HAJDU OTTÓ*

A szegénységi küszöb alá kerülés esélyének elemzése logisztikus regressziószámítás alkalmazásával

I. Bevezetés

A *szegénységi küszöb relatív* jellegű rögzítésének elterjedt módszerei a medián adott százalékát, vagy az alsó decilist, kvintilist, kvartilist adni meg közvetlenül küszöbértékként. A kvantilisek *robusztusak* az extrém „outlierek" tekintetében. Ugyanakkor, a különböző társadalmi rétegekben a küszöb szintje rétegspecifikus. Ha *medián* alapú a küszöbszint meghatározása, kézenfekvő a medián feltételes értékét rétegképző regresszor változókkal magyarázni, rétegspecifikus medián becslést kapva ezáltal. Viszont, mivel adott szegénységi dimenzió - jövedelem, fogyasztás, kiadás, vagyon – a szóródás tekintetében *heteroszkedasztikus*, logikus nem a regresszált medián valamely százalékát használni küszöbként, hanem egy alkalmas tau%-rendű kvantilist közvetlenül regresszálni alkalmas prediktorok alapján. A tanulmány első része rámutat a kvantilis regresszió specifikus alkalmazására, második része pedig a regressziós prediktorok problémáival foglalkozik.

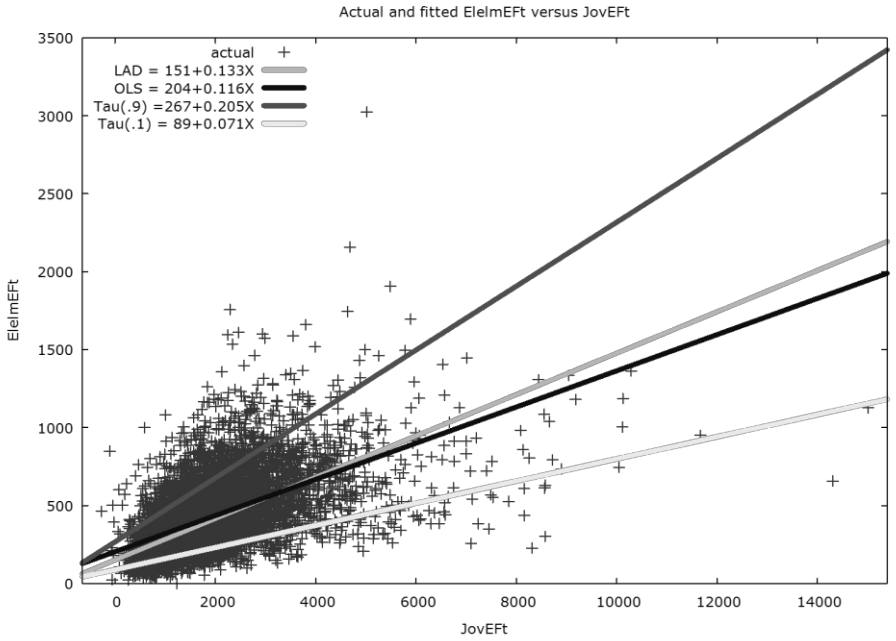
* egyetemi tanár, MTA doktora, Eötvös Loránd Tudományegyetem, Neumann János Egyetem

II. A kvantilis regresszió

Az alábbiakban *n*=8314 magyar háztartás adott évi *élelmiszer kiadásait* (EFt) ábrázoljuk az *éves jövedelmeik* (EFt) függvényében:

1. ábra

Élelmiszerkiadás vs. jövedelem „Engel-görbék”



A pontfelhő jellegzetességei: i) outlierek jelennek meg mind Jövedelem, mind Kiadás tekintetben, ii) a Kiadás terjedelme a jövedelmi szint emelkedésével tágul. Látható, hogy egyetlen regressziós egyenessel nem lehet leírni a pontfelhőt, és ha éppen a „centrális tendenciát” modellezzük, akkor az OLS egyenes alkalmazása nem megfelelő, mert az átlag érzékeny az *outlierekre*, és jelen adatfelhő outlierektől terhelt. Az egyre szélesedő pontfelhőt érdemes tehát kvantilisenként regresszálni, így megőrizzük az eloszlás extrém széleinek az információit is.

