HAJDU OTTÓ

A relatív deprivációs szegénységi küszöb rétegspecifikus, kvantilis regressziós becslése

Kivonat

A *szegénységi küszöb relatív* rögzítésének alapvető statisztikai módszere adott rendű kvantilist adni meg küszöbként, mivel a kvantilis *robusztus* az *outlier* értékekre. Különböző társadalmi rétegek küszöbei rétegspecifikusak, ezért a kvantilist kézenfekvő a rétegképző változók feltétele mellett regresszálni. Mindenki érezheti magát relatíve depriváltnak valamely „jószág" tekintetében a környezetéhez és vágyaihoz viszonyítva. A relatív depriváció szerint az emberek inkább a társadalom adott csoportjaihoz, és nem a társadalom egészéhez viszonyítják magukat. A tipikus szegénységi dimenziók *heteroszkedasztikusan* szóródnak, és erre tekintettel logikus nem a középértéket, hanem egy *Tau*-rendű kvantilist regresszálni az X prediktorok alapján. Ezt a *célt* szolgálja a *kvantilis regresszió*[11]. A küszöb alá kerülés esélyének a vizsgálatára az *egzakt logisztikus* regresszió módszere szolgál.

1 Bevezetés

A rétegképzés nyomán kialakulhatnak alacsony gyakoriságú, *ritka* elemszámú rétegek. A problémának a háztartás küszöb fölé/alá kerülés *valószínűségének* regressziós becslése szempontjából van jelentősége. Az adekvát regresszió a logisztikus regresszió, de ennek klasszikus Maximum Likelihood becslése csak nagymintás esetben rendelkezik kedvező mintavételi következtetési tulajdonságokkal[12]. Ha ez nem teljesül, javasoljuk az „*Exact Logistic Regression*" módszert ilyen esetekben[13].

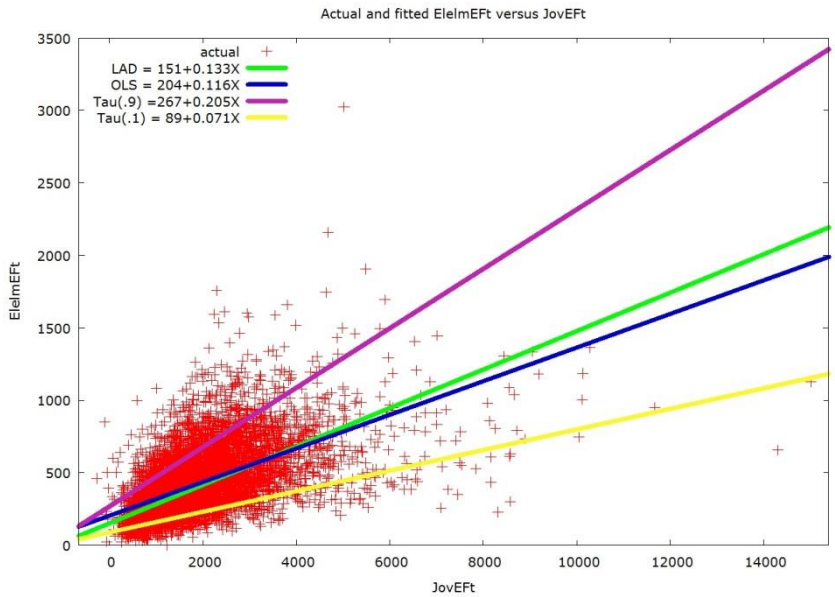
A tanulmány két részből áll. Az első rész rámutat a kvantilis regresszió alapvető előnyeire, mozzanataira, a második rész pedig a regressziós prediktorok rétegzésből fakadó problémáival és egzakt-logit megoldásaival foglalkozik[14].

[11] A módszer leírását, bevezetését lásd pl.: Koenker-Bassett (1978) vagy Koenker-Hallock (2001).
[12] Részletesen lásd Agresti (2002).
[13] www.cytel.com, a módszertan ismertetése itt olvasható: King-Zeng(2001), illetve King-Ryan(2002).
[14] Mindkét megközelítés itt olvasható: Hajdu (2017).

2 A kvantilis regresszió ábrázolása

Az *1. ábrán* 8314 magyar háztartás *éves élelmiszer kiadásait* (eFt) ábrázoljuk az *éves összes jövedelmeik* (eFt) függvényében adott évre.

1. ábra[15] Élelmiszerkiadás vs. Jövedelem „Engel-görbék”



A pontfelhő jellegzetességei: i) outlierek jelennek meg mind a Jövedelem, mind a 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 külön regresszálni, megőrizve így az eloszlás extrém széleinek az információit is.

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

- OLS: A várható, átlagos kiadást becsli: 204 + 0.116X
- LAD: Tau(0.5): A várható medián kiadást becsli: 151 + 0.133X

[15] Az ábra a http://gretl.sourceforge.net/ /Gretl for Windows ökonometriai programmal készült.

