Analýza ekvivalentného disponibilného príjmu slovenských domácností s využitím príkazov CONTRAST a LSMEANS v procedúre GLM

Martina Košíková[1], Erik Šoltés[2]

Abstrakt
Kvalita životnej úrovne obyvateľstva je posudzovaná prostredníctvom rôznych indikátorov, pričom v príspevku sme sa zamerali na ekvivalentný disponibilný príjem domácností z pohľadu faktorov ako sú napríklad status ekonomickej aktivity, úroveň vzdelania, zdravotný stav, typ domácností z hľadiska počtu členov zdieľajúcich jednu domácnosť, taktiež aj stupeň urbanizácie alebo kraj, z ktorého domácnosť pochádza.
Na základe databázy EU-SILC 2018 a využitím softvéru SAS Enterprise Guide aplikujeme metódy všeobecných lineárnych modelov na posúdenie vplyvu vybraných ukazovateľov na výšku ekvivalentného disponibilného príjmu domácností Slovenskej republiky. Zároveň sa pri analýze zameriame aj na využitie viacerých dostupných príkazov, prostredníctvom ktorých vieme získať ďalšie informácie popisujúce vzťahy medzi predmetným ukazovateľom a posudzovanými faktormi.

Kľúčové slová
EU-SILC, ekvivalentný disponibilný príjem, všeobecný lineárny model, marginálne stredné hodnoty

Abstract
The life quality of the population is assessed through different indicators, however in the article we focused on the equivalent disposable income of households in terms of factors such as economic activity, level of education, health status, type of household, urbanization and region. We apply the methods of general linear models to assess the effect of selected indicators on the amount of equivalent disposable income of Slovak households using the EU-SILC 2018 database and statistical software SAS Enterprise Guide. In the analysis, we will also use several available statements, through which we can obtain additional information describing the relationships between the indicator and the factors.

Key words
EU-SILC, equivalised disposable income, general linear model, least squares means

JEL classification
C12; C51; R29

1 Úvod

Výberové zisťovanie EU-SILC je vo všeobecnosti základom analýz životnej úrovne obyvateľstva. Sledované ukazovatele zobrazujú úroveň a štruktúru chudoby na národnej úrovni, konkrétne chudobu a sociálnu elimináciu z viacerých aspektov a dimenzií z hľadiska časového vývoja chudoby, typu domácností, zdravotného stavu, vzdelania a pod. (Kováčová a

[1] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra štatistiky, Dolnozemská cesta 1/b, 852 35 Bratislava, e-mail: martina.kosikova@euba.sk

[2] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra štatistiky, Dolnozemská cesta 1/b, 852 35 Bratislava, e-mail: erik.soltes@euba.sk.

Vlačuha, 2019) Slovensko sa z globálneho pohľadu už dlhodobo zaraďuje ku krajinám s neproporcionalitou medzi regiónmi, a to z toho dôvodu, že každý z regiónov sa rôzne ekonomicky rozvíja. Problémom je napríklad rozvoj infraštruktúry, budovanie dopravnej siete, nízka úroveň vzdelania alebo nedostatočne sa rozvíjajúca podnikateľská oblasť.

V súčasnej dobe je chudoba jeden z problémov ovplyvňujúci kvalitu života ľudí v krajine. Podľa výsledkov ŠÚ SR (2019) bolo v roku 2018 na Slovensku chudobou ohrozených až 16,3 % obyvateľstva (približne 872 – tisíc ľudí), čo znamená, že táto časť populácie trpela buď príjmovou chudobou, žila v materiálne znevýhodnenom prostredí, alebo ich pracovná intenzita bola nižšia ako 20 %. Hlavnou podstatou problému chudoby je však nepostačujúca výška príjmu, ktorým jednotlivec alebo domácnosť disponuje. Cieľom článku je na základe všeobecného lineárneho modelu vychádzajúceho z údajov EU-SILC 2018 analyzovať vplyv vybraných faktorov na ekvivalentný disponibilný príjem slovenských domácností. Osobitnú pozornosť venujeme posúdeniu rozdielov cieľovej premennej medzi krajmi SR, a to prostredníctvom analýzy marginálnych stredných hodnôt (Least squares means alebo LS-means) a kontrastnej analýzy.

