DOI: 10.12663/PsyHung.1.2013.1.2.4

ROBUSZTUS LINEÁRIS REGRESSZIÓ ALKALMAZÁSA PSZICHOLÓGIAI ELEMZÉSEKBEN

Takács Szabolcs[1] – Smohai Máté[2]
[1]Károli Gáspár Református Egyetem és Budapest Főváros Kormányhivatala Munkaügyi Központ,
[2] Eötvös Lóránt Tudományegetem – PPK
Levelező szerző: Takács Szabolcs E-mail: tretarkhon@gmail.com

Kivonat

Változók közötti kapcsolatok feltárása során általában az egyszerűbb (például lineáris) összefüggések között kezdjük el a keresést: tágabb értelemben gondolhatunk itt a t-próbára, a varianciaanalízisre (ANOVA), vagy akár a lineáris regresszióra (Pearson-féle korrelációs együtthatóra).

Azonban addig, amíg a t-próba és a varianciaanalízis esetén akár egy teljesen alap szintű programcsomag is képes a szóráshomogenitás megsértése esetén alkalmazandó robusztus tesztek elvégzésére, addig a lineáris regresszió esetén erre még a komolyabb, hoszszabb múltra visszatekintő programcsomagok sem kínálnak azonnali megoldást.

Dolgozatunkban bemutatunk egy lehetséges, egyszerű, könnyen elvégezhető eljárást a lineáris regresszióban található szóráshomogenitási probléma kezelésére. Fontos kiemelnünk, hogy e probléma megoldására számos módszer lehetséges: többek között másfajta regressziós eljárások alkalmazása, vagy akár a vizsgált változókon alkalmazott különböző transzformációk (például logaritmizálás). Nem célunk minden említett módszer áttekintése – helyette egy gyorsan elsajátítható, könnyen alkalmazható megoldást szeretnénk kínálni a fent vázolt probléma megoldására, mely jól illeszkedik a többi, hasonló elven működő módszer megoldásainak sorába.

Kulcsszavak: robusztus lineáris regresszió ▪ legkisebb négyzetek módszere ▪ általánosított legkisebb négyzetek módszere

Abstract

In the analysis of relationships among many variables one usually starts with the less complex methods: the t-test, ANOVA or linear regression with Pearson's correlation.

In many statistical software we can use robust tests for the case when the homogeneity of the variances is not assumed. In the current paper we demonstrate a simple modification for a robust linear regression method when we find heteroscedasticity in our data.

The paper explains our process step by step: it presents – with SPSS snytax – all steps of the algorythm. It is important to note that this is not the only way which can be applied when heteroscedasticity is not assumed (for example we could normalize the variables). We describe a method which is easy to apply and we show the analogy with robust ANOVA methods.

Keywords: robust linear regression ▪ ordinary least square method ▪ generalized least square method

BEVEZETÉS

A hagyományos statisztikai eljárásoknak (t-próba, ANOVA, lineáris regresszió) két általános feltétele van: a vizsgált változók normalitása és a szóráshomogenitás. Ezek mindegyikének megsértésére kisebb-nagyobb mértékben érzékenyek a hagyományos módszerek – bár nagyobb minták esetén a normalitás esetenként nem okoz akkora problémát [lásd pl. Vargha, 2008a]. A kétmintás t-próbának és a varianciaanalízisnek létezik többfajta, bizonyos szempontból összetettebb[1] korrekciós változata, például a Welch-féle korrekció (lásd pl. [Welch, 1947]). A korrekciós eljárások egy jelentős része szabadságfok korrekciós eljárás,[2] ilyen például az összetartozó minták esetén használt Geisser-Greenhouse (lásd pl. [Geisser, Greenhouse, 1958], vagy a Huynh-Feldt féle epszilon korrekció, lásd pl. [Huynh-Feldt, 1970], továbbá mindkét korrekciós eljárásra gyakorlati példát is láthatunk például Vargha [Vargha, 2008a] könyvében).