Az 1. ábra 4 regressziós egyenest ábrázol, rögzített *X* jövedelmi szintek mellett, és a becsült egyenesek rendre:

1. OLS: A várható, átlagos kiadást becsli: 204 + 0.116X
2. LAD: tau(0.5): A várható medián kiadást becsli: 151 + 0.133X
3. *tau*(0.1): A várható alsó decilis kiadást becsli: 89 + 0.071X
4. *tau*(0.9): A várható felső decilis kiadást becsli: 267 + 0.205X.[1]

A robusztus centrális-tendencia módszerként adódik a *medián* modellezése.[2]

[1] A várható felső decilis közelítése csak a teljesség igénye miatt került ábrázolásra.

Mikor a függő változó empirikus értékei a LAD regresszióval nem párhuzamosan alakulnak, hanem az *X* prediktor változó tekintetében szétnyílnak, zárulnak, kvadratikusak, akkor maga a centrális tendencia modell nem adekvát, és fölmerül az igény a függő változó eloszlásának valamely *tau-rendű feltételes kvantilisét* prediktálni. Míg a *centrális kiadás* leírására a *feltételes mediánt* modellezzük, addig az alacsony kiadások esetén a *feltételes alsó decilis* modellezése is egy járandó út. Bár „Outlier" kiadások hiánya esetén az OLS módszer lehet adekvát a centrális értékre, de a *nem medián* kvantilis értékek regresszálása ekkor is feladat marad a heteroszkedasztikus, volatilis kiadás okán.
Jelölje *diff* a regresszió eltérését az empirikus Y-értéktől: regresszió fölötti megfigyelés *pozitív diff* értéket, regresszió alatti megfigyelés pedig *negatív diff* értéket eredményez:[3]

$$\boldsymbol{diff_i = Y_i - \underbrace{Q_{TAU}|X_i}_{=\beta X_i}}$$

Ebben a *béta*X* regresszióban a *diff távolságok összegét minimáljuk*, ahol pozitív *diff* értékeknek nagyobb, mint 0.5 súlyt adva a regressziós egyenest *fölfelé* húzzuk el, míg negatív *diff* értékeknek nagyobb, mint 0.5 súlyt adva a regressziós egyenest az alsó szegmensbe húzzuk le. A szegénységi küszöb becslésekor ez utóbbi eset a cél.
A *tau*-regresszió súlyozott regresszió célfüggvénye általánosságban:

$$\sum_{i=1}^{n} \begin{cases} \tau * (diff > 0) \\ (\tau - 1) * (diff \leq 0) \end{cases} \rightarrow min$$

ahol pl. az alsó decilis modelljében *tau*=0.1 esetén a célfüggvény:

$$\sum_{i=1}^{n} \begin{cases} 0.1 * (diff > 0) \\ (-0.9 * (diff \leq 0) \end{cases} \rightarrow min$$

A magyarázó változók körét bővítettük a *specifikációs torzítás csökkentése* miatt, az 1.–2. táblák szerint. A „*kiadási határhajlandóságot*" vizsgálva (most lineáris esetben a parciális Jövedelem-koefficiens) a LAD medián becslés 73 Ft. Összevetve a „*csak jövedelem*" prediktor modellel, jelentős a specifikációs torzítottság: LAD esetben 0.133.

A kiemelt értékek adott X prediktor tekintetében (sorában) azt jelzik, hogy az adott magyarázó változó a megjelölt rendű kvantilis regresszió alkalmazásával szignifikánsan más eredményt mutat, mint másik rendű kvantilis regressziók alkalmazásával.

A becsült koefficiensekkel bármely réteg deprivációs küszöbszintje egyszerű X behelyettesítéssel kalkulálható, ahol a vizsgált X faktorok:

- Településtípus: Budapest/Nagyváros/Többi város,
- A háztartás mérete: Háztartás tagszáma, Lakásértéke, Gépkocsi futása,
- Üdülő: van/nincs,
- Foglalkoztatottság: Vállalkozók száma, Aktív keresők száma, Munkanélküliek száma, Eltartottak száma,
- Demográfiai jellemzők: Háztartásfő neme, Iskolai végzettsége, Kora,
- Háztartás jövedelme.

