doi:10.20312/dim.2016.07

Többváltozós lineáris regresszió a gyakorlatban

Pödör Zoltán
NymE Simonyi Károly Műszaki, Faanyagtudományi és Művészeti Kar,
Informatikai és Gazdasági Intézet
podor@inf.nyme.hu

ÖSSZEFOGLALÓ. Cikkünkben bemutatjuk a többváltozós lineáris regresszió gyakorlati alkalmazásának nehézségeit az eredmények felhasználhatósága szempontjából. A matematikai-statisztikai alapú eljárások által generált ideális modellek sokszor nem felelnek meg teljesen a vizsgált adatokat adó szakterület elvárásainak.

ABSTRACT. In this paper we introduce the difficulties of the practical application of the multivariable linear regression in terms of the result's usability. The mathematical-statistical based optimum models do not answer exactly the expectations of the examined data's area.

1. Bevezetés

A különböző mérési adatok közötti kapcsolatkeresés az adatfeldolgozás fontos feladata. Számtalan statisztikai módszer ismert ennek vizsgálatára az egyszerű kétváltozós lineáris kapcsolat elemzéstől a többváltozós nemlineáris kapcsolatkeresésen keresztül a már inkább az adatbányászat területére kimutató mesterséges neurális hálókig bezárólag. Azonban fontos azt is látni, hogy módszertani értelemben egyre bonyolultabb és számításigényesebb feladatokról van szó.

Jelen munkában a többváltozós lineáris regresszió gyakorlati alkalmazásával, alkalmazhatóságának különböző aspektusaival foglalkozunk. Kitérünk a módszertan előnyeire és a gyakorlatban tapasztalható hátrányaira is. Erdészeti jellegű adatok és alapvető meteorológiai paraméterek között végzett összefüggés vizsgálatokon keresztül mutatjuk be a módszertan felhasználásával kapcsolatos gyakorlati tapasztalatokat és nehézségeket.

Különösen a kapott eredmények értelmezésével, értelmezhetőségével kapcsolatosan felmerülő kérdésekre és problémákra fókuszálunk, mint például mennyire összeegyeztethető a matematikai értelemben legjobb modell az adott szakterület számára jó és elfogadható modellel. Mennyire lehet a különböző (ám azonos alapú) adatsorokra kapott eredményeket egységesen kezelni, a kapott összefüggéseket összehasonlítani.

A dolgozatban részletesen is bemutatjuk az alkalmazott módszertant, hiszen ennek ismerete fontos a fenti kérdések vizsgálatának szempontjából.

2. Alkalmazott módszertan, felhasznált adatok

A fejezetben bemutatjuk a többváltozós lineáris regresszió elméleti hátterét és a felhasznált erdészeti jellegű adatokat. Ez a regressziós módszer alkalmas többek között a

diszkrét adatsorok folytonossá tételére, adathiányok, kiugró adatok kezelésére, zajszűrésre és természetesen jövőbeni értékek előrejelzésére is a független paraméterek megfelelő rendelkezésre állása esetén.

2.1. Többváltozós lineáris regresszió

A regresszió számítás [3] lehetővé teszi, hogy lineáris kapcsolatot állítsunk fel egy függő és több független változó között, felépítve rájuk egy lineáris modellt (1).

$$y = b + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n, \tag{1}$$

ahol y a függő, $x_1, x_2, \dots, x_n$ a független változók, $b, a_1, a_2, \dots, a_n$ pedig a regressziós együtthatók.

A modellbe bevont független változók számának növelésével a modellt jellemző determinációs együttható (R^2) értéke minden egyes lépésben biztosan nem romlik, általában javul is valamennyit. Ez azt a tévképzetet keltheti bennünk, hogy a legjobb modell a legtöbb változó bevonásával érhető el. Ezzel szemben az optimális modell előállításához meg kell határoznunk azon változók minimális körét, melyek érdemi, statisztikailag is mérhető hatást fejtenek ki a függő változóra és egy-egy újabb változó bevonásával szignifikánsan javítjuk a modellt.