2 Ekvivalentný disponibilný príjem

Ekvivalentný disponibilný príjem predstavuje mieru príjmov všetkých osôb zdieľajúcich jednu domácnosť. Jeho výpočet spočíva v tom, že sa disponibilný príjem[3] domácností vydelí ekvivalentnou veľkosťou domácností[4] upravenej podľa vybranej stupnice ekvivalencie. Štatistické zisťovanie EU-SILC využíva na výpočet modifikovanú stupnicu OECD (Organisation for Economic Co-operation and Development), ktorá pripisuje všetkým členom domácností konkrétnu váhu, a to nasledovne:

- prvý dospelý člen domácnosti má pridelenú váhu 1,0,
- každý ďalší člen domácnosti vo veku viac ako 14 rokov má pridelenú váhu 0,5 a
- každej osobe vo veku menej ako 14 rokov je pridelená váha 0,3.

Podrobnejšie vysvetlenie riešenej problematiky nájdeme napríklad v Anyaegbu (2010), Atkinson, Rainwater a Smeeding (1995), Bellu (2005), Kováčová a Vlačuha (2019), OECD (2008), Šoltés a kol. (2018), UNECE (2011) a pod.

3 Všeobecné lineárne modely

Podstatou všeobecných lineárnych modelov (z angl. General Linear Models – *GLM*) je modelovanie číselnej vysvetľovanej premennej v závislosti od jednej alebo viacerých kvantitatívnych alebo kvalitatívnych vysvetľujúcich premenných.

Procedúra *GLM* v štatistickom analytickom systéme SAS, ktorú v príspevku využívame, je hlavne spojením regresnej analýzy a analýzy rozptylu. Programovací jazyk SAS poskytuje aj mnoho ďalších analýz prostredníctvom rôznych príkazov na identifikáciu vzájomných vzťahov medzi premennými. V prípade, že model obsahuje viackategoriálne vysvetľujúce premenné, zväčša nás zaujíma, či medzi jednotlivými úrovňami premennej existujú z pohľadu cieľovej premennej štatisticky významné rozdiely. Na identifikáciu rozdielov poskytuje procedúra *GLM* príkazy *MEANS* a *LSMEANS*, ktorých súčasťou sú testy slúžiace na viacnásobné porovnávanie stredných hodnôt (*post hoc testy*). V prípade, že je súbor vyvážený, sú výsledky príkazov *MEANS* a *LSMEANS* identické. Kým príkazom *LSMEANS* sa priemery odhadujú z modelu, príkazom *MEANS* sú priemery odhadované priamo z údajov (tzv.

[3] Disponibilný príjem domácnosti predstavuje celkový čistý príjem domácnosti, resp. všetky zdanené príjmy každého člena domácnosti získané z akýchkoľvek zdrojov.

[4] Ekvivalentná veľkosť domácnosti predstavuje súčet váh všetkých členov domácností.

aritmetické priemery). Príkaz *LSMEANS* je pri analýzach lineárnych modelov viac preferovaný, pretože upravuje výsledky o vplyvy kovariátov[5].

3.1 Využitie príkazov CONTRAST a LSMEANS v procedúre GLM

Všeobecné lineárne modely môžeme využiť na mnoho štatistických analýz, pričom jednou z nich je aj analýza marginálnych stredných hodnôt. Využitím príkazu *LSMEANS* vieme v procedúre *GLM* overiť významnosť diferencií v strednej hodnote vysvetľovanej premennej medzi dvojicami úrovní kategoriálnej vysvetľujúcej premennej.

```
LSMEANS effects </ options>;
```

Príkaz poskytuje informáciu o významnosti, resp. nevýznamnosti rozdielu stredných hodnôt cieľovej premennej pre dvojice obmien kategoriálnej vysvetľujúcej premennej. Ak však chceme overiť zhodu stredných hodnôt vysvetľovanej premennej vo viac ako dvoch kategóriách niektorého faktora, ktorý je v modeli zahrnutý ako vysvetľujúca premenná, tak na tento účel môžeme použiť príkaz *CONTRAST*.

```
CONTRAST 'label' effect values <…effect values> </ options>;
```

Základom úspešného spustenia príkazu je jeho korektné zostavenie, ktoré závisí od formulácie nulovej hypotézy. Dôležité je nulovú hypotézu zadefinovať tak, aby jej výsledná formulácia bola v tvare lineárnej kombinácie parametrov modelu, pretože koeficienty lineárnej kombinácie sa musia zadať do príkazu *CONTRAST*. Ak sa netestuje interakcia, tak postačuje, ak sa určí lineárna kombinácia posudzovaných stredných hodnôt. Súčet koeficientov vstupujúcich do príkazu musí byť nulový, inak sa príkaz v procedúre zanedbá a nespustí sa. Okrem toho treba poznamenať, že zápis a zoradenie koeficientov v príkaze závisí od poradia kategórií kategoriálnej vysvetľujúcej premennej v modeli. Pokiaľ nie je postupnosť koeficientov dodržaná, môže výsledok príkazu vykazovať chybný záver. Bližšiu špecifikáciu príkazov a využitia všeobecných lineárnych modelov nájdeme napríklad v publikáciách Bowley (2013), Howell (2010), Chen (2008), Lenth (2016), Littell a kol. (2010), Mangiafico (2015), Pituch a Stevens (2015), Rutherford (2001), Savarese a Patetta (2010) a pod.