[2] A LAD (Least Absolute Deviation) medián regresszió a medián abszolút érték minimum tulajdonságát használja a regressziós koefficiensek becslése érdekében.

[3] A *Tau*=0.5 rendű Q-kvantilis medián eset kiterjeszthető bármilyen más 0<*Tau*<1 kvantilis esetére a *Tau* paraméter megfelelő megválasztásával.

Az empirikus eredményeket az 1. és 2. táblák közlik. Az 1. táblázat a kvantilis regressziók becsült koefficienseit, a 2. táblázat pedig azok *p*-szignifikancia értékeiket (*p-value*) tartalmazza. Az 1. táblázat szerint:

1. A *tau*=0.5 LAD-medián, és az OLS-átlag marginális hatások (koefficiensek) jelentősen eltérnek egymástól, a *vállalkozók száma* prediktornál pedig az előjelben is különböznek.
2. A „const" tengelymetszet *tau* növelésével növekszik, és negatív előjelről indulva pozitív előjelűre vált át.
3. A DBpNvTv_3 dummy hatás *tau*=0.05 szinten markánsan pozitív, egyébként markánsan negatív!
4. Az „Üdülő van-e, vagy nincs" prediktor esetén a marginális koefficiens hatás egy viszonylag stabil negatív szintről *tau* extrém 0.9, 0.95-re való emelkedésével abszolút értékben igen nagy mértékben emelkedik, míg az egyik esetben negatív, az utolsó esetben viszont pozitív előjelű.
5. Az Akaike, Hannan-Quinn és Schwarz kritériumok egyaránt a *tau*=0.25 kvantilis regressziót preferálják.

Konkrét X-feltétel melletti szegénységi küszöb kalkulálását az Olvasóra bízzuk.

1. táblázat

A regressziós koefficiensek értékei, különböző kvantilisek mellett

Quantile estimates, using observations 1-8314								
tau =	0.05	0.1	0.25	0.5	0.75	0.9	0.95	OLS
Coefficient	Dependent variable: ElelmEFt (medián=372.3)							
const	-28.17	-7.97	7.38	36.18	89.35	129.53	171.15	46.39
DBpNvTvKo_1	-22.44	-19.23	-30.52	-27.98	-28.03	-35.94	-14.68	-31.65
DBpNvTvKo_2	-1.27	-2.11	-12.42	-0.01	-1.48	-26.26	-39.10	-8.61
DBpNvTvKo_3	4.50	-1.32	-3.78	-2.53	-6.66	-16.31	-31.12	-7.02
TLetszam	32.99	34.20	44.40	51.85	65.15	79.08	97.66	52.15
LakasMFt	0.09	0.37	0.75	0.69	1.05	1.91	1.50	0.76
GepKoEKm	0.72	0.96	0.86	1.25	1.83	2.95	3.28	1.32
UduloVan	-5.77	-8.48	-2.21	-3.90	-6.39	-27.90	17.78	-1.37
Vallalk	-5.65	-9.50	0.24	-6.60	10.77	24.99	22.24	2.98
AKeres	7.85	7.41	8.30	5.59	-0.20	-7.72	-15.49	4.83
Mnelkuli	-14.13	-18.78	-16.73	-10.86	-18.47	-33.19	-51.01	-17.16
Eltartott	6.05	10.65	9.58	11.71	7.50	-2.52	-14.45	13.29
HFneme	7.64	19.49	24.03	27.43	32.41	31.36	39.14	31.54
HFiskv	2.91	2.80	3.27	3.03	2.46	2.28	1.59	3.79
HFkora	0.81	0.70	0.69	0.58	0.36	0.36	-0.09	0.77
JovEFt	0.035	0.038	0.050	0.073	0.090	0.119	0.137	0.069
Akaike criterion	**109397**	**108308.1**	**107195.7**	**107618.8**	**110252.9**	**114461.6**	**117501.1**	
Hannan-Quinn	**109436**	**108346.5**	**107234.1**	**107657.2**	**110291.3**	**114500**	**117539.5**	
Schwarz criterion	**109509.8**	**108420.5**	**107308.1**	**107731.2**	**110365.3**	**114574.0**	**117613.5**	

A 2. táblából látható, hogy adott *tau-kvantilis* rend mellett a *p*-értékek jelentősen széthúzódnak – de adott esetben stabilak is maradnak prediktor függően, és pl. a LakásértékMFt esetében jelentős elhatárolódás tapasztalható.

