A lineáris regressziót befolyásoló esetek diagnosztikája

Dr. Zrínyi Miklós
PhD, a Debreceni Egyetem Egészségügyi Karának vendégtanára
E-mail: zrinyim@yahoo.com

Dr. Katona Éva
PhD, az I. sz. Belgyógyászati Klinika, Debreceni Egyetem Orvos- és Egészségtudományi Centrum egyetemi adjunktusa
E-mail: katona@internal.med.unideb.hu

Dr. Szántó Ildikó
az I. sz. Belgyógyászati Klinika, Debreceni Egyetem Orvos- és Egészségtudományi Centrum szakorvosjelöltje
E-mail: szanto@internal.med.unideb.hu

Dr. Páll Dénes
PhD, az I. sz. Belgyógyászati Klinika, Debreceni Egyetem Orvos- és Egészségtudományi Centrum egyetemi docense
E-mail: pall@internal.med.unideb.hu

A tanulmány a lineáris regresszió során alkalmazható diagnosztikai eljárásokról nyújt áttekintést, konkrét példákkal illusztrálva az egyes módszerek alkalmazhatóságát a gyakorlatban. A részletesen kifejtett diagnosztikai módszerek az ún. befolyásos vagy szélsőséges esetek azonosítására fókuszálnak. Ahogy példák is mutatják, az ilyen adatpontok eltávolítása az elemzési eljárásból további hangsúlyt ad a minta alapvető tulajdonságainak kimutatásához, és javítja a regressziós becslés pontosságát. A befolyásos esetek azonosításához a reziduumok vizsgálatát tartják a szerzők a legmegfelelőbb módszernek, de a többi kiegészítő módszer is fontos annak értelmezésében, hogy a kiindulási determinációs együttható (R^2) értéke milyen irányba és mértékben mozdul el.

TÁRGYSZÓ:
Lineáris regresszió.
Determinációs együttható.
Reziduumok.

Közismert, hogy noha a statisztikai elemzések eredményét akár jelentősen is befolyásolhatják a mintában rejtve maradó ún. „szélsőséges" vagy „befolyásos" esetek, ám a gyakorlati tapasztalatok azt mutatják, hogy az elemzést végzők megfeledkeznek a szélsőséges adatok azonosításáról, és szükség esetén az elemzésből való eltávolításukról. Többváltozós lineáris regressziós eljárások során egy-egy szélsőséges eset akár kritikusan is megváltoztatja a regressziós modell eredményét. Ez abban is megnyilvánulhat, hogy valamely független változó összefüggését a függő változóval szignifikánsnak mutatja, vagy ellenkező esetben az összefüggés hiányát látszik megerősíteni. Még gyakoribb a determinációs együttható (R^2) értékére kifejtett hatásuk; a regressziós modell prediktív képessége a szélsőséges esetek eltávolítása után számottevően javul. Az ilyen analitikai tévedések felvetik annak kockázatát, hogy helyes hipotézist utasítunk el, vagy az általunk kimutatott összefüggés nagyságrendje elmarad a valós értéktől. Hogy ez minél kevésbé történhessen meg, a szélsőséges esetek azonosítására szolgáló eljárások és azok helyes alkalmazása segítenek az ilyen csapdákat elkerülni.

Az, hogy a szélsőséges esetek azonosításának igénye mennyire széleskörű *Walsh* [2006] mutat jó példát. Az Egyesült Államok Élelmiszer- és Gyógyszer-felügyeleti Hatóságának (Food and Drug Administration – FDA) minőségbiztosítási irányelvei kifejezetten kötelezővé teszik a gyógyszergyárak számára, hogy statisztikai módszerekkel azonosítsák a gyártási devianciákat (szélsőséges adatokat). Ezzel megakadályozzák, hogy a fogyasztókhoz hibás termék kerülhessen. A szélsőséges esetek azonosítása a genetikai daganatkutatásban is hasznosítható. *Hu* [2008] olyan módszertani megoldást ismertet, amivel visszaszorítható a fals terápiás felfedezések száma. *Kumar*, *Kumar* és *Singh* [2008] az egészségügyi adatbázisokban való szélsőséges esetek felkutatásának fontosságára irányítják a figyelmet. Wisconsin állam emlőrák regiszterében végeztek a szélsőségek azonosítására irányuló adatbányászatot, aminek eredményeképpen könnyebbé és gyorsabbá vált a helyes klinikai diagnózis felállítása, és pontosabbá tették a daganatos betegek terápiás menedzselését. A szélsőségek azonosítása akár egy teljes egészségügyi rendszert is felölelhet. *Vidmar és szerzőtársai* [2011] arra mutatnak példát, hogy miként használja a szlovén Egészségügyi Minisztérium a szélsőségek azonosításának módszerét az egészségügyi rendszer minőségi indikátorainak összehasonlító elemzésére. Mivel jelen írás szerzői az egészségügy területén jártasak, a szemléltetés erre a területre szorítkozik, de a szélsőségek azonosításának hasznossága az élet bármely más területén is hasonlóan értékes lehet.