4 Vplyv vybraných faktorov na ekvivalentný disponibilný príjem domácnosti

Z údajov získaných z databázy EU-SILC 2018 sme vytvorili výberový súbor obsahujúci okrem disponibilného príjmu domácností aj tie faktory, ktoré by podľa nášho názoru mohli mať vplyv na jeho výšku. Spojitou vysvetľovanou premennou bude spomínaný ekvivalentný disponibilný príjem (*EDP*) a kvalitatívnymi vysvetľujúcimi premennými, ktoré sme sa rozhodli do analýzy zaradiť, budú: status základnej ekonomickej aktivity (*EAS*), najvyššie dosiahnuté vzdelanie (*EDUCATION*), typ domácnosti (*HT*), rodinný stav (*MARITAL_STATUS*), všeobecné zdravie (*HEALTH*), stupeň urbanizácie (*URBANISATION*) a kraj (*REGION*).

Uvedená tabuľka (tab. 1) obsahuje základné informácie o vstupných premenných, ako aj ich označenie, ktoré v tomto článku budeme používať. Každá z vysvetľujúcich premenných, zaradených do modelu je kategoriálna a obsahuje niekoľko obmien. Vzhľadom na nízku početnosť istých kategórii, sme najpodobnejšie z nich zlúčili a vytvorili jednu spoločnú kategóriu. Pri tomto type premenných je vždy jedna z kategórii zvolená ako referenčná, tzv. porovnávajúca. Referenčné kategórie sme zvolili tak, aby neboli marginálne, inak povedané, aby ich zastúpenie v slovenských domácnostiach nebolo veľmi nízke.

[5] Kovariáty predstavujú vplyv určitej vysvetľovanej premennej, ktorá nie je pre analýzu zaujímavá. Vplyv tejto premennej neinterpretujeme, iba ho berieme do úvahy pri posudzovaní vplyvu ostatných premenných.

Tab. 1: Základné informácie o vstupných premenných

Pôvodné premenné	Označenie
EQ_INC20 - Ekvivalentný disponibilný príjem	**EDP**
RB210 - Status základnej ekonomickej aktivity	**EAS**
pracujúci	**at_Work (referenčná k.)**
nezamestnaný	Unemployed
osoba na dôchodku	Retired
iná neaktívna osoba	Inactive_person
PE040 - Najvyššie dosiahnuté vzdelanie	**EDUCATION**
Primárne, nižšie sekundárne vzdelanie	Less_than_Secondary
Vyššie sekundárne vzdelanie	Upper_Secondary
Post-sekundárne vzdelanie (nie terciárne)	Post_Secondary
Terciárne vzdelanie I. stupňa	Tertiary_1
Terciárne vzdelanie II. a III. Stupňa	**Tertiary_2_3 (referenčná k.)**
HT - Typ domácnosti	**HT**
Jednočlenná domácnosť	1Adult
Domácnosť 2 dospelých, obaja vo veku do 65+	2Adult_0Ch
Domácnosť 2 dospelých, aspoň 1 vo veku 65+	2A_1R
Iné domácnosti bez závislých detí	Other_0Ch
Domácnosť 1 rodiča aspoň s 1 závislým dieťaťom	1A_at_least_1Ch
Domácnosť 2 dospelých s 1 závislým dieťaťom	2A_1Ch
Domácnosť 2 dospelých s 2 závislými deťmi	**2A_2Ch (referenčná k.)**
Domácnosť 2 dospelých s 3+ závislými deťmi	2A_at_least_3Ch
Iné domácnosti so závislými deťmi	Other_with_Ch
PB190 - Rodinný stav	**MARITAL_STATUS**
slobodný/á	Never_married
ženatý/vydatá	**Married (referenčná k.)**
vdovec/vdova	Widowed
rozvedený/á	Divorced
PH010 - Všeobecné zdravie	**HEALTH**
Veľmi dobré, dobré	**Good (referenčná k.)**
Priemerné	Fair
Zlé, veľmi zlé	Bad
DB100 - Stupeň urbanizácie	**URBANISATION**
Územie s hustým osídlením	**Dense (referenčná k.)**
Územie s priemerne hustým osídlením	Intermediate
Územie s riedkym osídlením	Sparse
KRAJ - Kraj	**REGION**
Bratislavský	**BA (referenčná k.)**
Trnavský	TT
Trenčiansky	TN
Nitriansky	NR
Žilinský	ZA
Banskobystrický	BB
Prešovský	PO
Košický	KE