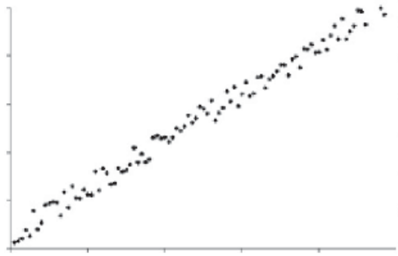
A különböző eljárások bemutatásához, elemzésekhez a www.ropstat.com oldalon található ROPstat programcsomag OECD_2003.msw adatállományát használjuk.

Az általunk alkalmazott robusztus eljárás másfajta megközelítést használ: súlyfüggvény segítségével hajt végre korrekciót. A kiugró (a regressziós egyenestől távol elhelyezkedő) mintaelemeket kisebb, míg a regressziós egyeneshez közeli mintaelemeket nagyobb súllyal vesszük figyelembe. Ha úgy tetszik, az eljárás a trimmeléssel rokon eljárásnak tekinthető – azonban a trimmeléssel ellentétben, ebben az esetben minden mintaelemet továbbra is figyelembe veszünk – csak eltérő mértékben.

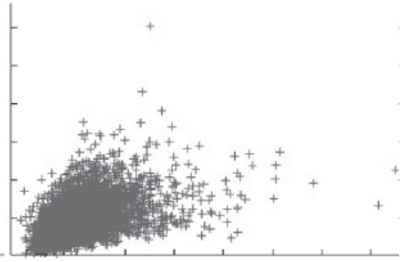
Az elemzéseket SPSS segítségével hajtottuk végre. A program output elemeit változtatás nélkül mutatjuk be – ezt abbéli megfontolásból tesszük, hogy ha e tanulmány olvasója kipróbálja az általunk végrehajtott eljárást, úgy ellenőrizni tudja az eredményeit. Minden fejezet végén a fejezethez használt számítások SPSS parancssorát is megadjuk, hogy az adatállomány letöltése után reprodukálhatóak legyenek a számításaink.

HOMOSZKEDASZTICITÁS TESZTELÉSE

A homoszkedaszticitás (a lineáris regressziónál feltételként szereplő szóráshomogenitás) azt jelenti, hogy a regressziós hibának (reziduálisnak) azonos mértékűnek kell lennie a magyarázó (független) változó bármely szintje mellett (ez hasonlóságot mutat a kétmintás t-próba és varianciaanalízis szóráshomogenitásával, hiszen a definíció formálisan azonos). Az alábbi ábrákon egy-egy tipikusabb megjelenését látjuk a homoszkedaszticitásnak és a heteroszkedaszticitásnak.



1. ábra Példa homoszkedaszticitásra



2. ábra Példa heteroszkedaszticitásra

A homoszkedaszticitást azonban másként kell tesztelnünk, mint a másik két eljárásban a szóráshomogenitást, ugyanis mint már említettük, a regresszió alkalmazásakor nem áll rendelkezésünkre a Levene-próba. Helyette azt ellenőrizzük, hogy a reziduálisnak – egészen pontosan annak abszolút értékének vagy négyzetének[3] – van-e bárminemű lineáris kapcsolata a független változóval (magyarázó változóval).

Ez más megközelítésben azt jelenti, hogy a regressziós modell hibája függ-e a magyarázó változó nagyságszintjétől? Itt – maradva annál a célnál, hogy egy egyszerű módszert adhassunk az alkalmazó kezébe – korrelációs együtthatót fogunk tesztelni:

1. Amennyiben a reziduális négyzetének (vagy abszolút értékének) a korrelációja a független változóval szignifikánsan eltér 0-tól, úgy korrekcióra van szükség a lineáris regresszióban (csakúgy, mint a kétmintás t-próba és varianciaanalízis esetén, ha sérül a szóráshomogenitás).
2. Amennyiben a reziduális négyzetének (vagy abszolút értékének) a korrelációja szignifikánsan nem tér el 0-tól, úgy nem alkalmazunk korrekciót. Fontos azonban kiemelnünk, hogy ebben az esetben is sérülhet a homoszkedaszticitás (hiszen a korrelációs együtthatóval kizárólag lineáris összefüggéseket tesztelünk, más típusú összefüggés fennállhat a korreláció 0 volta esetén is).