- Tau(0.1): A várható alsó decilis kiadást becsli: 89 + 0.071X
- Tau(0.9): A várható felső decilis kiadást becsli: 267 + 0.205X

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 pl. a *feltételes alsó decilis,* míg magas jövedelmek esetén a *felső decilis* modellezése egy járandó út. Bár „Outlier" kiadások hiánya esetén az OLS módszer lehetne adekvát a centrális tendencia értékének leírására, de a *nem-medián* kvantilis értékek regresszálása ekkor is feladat marad a heteroszkedasztikusan, volatilisen alakuló 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:

$$diff_i = Y_i - \underbrace{Q_{\tau au} \mid X_i}_{=\beta X_i}$$

Ebben a *béta·X* regresszióban a *diff távolságok összegét minimáljuk*, ahol pozitív *diff* értékeknek nagyobb mint 0.5 súlyt adva a regressziós egyenest a medián regresszió *fölé* húzzuk el, míg negatív *diff* értékeknek nagyobb mint 0.5 súlyt adva a regressziós egyenest a medián regresszió alatti szegmensbe húzzuk le. 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:

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

ahol értelemszerűen a *Tau*=0.9 esetén a célfüggvény a felső decilis modelljét eredményezi.

A magyarázó változók körét bővítettük a *specifikációs torzítás csökkentése* miatt, az *1.-2. táblázatok* szerint. Mint elemzési célt, a „*kiadási határhajlandóságot*" vizsgálva (ez most lineáris esetben a parciális Jövedelem-koefficiens) a LAD medián becslés 73 Ft. Összevetve a „*csak jövedelem*" prediktor modellel, a specifikációs torzítás jelentős, mert LAD esetben ez 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 regresszióval.

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

- Településtípus: Budapest, Nagyváros, Többi város, Községek
- A háztartás mérete: Háztartás tagszáma (Fő[16]), Lakásértéke (MFt), Gépkocsi éves futása (EKm)
- Üdülő: van/nincs (1;0)
- Foglalkoztatottság: Vállalkozók száma, Aktív keresők száma, Munkanélküliek száma, Eltartottak száma
- Demográfiai jellemzők: Háztartásfő neme, Iskolai végzettsége (1-13PhD), Életkora
- Háztartás jövedelme

Az empirikus eredményeket az *1.* és a *2. táblázatok* közlik. Az *1. táblázat* a kvantilis regressziók becsült koefficienseit, míg a *2. táblázat* azok *p*-szignifikancia értékeiket (*p-value*) tartalmazza[17].

[16] Lehetne valamely definíció szerinti fogyasztási egység is.

[17] Egy konkrét X-háztartás lehet például: Községi, 5_tagú, 10_MFT-lakásérték, 10_Ekm éves gépkocsi futású, Nincs üdülő, 0 fő_vállalkozó, 1 fő_aktív, 1 fő_munkanélküli, 3 fő_eltartott, 40_éves Férfi háztartásfő, 10_Iskola, 1500EFt éves jövedelem. A konkrét X-feltétel melletti háztartás szegénységi küszöbének 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	

Az *1. táblázat* szerinti főbb konklúziók:

1. A *Tau*=0.5 LAD-medián vs. 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, tehát a konkrét adatállomány leírására leginkább az alsó 25% szegénységi küszöb áll a legközelebb. Ez a szó szoros értelmében a szegénységi küszöb becslése.

A *2. táblázatból* 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, amelyek 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

3 A küszöb alá kerülési esély kismintás problémái[18]

A logisztikus regresszió a klasszifikálás alapvető módszere, alkalmazása a *deprivációvizsgálatban* is kézenfekvő[19]. Mivel esetünkben a függő változó "Depriválté/Nemdepriválté" kimenetű, tehát a *dichotom* módszer alkalmazandó, ahol a függő változó bináris eloszlása ismeretében a regressziós paraméterek becslésére a Maximum Likelihood módszer adó-

[18] A szerző ez irányú alkalmazását lásd: Hajdu (2006).
[19] A módszertani alapmű: Agresti (2002).

dik, azonban ennek kedvező tulajdonságai (minimum variancia, konzisztencia) csak *nagymintás* esetben, aszimptotikusan érvényesek. A deprivációs-szegénységi küszöb klasszifikálása ugyanakkor a kismintás következtetés tipikus esete, mikor a küszöb alá kerülés adott rétegen belül ritka esemény[20]. Háztartástípus szerinti rétegzés esetén a kismintás, ritka gyakoriságú rétegek becslési esete szükségszerű adottság. Erre megoldás az egzakt logisztikus regresszió (ELR) alkalmazása[21].