Zdroj: EU-SILC 2018, vlastné spracovanie

Samotnú analýzu sme realizovali vytvorením dvoch modelov. Súčasťou prvého modelu budú všetky premenné tvoriace vstupnú databázu a druhý model bude na základe zistených okolností modifikáciou prvého modelu. Na detailnejšie posúdenie vplyvu faktora *REGION* využijeme kontrastnú analýzu. Tú budeme realizovať príkazom *CONTRAST* v rámci procedúry *PROC GLM*, pričom význam a využitie príkazu *CONTRAST* vysvetlíme priamo pri jeho aplikácii.

4.1 Všeobecný lineárny model EDP domácností s pôvodnými premennými

Na overenie významnosti lineárnych modelov používame analýzu rozptylu vysvetľovanej premennej. Konečné rozhodnutie o prijatí alebo zamietnutí tvrdenia závisí od vypočítanej *F*-štatistiky. Pokiaľ hodnota testovacej štatistiky presiahne hranicu kritickej hodnoty, zamietneme nulovú hypotézu. Keďže výstupy v štatistickom softvéri neposkytujú informáciu o kritickej hodnote, na overenie významnosti môžeme použiť *p*-hodnotu, ktorá predstavuje najnižšiu hladinu významnosti, pri ktorej je možné nulovú hypotézu zamietnuť.

Tab. 2: Test štatistickej významnosti vplyvu vybraných premenných na vysvetľovanú premennú

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EAS	3	3108103329	1036034443	180.46	<.0001
HT	8	6245966277	780745785	136.00	<.0001
EDUCATION	4	2242962036	560740509	97.67	<.0001
MARITAL_STATUS	3	666275740	222091913	38.69	<.0001
URBANISATION	2	207150861	103575431	18.04	<.0001
HEALTH	2	190166421	95083211	16.56	<.0001
REGION	7	373954801	53422114	9.31	<.0001

Zdroj: EU-SILC 2018, vlastné spracovanie v SAS Enterprise Guide

Prostredníctvom kvantifikovaného *F*-testu vplyvu jednotlivých vysvetľujúcich premenných zaradených do modelu na vysvetľovanú premennú *EDP* (ekvivalentný disponibilný príjem) na zvolenej hladine významnosti 0,05 zisťujeme, že každá z vysvetľujúcich premenných zahrnutých do modelu má štatisticky významný vplyv na *EDP*. Je zrejmé, že vplyv premenných nie je rovnaký, preto sa veľkosť vplyvu posudzuje na základe vypočítanej testovacej štatistiky. Keďže bolo použité testovanie prostredníctvom Fisherovej štatistiky, ktorá nadobúda iba kladné hodnoty, tak platí, že čím je hodnota tejto štatistiky vyššia, tým väčší vplyv má daná premenná na vysvetľovanú premennú.

Aby boli výsledky prehľadnejšie, tak sme vo výstupe v tab. 2 premenné postupne zoradili podľa veľkosti ich vplyvu. Najväčší vplyv na ekvivalentný disponibilný príjem má práve *EAS* (status ekonomickej aktivity), ďalej *HT* (typ domácnosti), *EDUCTATION* (vzdelanie) a najnižší vplyv má premenná *REGION* (kraj).

Tab. 3: Tabuľka p-hodnôt pre test štatistickej významnosti zhody marginálnych stredných hodnôt EDP domácností v závislosti od kraja

Least Squares Means for effect REGION Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: EDP								
i/j	**BB**	**KE**	**NR**	**PO**	**TN**	**TT**	**ZA**	**z_BA**
BB		0.2657	0.5349	0.2411	0.2477	0.0439	0.7799	<.0001
KE	0.2657		0.6308	0.0236	0.9569	0.3440	0.1716	<.0001
NR	0.5349	0.6308		0.0779	0.5961	0.1617	0.3765	<.0001
PO	0.2411	0.0236	0.0779		0.0216	0.0019	0.3857	<.0001
TN	0.2477	0.9569	0.5961	0.0216		0.3761	0.1584	<.0001
TT	0.0439	0.3440	0.1617	0.0019	0.3761		0.0249	<.0001
ZA	0.7799	0.1716	0.3765	0.3857	0.1584	0.0249		<.0001
z_BA	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	