A homoszkedaszticitás tesztelésének (ahogy a szóráshomogenitás tesztelésének is) számos eljárása ismert, például könnyen használható eljárást ismertet Breusch és Pagan [Breusch, 1979]. Azonban egy olyan általános, de mégis egyszerűbb eljárást kívánunk most bemutatni, melyet a programcsomagon belül, külön eljárások megírása nélkül is el tudunk végezni – tehát nincsen szükség más program, például EXCEL© alkalmazására.

Ahhoz, hogy a homoszkedaszticitást teszteljük, elegendő annyit tennünk, hogy a reziduális négyzetének, valamint a független változónak a korrelációját meghatározzuk. Amennyiben szignifikánsan különbözik 0-tól a reziduális négyzetének és a független változónak a Pearson-féle korrelációs együtthatója, úgy korrekcióra van szükség.[4]

SZÁMÍTÁSI EREDMÉNYEK AZ OECD2003 ADATÁLLOMÁNYON

Az OECD2003 adatállományon[5] azt teszteljük a továbbiakban, hogy a diákok morálja és a matematikai teljesítmény között milyen lineáris összefüggés írható fel. Miként tudunk előrejelzést adni a matematika teljesítményre a diákok moráljának ismeretében?

Amennyiben elvégezzük a lineáris regressziós elemzést, a következő eredményre jutunk:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,360[a]	,130	,129	80,8042558

a. Predictors: (Constant), Diákok morálja - OECD átlag 0, OECD szórás 1

b. Dependent Variable: Matematika, átlag - OECD átlag 500, OECD szórás 100

1. táblázat A diákok morálja és a matematika teljesítmény közötti korrelációs és determinációs együttható

A modell magyarázó ereje 12-13%-os (korrigált determinációs együttható 12,9%), mely szignifikánsan különbözik 0-tól az alábbi teszt alapján:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	938198,093	1	938198,093	143,690	,000[b]
	Residual	6281213,300	962	6529,328		
	Total	7219411,393	963			

a. Dependent Variable: Matematika, átlag - OECD átlag 500, OECD szórás 100

b. Predictors: (Constant), Diákok morálja - OECD átlag 0, OECD szórás 1

2. táblázat A korrelációs és determinációs együttható szignifikanciájának tesztelése

Az F(1,962)=143,69, SIG=0,000 érték arról biztosít minket, hogy a diákok moráljának segítségével szignifikáns modell alkotható a matematikai teljesítmény előrejelzésére.

A regressziós modell – melyet majd korrigálni fogunk – a következő táblázatból kiolvasható:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	502,430	,139		3627,222	,000
	Diákok morálja - OECD átlag 0, OECD szórás 1	31,479	,159	,988	197,666	,000

a. Dependent Variable: Matematika, átlag - OECD átlag 500, OECD szórás 100

b. Weighted Least Squares Regression - Weighted by Súlyozás a súlyozott regresszióhoz

3. táblázat A diákok morálja és a matematika teljesítmény közötti lineáris kapcsolatot leíró regressziós modell

Elmondható, hogy az alábbi lineáris összefüggés adható meg a matematikától való aggódás és a matematika teljesítmény között:

Teljesítmény = 502,380+31,458*Morál

Ez tehát azt jelenti, hogy minél jobb a morál a diákok között, annál jobb teljesítményre képesek.

Ha megvizsgáljuk a reziduális négyzete és a független (magyarázó) változó közötti korrelációs együtthatót, akkor az alábbi összefüggések mutatkoznak:

Correlations

		Diákok morálja - OECD átlag 0, OECD szórás 1	res_sq	Unstandardized Residual
Diákok morálja - OECD átlag 0, OECD szórás 1	Pearson Correlation	1	,135**	,000
	Sig. (2-tailed)		,000	1,000
	N	964	964	964
res_sq	Pearson Correlation	,135**	1	-,115**
	Sig. (2-tailed)	,000		,000
	N	964	964	964
Unstandardized Residual	Pearson Correlation	,000	-,115**	1
	Sig. (2-tailed)	1,000	,000	
	N	964	964	964

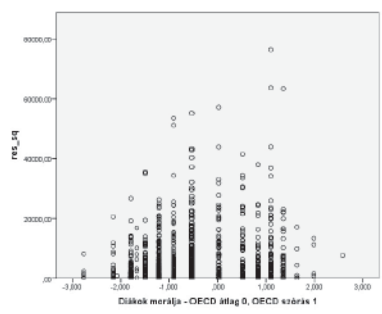
**. Correlation is significant at the 0.01 level (2-tailed).