Ezért olyan megoldásra kell törekedni, mely során a modellbe csak a minimális számú, 0-tól szignifikánsan eltérő együtthatójú, egymással minél kevésbé összefüggő magyarázó változó kerüljön bevonásra, mégpedig úgy, hogy a kapott modell még megfelelő biztonsággal írja le a vizsgált folyamatot. Összegezve, a modellépítés során optimális egyensúlyra törekszünk a gazdaságosság és a jó közelítés között.

A feladat megoldására különböző technikák léteznek, az egyik ismert és elterjedt módszertant a lépésenkénti regressziós technikák jelentik. A lépésenkénti regressziós technikáknak [1] alapvetően három típusát szokták megkülönböztetni:

- forward selection,
- backward elimination,
- stepwise regression.

A módszerek mindegyikének alapötlete, hogy egyesével vizsgáljuk a lehetséges változókat és egyenként döntjük el, hogy az adott változóra szükség van-e az épülő modellben. Annak eldöntésére, hogy egy változó beépítése a modellbe szignifikáns javulást hoz-e az eggyel korábbi állapothoz képest, F-próbát használunk (2). Annak vizsgálatára, hogy egy beépítendő változó együtthatója a modellben szignifikánsan eltér-e 0-tól t-próbát alkalmazhatunk.

$$F = \frac{\left(r^2_{y.1,2,\dots,p} - r^2_{y.1,2,\dots,p-1}\right)}{\frac{1 - r^2_{y.1,2,\dots,p}}{n-p-1}} \tag{2}$$

A három, említett módszer esetében a megközelítés irányában van különbség. A *forward selection* során egyesével vesszük a lehetséges magyarázó változókat, és döntjük el, hogy beépítésre kerüljön-e vagy sem. Így a modell kezdetben egyetlen független változót sem tartalmaz, majd minden egyes iterációban egy-egy elemmel bővülhet ez a halmaz, bevonva azt a változót, mely a legerősebb kapcsolatot mutatja a függő változóval. A *backward selection* ennek éppen az ellentettje. A kezdő lépésben minden lehetséges független változót bevonunk a modellbe, majd az egyes iterációs lépésekben egyesével hagyjuk el azokat a változókat, melyek a legkevésbé gyakorolnak hatást a függő változóra. *Stepwise* módszer pedig a fenti két eljárás ötvözete. Egy-egy iterációs lépésben bevonunk egy új változót, mely szignifikáns javulást okoz a modellben, majd vizsgáljuk, hogy a már bevont változók közül el tudunk-e hagyni úgy, hogy az ne okozzon statisztikailag mérhető romlást a modell jóságát tekintve.

Az illesztés, modellezés során az illeszkedés jóságának mérésére leggyakrabban használt mutató az R^2 determinációs együttható. Az R^2 mellett olyan mutatók alkalmazása is célszerű, amelyek figyelembe veszik a becslés során a bevont változók számát is, és ezáltal a kevés számú paramétert tartalmazó modelleket – még ha kevésbé tűnnek is pontosnak, mint a több paramétert tartalmazó társaik – versenyképessé teszik a több változót, illetve paramétert tartalmazó modellekkel. A legegyszerűbb ilyen mutató a Theil-féle, szabadságfokkal korrigált determinációs együttható (3), ahol n az összes lehetséges paraméterek, p pedig a modellbe bevont paraméterek számát jelöli.

$$\hat{R}^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2) \qquad (3)$$

2.2. Felhasznált adatok

Néhány magyarországi, adott helyről származó erdészeti jellegű adatsorokat vizsgáltunk, mint függő paramétereket, illetve az ezekhez köthető helyi mérésű meteorológiai adatsorokat, mint függő paramétereket. A függő paraméterek éves bontásúak és 1994-2010 vonatkozásában állnak rendelkezésre. A meteorológiai adatok pedig helyi mérésekből származó, havi bontású átlaghőmérséklet és csapadékösszeg adatok. A gyakorlati felhasználhatóság szempontjából fontos hangsúlyozni, hogy ez az a két paraméter, ami gyakorlatilag az ország bármely pontjára, tetszőleges időszakra elérhető és használható.

