iskolázottabbak – vagyis olyan tulajdonságokkal rendelkeznek, amelyek önmagukban megkönnyítik az elhelyezkedést. Ahhoz, hogy a képzés tényleges hatását megállapítsuk, mindezeket a tulajdonságokat kontrollváltozóként be kell vonni az

elemzésbe. Ezzel azonban – a magyarázó változók körének kibővítésével – egyszersmind a determinációs együttható értékét is minden valószínűség szerint jócskán megnöveljük, vagyis modellünk – pusztán az R^2 nagysága alapján ítélve – igencsak jónak látszik.

Képzeljük most el, hogy a részvétel a programban nem önkéntes, hanem *randomizálást* alkalmazva a véletlenre bízzuk annak eldöntését, hogy az állástalanok közül ki kerül a képzésben részesülők csoportjába. Ebben az esetben a programban részt vevők és az abból kimaradók összetétele minden lehetséges szempontból nagyjából azonos lesz – körülbelül ugyanannyi lesz a fiatalok és az idősek, az iskolázottak és az iskolázatlanok aránya, és ugyanígy durván azonos lesz azoknak az aránya, akik eleve nagyobb igyekezettel próbálnak elhelyezkedni. Mi következik ebből? Az, hogy a program tényleges hatásának megállapítása szempontjából ezúttal nincs szükség a korábban használt kontrollváltozókra, hiszen most sem az életkor, sem az iskolázottság, sem semmilyen más tulajdonság nem korrelál a képzésben való részvétellel.[8] Ez azonban – a kontrollváltozók kihagyása – egyszersmind azt is jelenti, hogy az R^2 értéke valószínűleg lényegesen alacsonyabb lesz, mint az előző esetben, amikor maguk a munkanélküliek döntötték el, részt vesznek-e a programban. De vajon mondhatjuk-e azt ennek alapján, hogy ez a második modell rosszabb, kevésbé „tökéletes", mint az első? Aligha; sőt, minden bizonnyal épp az ellenkezője az igaz, hiszen az oksági összefüggések feltárása szempontjából a randomizált vizsgálatoknál nehéz tökéletesebbet elképzelni.

A nagy R^2 azonosítása a „tökéletes" modellel egy másik szempontból is alapvetően hibás. A determinációs együttható növelésének lázában a kutatók a regressziós modellt gyakran *az adatpontok véletlenszerű ingadozásaihoz illesztik* (Kennedy 1992: 70), figyelmen kívül hagyva, hogy minden adathalmaz csupán minta, egyike a számtalan lehetséges adathalmaznak. Ha történetesen másik adathalmazt figyeltünk volna meg, akkor – a véletlen szeszélye folytán – az adatpontok eloszlása némileg más képet mutatna, és ehhez az eloszláshoz már aligha illeszkedne ugyanolyan jól a modellünk. Akkor hát keressünk másik modellt, ami ehhez az adathalmazhoz hibátlanul illeszkedik? De még újabb mintához már ez a modell sem illeszkedne teljesen – és így tovább a végtelenségig. Nem sokat ér az a „tökéletes" modell, az a nagy R^2 , ami csak egyetlen konkrét mintára érvényes. A modell illesztése során mindig csak addig a mértékig érdemes teljességre, tökéletességre törekednünk, ameddig az adatpontok még a vizsgált jelenségben rejlő törvényszerűséget tükrözik – azt, ami mintáról mintára nagyjából állandó –, nem pedig a puszta esetlegességet, a véletlen ingadozást. Ez is csak azt a régi bölcsességet igazolja, hogy a kevesebb néha több.

Ezt a bölcsességet hagyják figyelmen kívül egyebek között azok, akik sportot űznek a minél pontosabb görbeillesztésből. Ők nem elégszenek meg az egyenessel, hanem másodfokú görbével próbálkoznak; majd a másodfokú görbét felcserélik harmadfokúra; aztán a harmadfokút egy negyedfokúra; míg végül eljutnak az n–1-ed fokú görbéig, amely az n számú adatpont mindegyikén átmegy, vagyis tökéletes illeszkedést, csodálatosan magas R^2-et nyújt – csak éppen teljesen értelmetlen, mivel kizárólag az adott mintát, az éppen megfigyelt n esetet képviseli, és így semmi értéke nincs „annak az összefüggésnek a feltárásában, amely valószínűleg érvényes

abban a sokaságban, amelyből a mintában szereplő megfigyeléseket vettük" (Ezekiel–Fox 1970: 119; lásd még Lieberson 1985: 93).