Bár a befolyásos esetek hatása a kisebb elemszámú minták esetében még hangsúlyosabb lehet, fontos a szélsőségek azonosítása a nagy elemszámú mintákban is.

A befolyásos esetek hatásának vizsgálatára kevesebb figyelmet szentelő szerzők hajlamosak feltételezni, különösen a nagy elemszámú mintáknál – a nagy számok törvénye miatt –, hogy a szélsőséges esetek hatása a minta egészére nem számottevő. Ez az elképzelés már egyváltozós elemzések során is félrevezető lehet, de a többváltozós modellekben komoly eltéréseket hozhat létre, ami az eredmények és a konklúziók értelmezésében is megmutatkozhat. Írásunk hátralévő részében *Páll és szerzőtársai* [2004] által felvett nagy elemszámú minta elemzésén keresztül mutatjuk be azokat a választási lehetőségeket, amelyek segítenek a szélsőségek okozta eltéréseket korrigálni. Ebben a vizsgálatban a szerzők 10 359 debreceni középiskolás tanuló vérnyomását mérték meg, kiegészítve a vérnyomást befolyásoló tényezők felmérésével (életkor, testmagasság, testsúly stb.).

1. A befolyásos eset fogalma

Az angol nomenklatúra alapján „outliernek", azaz kirívónak, szélsőségesnek nevezzük azt az adatpontot, amely aránytalanul nagy torzulást okoz a regressziós egyenes becslésében. A kirívó eset pontos meghatározására több alternatíva is létezik. *Fox* [1997] szerint kirívó eset az, aminek az y értéke, x értékének függvényében, szokatlanul eltér a többi adatponttól. *Cohen és szerzőtársai* [2003] szerint olyan atipikus adatpontról beszélünk, amely az adathalmaz többi eleméhez nem illeszkedik megfelelően, és mintha más populációból származna a megfigyelés. *Bobko* [2001] a szélsőséges esetet a függő vagy független változók kiugró értékeiként definiálja, azaz, olyan esetnek tekinti, amely a regressziós illeszkedéstől jóval messzebb került, mint a többi adatpont által kifejtett trend. Végül pedig *von Eye* és *Schuste* [1998] a kirívó esetet olyan adatpontnak gondolja, amely extrém messze került a függő változó átlagától, vagy olyan torzító hatást fejt ki a regressziós egyenes becslésére, ami alapjaiban változtatja meg annak pontosságát.

Nurumnabi és *Nasser* [2008] szerint az említett helyzetek akkor állhatnak elő, ha az adatpontokban valamilyen hiba lép fel. Az extrém adatpontnak lehetnek természetes okai is, például a normális eloszlástól egy-egy eset szélsőséges, de valós módon különbözik (nagyon magas vagy nagyon elhízott egyén is kerül az átlagos mintába). A leggyakoribb ok azonban a mérési hiba, ami a kirívó esetek megjelenésével járhat. Előfordul a hibás adatfelvitel, kódolás is, ami a helyes adatokat értelemszerűen eltorzítja, ezáltal okozva zavart a becslési folyamatban. Kirívó esetet okozhat olyan rejtett változó jelenléte is, ami befolyást gyakorol a függő és független változók kapcsolatára, de a mérés során erre a befolyásoló hatásra nem gondoltunk (vagy közvetlen nem mértük). Végül, de nem utolsó sorban, az is előfordulhat, hogy a feltételezett össze-

függések nem lineáris jelleget mutatnak. Az ilyen adatpontok azonban a lineáris összefüggésre épülő regressziós folyamatban olyan torzításokat okoznak, amelyek kiszűrése nélkül a regressziós egyenes vagy felület becslése pontatlanná válik.

Összefoglalva, kirívó adatpont az, ami számszerűen távol esik az adathalmazunk többi megfigyelésétől. Kiszűrésük azért fontos, mert *1.* drámai hatással lehetnek a regressziós elemzés eredményére, különösen, ha a mintanagyságunk, elemszámunk alacsony; *2.* torzíthatják a determinációs együttható és a regressziós együtthatók értékét; *3.* korlátozzák az adatokból levonható helyes következtetéseket, azok értelmezését.

2. A befolyásos eset azonosításának diagnosztikai módszerei

A regressziós diagnosztika leggyakrabban használt módszerei a reziduumok elemzéséből állnak. Reziduumnak a megfigyelt értéknek számítottól való eltérését nevezzük (mérési hiba: e) (*Pedhazur* [1982]). Ha a reziduumok nagyságrendje függ x nagyságától, az azt jelzi, hogy a hiba szórása nem állandó. Egy megfigyelést akkor nevezünk kirívónak, ha az adott x érték mellett y értéke kiugró, és így a reziduum értéke különösen nagy (összehasonlítva a többi adatpontéval). Az ilyen adatpontok kiszűrésének módja a reziduális diagnosztika. A i-edik adatpont nem standardizált reziduumának meghatározása:

$$e_i = y_i - \hat{y}_i , \qquad /1/$$

ahol y_i az i-edik adatpont mért és $\hat{y}_i$ az i-edik adatpont becsült értéke.