Az alap meteorológiai adatok mellett a CReMIT [2] módszer alkalmazásával speciális, időszaki adatokat is képeztünk előző év januárjától adott októberéig, legfeljebb 6 hónapnyi szélességű időablakok alkalmazásával. Ez lehetővé teszi a különböző időszakok és az előző év meteorológiai hatásainak figyelembe vételét is a szimpla havi adatok mellett.

3. Eredmények

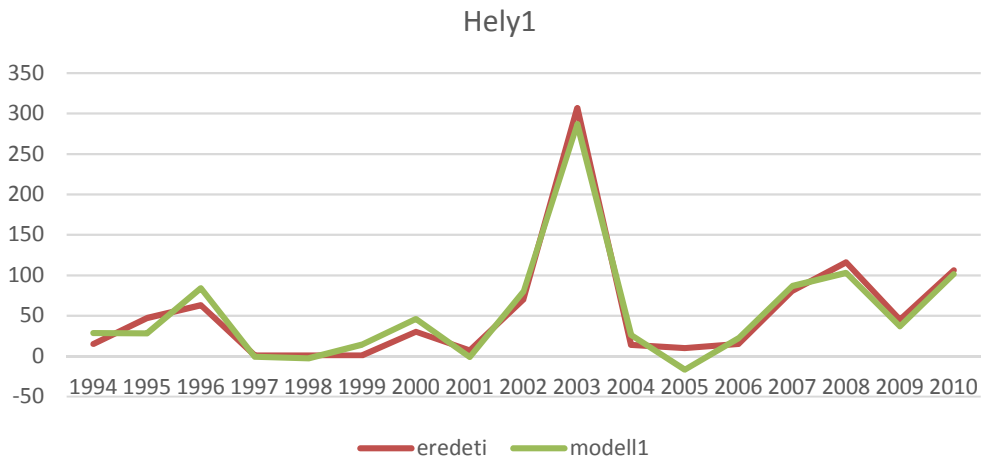
Az eredmények kapcsán bemutatunk több illesztési eredményt, a hozzájuk tartozó modelleket. Ugyanazon jellegű erdészeti adatsorra több helyen is elvégeztük az illesztést. Megvizsgáltuk a modelleket több szempontból is: mennyire hasonlíthatóak össze egymással a különböző földrajzi területen, de azonos adatsorokra kapott modellek, mennyire hangolható össze a matematikai értelemben legjobb modell a szakmai szempontokkal. Készíthető-e univerzálisan jó modell több, hasonló adatsorra, van-e ennek létjogosultsága egyáltalán?

Négy adatsorra mutatjuk be az eredményeket. Ahogy majd látni fogjuk ezek közül 3 elég hasonló karakterisztikát mutat, egy pedig jelentősen eltér már az alapadatok tekintetében is. A három homogén adatsor esetében 2003-ben tapasztalható egyöntetűen egy kiugró adat, ez a negyedik adatsor esetében egyáltalán nem mutatható ki, így itt eleve nem is várnánk feltétlenül a másik háromhoz hasonló eredményeket.