Zdroj: EU-SILC 2018, vlastné spracovanie v SAS Enterprise Guide

Tab. 4: Bodové a intervalové odhady marginálnych stredných hodnôt EDP domácností v závislosti od kraja

REGION	**EDP LSMEAN**	**95% Confidence Limits**	
BB	6324.816325	6076.853897	6572.778754
KE	6467.246154	6220.781937	6713.710371
NR	6404.727311	6148.306264	6661.148358
PO	6176.021632	5925.696283	6426.346981
TN	6474.352901	6213.326930	6735.378872
TT	6595.401305	6324.171780	6866.630830
ZA	6288.670154	6032.242922	6545.097385
z_BA	7134.578127	6885.058976	7384.097279

Zdroj: EU-SILC 2018, vlastné spracovanie v SAS Enterprise Guide

Na základe vyššie uvedeného výstupu (tab. 3) môžeme vidieť štatisticky nevýznamný rozdiel v ekvivalentnom disponibilnom príjme medzi niektorými dvojicami krajov, konkrétne medzi tými, ktorých *p*-hodnota je vyššia ako zvolená hladina významnosti 0,05. V takomto prípade, kedy neuvažujeme nad vylúčením premennej z modelu, je vhodným riešením niektoré podobné kategórie zlúčiť a vytvoriť novú premennú, ktorá pôvodnú premennú v modeli nahradí. Prvú kategóriu novovytvorenej premennej budú tvoriť domácnosti žijúce v kraji s najvyšším rozdielom v *EDP* oproti Bratislavskému kraju (tab. 4). Predpokladáme, že by sme mohli zlúčiť tri kraje, konkrétne Banskobystrický (*BB*), Prešovský (*PO*) a Žilinský (*ZA*) kraj, medzi ktorými sú malé diferencie, ktoré sú pre jednotlivé dvojice štatisticky nevýznamné. Ďalej uvažujme o zlúčení Košického (*KE*), Nitrianskeho (*NR*), Trenčianskeho (*TN*) a Trnavského kraja (*TT*). Samostatnú a zároveň referenčnú kategóriu ponecháme nezmenenú a budú ju tvoriť domácnosti žijúce v Bratislavskom kraji (*BA*). Opodstatnenosť zlúčenia uvedených krajov overíme v procedúre *PROC GLM* s využitím príkazu *CONTRAST*.

4.1.1 Analýza stredných hodnôt EDP s využitím príkazu CONTRAST

Základom správneho použitia príkazu *CONTRAST* v procedúre *PROC GLM* je jeho korektné zostavenie. Jeho hlavnou súčasťou sú však koeficienty, ktorých hodnoty závisia od toho, ako je zadefinovaná nulová hypotéza. V prípade príkazu *CONTRAST* nulová hypotéza musí byť vyjadrená ako lineárna kombinácia parametrov, ktorá sa musí rovnať nule. Keďže neuvažujeme o interakcii [bližšie pozri (Littell a kol., 2010)], nulové hypotézy postačuje vyjadriť ako lineárne kombinácie testovaných stredných hodnôt, tak ako to uvádza tab. 5. Príkaz *CONTRAST* produkuje hodnotu testovacej *F*-štatistiky a *p*-hodnotu, na základe ktorých vieme rozhodnúť o výsledku testu. Výhodou je, že počet príkazov *CONTRAST* nie je v procedúre limitovaný a umožňuje súčasne testovať niekoľko hypotéz. Takéto simultánne testovanie bude nevyhnutné na získanie korektných záverov o zhode stredných hodnôt *EDP* domácností vo viacerých krajoch SR.

Tab. 5: Odhad parametrov regresného modelu EDP domácností

Skupiny krajov	Testované hypotézy	Formulácia príkazu CONTRAST
BB, ZA, PO	$H_0: \mu_{BB} - \mu_{ZA} = 0$	`contrast 'REGION BB=ZA=PO simult.' REGION 1 0 0 0 0 0 -1 0, REGION 0.5 0 0 -1 0 0 0.5 0;`
	$H_0: 0{,}5 * \mu_{BB} + 0{,}5 * \mu_{ZA} - \mu_{PO} = 0$	
KE, TN, NR, TT	$H_0: \mu_{KE} - \mu_{TN} = 0,$	`contrast 'REGION KE=TN=NR=TT simult.' REGION 0 1 0 0 -1 0 0 0, REGION 0 0.5 -1 0 0.5 0 0 0, REGION 0 1 1 0 1 -3 0 0;`
	$H_0: 0{,}5 * \mu_{KE} + 0{,}5 * \mu_{TN} - \mu_{NR} = 0,$	
	$H_0: \mu_{KE} + \mu_{TN} + \mu_{NR} - 3 * \mu_{TT} = 0$	