A táblában kiemelten szerepelnek azon szignifikancia *p*-értékek, melyek markánsan különböznek az adott prediktor más *tau*-szinten nyert *p*-értékektől.

2. táblázat

A kvantilis regresszió becsült koefficienseinek szignifikancia (p) értékei

Dependent variable: ElelmEFt										
Változó	Tau=		0.05	0.1	0.25	0.5	0.75	0.9	0.95	OLS
		p-value								
const			0.02	0.54	0.42	0.00	<0,00001	<0,00001	<0,00001	0.00
DBpNvTvKo_1			0.00	0.00	<0,00001	<0,00001	0.00	0.00	0.27	<0,00001
DBpNvTvKo_2			0.82	0.72	0.00	0.999	0.81	0.02	0.00	0.11
DBpNvTvKo_3			0.37	0.80	0.31	0.55	0.23	0.10	0.00	0.15
TLetszam			<0,00001	<0,00001	<0,00001	<0,00001	<0,00001	<0,00001	<0,00001	<0,00001
LakasMFt			0.74	0.23	0.00	0.00	0.00	0.00	0.02	0.01
GepKoEKm			0.00	0.00	<0,00001	<0,00001	<0,00001	<0,00001	<0,00001	<0,00001
UduloVan			0.51	0.36	0.73	0.59	0.51	0.11	0.35	0.87
Vallalk			0.31	0.11	0.95	0.16	0.08	0.02	0.06	0.58
AKeres			0.01	0.02	0.00	0.03	0.95	0.20	0.02	0.10
Mnelkuli			0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00
Eltartott			0.12	0.01	0.00	0.00	0.08	0.75	0.09	0.00
HFneme			0.12	0.00	<0,00001	<0,00001	<0,00001	0.00	0.00	<0,00001
HFiskv			0.00	0.00	<0,00001	<0,00001	0.00	0.14	0.35	<0,00001
HFkora			<0,00001	0.00	<0,00001	0.00	0.06	0.29	0.81	<0,00001
JovEFt			<0,00001	<0,00001	<0,00001	<0,00001	<0,00001	<0,00001	<0,00001	<0,00001

III. Logisztikus regresszió alkalmazása a szegénységi prediktorok szelektálásában

A logisztikus regresszió a klasszifikálás egyik alapvető módszere, így alkalmazása *a szegénység mérésében* is kézenfekvő. Mikor a függő változó „Igen/Nem”, „Szegény/Nemszegény” kimenetű, mint esetünkben, akkor a dichotom regresszió alkalmazandó. A függővváltozó eloszlása ismeretében a regressziós paraméterek becslésére a maximum likelihood (ML) módszer alkalmas, de kedvező tulajdonságai (minimum variancia, konzisztencia) csak *nagymintás* esetben, aszimptotikusan érvényesek. A szegénységi küszöb szerinti klasszifikálás a kismintás, ritka-esemény következtetés tipikus esete, ha a küszöb alá kerülés adott rétegen belül ritka esemény. Háztartástípus-szerinti rétegzés esetén a kismintás becslés *esetileg* szükségszerű adottság.[4]

Ha az aszimptotikus ML becslés nem létezik, az ELR módszerrel[5] akkor is következtetni tudunk a regressziós paraméterekre. Az alábbiakban a releváns szegénységi

[4] A ritka, kismintás, „Igen” esemény kezelését az egzakt permutációin alapuló *egzakt logisztikus* regresszió (ELR) szolgálja. Az ELR eljárás a regressziós paraméterek *elégséges statisztikáinak* az egzakt, feltételes, permutációs eloszlásán alapuló módszere.

[5] Exact Logistic Regression: www.cytel.com

regressziós prediktor változók szelektálására helyezzük a hangsúlyt, mikor a kiválasztás a *p*-value kritérium alapján történik, tehát a korrekt *p*-érték kalkulálása kulcskérdés! A társadalmi-gazdasági indikátorok háztartások sokaságát rétegzik, adott rétegben a mintavétel során *kicsiny méretű, kiegyensúlyozatlan*, hasonló csoportok kialakulása reális helyzet. Ez esetben az „egzakt" következtetés korrekt *p*-értéket, és konfidencia intervallumot ad a kérdéses paraméterekre.