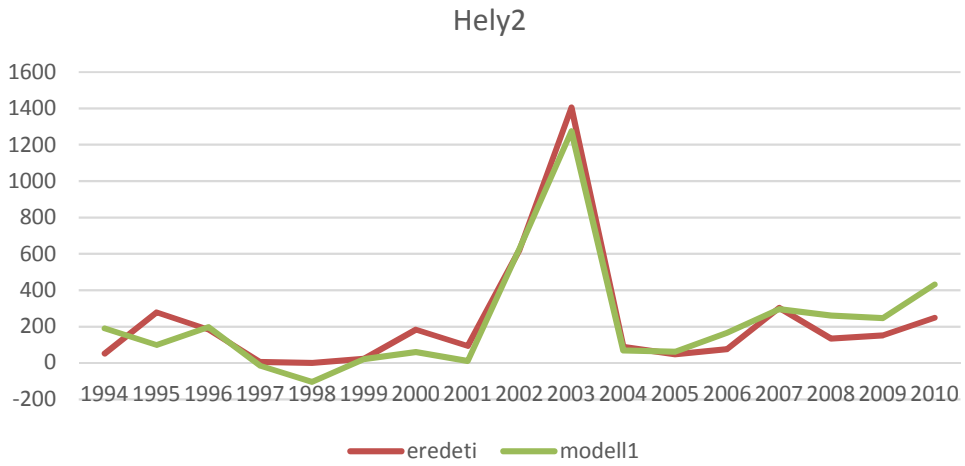
A vizsgálatokat az ingyenes R szoftverrel [4] végeztük, illetve a regressziós modell megvalósítása saját fejlesztés alapján történt, hogy olyan részletinformációk is megjelenhessenek a folyamat során, amit egyéb szoftverekben nem, vagy csak nehezen lehet elérni. Így a kimenetben teljesen nyomon követhető a modellépítés folyamata, az egymás után bevonásra kerülő változók, az aktuális modell együtthatói és a korrigált determinációs együttható értékei. A modellek vizsgálata 90%-os szignifikancia szinten történt, ez definiálta egyúttal azt is, hogy melyik lépésben áll le a modellépítés folyamata.

Az alábbiakban, az 1.-4. ábrák mutatják az eredeti függő paraméterek adatait és az illesztette, többváltozós regressziós modell által generált adatokat. Az x-tengelyen az évek, az y-tengelyen pedig a vizsgált erdészeti adatok értékei látszódnak. Jól érzékelhető, hogy az első három esetben az illeszkedés meglehetősen jó, míg a negyedik esetben elég gyenge. Ezt a

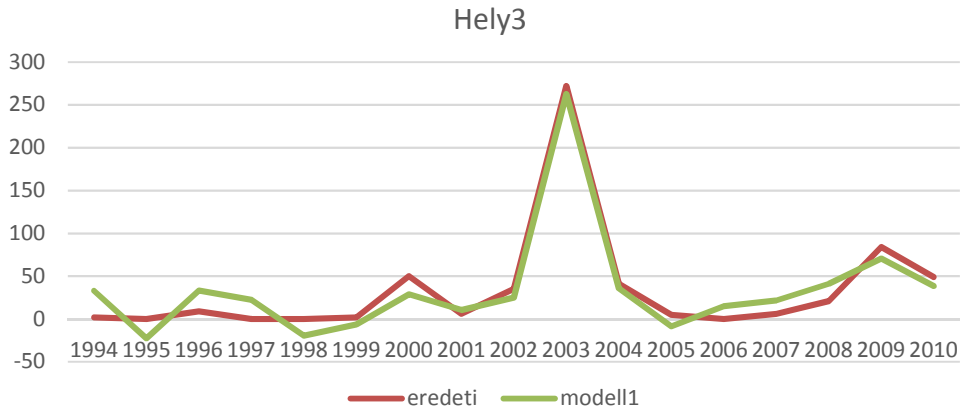
kapott korrigált determinációs együtthatók is megerősítették, azonban ennek részletezésére most nem térünk ki.