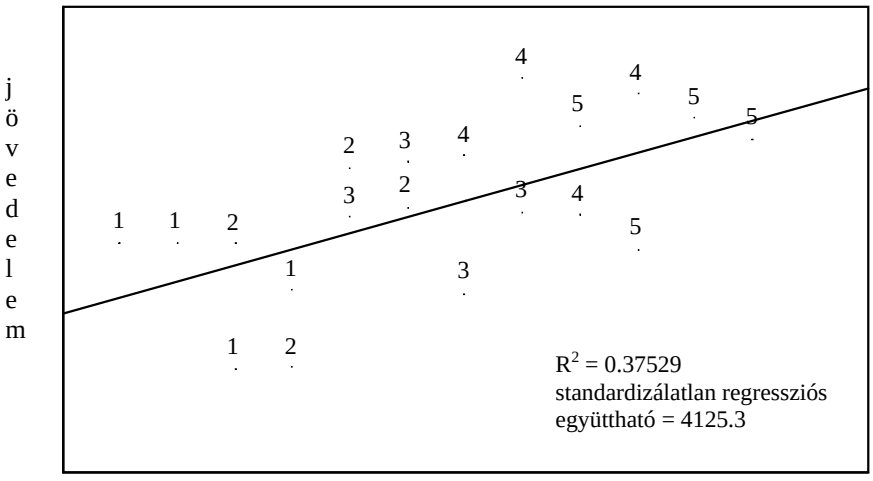
Még egy dolgot érdemes ezen a ponton megemlíteni. Korábban arról beszéltünk, hogy azokban a kutatásokban, amelyeknek célja egy jelenség előrejelzése, a nagy R^2 általában örvendetes tény, és valóban, a legtöbb tankönyv a sikeres előrejelzés feltételeként említi a determinációs együttható magas értékét (például Lewis-Beck 1993: 16). Amikor azonban a nagy R^2 pusztán annak eredménye, hogy modellünket az éppen megfigyelt adatok esetlegességeihez igazítottuk – vagy, ahogyan sokszor nevezik, tőkét kovácsoltunk a véletlenből (Kennedy 1992: 70) –, akkor a nagy R^2 egyáltalán nem feltétlenül garantálja, hogy a modell az adott konkrét mintán kívül is ugyanolyan tökéletes lesz. Sőt, Mayer elemzései éppenséggel azt bizonyítják, hogy „amennyiben olyan hipotézisek érdekelnek bennünket, amelyek a minta által felölelt időszakon túl is érvényesek, akkor az illeszkedés mutatói igen gyenge iránymutatást jelentenek csupán" (Mayer 1975: 882).

Azt, hogy mennyire gyenge lehet ez az iránymutatás, megtudhatjuk Lieberson (1985: 97–99) találó példájából. Képzeljük el, hogy nagy számú szabályos pénzdarabot dobunk fel, mindegyiket egymás után tízszer. Ha megszámoljuk, a tízből hány alkalommal kaptunk „fej"-et, az eredmény érménként változó lesz. Lesznek pénzdarabok, amelyek esetében a „fej"-ek száma csupán kettő vagy három – az elméletileg várt öt helyett –, lesznek azonban olyanok is, amelyek esetében nyolc, kilenc, sőt akár tíz „fej"-et kapunk. Tegyük föl, hogy megpróbáljuk megmagyarázni ezt az ingadozást; azt, hogy a „fej"-ek száma egyes érméknél miért olyan alacsony, másoknál pedig miért olyan magas. Ha elég kitartóak és türelmesek vagyunk, rábukkanhatunk a pénzdaraboknak azokra az egyedi vonásaira, amelyek összefüggenek a „fej"-ek számával. Ilyen vonás lehet például az, hogy mikor készült az adott érme, hol gyártották, a számos pénzdarab közül hányadikként dobtuk fel stb. Bármily szorgalmasak vagyunk is azonban, bármennyi tulajdonságot veszünk is figyelembe, erőfeszítésünknek az égvilágon semmi értelme: azok az érmék ugyanis, amelyek az általunk elvégzett dobássorozatban nagy számú „fej"-et eredményeztek, és amelyeknek a tulajdonságait oly lázasan kutattuk, újabb sorozatban *pontosan ugyanakkora* valószínűséggel eredményeznek nagy számú „fej"-et, mint azok a pénzdarabok, amelyek esetében az első körben a „fej"-ek száma igen alacsony volt. Míg tehát magyarázó modellünk kiválóan illeszkedik az adott konkrét dobássorozat eredményéhez, az érmék tulajdonságainak szerepét, előrejelző képességét illetően teljesen értéktelen. Mindennek alapján Lieberson joggal vonja le a következtetést, hogy a „megmagyarázandó variancia" szükséges mértékét alkalmasint túl is lehet becsülni, és ez a túlbecsülés kedvezőtlen következményekkel járhat. Egyebek között arra ösztönzi a kutatót, hogy *ad hoc* magyarázatok kitalálásával növelje az R^2 értékét, vagyis olyan eljárásra csábít, aminek hosszú távon nincs semmi haszna.