Módszertani vetületben tekintsük a bináris véletlen változókat, ahol az Y_i megfigyelés háztartást azonosít. A *response* Y_i változó az „1" értéket veszi fel küszöb alatti háztartás esetén, egyébként értéke zéró. Az Y_i változónak megfelelően, a regressziós prediktorok (p' 1) rendű kovariánsa $x_i=(x_{i1}, x_{i2},\ldots, x_{ip})'$.

Jelölje p_x a $Pr(Y=1|x)$ feltételes valószínűséget. A $p_x/(1- p_x)$ A *odds*-arány alapján az „1" esemény feltételes valószínűsége:

$$p_x = \frac{p_x/(1-p_x)}{1+p_x/(1-p_x)} = \frac{odds_x}{1+odds_x}$$

Ha p_x meghaladja a rögzített *C* kritikus értéket, akkor az előrejelzés $\hat{Y}=1$ egyébként előrejelzés $\hat{Y}=0$.[6] Rétegzett a modell, mikor minden réteget egy rétegspecifikus konstans jellemez, de *közös „meredekség"* paraméterrel.

A mintavételi következtetés három módja áll rendelkezésre: a feltétel nélküli likelihood, a feltételes likelihood, és a feltételes *egzakt* következtetés. Az *R* visszautasítási tartomány megválasztása az egzakt teszt típusának a megválasztásán múlik. Erre három módszert tekintünk:

i) exact conditional scores teszt (akár aszimptotikus, akár egzakt variancia alapú), ii) exact conditional probability teszt, iii) exact likelihood ratio teszt.

Az *exact conditional scores teszt* esetén az *R* régiót a teszt statisztika mindazon értékei alkotják, melyek nagyobb-egyenlők, mint a teszt statisztika megfigyelt értéke. Az *exact conditional probability teszt* esetén, az *R* régiót a teszt statisztika mindazon értékei alkotják, melyek valószínűsége kisebb-egyenlő, mint a teszt statisztika megfigyelt értékének a valószínűsége. Az *exact likelihood ratio teszt* esetén, az *R* régiót a teszt statisztika mindazon értékei alkotják, melyek LR értékei nagyobb-egyenlőek, mint a megfigyelt adat LR értéke.

A különbség az UMLE és a CMLE következtetés között, hogy míg UMLE igényli a H_1: $\boldsymbol{\beta}_2$ zavaró paraméter becslését is, addig CMLE kontroll alatt tartja, és csak $\boldsymbol{\beta}_1$ becslésére koncentrál. Hipotézisteszteléskor a

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$$

null-hipotézis tesztelésére a *scores* statisztika[7], a *likelihood ratio* statisztika és a *Wald* statisztika áll rendelkezésre. Mindhárom aszimptotikusan Chi^2 eloszlású *df* szabadsági fokkal H_0 érvénye mellett, ahol *df* az alkalmazott megszorítások száma. Hangsúlyozzuk, hogy a *scores statisztika* nem igényli a *full* modell MLE becslését, csak a restriktív modell becslésén alapul. *Ez azt eredményezi, hogy a scores statisztika létezhet akkor is, mikor a full modell MLE becslése nem létezik.*

---

[6] A logisztikus regresszió szerint az *odds logaritmusa* a **x** prediktorok lineáris függvénye: $\log(odds_x)=\boldsymbol{\beta}\mathbf{x}$, ahol $\boldsymbol{\beta}$ az ismeretlen paraméterek ($1'p$) vektora.

[7] Másképp Lagrange-Multiplier teszt statisztika.

Tekintsük a legalább hattagú budapesti háztartásokat, adott évben.[8] A medián jövedelem 60 százaléka alatti háztartásokat kezeljük szegényként.[9] A *szegényvolt* a *Poverty*={0,1} bináris *response* változóban kódolt, ahol „1" szegény háztartást jelöl.