4. táblázat A heteroszkedaszticitás egyik lehetséges tesztelése Pearson-féle korreláció segítségével

E fenti összefüggés azt mutatja, hogy reziduális négyzete és a diákok morálja között szignifikáns összefüggés van, azaz a diákok moráljának növekedésével a hibák négyzetének növekedése jár együtt – ez azonban definíció szerint azt jelenti, hogy a diákok moráljának különböző szintjei mellett a hibák négyzetének mértéke eltérő: heteroszkedaszticitás mutatkozik az adatokban.

Ez azt jelenti, hogy korrekciós eljárást kell alkalmazni: a meglévő lineáris összefüggésben, egészen pontosan a modell illesztésében nem lehetünk biztosak. A következő fejezetben megmutatjuk, hogy ilyen helyzetben milyen korrekciós módszert alkalmazhatunk.

SPSS parancssor a homoszkedaszticitás teszteléséhez



3. ábra A diákok moráljának növekvő értékei mellett a reziduálisok négyzetei is növekményt mutatnak

```
*** Lineáris regresszió a matematika teljesítmény és diákok
moráljának kapcsolatára, reziduális mentése ***

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT MEAN _ M
  /METHOD=ENTER STMORALE
  /SAVE RESID.

*** Reziduális négyzetének mentése és a homoszkedaszticitás
tesztelése ****

COMPUTE res2=RES _ 1 ** 2.
VARIABLE LABELS res2 ‹res _ sq›.
EXECUTE.

CORRELATIONS
  /VARIABLES=STMORALE res2
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

MÁSFAJTA MÓDSZER A SZÓRÁSHOMOGENITÁS (HOMOSZKEDASZTICITÁS) TESZTELÉSÉRE

Szintén alkalmazható eljárás az, hogy a független (magyarázó) változóval övezeteket hozunk létre, majd varianciaanalízis segítségével vizsgáljuk, hogy a reziduális négyzete miként viselkedik a függő változó különböző övezeteiben.

A varianciaanalízis során azt is kiszámítjuk, hogy a különböző övezetekben van-e a reziduálisok négyzeteinek átlagában valami eltérés (mely övezetben emelkedik meg például a reziduálisok eltérése, vagy mely övezetben a legkisebbek az eltérések).

Jelen esetben a diákok morálját 4 övezetre osztottuk: miután a nemzetközi átlaga e változónak 0 és szórása 1, ezért a 4 kialakított övezet: -1 alatt, -1 és 0 között, 0 és 1 között, illetve 1-nél nagyobb értékek övezetei.

Ezek után megvizsgáltuk a reziduálisok négyzeteit, hogy tapasztalunk-e bárminemű eltéréseket a különböző övezetekben. A kapott eredmények szintén azt támasztották alá, hogy a reziduálisok négyzetei a diákok moráljának különböző értékei mentén eltérő viselkedést mutatnak.

Descriptives

res_sq

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
-2,00	327	4772,6822	5992,56829	331,38978	4120,7498	5424,6145	,16	35568,37
-1,00	306	6825,4049	9157,07554	523,47517	5795,3250	7855,4849	,00	55179,96
1,00	197	7697,2308	9899,62741	705,31925	6506,2416	9288,2200	,03	57104,07
2,00	134	8031,4766	12230,54655	1056,55830	5941,6452	10121,3080	12,72	76412,50
Total	964	6515,7814	9005,13244	290,03579	5946,6064	7084,9565	,00	76412,50

Már a leíró statisztikákból is látszik, hogy az övezetek mentén mind az átlagok, mind a szórások között eltéréseket lehet tapasztalni, mely eltérések az alábbiak szerint szignifikánsak is:

	F	df1	df2	SIG
Levene-teszt eredménye a szóráshomogenitásra	14,283	3	960	0,000
Hagyományos VA eredménye az átlagok egyenlőségének tesztelésére	7,150	3	960	0,000
Welch-féle robusztus VA az átlagok egyenlőségének tesztelésére	8,878	3	383,854	0,000
Brown-Forsythe-féle VA az átlagok egyenlőségének tesztelésére	5,932	3	492,189	0,001

5. táblázat A reziduálisok négyzeteinek viselkedésére vonatkozó tesztek összefoglaló táblázata

A Levene-teszt alapján elmondható, hogy szignifikáns eltéréseket tapasztaltunk a varianciák között (SIG = 0,000), ráadásul a reziduálisok négyzeteinek átlaga is szignifikáns eltérést mutat (mind a hagyományos, mind a robusztus tesztek alapján, SIG = 0,00 mindhárom esetben).

Ez a második fajta tesztelés is megerősíti, amit már a korreláció alapján kijelentettünk: a diákok moráljának különböző értékei mentén a reziduálisok négyzetei különböző viselkedést mutatnak (mind az átlagokban, mind a szórásokban különbségeket tapasztaltunk).

SPSS syntax a másodikfajta tesztelés megvalósítására