Ahhoz, hogy a diagnosztika elvégezhető legyen, standardizálnunk kell a közönséges reziduumokat, ami valójában a standard hibával való elosztást jelenti (így a reziduumok átlaga 0, a szórásuk egységnyi lesz). A standard reziduumok definíciója tehát:

$$e_i^{\text{stand}} = e_i / SE(e_i) , \qquad /2/$$

ahol $SE(e_i)$ a mintából való becslésből származik.

Hair és szerzőtársai [1995] amellett érvelnek, hogy standardizálás nélkül nem tudnánk eldönteni, vajon egy reziduum nagysága kirívó-e, ha nem ismerjük azok eloszlását, amihez viszonyítani tudunk. A standardizálás eredményeképpen viszont bizonyítható, hogy ha a regressziós modell maradéktagja (ε) normális eloszlású, akkor a stan-

dardizált reziduumok megközelítőleg Student-féle t-eloszlást követnek (*Pedhazur* [1982]). Ennek alapján kirívó esetnek tekinthetjük azokat a megfigyeléseket, amelyek *standardizált reziduuma kívül esik a t-eloszlás választott (például 95 százalékos) határain*, azaz a standardizált reziduum értéke a ±1,96 értékét meghaladja.

Az ún. studentizált reziduumok vagy „törölt" studentizált reziduumok (deleted studentized residuals) használata a bemutatott módszer kiterjesztésének tekinthető. A felvetés szerint a kirívó esetek is befolyásolják y számított értékeit, így minden pontra úgy számoljuk ki a reziduumot, hogy a regressziós felület meghatározásakor az adott pontot a számításból kihagyjuk. Vagyis:

$$e_{i(-i)} = y_i - \hat{y}_{i(-i)}, \qquad /3/$$

ahol $\hat{y}_{i(-i)}$ az i-edik pont törlését követően számított i-edik pontbeli érték. A standardizálás a már ismert módszerrel történik:

$$e^{\text{stand}}_{i(-i)} = e_{i(-i)} \Big/ SE\left(e_{i(-i)}\right). \qquad /4/$$

A studentizált reziduumok a reziduumokhoz hasonlóan t-eloszlást mutatnak $k - p - 2$ szabadságifokkal, ahol k az adatpontok; p a független változók száma. Akárcsak a standard reziduumok esetében, egy pontot akkor tekinthetünk kirívónak, ha a studentizált reziduumok nagyobbak, mint az eloszlás választott (például 95 százalékos) kritikus értéke, azaz a ±1,96 értékét meghaladják.

A másik lehetséges diagnosztikai módszer a befolyásos pontok hatóerejének (leverage) vizsgálata. Hatóerő alatt egy adott pont (x) értékének távolságát értjük a minta x értékeinek átlagától (*Pedhazur* [1982]). Minél nagyobb ez a távolság a minta középpontjától, annál nagyobb lehet a kérdéses pont hatóereje. Másképpen fogalmazva, a középponttól távol eső pont „elhúzza" a pont irányába a regressziós felületet, ezzel „torzítva" a felület becslését. A nagy hatóerejű pontok befolyásolhatják – bár nem szükségszerűen – a regressziós paraméterek becslését. *Nurumnabi* és *Nasser* [2008] megjegyzi, hogy egy kirívó eset nem feltétlen válik torzító ponttá, mert esetleg kicsi a hatóereje. Egy torzító pont nem szükségszerűen lesz kirívó eset, ha kicsi a reziduuma, így statisztikailag a t-eloszlás választott kritikus ékén belül marad. Egy nagy hatóerejű pont torzító hatása végső soron az y értékére gyakorolt hatástól függ. A gyakorlatban az ún. részleges hatóerő (partial leverage – PL) számításával határozzuk meg egy megfigyelés befolyásos jellegét. Képlettel kifejezve:

$$\left(PL_j\right)_i = \left(e_{x_j}\right)^2_i \Big/ \sum_{k=1}^{n} \left(e_{x_j}\right)^2_k, \qquad /5/$$

ahol j a j-edik független változó, i az i-edik megfigyelés és e_{x_j} a reziduumok, amelyeket úgy kapunk, hogy x_j-t függő változóként használva, a fennmaradó független változókkal becsültetjük meg.

Mindezekkel összhangban, több dimenzióban a hatóerőt mindig az adatok adott irányú szóródásához képest mérjük, amire az e_{x_j} reziduumok vizsgálata ad megoldást. A *PL* kritikusérték meghatározása javaslat alapján $2p/n$ képlettel történik, ahol p a független változók száma, n a mintanagyság (*Hair et al.* [1995]). Azok az adatpontok, amelyek *PL* értékei a $2p/n$ kritikusértéken kívül esnek, „nagy hatóerővel" bíró, torzító pontoknak minősülnek, és a regressziós becslésből eltávolításuk javasolt.