A becslési eredmények a 3. táblában, a prediktorok eloszlásai pedig a 4. táblában láthatók.[10]

Elsőként a háztartásfő nemét véve mint egyedi prediktor változót (Modell 1), a „Nő" egy perfekt prediktor, így az MLE nem létezik (ezt jelzi a ? jel) miközben az MUE pontbecslés és az egyoldali CI elérhető. CI felső határa +INF, mert a zéró gyakoriság megjelenik a Nem terjedelmének alsó extrém értékénél, vagyis a Nőknél, mikor Nem=0.

Szemben ezzel, tekintsünk egy másik bináris prediktort, nevezetesen, hogy van-e tartósan beteg a háztartásban: „1:van", „0: nincs" (Modell 2).

A konklúziók hasonlóak a fentiekhez azon kivétellel, hogy CI alsó határa (–INF), mivel a zéró frekvencia megjelenik a *tartósan beteg jelenlét* terjedelmének felső extrém értékénél.

Kategóriák összevonása is befolyásolhatja az MLE létezését. Tekintsük ugyanis a háztartásfő iskolai végzettségét mint egyedi prediktort (Modell 3).[11] Látható, hogy mind az MLE mind a CMLE létezik, a tény ellenére, hogy zéró gyakoriságok csak az eloszlás alsó szélén jelennek meg. Azonban, összevonva a végzettség szinteket három kategóriába az MLE már nem létezik, ahogy ez a Modell 4 alatt látható.

A relatíve magas mintaméret ellenére – a minta kiegyensúlyozatlan volta (a szegény/nem szegény arány 642/6895) miatt – várható lenne, hogy az aszimptotikus és az egzakt *p*-értékek jelentősen különböznek. Vegyük a *munkanélküli személyek számát* a háztartásban mint egyedüli prediktort (Modell 5). Esetünkben ez nem történik meg, mert a munkanélküliek száma bármely szinten szignifikáns, és a pont és intervallum becslések értékei teljesen hasonlók.

Az *eltartott személyek száma* tekintetében Modell 6 mutatja, hogy az egzakt *p*-value jelentősen különbözhet a feltétel nélküli megfelelőjétől. Bár az eltartottak száma példánkban semmilyen megszokott szinten nem releváns, de extrém kritikus szintet alkalmazva a két módszer eltérő konklúzióra vezetne. E jelenség bármely prediktor esetén előállhat, réteg függvényében.

3. táblázat

Paraméterbecslés, mikor az MLE nem létezik

Modell 1	Type	Beta	SE(Beta)	Type	95%CI Lower	95%CI Upper	$2 * p_1 = p_2$
Const	MLE	?	?	Asymptotic	?	?	?
Nem	MLE	?	?	Asymptotic	?	?	?
	MUE	4.481	NA	Exact	2.804	+INF	1.094e-024
Modell 2							
Const	MLE	?	?	Asymptotic	?	?	?
Tartósan beteg	MLE	?	?	Asymptotic	?	?	?
	MUE	-5.29	NA	Exact	-INF	-3.614	5.809e-052

[8] KSH, Háztartási Költségvetési Felvétel, 2003.

[9] Az egy fogyasztási egységre jutó medián jövedelem 2003-ban 754.000 HUF, ahol 1, 0.7 és 0.5 az első és a további felnőtteket, majd a gyermekeket reprezentálja.

[10] A számítások a LogXact 7 programmal készültek (www.cytel.com).

[11] Az iskolai végzettség score (kód) teljes terjedelme: [1,2,...,13] ahol 13 PhD fokozatot jelöl.

Modell 3							
Const	MLE	-8.522	0.3566	Asymptotic	-9.221	-7.823	2.493e-051
Iskola-score	MLE	0.5927	0.03053	Asymptotic	0.5328	0.6525	2.327e-043
	CMLE	0.5926	0.03053	Exact	0.534	0.6547	3.763e-202
Modell 4							
Const	MLE	?	?	Asymptotic	?	?	?
Iskolai végzettség	MLE	?	?	Asymptotic	?	?	?
	MUE	7.092	NA	Exact	5.418	+INF	6.977e-257

NA: not applicable, ?: does not exist, INF: infinite, e: exponent.