Zdroj: vlastné spracovanie

Zhodu stredných hodnôt trojice najpodobnejších krajov (*BB, ZA, PO*) overíme prostredníctvom simultánneho testovania nulových hypotéz, ktorých formuláciu upravíme tak, aby sme dostali tvar lineárnej kombinácie a tým pádom získali hodnoty koeficientov, ktoré sú potrebné pre zostavenie príkazu. Keďže príkaz *CONTRAST* v podstate testuje len zhodu dvoch stredných hodnôt, tak hypotézu

$$H_0: \mu_{BB} = \mu_{ZA} = \mu_{PO}$$

prepíšeme do dvoch nulových hypotéz napríklad takto

$$H_0: \mu_{BB} = \mu_{ZA} \qquad H_0: \mu(\mu_{BB;}\, \mu_{ZA}) = \mu_{PO}$$

a budeme ich testovať simultánne. Najskôr ich však prepíšeme do tvaru lineárnej kombinácie, tak ako to je uvedené v tab. 5. Prvá hypotéza obsahuje dva parametre a zároveň aj dve hodnoty koeficientov, ktoré zapíšeme do príkazu *CONTRAST*. Hodnotu 1 zapíšeme pre Banskobystrický kraj a hodnotu -1 pre Žilinský kraj. Druhá hypotéza pozostáva z lineárnej kombinácie troch parametrov, pričom hodnotu 0,5 majú parametre Banskobystrického a Žilinského kraja a hodnotu -1 má Prešovský kraj. Zvyšné koeficienty budú nulové. Správne naformulované hypotézy si vieme opäť overiť prostredníctvom súčtu všetkých koeficientov, ktorého hodnota musí byť nulová. Finálna syntax príkazu *CONTRAST* je uvedená v poslednom stĺpci v tab. 5.

Predpokladanú zhodu stredných hodnôt štvorice krajov (*KE, TN, NR, TT*) overíme obdobným spôsobom ako v predchádzajúcom prípade. Princíp zostavenia príkazu spočíva v naformulovaní troch hypotéz, následne upravených do tvaru lineárnej kombinácie. Z vyššie uvedených nulových hypotéz (tab. 5), ktoré testujeme simultánnym testom, prvé dve ponecháme nezmenené a tretiu môžeme upraviť aj na tvar:

$$H_0: \frac{\mu_{KE}+\mu_{TN}+\mu_{NR}}{3} = \mu_{TT},$$

$$H_0: 0{,}3333 * \mu_{KE} + 0{,}3333 * \mu_{TN} + 0{,}3333 * \mu_{NR} - \mu_{TT} = 0.$$

Pred samotnou formuláciou príkazu je potrebné upraviť koeficienty z hypotézy. V prípade, že by sme koeficienty ponechali nezmenené a v nasledovnom tvare ich vložili do príkazu *CONTRAST*, tento príkaz sa v procedúre *GLM* nezrealizuje. Problémom by bol súčet koeficientov, pretože by sa nerovnal nule. Práve preto je potrebná úprava koeficientov pre tretiu nulovú hypotézu na taký tvar, aby bol výsledný súčet nulový, a to napríklad:

$$H_0: 0{,}3333 * \mu_{KE} + 0{,}3333 * \mu_{TN} + 0{,}3334 * \mu_{NR} - \mu_{TT} = 0$$

Aj keď výsledky týchto testov by boli v obidvoch prípadoch totožné, uprednostňuje sa zostavenie hypotézy v takom tvare, aby sa koeficienty v príkaze nemuseli upravovať. V takomto prípade vynásobíme testovanú rovnosť $\frac{\mu_{KE}+\mu_{TN}+\mu_{NR}}{3} = \mu_{TT}$ hodnotou 3 a po jednoduchej úprave získame lineárnu kombináciu uvedenú v poslednom riadku tab. 5. Vďaka takejto úprave nebude potrebné v príkaze *CONTRAST* koeficienty zaokrúhľovať.