```
   *** Diákok moráljának újrakódolása a szórás alapján ***
   *** Szórások alapján hoztunk létre övezeteket **

   RECODE STMORALE (Lowest thru -1=-2) (-1 thru 0=-1) (0 thru
1=1) (1 thru Highest=2)
     (ELSE=SYSMIS) INTO morale _ rec.
   VARIABLE LABELS morale _ rec ‚Diákok morálja, övezetesítés-
sel szórás alapján’.
   EXECUTE.
```

```
   *** Vizsgáljuk meg a reziduális négyzetének viselkedését
az így létrehozott 5 övezetben ***

   ONEWAY res2 BY morale _ rec
   /STATISTICS DESCRIPTIVES HOMOGENEITY BROWNFORSYTHE WELCH
   /MISSING ANALYSIS
   /POSTHOC=TUKEY BTUKEY BONFERRONI GH ALPHA(0.05).
```

EGY LEHETSÉGES ELJÁRÁS A SZÓRÁSHOMOGENITÁS (HOMOSZKEDASZTICITÁS) MEGSÉRTÉSE ESETÉN

A heteroszkedaszticitás megjelenése esetén az általánosan bevett, legkisebb négyzetek módszere helyett súlyozott legkisebb négyzetek módszerét lehet alkalmazni. Mindkét módszert megtalálhatjuk például Myers összefoglaló könyvében [Myers, 2000]. A módszerek közötti különbséget az alábbiak szerint tudjuk vázlatosan összefoglalni.

LEGKISEBB NÉGYZETEK ELVE

Tegyük fel, hogy adott egy Y függő és egy X független változó. Az Y = a + bX + e egyenlettel való közelítés szerint az Y* = a + bX egyenlet a becsült értékeket határozza meg bármely X mért érték esetén.

Ekkor minimalizáljuk az alábbi kifejezést:

$$\sum_{i=1}^{n}(Y_i-a-bX_i)^2$$

A fenti összegzés minimalizálása adja a klasszikus lineáris regresszió megoldását. E minimalizálás tehát nem más, mint a reziduálisok összességének minimalizálása – azt a regressziós egyenletet keressük, amely mentén a teljes, összes eltérés minimális. Ez azt jelenti, hogy a négyzetes eltérés nem más, mint egy-egy személyre vonatkozó becslés hibája, ezek összegét szeretnénk minimalizálni azzal a feltétellel, hogy a becslő függvény csak lineáris lehet. A regressziós egyenes tehát így nem más, mint a minimális hibát elérő egyenes – tehát az összes egyénre vonatkozó becsléseket nézve ezen egyenes adja a legkevesebb összes hibát.

SÚLYOZOTT LEGKISEBB NÉGYZETEK ELVE

A fenti jelölése mellett a súlyozott legkisebb négyzetek módszeréhez bevezetjük az

$$s_t=Y_t-a-bX_t$$

reziduálisokat[6]. Ekkor a súlyozott legkisebb négyzetek módszere szerint a minimalizálandó mennyiség:

$$\sum_{t=1}^{n}W_i(Y_i-a-bX_i)^2$$

A W_t súlyok optimális megválasztása az alábbiak szerint történhet:

$$W_t=1/s_t^2$$

mely úgy interpretálható, hogy azon elemeket, amelyek távol lennének az összefüggést legjobban leíró egyenestől (nagy a reziduális), azokat kisebb súllyal, míg az egyeneshez közel lévő egyedeket nagyobb súllyal vesszük figyelembe.

A fenti súlyozás különbözik a trimmeléstől: a trimmelés során ugyanis egyes elemeket egyáltalán nem veszünk figyelembe – míg ebben az esetben minden mintaelemet figyelembe veszünk, legfeljebb kisebb súllyal[7].