Torzító pontok azonosításához az ún. Cook-féle *D*-statisztikát (vagy -távolságot – Cook's distance) is kiszámíthatjuk az egyes megfigyelésekre. Ez egy olyan standardizált index, ami azt méri, hogy a regressziós együtthatók hogyan változnak meg, ha az adott adatpontot töröljük (*Hair et al.* [1995]). Nagy reziduumokkal vagy hatóerővel rendelkező adatpontok jelentősen torzíthatják a regressziós együtthatók becslését, amire a Cook-féle *D*-statisztika felhívhatja a figyelmünket. Ennek értékét a következő képlettel határozhatjuk meg:

$$D_k = \sum_{i=1}^{n} \left(\hat{y}_i - \hat{y}_{i(k)} \right)^2 \Big/ p \cdot MSE \,, \qquad /6/$$

ahol $\hat{y}_i$ az i megfigyeléshez tartozó regressziós becslés értéke; $\hat{y}_{i(k)}$ az i megfigyeléshez tartozó újraillesztett becslés értéke úgy, hogy az k-adik megfigyelés törlésre került; p a független változók száma és *MSE* a regressziós átlagos négyzetes hiba. Ahhoz, hogy egy adatpontot torzító pontnak tekintsünk, $D_k > 1$ kritikusértéket javasolnak kisebb elemszámú minták esetében, nagyobb elemszámnál $D_k > 4/n$ képlettel kell számolnunk, ahol n a mintanagyság (*Hair et al.* [1995]).

Végezetül, ismert az ún. *DFFITS*-eljárás is, amely azt mutatja meg, hogy mennyire befolyásos egy adatpont a regressziós becslés folyamán. Az eljárást először 1980-ban publikálták (*Belsley–Kuh–Welsch* [1980]). Az eljárás azt mutatja meg, hogy miként változnak a regressziós együtthatók az i-edik adatpont becslésekor, ha az adatpontot kihagyjuk a regressziós modellből:

$$DFFITS_i = \left(\hat{y}_i - \hat{y}_{i(-i)} \right) \Big/ s_{(-i)} \cdot \sqrt{h_{ii}} \,, \qquad /7/$$

ahol $\hat{y}_i$ és $\hat{y}_{i(-i)}$ az i-edik pont regressziós becslésének értékei, amikor az i-edik pont szerepel, illetve nem szerepel a regressziós folyamatban; $s_{(-i)}$ a standard becs-

lési hiba az *i*-edik pont nélkül; és h_{ii} az *i*-edik pont hatóereje. A *DFFITS*-diagnosztika nagyon hasonlít a studentizált reziduumokhoz, olyannyira, hogy a *DFFITS* tulajdonképpen a studentizált reziduum és a hatóerő (leverage) szorzataként is felfogható. Mivel a studentizált reziduumok követik a *t*-eloszlást, így a kritikus érték (a *t*-eloszlás választott, 95 százalékos határa) 1,96-nak adódik, amit általában 2-re kerekítenek. A hatóerő kiszámítása $\sqrt{p/n}$ képlettel történik, ahol p a független változók száma, n a mintanagyság. A *DFFITS* kritikus érték meghatározása tehát

$$DFFITS > 2\sqrt{p/n}$$

képlet alapján történik (*Hair et al.* [1995]).

3. A befolyásos esetek azonosítása és az eredményre kifejtett hatásuk

A bemutatott módszereket egy-egy példával illusztráljuk. Tételezzük fel, hogy a serdülőkori fiatalok systolés vérnyomását (systolésRR) a testsúly (tsúly), a testmagasság (magasság) és a serdülőkorú neme (nem) egyaránt meghatározza. Ennek a feltevésnek az eldöntésére többváltozós regressziós modellt alkalmaztunk.[1] A 2. táblázat alapján a modell szignifikánsnak bizonyult ($F = 990{,}16$; $p < 0{,}001$), mindhárom független változó szignifikáns módon felelt a fiatalkori systolés vérnyomás kialakulásáért. A három független változó 23 százalékban adott magyarázatot a systolés vérnyomás alakulására ($R^2 = 0{,}23$), ami az egészségtudományi kutatások esetében jó eredménynek tekinthető. (Lásd az 1. táblázatot.) A systolés vérnyomás kialakulásában, a béta értékek összehasonlítása alapján, a testsúly volt a meghatározó, amit a nemek közötti eltérés követett. Ha a függő változóban okozott konkrét változás mértékét vizsgáljuk (*b* súlyok), két vizsgálati alany közötti 10 kg súlybeli eltérés 3,92 Hgmm-rel magasabb systolés vérnyomást eredményezett. A systolés vérnyomásban tapasztalható legnagyobb változást a nemek közötti különbség idézte elő: a serdülőkori fiúkhoz képest a lányok systolés vérnyomása átlagban 8,06 Hgmm-rel bizonyult alacsonyabbnak. A testmagasság a systolést vérnyomást fordított módon határozta meg: két vizsgálati alany közötti 10 cm-es magasságbeli különbség átlagban 0,85 Hgmm-rel csökkentette a systolés vérnyomás értékét.