4. táblázat

A prediktor változók gyakorisági eloszlásai

Nem	*Poverty*=0	*Poverty*=1	*Total*
0: Nő	601	0	601
1: Férfi	6294	642	6936
Tartósan beteg			
0: nincs	5678	642	6320
1: van	1217	0	1217
Iskola-score			
3	1009	0	1009
5	1573	0	1573
7	370	0	370
8	1383	0	1383
11	545	355	900
12	1809	126	1935
13	206	161	367
Iskolai végzettség			
1	1009	0	1009
2	3326	0	3326
3	2560	642	3202
Munkanélküliek száma			
0	6459	516	6975
1	436	0	436
2	0	126	126
Gazdasági aktivitás			
111 típus	681	0	681
112 típus	986	161	1147
113 típus	410	0	410
114 típus	656	0	656
115 típus	1520	0	1520
117 típus	996	0	996
121 típus	609	481	1090
122 típus	601	0	601
232 típus	436	0	436
Összesen	*6895*	*642*	*7537*

5. táblázat

Paraméterbecslés, mikor az MLE létezik

Modell 5	*Type*	*Beta*	*SE(Beta)*	*Type*	*95%CI Lower*	*95%CI Upper*	*2*1-sided=p_2*
Const	MLE	-2.642	0.04741	Asymptotic	-2.735	-2.549	3.92e-085
Munkanélküliek	MLE	1.491	0.07773	Asymptotic	1.339	1.644	6.443e-043
	CMLE	1.491	0.07772	Exact	1.336	1.647	3.333e-073
Modell 6	*Type*	*Beta*	*SE(Beta)*	*Type*	*95%CI Lower*	*95%CI Upper*	*2*1-sided=p_2*
Const	MLE	-1.459	0.4144	Asymptotic	-2.271	-0.6464	0.000432
Eltartottak	MLE	-0.0877	0.1012	Asymptotic	-0.286	0.1106	0.386
	CMLE	-0.0876	0.1011	Exact	-0.2912	0.1158	0.4143

Elemezzük újra a *munkanélküliek száma a háztartásban* prediktor hatását, de most úgy, hogy a háztartás gazdasági aktivitását – mint rétegképző változót – kontroll alatt tartjuk (Modell 7). Számos réteg képezhető a munkanélküliek számának és a háztartásfő gazdasági aktivitásának a kombinálásával. Kiemelendő, hogy az alkalmazott rétegzés után MLE nem adható, de az egzakt MUE létezik, és az egzakt *p*-érték a táblában mutatja, hogy a „*Munkanélküliek száma*" továbbra is szignifikáns bármely szokásos szinten. Figyeljük meg, hogy mind a tengelymetszet, mind a rétegspecifikus konstansok eliminálódtak a becslésből.

6. táblázat:

Rétegzés a háztartásfő gazdasági aktivitása szerint

Modell 7	*Type*	*Beta*	*SE(Beta)*	*Type*	*95% CI Lower*	*95% CI Upper*	*2*1-sided=p_2*
Munkanélküliek	MLE	?	?	Asymptotic	?	?	?
	MUE	2.868	NA	Exact	2.023	+INF	9.471e-050
Modell 8							
Iskola-Score	MLE	-0.2139	0.09588	Asymptotic	-0.4018	-0.02596	0.0257
	CMLE	-0.2139	0.09588	Exact	-0.4065	-0.02154	0.02889

A táblázat újra tekinti a háztartásfő iskolai végzettségének 13 fokozatú változóját, de most a rétegzett módon. Bár mind az MLE mind a CMLE létezik, de a prediktor 2% szinten már nem szignifikáns, sőt a koefficiensek előjelei is megváltoztak. A tengelymetszet és a specifikus konstansok most is eliminálódtak a becslésből. Az eddigiekben csak a 2*1-sided típusú *p*-value került alkalmazásra, a konzisztenciát biztosítandó a 95% CI határokkal. Azonban az egzakt *p*-érték változik a teszt statisztika speciális *scores*, *likelihood ratio* vagy *Wald* választásától függően is. Különösen akkor, ha a mintaméret extrém alacsony. Az alábbiakban ezt a problémát illusztráljuk. Két prediktorra vonatkozóan az egzakt tesztek eredményeit a „háztartásfő életkora", majd a „háztartás