SÚLYOZOTT LEGKISEBB NÉGYZETEK MÓDSZERÉNEK ALKALMAZÁSA

A módszer alkalmazása egyszerű. A regresszió során mentsük el a reziduálisokat, majd azokra alkalmazzuk a fenti transzformációt (vegyük a reciprokának a négyzetét), mely a regressziós eljárás súlyfüggvényét adja.

Ezek után alkalmazzuk a regressziós eljárást úgy, hogy súlyfüggvényként e fent definiált új változót jelöljük meg.

Az eredmények az alábbi táblázatokban foglalhatók össze:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,988[a]	,976	,976	1,0006627

a. Predictors: (Constant), Diákok morálja - OECD átlag 0, OECD szórás 1

6. táblázat Súlyozott legkisebb négyzetekkel számított lineáris regresszió korrelációs és determinációs együtthatója

A modell illesztése ilyenkor értelemszerűen lényegesen jobb lesz: a súlyozott módszerrel számított variancia drasztikusan magasabb lesz, hiszen a kiugró értékeket lényegesen kisebb súllyal, míg az egyeneshez jól illeszkedő pontokat kifejezetten nagy súllyal vesszük figyelembe.

ANOVA[a,b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	39123,602	1	39123,602	39071,801	,000[c]
	Residual	963,275	962	1,001		
	Total	40086,877	963			

a. Dependent Variable: Matematika, átlag - OECD átlag 500, OECD szórás 100

b. Weighted Least Squares Regression - Weighted by Súlyozás a súlyozott regresszióhoz

c. Predictors: (Constant), Diákok morálja - OECD átlag 0, OECD szórás 1

7. táblázat A súlyozott legkisebb négyzetek módszere esetén kapott korrelációs és determinációs együttható szignifikanciájának vizsgálata

A súlyozott módszer szerinti determinációs és korrelációs együttható várhatóan (miután már az eredeti, súlyozás nélküli módszer is erősen szignifikáns volt) szintén szignifikáns lesz.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	502,380	2,874		174,776	,000
	Diákok morálja - OECD átlag 0, OECD szórás 1	31,458	2,624	,360	11,987	,000

a. Dependent Variable: Matematika, átlag - OECD átlag 500, OECD szórás 100

8. táblázat A súlyozott legkisebb négyzetek módszerének eredménytáblája

A módosított, súlyozott módszerrel számított modell kicsit módosított csupán az együtthatókon (a hagyományos módszer szerinti eredményhez képest):

*Teljesítmény=502,430+31,479*Morál*

Fontos észrevennünk: ahogyan adott esetben a kétmintás t-próba, vagy az ANOVA elemzés során sem feltétlenül kapunk a korrekciók utána a hagyományos elemzés eredményeitől lényegesen különböző adatokat, most sem történt ez meg. A korrekciós eljárásokra nem is azért van szükség, mert lényegesen különböző eredményekre számítunk, hanem azért, hogy a számításainkat az elérhető legkorrektebb módon tudjuk elvégezni, így a lehető legkisebb esélyt adva annak, hogy fals eredményekre, következtetésekre jutunk.

SPSS PARANCSSOR A SÚLYOZOTT LEGKISEBB NÉGYZETEK MÓDSZERÉNEK ALKALMAZÁSÁRA