[1] Az elemzésekhez az SPSS szoftvercsalád 11.5 verzióját használtuk. A táblázatokat az internetes Melléklet tartalmazza (www.ksh.hu/statszemle).

1. táblázat

A modell

Modell	R	R^2	Korrigált R^2	Becsült standard hiba
1. Alapmodell	0,483	0,233	0,233	12,40578
2. Standardizált reziduumok	0,536	0,288	0,287	10,73747
3. Torzító pontok	0,450	0,202	0,202	12,34964
4. Cook-féle távolság	0,514	0,265	0,264	11,02324
5. DFFITS	0,501	0,251	0,250	10,99659

Megjegyzés. Magyarázóváltozó: fiú/lány, TSÚLY, MAGASSÁG.

2. táblázat

ANOVA

Modell	Forrás	Négyzetösszeg	Szabadságfok	Átlagos négyzetösszeg	F	p-érték
1.	Regresszió	457168,426	3	152389,475	990,164	0,000
	Maradék	1505789,851	9784	153,903		
	Összesen	*1962958,277*	*9787*			
2.	Regresszió	434724,289	3	144908,096	1256,865	0,000
	Maradék	1077069,826	9342	115,293		
	Összesen	*1511794,115*	*9345*			
3.	Regresszió	345360,584	3	115120,195	754,819	0,000
	Maradék	1360269,132	8919	152,514		
	Összesen	*1705629,716*	*8922*			
4.	Regresszió	406049,369	3	135349,790	1113,881	0,000
	Maradék	1129088,317	9292	121,512		
	Összesen	*1535137,686*	*9295*			
5.	Regresszió	377589,862	3	125863,287	1040,838	0,000
	Maradék	1128593,107	9333	120,925		
	Összesen	*1506182,969*	*9336*			

Megjegyzés. Magyarázóváltozó: fiú/lány, TSÚLY, MAGASSÁG. Függő változó: systolés RR.

3. táblázat

Együtthatók

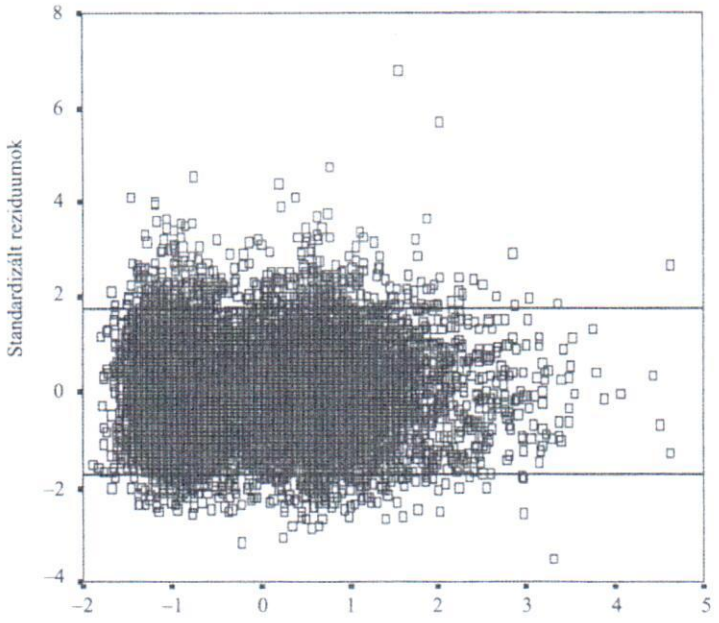
Modell	Magyarázó-változó	Nem standardizált együttható		Standardizált együttható β	*t*	*p*-érték
		B	Standard hiba			
1.	Állandó	119,412	3,515		33,975	0,000
	Tsúly	0,392	0,014	0,324	27,944	0,000
	Magasság	–0,085	0,021	–0,053	–4,059	0,000
	Fiú/lány	–8,066	0,317	–0,285	–25,427	0,000
2.	Állandó	118,207	3,100		38,131	0,000
	Tsúly	0,381	0,012	0,350	30,620	0,000
	Magasság	–0,077	0,018	–0,053	–4,159	0,000
	Fiú/lány	–8,189	0,281	–0,322	–29,187	0,000
3.	Állandó	126,390	4,337		29,140	0,000
	Tsúly	0,385	0,020	0,260	19,292	0,000
	Magasság	–0,118	0,026	–0,068	–4,459	0,000
	Fiú/lány	–8,620	0,356	–0,312	–24,207	0,000
4.	Állandó	119,204	3,305		36,067	0,000
	Tsúly	0,379	0,014	0,321	26,901	0,000
	Magasság	–0,080	0,020	–0,054	–4,038	0,000
	Fiú/lány	–8,281	0,293	–0,322	–28,251	0,000
5.	Állandó	113,757	3,230		35,223	0,000
	Tsúly	0,333	0,013	0,297	24,966	0,000
	Magasság	–0,040	0,019	–0,027	–2,059	0,040
	Fiú/lány	–7,903	0,289	–0,311	–27,353	0,000

Megjegyzés. Függő változó: systolés RR.