Az R^2 és a megfigyelések aggregálása

Bizonyára sokaknak feltűnt már, hogy azokban a vizsgálatokban, amelyek régiókat vagy országokat hasonlítanak össze egymással, az R^2 értéke rendszerint lényegesen magasabb, mint az egyének megkérdezésén alapuló kérdőíves kutatásokban. Ennyivel okosabbak lennének a területi elemzéseket végzők, mint azok, akik a *survey* módszerét választják? Ennyivel jobb, tökéletesebb modelleket tudnának felállítani? A kérdés bonyolult, az azonban egyértelmű, hogy önmagában a magasabb R^2 nem bizonyítja ezt. Ez ugyanis alapvetően nem a kutató képességeinek, hanem az adatok aggregálásának köszönhető: amikor az egyénekre vonatkozó megfigyeléseket csoportokba vonjuk össze, és az eredetiek helyett ezekkel a csoportosított adatokkal dolgozunk, az adatpontok általában a korábbinál jobban „rásimulnak" a regressziós egyenesre, növelve ezzel a determinációs együttható értékét. Az aggregálásnak ezt a hatását szemlélteti az alábbi két, hipotetikus adatokon alapuló rajz. Az *1. ábra* 5 különböző régióban lakó 20 egyén iskolai végzettségének és jövedelmének az adatait tartalmazza; az adatpontok melletti számok a lakóhelyet – a régió sorszámát – jelölik.

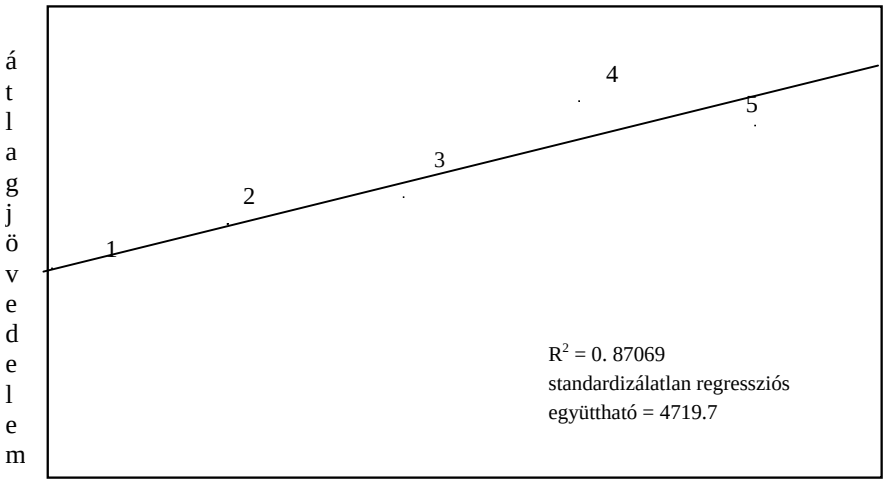
1. ábra
Egyénekre vonatkozó adatok



Látható, hogy az ugyanabban a régióban élők jövedelme különbözik egymástól; az azonos sorszámot viselő egyénekre vonatkozó adatok szóródnak az adott régió átlaga körül. Ez a szóródás „tűnik el" akkor, amikor az adatokat régiónként aggregáljuk, s az egyéni adatok helyett a régiók átlagait használjuk. Ennek eredménye pedig az a rendkívül szoros illeszkedés, amit a *2. ábra* mutat, és amit az R^2 ma-

gas értéke (0.87) is tükröz. (Érdemes megjegyezni, hogy miközben a determinációs együttható két és félszeresére nőtt, a standardizálatlan regressziós együttható alig változott. Általában elmondható, hogy ez utóbbi mutató kevésbé érzékeny az adatok aggregálására.)[9]