Mikor a nagymintás, aszimptotikusan érvényes Maximum Likelihood becslés *nem is létezik*, az ELR módszerrel akkor is következtetni tudunk a regressziós paraméterekre. A következőkben a releváns deprivációs prediktor változók *korrekt* szelektálására helyezzük a hangsúlyt, mikor a kiválasztás a *p*-kritérium alapján történik, tehát a korrekt *p*-érték kalkulálása kulcskérdés! Mint korábban említettük, a társadalmi-gazdasági indikátorok háztartások sokaságát rétegzik, ezért adott rétegben a mintavétel során *kicsiny, kiegyensúlyozatlan*[22], hasonló csoportok kialakulása reális helyzet. Ez esetben az ún. „egzakt" következtetés ad korrekt *p*-értéket, és konfidencia intervallumot a kérdéses paraméterekre[23].

Az alábbiakban néhány probléma-felvetést nyújtunk praktikus példákon, de a számítási eredményeket terjedelmi okból mellőzzük.

Tekintsük az Y_i=[1;0] bináris véletlen változókat, ahol az „*i*" megfigyelés háztartást, *Y* pedig kiadást azonosít. A *response* Y_i változó az „1" értéket veszi fel küszöb alatti háztartás esetén, egyébként értéke zéró. A regressziós *béta* paraméterekre való mintavételi következtetés három módja áll rendelkezésre: a feltétel nélküli likelihood (UMMLE), a feltételes likelihood (CMLE), és a feltételes *egzakt* következtetés[24].

Az *R* „rejection" visszautasítási tartomány megválasztása az egzakt teszt típusának a megválasztásán múlik. Erre három módszert tekintünk: exact conditional scores teszt (akár aszimptotikus, akár egzakt variancia alapú), exact conditional probability teszt, exact likelihood ratio (LR) teszt[25].

A különbség az UMLE és a CMLE következtetés között, hogy míg UMLE igényli a H_1 zavaró paraméter becslését is, addig CMLE kontroll alatt tartja. A $H_0 : \beta_1 = 0$ hipotézis teszte-

[20] King-Zeng (2001).

[21] A ritka, kismintás, „Igen" esemény kezelését az egzakt permutáción alapuló *egzakt logisztikus* regresszió (ELR) szolgálja. Az ELR eljárás a regressziós paraméterek *elégséges statisztikáinak* az egzakt, feltételes, permutációs eloszlásán alapuló módszere. Lásd: Exact Logistic Regression: www.cytel.com.

[22] Kiegyensúlyozatlan a minta akkor, ha az Igen esetek aránya jelentősen eltér a Nem esetekétől.

[23] Hirji, K.F.-Mehta, C.R.-Patel, N.R.(1987).

[24] Garthwaite-Jolliffe (1995).

[25] Az *exact conditional scores teszt* esetén az *R* régiót a teszt statisztika mindazon értékei alkotják, melyek nagyobbak vagy 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 vagy 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 nagyobbak vagy egyenlők, mint a megfigyelt adat LR értéke.

lésére a *scores* statisztika[26], 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.*

Tekintsük a legalább hattagú budapesti háztartásokat, adott évben[27]. A medián jövedelem 60 százaléka alatti háztartásokat kezeljük szegényként[28]. A *szegényvolt* a *Poverty*={0,1} bináris *response* változóban kódolt, ahol „1" szegény háztartást jelöl. Például a háztartásfő nemét véve mint egyedi prediktor változót, legyen a "Nő" egy perfekt prediktor, így az MLE nem létezik, 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". 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. Mind az MLE mind a CMLE létezik, azon tény ellenére, hogy zéró gyakoriságok csak az eloszlás alsó szélén jelennek meg. Azonban a végzettség szinteket három kategóriába összevonva az MLE már nem létezik.

Hivatkozások

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

Garthwaite, P.H., Jolliffe,I.T., Jones,B. (1995): *Statistical Inference*. Prentice Hall

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

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

Hirji,K.F., Mehta,C.R., Patel,N.R. (1987): *Computing distributions for exact logistic regression*. JASA, 82, pp. 1110-1117.

Koenker,R., Bassett,Jr.G. (1978): Regression Quantiles. *Econometrica*, Vol. 46, No.1. pp. 33-50.

[26] Másképp Lagrange Multiplier teszt statisztika.

[27] KSH, Háztartási Költségvetési Felvétel, 2003.

[28] 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.

Koenker, R., Hallock,K.F. (2001): Quantile Regression. *Journal of Economic Perspectives*, Vol. 15, No. 4, pp. 143-156.

King,G., Zeng,L. (2001): Logistic Regression in Rare Events Data, *Political Analysis*, 9, pp. 137-163.

King,E.N., Ryan,T.P. (2002): A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression. *The American Statistician*, August, Vol. 56, No. 3, pp. 163-170.