Az 1. modellt tekinthetjük az alapmodellnek, amelyből kiindulva vezetjük le a befolyásos esetek hatását a regressziós modell egészének változására (R, R^2 és regressziós együtthatók).

A standard reziduumok (y tengely) és az y (függő változó) becsült értékeinek (x tengely) grafikus ábrázolásával arra kaphatunk választ, hogy a mintánkban tapasztalható-e egyáltalán befolyásos esetek jelenléte. Az ábrán feltüntetett két, az x tengellyel párhuzamos egyenes, az ajánlás szerinti (*Hair et al.* [1995]) reziduum értékek szórásának kritikus ékeit tünteti fel (± 1,96). Az ezen túl elhelyezkedő reziduumok befolyásos vagy szélsőséges adatpontokat jelölnek. Az ábrán jól kivehetően, mindkét kritikus értéken túl találunk ilyen szélsőséges pontokat, azonban ez alapján nem tudjuk egyértelműen beazonosítani, hogy a minta melyik konkrét eleméről van szó. Ehhez az előbbiekben javasolt módszerekhez kell folyamodjuk.

Függő változó: systolés vérnyomás



Az általunk elsőként használt diagnosztikai eljárás a studentizált reziduumok vizsgálata volt. Az ajánlás szerint (*Hair et al.* [1995]) azokat az eseteket tekintettük szélsőségesnek, amelyek a ± 1,96 értéknél nagyobbnak bizonyultak. (Lásd a 4. táblázatot.)

A reziduum diagnosztikáját követően 443 szélsőséges eset eltávolítására került sor. Az új modell szignifikáns maradt ($F = 1256{,}86$; $p < 0{,}001$), a három független változó továbbra is szignifikáns módon határozta meg a systolés vérnyomás értékét. Ami azonban változott az alapmodellhez képest, a determinációs együttható értéke javult (0,23-ról 0,28-ra). A szélsőséges esetek kivonásával a becslési pontosságunk 5 százalékkal növekedett, az új modellben a független változók már 28 százalékban magyarázzák a systolés vérnyomás kialakulását.

4. táblázat

Standardizált reziduum

Megnevezés	*N*	Minimum	Maximum	Átlag	Szórás
Standardizált reziduum	9788	−3,51613	6,76088	−0,0000033	1,00005434

A következő diagnosztikai módszer a megfigyelések hatóerejének (leverage points) vizsgálata volt. (Lásd a 3. modellt és a 5. táblázatot.) Vagyis annak azonosítása, hogy a mintánkban vannak-e nagy hatóerejű torzító pontok. A leírtak alapján torzító pont volt az, amelynek értéke az ajánlás szerinti (*Hair et al.* [1995]) a kritikus értéken kívül esett, azaz ($2p/n$ képlettel számolva: 6/9788) 0,0006-nál nagyobb volt.

5. táblázat

Torzító pontok eljárás

Megnevezés	*N*	Minimum	Maximum	Átlag	Szórás
Torzító pontok	9961	0,00010	0,00513	0,0003066	0,00031468

Összesen 1039 torzító pontnak ítélhető esetszám került eltávolításra a diagnosztikai eljárás eredményeként. Az új regressziós modell továbbra is szignifikáns ($F = 754{,}82$; $p < 0{,}001$) maradt az alapmodellhez hasonlóan, és a független változók is szignifikáns módon határozták meg a systolés vérnyomást. Ami azonban az alapmodellhez és a studentizált reziduumok vizsgálatával végzett modellhez képest változás, hogy a torzító pontok diagnosztikája rontott a determinációs együttható becslésén, az alapmodellhez képest 3 százalékponttal, a studentizált reziduumok modelljéhez képest 8 százalékponttal csökkent a modell magyarázó képessége.

A torzító pontok kereséséhez egy másik diagnosztikai módszert is használtunk, a Cook-féle távolságot (vagy *D*-statisztikát) számoltuk ki az egyes megfigyelésekre. Torzító pont, különösen nagy elemszámú minta esetén, az ajánlás szerint (*Hair et al.* [1995]) minden olyan érték, amely kívül esik a $4/n$ kritikus értéken, azaz (4/9788) nagyobb mint 0,0004. A 6. táblázat a Cook-féle távolság leíró statisztikáját mutatja be. (Lásd a 4. modell értékeit.)