2. ábra
Aggregált adatok



Az aggregálás imént bemutatott hatása mögött általánosabb összefüggést ismerhetünk fel. A determinációs együttható értékét döntően meghatározza, hogy mekkora azoknak az *egyéb tényezőknek* a súlya, szerepe, amelyek szintén befolyásolják a függő változót, ám nem korrelálnak az általunk vizsgált magyarázó változóval (Darlington 1990: 19). Ha ezeknek az egyéb tényezőknek – amelyeket a regressziós modell hibatagjában foglalunk össze, és amelyeket az elemzés során „zavaró változókként" kezelünk – csökken a súlya, akkor,– feltéve, hogy minden más változatlan, az R^2 értéke nő. Az adatok aggregálása az előző példában éppen ilyen csökkenést eredményezett: az egyes egyénekre vonatkozó megfigyelések régiónkénti átlagolásával mintegy kiszűrtük vagy közömbösítettük a jövedelmet meghatározó számtalan tényező jelentős részét (Blalock 1964: 99–101, 112–114).[10]

Befejezés

Áttekintve a determinációs együtthatóval kapcsolatos különféle értelmezéseket, rávilágítva e mutató fogyatékosságaira, befejezésül hasznos lehet szemügyre venni egy olyan formulát, amely mintegy összefoglaló képet nyújt az R^2-et befolyásoló tényezőkről, és ezáltal segíthet jobban megérteni e mutató természetét.[11] Ehhez első lépésként idézzük fel az (1) egyenlőséget:

$$\sum(\hat{Y} - \bar{Y})^2 = b_1^2 * \sum(X - \bar{X})^2$$

Emlékezzünk, ennek az egyenlőségnek a bal oldala nem egyéb, mint a regressziónak tulajdonítható eltérésnégyzet-összeg, vagyis az R^2 számlálója.

Ismeretes, hogy a teljes eltérésnégyzet-összeg – tehát az R^2 nevezője – két részből, a regressziónak tulajdonítható és a maradék vagy reziduális négyzetösszegből áll:

$$\sum(Y - \bar{Y})^2 = \sum(\hat{Y} - \bar{Y})^2 + \sum(Y - \hat{Y})^2 \qquad 2$$

Helyettesítsük most be a (2) egyenlőségbe az (1) egyenlőséget:

$$\sum(Y - \bar{Y})^2 = b_1^2 * \sum(X - \bar{X})^2 + \sum(Y - \hat{Y})^2$$

Mindezek alapján a determinációs együtthatót a következőképpen írhatjuk fel:

$$R^2 = \frac{b_1^2 * \sum(X - \bar{X})^2}{b_1^2 * \sum(X - \bar{X})^2 + \sum(Y - \hat{Y})^2}$$

Szavakkal ezt így fogalmazhatjuk meg:

$$R^2 = \frac{\text{hatás nagysága}^2 * \text{X szóródása}}{\text{hatás nagysága}^2 * \text{X szóródása} + \text{illeszkedés szorossága}}$$

Ebből jól látható, hogy a determinációs együtthatóban háromféle tényező keveredik: a magyarázó változó hatása, ennek a változónak a szóródása, és végül a regressziós modell illeszkedésének a „jósága” vagy szorossága. Éppen mert egyszerre ennyi különböző tényezőtől függ, az R^2 ezek egyikének mérésére sem igazán alkalmas. Sem a magyarázó változó hatását, annak nagyságát nem tükrözi pontosan, sem pedig

a regressziós modell illeszkedését. Mindkét feladatra jobb mutatók állnak rendelkezésünkre: a hatás nagyságának mérésére a standardizálatlan regressziós együttható, az illeszkedésére pedig a becslés standard hibája. Mindezek fényében az a tisztelet, ami a determinációs együtthatót rendszerint övezi, nem tűnik megalapozottnak; népszerűségét ez a mutató alighanem inkább retorikai értékének, mintsem tényleges teljesítményének köszönheti.