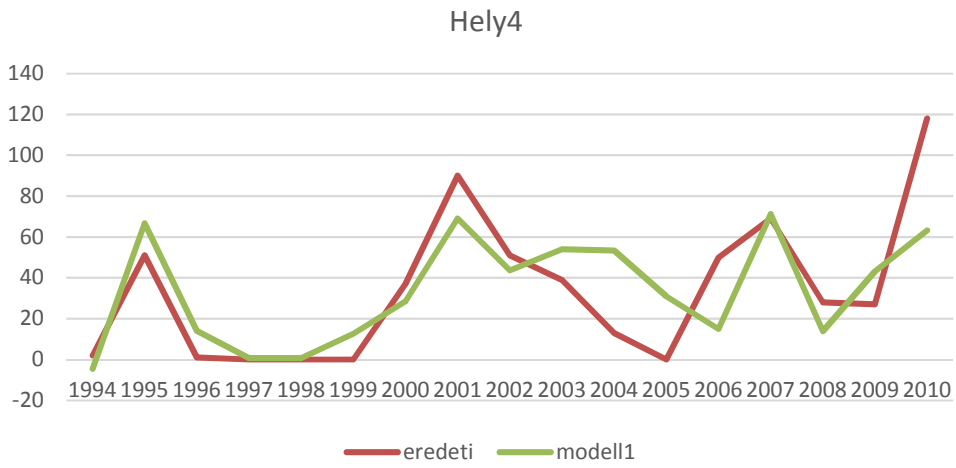
1. ábra. Első helyszín, eredmények



2. ábra. Második helyszín, eredmények



3. ábra. Harmadik helyszín, eredmények



4. ábra. Negyedik helyszín, eredmények

Az egyes helyekhez az alábbi többváltozós regressziós modellek álltak elő:

Modellbe bevont paraméterek listája									
	paraméter1	***paraméter2***	***paraméter3***	***paraméter4***	***paraméter5***	***paraméter6***	***paraméter7***	***paraméter8***	***paraméter9***
modell1	Temp_a5_a7	Temp_a1_a6	Temp_p8_p8	Prec_p12_a6	Prec_p3_p3	Temp_p5_p6	Prec_p9_p11	Prec_p7_a1	Prec_p3_p4
modell2	Temp_a5_a8	Temp_a1_a5	Temp_p6_p7	Prec_a4_a5	Prec_a2_a3	Temp_p11_p11	Prec_p3_p3	Temp_a7_a7	Prec_a9_a10
modell3	Temp_a5_a8	Temp_a1_a6	Temp_p6_p6	Prec_p6_p7	Temp_p8_p8	Prec_p4_p4	Temp_a3_a3	Prec_a5_a7	Prec_p4_p8
modell4	Temp_p7_p11	Temp_p4_p5	Temp_p6_p6	Prec_p7_p8	Prec_a2_a2	Prec_a6_a6	Prec_p8_a1	Prec_p3_p7	Temp_a7_a7

1. táblázat. A négy helyszín modelljeinek paraméterlistája

Az 1. táblázatban látható a négy vizsgált terület modellje, amely tartalmazza a modellbe bevont paraméterek neveit, melyeket az alábbi módon értelmezünk: *Temp*, illetve *Prec* előtag utal arra, hogy hőmérséklet vagy csapadék adatról van-e szó. Az ezt követő két érték a vizsgált ablak kezdete és vége hónapban megadva, ahol *a* az aktuális, *p* az előző évre utal. Így például *Temp_p9_a2* az előző év szeptemberétől adott év februárjáig tartó időszak átlaghőmérséklete. Most csak ezeket az értékeket jelenítjük meg, eltekintünk a modellek részletesebb adatainak (együtthatók, determinációs együttható) megjelenítése.