6. táblázat

Cook-féle távolság

Megnevezés	*N*	Minimum	Maximum	Átlag	Szórás
Cook-féle távolság	9788	0,00000	0,00944	0,0001039	0,00023896

A diagnosztikai eljárást követően 493 torzító pontnak megfelelő eset került törlésre. Az előzőkhöz hasonlóan a regressziós modell szignifikáns maradt ($F = 1113{,}89$; $p < 0{,}001$) és a független változók mindegyike szignifikáns módon ha-

tározta meg a systolés vérnyomást. Az új modell a determinációs együttható becslését javította (0,23-ról 0,265-re), azonban a három független változó csak 26,5 százalékban adott magyarázatot a systolés vérnyomás alakulására, szemben a studentizált reziduumok vizsgálatát követő 28 százalékos eredménnyel.

Utolsóként a *DFFITS*-eljárás lefolytatását végeztük el. (Lásd az 5. modellt és a 7. táblázatot.) Itt az adott megfigyelés hatását úgy mértük, hogy mennyit változtak a regressziós együtthatók abban az esetben, ha az adott megfigyelést töröltük. Az ajánlás szerinti (*von Eye–Schuste* [1998]) kritikus érték, ami felett torzító pontnak tekintünk egy megfigyelést: $2\sqrt{p/n}$ ($2\sqrt{3/9788}$), azaz a 0,035-nél nagyobb értékeket vesszük figyelembe.

7. táblázat

DFFIT-eljárás

Megnevezés	*N*	Minimum	Maximum	Mean	Std. Deviation
DFFIT	8446	–0,11070	0,08972	–0,0008864	0,01774439

A *DFFITS*-eljárás alapján összesen 452 esetet töröltünk az elemzésből. Akárcsak az eddigiekben, a modell és az egyes független változók szignifikánsak voltak. Az alapmodellhez képest az R^2 értéke 2 százalékponttal ismét javult (0,23 százalékról 0,25-re). Ebben a modellben tehát a független változók a systolés vérnyomásban létrejött variancia 25 százalékára szolgáltak magyarázattal.

4. Következtetések

A bemutatott diagnosztikai eljárások lefolytatását követően részben ellentmondásosnak tűnő következtetésre jutottunk. Az alap regressziós modell determinációs együtthatók becslése a befolyásos esetek eltávolítását követően először javult, egy esetben rosszabb lett, illetve kis mértékben pozitívan változott. Szembetűnő, hogy három diagnosztikai módszer (studentizált reziduumok, Cook-távolság és *DFFITS*-eljárás) eredménye azonos irányba mutat, és közel helyezkedik el egymáshoz, ami azt az elképzelést erősíti, hogy az alapmodellhez képest azt a modellt kell elfogadnunk, amelynél az R^2 értéke a legnagyobb. A torzító pontok vizsgálatára használt módszer viszont a kiindulási érték csökkentését sugallta. Ezt az eredményt azért vetettük el, mert a diagnosztikai teszt a valid mintanagyságból olyan jelentős adathal-

mazt távolított el (befolyásos esetnek diagnosztizálva ezeket), amely a vizsgálatra alkalmas minta tulajdonságait jelentősen megváltoztatta. Feltételezzük, hogy ötszáz megfigyelés alatti esetszám-csökkenés alapvető tulajdonságaiban nem változtatta meg a mintánkat. Ezer feletti adat eltávolításával azonban olyan megfigyelések is kikerültek az elemzés köréből, amelyek egyébként a minta valós tulajdonságainak részét kellett képezzék. Kétszer annyi esetszám eltávolítása már nem javította, inkább rontotta a modell R^2 értékét. Így a torzító pontok azonosításának eredményét nem tartottuk megfelelően értékelhetőnek.

Helyes iránymutatást a reziduumok grafikus ábrázolása jelentett, amely egyértelműen azt mutatta, hogy a mintában vannak egészen szélsőséges esetek, amelyek standardizált reziduuma kívül esik a t-eloszlás választott 95 százalékos határain, azaz a ± 1,96-os kritikus értéken. Azonban a kritikus értékeken belül is előfordultak megfigyelések, amelyek az x tengelyen mérve a 4-es, 6-os érték körül találhatók. Az egyes módszerek eredménye közötti különbség részben abból is adódhat, hogy a kritikus értékeken belül előforduló befolyásos eseteket mennyire észleli és azonosítja torzító pontként. Ahogy erre már utaltunk, egy szélsőséges eset nem feltétlen torzító pont (ha kicsi a hatóereje). Egy torzító pont ugyanúgy nem feltétlen befolyásos eset (amennyiben kicsi a reziduuma). Az, hogy egy nagy hatóerejű pont torzító-e vagy sem, végső soron az y koordinátájának értékétől függ.