```
*** Súlyváltozó meghatározása – a reziduálisok reciprokának négyzete ***
COMPUTE weight=1 / res2.
VARIABLE LABELS weight ‚Súlyozás a súlyozott regresszióhoz’.
EXECUTE.
```

```
*** Súlyozott legkisebb négyzetek módszerének egyik lehet-
séges alkalmazása

REGRESSION
  /MISSING LISTWISE
  /REGWGT=weight
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT MEAN _ M
  /METHOD=ENTER STMORALE.
```

ÖSSZEFOGLALÓ

A lineáris modellek szóráshomogenitási feltételét bizonyos eljárásokban könynyen kezelni tudjuk, hiszen a statisztikai programcsomagok általában külön kérés nélkül is kiszámítják mind a rájuk vonatkozó teszteket, mind a sérülésük esetén alkalmazandó robusztus eljárásokat.

Lineáris regresszió esetén a szóráshomogenitás (homoszkedaszticitás) feltételének ellenőrzésére számos eljárás létezik, melyek közül talán a legegyszerűbb, leggyorsabb módszert ismertettük jelen tanulmányunkban. E robusztus eljárás korrigálja a varianciákban (reziduálisokban) fellépő különbségeket – és ezek figyelembevételével számítja ki a lineáris regresszió együtthatóit. Továbbá mindezt úgy tesszük, hogy – ellentétben például a trimmeléssel, nem csökkentjük a minta elemszámát, minden mintaelemet figyelembe veszünk (igaz, eltérő súlyokkal).

A módszerek után minden esetben szerepeltetjük a minta adatállományon alkalmazható SPSS parancssort is, melynek segítségével az olvasó maga is reprodukálhatja bemutatott számításainkat.

JEGYZETEK

1 A Welch-féle korrekció a szabadságfokon és a próbastatisztikán is módosítást hajt végre, ezért tekinthető összetettebbnek.

2 Ellentétben a Welch-féle korrekcióval, a Geisser-Greenhouse és Huynh-Feldt-korrekciók a szabadságfokon módosítanak, a próbastatisztikát változtatások nélkül hagyják.

3 A varianciák egyenlőségének tesztelésével való analógia fenntartása érdekében az abszolút érték helyett minden esetben a négyzetet fogjuk tesztelni.

4 Fontos kiemelni, hogy ez egy igazán gyors teszt, de közel sem teljes körű!

5 Az adatok letölthetők a www.ropstat.com oldalról, a ROPstat programcsomag beépített adatállományát használjuk az elemzéshez.

6 A korábbiak alapján e reziduálisoknak kell bármely Y esetén azonos mértékűeknek lenniük a homoszkedaszticitási feltétel szerint.

7 Bizonyos szempontból hasonló megfontolás található a Wilcox-féle robusztus korrelációs együttható alkalmazásánál, lásd (Wilcox, 1994)

BIBLIOGRÁFIA

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, *47*(5), 1287-1294.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics, 29*, 885-891.

Huynh, H. & Feldt, L. S. (1970). Conditions under which mean square ratio sin repeated measurement designs have exact F-distributions. *Journal of the American Statistical Association, 65*, 1582-1589.

Myers, R. H. (1990). *Classical and modern regression with applications* (Vol. 2). Belmont, CA: Duxbury Press.

Vargha A. (2008a): *Matematikai statisztika – Pszichológiai, nyelvészeti és biológiai alkalmazásokkal* (2nd ed. pp.364.). Budapest: Pólya Kiadó.

Vargha A. (2008b). Új statisztikai módszerekkel új lehetőségek: a ROPstat a pszichológiai kutatások szolgálatában. *Pszichológia, 28*(1), 81-103.

Welch, B. L. (1947). The generalization of „Student's" problem when several different population variances are involved. *Biometrika, 34*(1-2), 28-35.

Wilcox, R. R. (1994). The percentage bend correlation coefficient. *Psychometrika, 59*(4), 601-616.