Tab. 6: Výstup príkazu CONTRAST

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
REGION BB=ZA=PO simult.	2	8524971.88	4262485.94	0.74	0.4760
REGION KE=TN=NR=TT simult.	3	11577887.00	3859295.67	0.67	0.5690

Zdroj: EU-SILC 2018, vlastné spracovanie v SAS Enterprise Guide

Formuláciou hypotéz a následným skonštruovaním príkazov pridaných do procedúry *GLM*, sme získali výstup, prostredníctvom ktorého verifikujeme stanovené predpoklady. Keďže podstatou nasledovných príkazov bolo simultánne testovanie dvoch a troch nulových hypotéz, počet stupňov voľnosti sa rovná 2 a 3 (tab. 6). Prijatie tvrdenia je opäť ovplyvnené výslednou hodnotou testovacej štatistiky, resp. jej *p*-hodnotou. V obidvoch prípadoch *p*-hodnoty prekračujú hladinu významnosti 0,05, a tak prijímame nulovú hypotézu o zhode stredných hodnôt ekvivalentných disponibilných príjmov domácností žijúcich v Košickom, Trenčianskom, Nitrianskom a Trnavskom kraji a zhodu stredných hodnôt *EDP* domácností v Banskobystrickom, Žilinskom a Prešovskom kraji. Keďže tab. 3 potvrdila štatisticky významne rôzny *EDP* v domácnostiach z Bratislavského kraja ako v ostatných krajoch, tak Bratislavský kraj ostal samostatný, a teda sme ho nezlučovali so žiadnym iným krajom.

4.2 Všeobecný lineárny model EDP domácností s modifikáciou premennej REGION

Vzhľadom na to, že sa nám pri predchádzajúcom modeli potvrdila zhoda stredných hodnôt pri niektorých kategóriách premennej *REGION*, rozhodli sme sa pôvodný model modifikovať. Pôvodnú premennú *REGION* sme nahradili novovytvorenou premennou, ktorá pozostáva zo zlúčených kategórií. Rovnako aj druhý model zahŕňa vysvetľujúce premenné, ktoré sú kategoriálne a obsahom každej z nich je rôzny počet obmien (kategórii). Z každej kategórie bola vytvorená nová premenná, tzv. umelá premenná, ktorá v modeli zastupuje túto kategóriu. Treba poznamenať, že interpretácia každej z umelých premenných spočíva v porovnávaní s referenčnou kategóriou za podmienok ceteris paribus.

Tab. 7: Odhad parametrov regresného modelu EDP domácností

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	9481.546204	B	146.6802819	64.64	<.0001
EAS Unemployed	-3681.753212	B	191.6736422	-19.21	<.0001
EAS Inactive_person	-1911.644476	B	155.8281512	-12.27	<.0001
EAS Retired	-1447.941499	B	94.0251839	-15.40	<.0001
EAS z_at_Work	0.000000	B	.	.	.
EDUCATION Less_than_Secondary	-2455.414775	B	134.0537186	-18.32	<.0001
EDUCATION Upper_Secondary	-1543.097927	B	91.3813086	-16.89	<.0001
EDUCATION Post_Secondary	-1093.980567	B	229.8267496	-4.76	<.0001
EDUCATION Tertiary_1	-640.361140	B	260.3659282	-2.46	0.0139
EDUCATION z_Tertiary_2_3	0.000000	B	.	.	.
HT 2A_at_least_3Ch	-1020.156597	B	240.7340041	-4.24	<.0001
HT 1A_at_least_1Ch	-487.521722	B	243.5134627	-2.00	0.0453
HT 1Adult	469.058001	B	159.7862152	2.94	0.0033
HT 2A_1Ch	845.775722	B	161.5971026	5.23	<.0001
HT Other_with_Ch	1571.330326	B	146.0674535	10.76	<.0001
HT 2A_1R	1965.930936	B	155.2844796	12.66	<.0001
HT 2Adult_0Ch	2649.847558	B	145.1925077	18.25	<.0001
HT Other_0Ch	3476.111937	B	140.9270957	24.67	<.0001
HT z_2A_2Ch	0.000000	B	.	.	.
MARITAL_STATUS Never_married	-287.260518	B	116.5742666	-2.46	0.0138
MARITAL_STATUS Divorced	-168.106876	B	114.5800398	-1.47	0.1424
MARITAL_STATUS Widowed	947.771465	B	112.7400873	8.41	<.0001
MARITAL_STATUS z_Married	0.000000	B	.	.	.
HEALTH Fair	-466.436729	B	81.2244701	-5.74	<.0001
HEALTH Bad	-449.808048	B	100.4184521	-4.48	<.0001
HEALTH z_Good	0.000000	B	.	.	.
URBANISATION Sparse	-588.971764	B	95.0419907	-6.20	<.0001
URBANISATION Intermediate	-438.169418	B	93.1042336	-4.71	<.0001
URBANISATION z_Dense	0.000000	B	.	.	.
REGION BB_PO_ZA	-856.870889	B	110.6978175	-7.74	<.0001
REGION KE_NR_TN_TT	-642.656998	B	105.8338041	-6.07	<.0001
REGION z_BA	0.000000	B	.	.	.