Az előzőkben hivatkozott szerzők abban egyet értenek, hogy amennyiben az alkalmazott módszerek között eltérések adódnak, a reziduumok vizsgálatából származó eredményeket tekintsük irányadónak (*Nurumnabi–Nasser* [2008], *Bollen–Jackman* [1990], *Belsley–Kuh–Welsch* [1980]). További lehetőség, hogy az egyes diagnosztikai eljárások által kölcsönösen torzító pontnak azonosított eseteket kiemeljük az elemzésekből, míg a többit érintetlenül hagyjuk. Adatbázisunk túlságos megnyirbálása azonban ellentétes változásokat is okozhat az eredmények szempontjából, ahogy ezt elemzésünkben is bemutattuk. Mindezeket figyelembe véve úgy véljük, hogy a kiindulási regressziós modellünk eredménye korrekcióra szorul, a determinációs együttható értékét a studentizált reziduumokkal végzett vizsgálat eredményével módosítottuk. A serdülőkori systolés vérnyomás kialakulásában szerepet játszó testsúly, testmagasság és a serdülőkorú nemének hármas kölcsönhatása nem 23, hanem 29 százalékban ad magyarázatot a systolés vérnyomásban tapasztalható varianciára. Ez azt is jelenti, hogy 71 százalék olyan megmagyarázatlan varianciát tapasztaltunk a systolés vérnyomásban, amiért vélhetően más, a regressziós modellünkben nem mért és nem szereplő tényezők feleltek.

Mindezek felhívják a figyelmet arra, hogy a befolyásos esetek azonosítása, jelentős mintanagyság mellett is fontos analitikai feladat. Ahogy példáink is mutatták, a befolyásos esetek eltávolítása az elemzési eljárásból akár további hangsúlyt adhat a minta alapvető tulajdonságainak kimutatásához. A befolyásos esetek azonosításához a reziduumok vizsgálatát javasoljuk, de ahogy ezt láttuk, a többi kiegészítő módszer

is fontos annak értelmezésében és megerősítésében, hogy a kiindulási értéket milyen irányba szükséges (ha szükséges) korrigálni.

Irodalom

BELSLEY, D. A. – KUH, E. – WELSCH, R. E. [1980]: *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley Series in Probability and Mathematical Statistics.* John Wiley & Sons. New York.

BOBKO, P. [2001]: *Correlation and Regression: Applications for Industrial Organizational Psychology and Management.* Sage Publications. Thousand Oaks.

BOLLEN, K. A. – JACKMAN, R. W. [1990]: Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. In: *Fox, J. – Long, J. S.* (eds.): *Modern Methods of Data Analysis.* Sage. Newbury Park. pp. 257–291.

COHEN, J. – COHEN, P. – WEST, S. G. – AIKEN, L. S. [2003]: *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates. Mahwah.

FOX, J. [1997]: *Applied Regression Analysis, Linear Models, and Related Methods.* Sage Publications. Thousand Oaks.

HAIR, J. F. – ANDERSON, R. E. – TATHAM, R. L. – BLACK, W. C. [1995]: *Multivariate Data Analysis.* Prentice-Hall.

HU, J. [2008]: Cancer Outlier Detection Based on Likelihood Ratio Test. *Bioinformatics.* Vol. 24. No. 19. pp. 2193–2199.

KUMAR, V. – KUMAR, D. – SINGH, R. K. [2008]: Outlier Mining in Medical Databases: An Application of Data Mining in Health Care Management to Detect Abnormal Values Presented in Medical Databases. *International Journal of Computer Science and Network Security.* Vol. 8. No. 8. pp. 272–277.

NURUMNABI, A. A. M. – NASSER, M. [2008]: Multiple Outliers Detection: Application to Research and Development Spending and Productivity Growth. *BRAC University Journal.* Vol. 5. No. 2. pp. 31–39.

PÁLL D. – KATONA É. – ZRÍNYI M. – ZATIK J. – PARAGH GY. – FÜLESDI B. [2004]: A serdülőkori vérnyomást befolyásoló tényezők: Debrecen Hypertension Study. *Lege Artis Medicinae.* 14. évf. 8–9. sz. **591–597. old.**

PEDHAZUR, E. J. [1982]: *Multiple Regression in Behavioral Research. Explanation and Prediction.* Harcourt Brace. New York.

VIDMAR, G. – BLAGUS, R. – STRELEC, L. – STEHLÍK, M. [2011]: Business Indicators of Healthcare Quality: Outlier Detection in Small Samples. *Applied Stochastic Models in Business and Industry.* Vol. 28. No. 3. pp. 282–295.

VON EYE, A. – SCHUSTE, C. [1998]: *Regression Analysis for Social Sciences.* Academic Press. San Diego.

WALFISH, S. [2006]. Review of Statistical Outlier Methods. *Pharmaceutical Technology.* 2. November. http://pharmtech.findpharma.com/pharmtech/IT/A-Review-of-Statistical-Outlier-Methods/ArticleStandard/Article/detail/384716

Summary

This paper discusses diagnostic techniques concerning the identification and removal of outliers in linear regression. Examples illustrate each method applied in the analytical practice and focus on identifying outliers in a large sample. Removing extreme data points, as demonstrated in our examples, helped exemplify the underlying nature of our data, and improved the prediction and the goodness of fit (R^2). Using regression residuals showed the best method to determine outliers in our sample, but, as demonstrated with additional techniques, more detection approaches need to be employed to conclude whether the best outlier filtering is applied.