Jegyzetek

[1] A szakirodalom gyakran különbséget tesz r^2 és R^2, „egyszerű" és többszörös determinációs együttható között. Mivel mondanivalóm egyformán vonatkozik mindkét mutatóra, fölöslegesnek ítéltem e megkülönböztetés hangsúlyozását, és az „R^2", illetve a „determinációs együttható" kifejezéseket felváltva, azonos értelemben használtam. Ez a némi pongyolaság – úgy gondolom – nem okoz majd félreértést, viszont gördülékenyebbé teszi a szöveget.

[2] Az előrejelzés problémakörén belül speciális esetnek tekinthető az a bizonyos fokig módszertani jellegű feladat, amikor egy változó valamilyen okból hiányzó értékeit igyekszünk pótolni más változóknak és az e változók hatását kifejező regressziós együtthatóknak a felhasználásával. A regresszióelemzésnek erre a fajta alkalmazására példa a foglalkozások presztízspontszámának meghatározása a foglalkozások egyéb jellemzői alapján (Loether–McTavish 1980: 362–363), de az ún. kisterületi becslésnél is találkozunk ezzel a megközelítéssel (Marton 1985: 68–69; Ericksen 1973).

[3] Ezt még azok a szerzők is elismerik, akik egyébként védelmükbe veszik a determinációs együtthatót. Lewis-Beck és Skalaban például így fogalmaz: „amikor a kutató X [változó] hatására kíváncsi, az R^2 -nek kevés haszna van. Ebben az esetben a figyelmet a megfelelő regressziós együtthatóra és annak standard hibájára kell fordítani" (Lewis-Beck–Skalaban 1991: 169).

[4] Az egyenlőség bizonyításához először is írjuk föl a regressziós egyenletet:

$$\hat{Y} = b_0 + b_1 * X$$

ahol $\hat{Y}$ a függő változó becsült értéke, X a magyarázó változó, b_0 és b_1 pedig a regressziós együtthatók. Mivel

$$b_0 = \overline{Y} - b_1 \overline{X}$$

ahol $\overline{X}$ és $\overline{Y}$ a magyarázó, illetve a függő változó átlaga, ezért

$$\hat{Y} = (\overline{Y} - b_1 \overline{X}) + b_1 * X$$

Emeljük ki a b_1 együtthatót, Y átlagát pedig vigyük át a bal oldalra:

$$(\hat{Y} - \overline{Y}) = b_1 (X - \overline{X})$$

Végül emeljük négyzetre és összegezzük minden megfigyelésre az egyenlőség mindkét oldalát:

$$\sum(\hat{Y} - \bar{Y})^2 = b_1^2 * \sum(X - \bar{X})^2$$

[5] Egy dichotóm változó varianciája ugyanis nem más, mint a két kategória relatív gyakoriságának a szorzata; ez a szorzat pedig akkor maximális, ha az összeszorzandó relatív gyakoriságok azonosak.

[6] Ennek feltétele azonban a homoszkedaszticitás, vagyis az, hogy a hiba szórása a magyarázó változó minden értéke esetében azonos legyen.

[7] Mindazonáltal, ha a becslési hibának közvetlen gyakorlati jelentősége van, akkor a regressziós becslés standard hibája nem szükségképpen a legjobb választás. Ez a mutató ugyanis a megfigyelt és a becsült értékek közötti eltérések négyzetén alapul, és ily módon nagyobb súlyt ad a nagyobb, és kisebb súlyt ad a kisebb eltéréseknek. Elképzelhető azonban, hogy a becslési hibák gyakorlati következményei – például a velük járó költségek – szempontjából minden hiba egyformán lényeges; ha ez a helyzet, akkor az eltérések négyzete helyett indokoltabb lehet azok abszolút értékét használni. (A négyzetes és az abszolút hibák közötti választás kérdéséről bővebben lásd Berk 1986; az előrejelzési hibák költségeinek figyelembevételéről általában pedig lásd Goodman 1966.)

[8] Más kérdés, hogy a program hatását kifejező regressziós együttható standard hibájának csökkentése érdekében a randomizálás ellenére is hasznos lehet e kontrollváltozók szerepeltetése, ez ugyanis mérsékli a reziduális szórást, ezen keresztül pedig a standard hibát.