A táblázatban megjelenő, bevont paraméterek sorrendje egyúttal a modellbe való bekerülés sorrendjét is tükrözi. Figyelembe véve, hogy az első három vizsgált alap adatsor meglehetősen homogén az eredmények tekintetében is egységes képet várnánk. A hasonlóságot itt most az egyszerűség kedvéért a bevont paraméterek előfordulásával és bekerülési sorrendjével értelmezzük. Ugyan felfedezhetőek hasonlóságok az első 3 modellben, különösen az első három bevont paraméter tekintetében, de az azt követő elemek már meglehetősen vegyes képet mutatnak mind a bevonás sorrendjének, mind az előfordulásnak a tekintetében. Ennek oka vélhetően abban keresendő, hogy a modellezés pusztán matematikai megközelítést használ, azaz egy adott lépésben azt az elemet vonja be az épülő modellbe, ami lokálisan a legjobb javulást hozza statisztikai értelemben. Ugyanakkor elképzelhető, hogy statisztikailag egy kicsit gyengébb javulást hozó paraméter bevonásával egy adott lépésben homogénebb és/vagy szakmai értelemben jobb modellsereg lenne létrehozható. Azonban esetleges szakmai szempontok figyelembe vételét a matematikai megközelítés ebben a formában nem teszi lehetővé.

A negyedik adatsor karakterisztikájában is eltér az előző háromtól, így a kapott modell is teljesen más elemeket tartalmaz, nem vagy nehezen hasonlítható össze az előző modellekkel.

Felmerülhet itt is a kérdés, hogy ha nem pusztán matematikai oldalról közelítjük a kérdést kialakítható-e egy kicsit gyengébb, de a másik háromhoz jobban hasonlító modell.

A nem pusztán matematikai, statisztikai szempontok figyelembe vételének fontosságát támasztja alá az is, hogy specifikus, szakmai feladatok esetében sok esetben jobban értelmezhető egy matematikai értelemben gyengébb, de szakmailag jobban magyarázható modell.

4. Összefoglaló

A változók közötti kapcsolatkeresés az adatelemzés egy fontos feladata. Ennek egyik, a statisztikából jól ismert módszere a regresszió, illetve annak egy speciális változata a többváltozós lineáris regresszió. Erdészeti jellegű példákon keresztül felhívtuk a figyelmet a pusztán matematikai alapokon nyugvó modellalkotás lehetséges kérdéseire, problémáira elsősorban az adatokat szolgáltató szakterület szempontjából. Biztos-e, hogy a matematikailag legjobb modell szakmailag is a legjobb? Összehasonlíthatóak-e azonos jellegű adatsorokra kapott egyedi modellek? Várhatunk-e azonos, hasonló eredményeket ilyen adatsorokra?

A bemutatott példák kapcsán azt a tapasztalatot fogalmazhatjuk meg, hogy a matematikai alapokon nyugvó modellek nem biztos, hogy szakmai értelemben is a legjobbak. Sok esetben előfordulhat, hogy egy statisztikai értelemben gyengébb modell szakmailag sokkal jobban magyarázható adott szakterületen, mint a matematikailag legerősebb. Ugyanakkor a szakmai szempontok bevonása közvetlen módon az ilyen jellegű modellezésben nehéz, ha egyáltalán lehetséges.

Ez megnehezíti az egységes modellek kialakítását is. Feltételezhető, hogy univerzálisan jó modell általában nem feltétlenül alakítható ki még ugyanolyan jellegű adatsorok esetében sem, így lokális optimumokra tudunk törekedni matematikai oldalról.

Irodalomjegyzék

[1] **Montgomery, D. C., Peck, A. E., Vining, G. G**., Introduction to linear regression analysis (fifth edition). Published by John Wiley & Sons, (2012), p. 672.
[2] **Pödör, Z., Edelényi, M., Jereb, L**., Systematic Analysis of Time Series – CReMIT. Infocommunication Journal, VI(1) (2014), 16-22.
[3] **Spiegel, M. R.**, Statisztika, Elmélet és gyakorlat, PANEM-McGraw-Hill, Budapest, (1995), p. 546.
[4] https://www.r-project.org/**,** utolsó megtekintés 2016.10.28.