korábban, valaha elszenvedett-e szegénységet" kérdések/válaszok érdekesek. A mintát leszűkítettük a *6 főnél több tagú, budapesti, férfi háztartásfős* háztartásokra. A táblázat mutatja, hogy az *életkor (Age) esetén* csak a *score* teszt létezik az aszimptotikus tesztek között, de a *p*-értéke 5% döntési szinten más döntésre vezet. Bár az egzakt teszt *p*-értékek most speciálisan azonosak (p=0.07143 egyaránt) ez általában nem szükségszerű. Míg a *p-mid* value az Exact Likelihood Ratio teszt esetén 5% szinten a null hipotézist elutasítja, addig a többi egzakt teszt elfogadja azt. A „*Poverty Ever Before*" kérdés esetén 5% döntési szinten az *Exact Probability Test* döntése eltér a többi típusú egzakt tesztétől, és mind a *p*-value mind a *p-mid* value értékek lényegesen eltérnek.

7. táblázat

Egzakt teszt eredmények

A teszt típusa	*Statistics*	*DF*	*p-value*	*p-mid*
	H_0: Beta_Age=0			
Score	4.317	1	0.03774	NA
Likelihood Ratio	?	?	?	?
Wald	?	?	?	?
Exact Score_asy	4.317	NA	0.07143	0.05357
Exact Score	3.777	NA	0.07143	0.05357
Exact Probability	0.03571	NA	0.07143	0.05357
Exact Likelihood Ratio	8.997	NA	0.07143	0.03571
	H_0: Beta_Poverty Ever Before=0			
Score	6.107	1	0.01347	NA
Likelihood Ratio	?	?	?	?
Wald	?	?	?	?
Exact Score	5.343	NA	0.03571	0.01786
Exact Probability	0.03571	NA	0.07143	0.05357
Exact Likelihood Ratio	8.997	NA	0.03571	0

IV. Konklúziók

A szegénységi küszöb definiálása, majd értékének megadása társadalmi, gazdasági okon át érzékeny feladat, ennek során alapvető hipotézisünk, hogy tekintet nélkül a jövedelmi szintjére, mindenki átérzi a saját relatív szegénységi küszöbét. Ez a szint rétegspecifikus társadalmi, gazdasági, demográfiai bontásban. Rétegen belül ritka esemény lehet a küszöb alá kerülés ténye, és a rétegen belüli alacsony almintaméret is tesztelési problémát okozhat statisztikailag. Jelen tanulmány e kérdésekre keresi a választ. A *szegénységi küszöb relatív* jellegű rögzítésének elterjedt módszere a medián adott százalékát, vagy vala-

mely nevezetes kvantilist adni meg küszöbértékként. A kvantilisek *robusztusak* az extrém *outlierek* tekintetében, ugyanakkor, a különböző társadalmi rétegekben mint küszöbszintek rétegspecifikusan regresszálhatók. Viszont, mivel adott szegénységi dimenzió – jövedelem, fogyasztás, kiadás, vagyon – a szóródás tekintetében *heteroszkedasztikus* alakulású, logikus egy alkalmas tau%-rendű kvantilist közvetlenül regresszálni alkalmas prediktorok alapján. A tanulmány ezen kérdéseket tárgyalja alapvetően.

Irodalomjegyzék

AGRESTI, A. (2002): *Categorical Data Analysis*, 2nd Edition, Wiley.

GARTHWAITE, P.H.; JOLLIFFE,I.T; JONES,B. (1995): *Statistical Inference*. Prentice Hall.

CHRISTMANN, A.; ROUSSEEUW,P.J. (2001): *Measuring overlap in logistic regression*. Computational Statistics and Data Analysis, 37, 65–75. pp.

HAJDU, O.: *A szegénység statisztikai mérése. Egy új, többváltozós módszertan*. GlobeEdit, Saarbrücken, 2017.

HAJDU, O. (2006): *Exact inference on poverty predictors based on logistic regression approach*, Hungarian Statistical Review, special number 10. Vol.84 134–147. pp.

HIRJI,K.F.; MEHTA,C.R.; PATEL,N.R. (1987): *Computing distributions for exact logistic regression*. JASA, 82. 1110–1117. pp.

HIRJI,K.F.; TSIATIS,A.A.; MEHTA,C.R. (1989): *Median unbiased estimation for binary data*. The American Statistician, 43. 7–11. pp.