[9] Ez azonban nem jelenti azt, hogy az aggregálás sohasem befolyásolja a standardizálatlan regressziós együttható értékét. Amennyiben az adatok csoportosítása nyomán specifikációs hiba jön létre, ez a mutató is torzul. Az aggregálásnak a különféle statisztikai mutatókra gyakorolt hatásáról bővebben lásd például Blalock 1964; Langbein–Lichtman 1978; Hanushek et al. 1974.

[10] Egy másik módja annak, hogy a „zavaró változók" szerepét mérsékeljük, s ezáltal a vizsgált magyarázó változó relatív súlyát, fontosságát növeljük, a függő változó pontosabb mérése.

[11] Az alábbi levezetéshez az ötletet Christopher Achen (1982: 63) tanulmánya adta.

Hivatkozások

Achen, Ch. 1982. *Interpreting and Using Regression*. Beverly Hills–London: Sage Publications

Berk, R. A. 1986. *How Aapplied Sociology Can Ssave Basic Sociology*. Unpublished manuscript.

Blalock, H. 1964. *Causal Inferences in Nonexperimental Research*. Durham, N. C.: University of North Carolina Press

Cook, Th.–D. T. Campbell 1979. *Quasi-Experimentation. Design and Analysis Issues for Field Settings*. Boston etc.: Houghton Mifflin Co.

Cramer, J. S. 1987. Mean and Variance of R^2 in Small and Moderate Samples. *Journal of Econometrics*, 35, 253–266.

Darlington, R. 1990. *Regression and Linear Models*. New York etc.: McGraw–Hill Publishing Co.

Elster, J. 1997. *A társadalom fogaskerekei*. Osiris Kiadó

Ericksen, E. P. 1973. A Method for Combining Sample Survey Data and Symptomatic Indicators to Obtain Estimates for Local Areas. *Demography*, 10, 137–160.

Ezekiel, M.–K- Fox 1970. *Korreláció- és regresszió-analízis. Lineáris és nem-lineáris módszerek*. Budapest: Közgazdasági és Jogi Könyvkiadó

Glenn, N. D.–B. A. Shelton 1983. Pre-Adult Background Variables and Divorce: a Note of Caution about Overreliance on Explained Variance. *Journal of Marriage and the Family*, 45: 405–410.

Goodman, L. 1966. Generalizing the Problem of Prediction. In: P. F. Lazarsfeld–M. Rosenberg (eds.) *The Language of Social Research*. 5th ed., Toronto, 277–281.

Hanushek, E. A. et al. 1974. Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy. *Political Methodology*, 1, 89–107.

Kennedy, P. 1992. *A Guide to Econometrics*. Oxford, UK.–Cambridge, USA Blackwell Publishers

King, G. 1986. How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science*, 30, 666–687.

– 1991. „Truth" is Stranger than Prediction, more Questionable than Causal Inference. *American Journal of Political Science*, 35, 1047–1053.

Langbein, L. I.–A. J. Lichtman 1978. *Ecological Inference*. Beverly Hills–London: Sage Publications

Lewis-Beck, M. 1978. Stepwise Regression: a Caution. *Political Methodology*, 5, 213–240.

– 1993. Applied Regression: an Introduction. In: M. Lewis-Beck (ed.) *Regression Analysis. International Handbooks of Quantitative Applications in the Social Sciences*,. 2. London–Thousand Oaks, CA–New Delhi: Sage Publications

Lewis-Beck, M.–A. Skalaban 1991. The R-Squared: Some Straight Talk. *Political Analysis*, 2, 153–171.

Lieberson, S. 1985. *Making it Count. The Improvement of Social Research and Theory*. Berkeley–Los Angeles–London: University of California Press

Loether, H. J.–D. G. McTavish 1980. *Descriptive and Inferential Statistics: an Introduction*. Boston etc.: Allyn and Bacon, Inc.

Marton Á. (szerk). 1985. *Területi és egyéb szempontok szerint részletezett statisztikai mutatószámok becslése*. Budapest: Központi Statisztikai Hivatal

Mayer, T. 1975. Selecting Economic Hypothesis by Goodness of Fit. *Economic Journal,* 85, 877–883.

Pedhazur, E. 1982. *Multiple Regression in Behavioral Research*. 2nd ed. Forth Worth etc.: Harcourt Brace Jovanovich College Publishers

Weisberg, S. 1985. *Applied Linear Regression*. 2nd ed. New York etc.: John Wiley & Sons