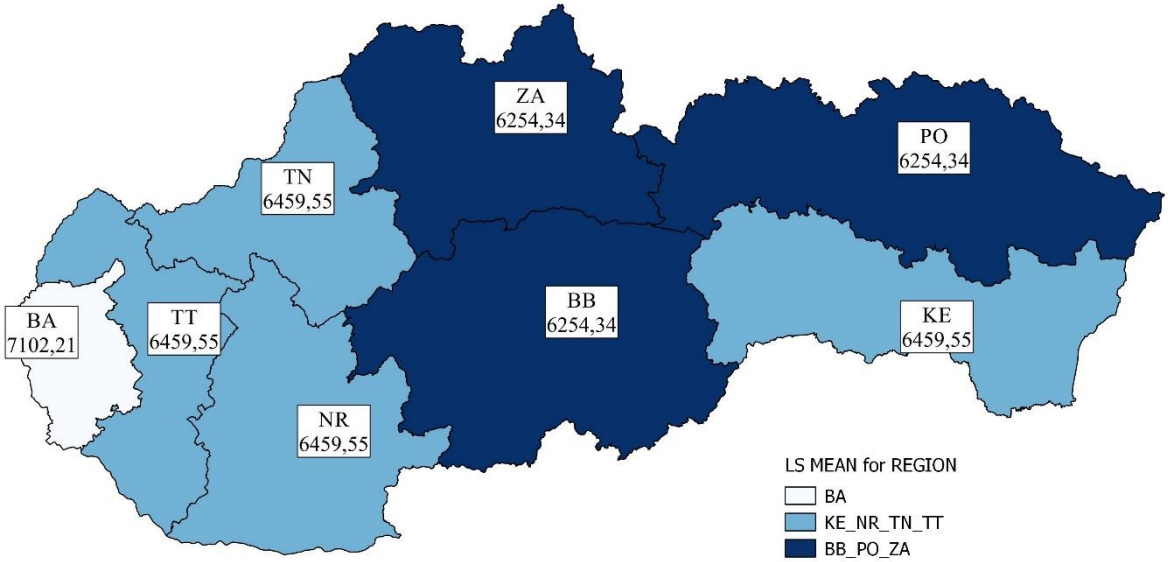
Zdroj: EU-SILC 2018, vlastné spracovanie v SAS Enterprise Guide

Novovytvorený model sa oproti modelu s pôvodnými premennými z hľadiska štatistickej významnosti regresných koeficientov veľmi nezmenil. Zmena modelu prispela k tomu, že ostal nevýznamný už len jeden regresný koeficient, ktorého *p*-hodnota prekračuje hladinu významnosti 0,05 ($p = 0{,}1424$). Znamená to, že pri fixovaní ostatných faktorov nie je na hladine významnosti 0,05 štatisticky významný rozdiel v priemernom *EDP* domácností, na ktorých čele stojí rozvedená osoba a v priemernom *EDP* domácností, na ktorých čele stojí osoba žijúca v manželskom zväzku.

V prípade ostatných kategórií (a neplatí to len pre rodinný stav – *MARITAL_STATUS*) sme odhalili štatisticky významný rozdiel priemerného *EDP* v týchto kategóriách od priemerného *EDP* v referenčnej kategórii príslušného faktora. Poznamenajme, že referenčné kategórie majú hodnotu regresného koeficienta na úrovni 0.

Premenná *REGION* spôsobovala v predchádzajúcom modeli problém, pretože sa niektoré úrovne navzájom neodlišovali. Problém sme vyriešili úpravou premennej, konkrétne zlúčením najpodobnejších kategórii. Na hladine významnosti 0,05 preto môžeme potvrdiť predpoklad, že výška disponibilného príjmu domácností jednotlivých kategórii je štatisticky významne rozdielna ako v prípade domácností patriacich do referenčnej kategórie. Ročný ekvivalentný disponibilný príjem domácností žijúcich v Banskobystrickom, Prešovskom a Žilinskom kraji je v priemere o 856,87 EUR (tab. 7) nižší ako *EDP* domácností pochádzajúcich z Bratislavského kraja. Domácnosti žijúce v Košickom, Nitrianskom, Trenčianskom a Trnavskom kraji mali v roku 2018 v priemere o 642,66 EUR (tab. 7) vyšší ročný *EDP* ako domácnosti v zhluku týchto krajov: Banskobystrický, Prešovský a Žilinský kraj. Tak ako sme už uvideli, tieto interpretácie predpokladajú, že ostatné faktory zahrnuté v modeli ostávajú na konštantnej úrovni.

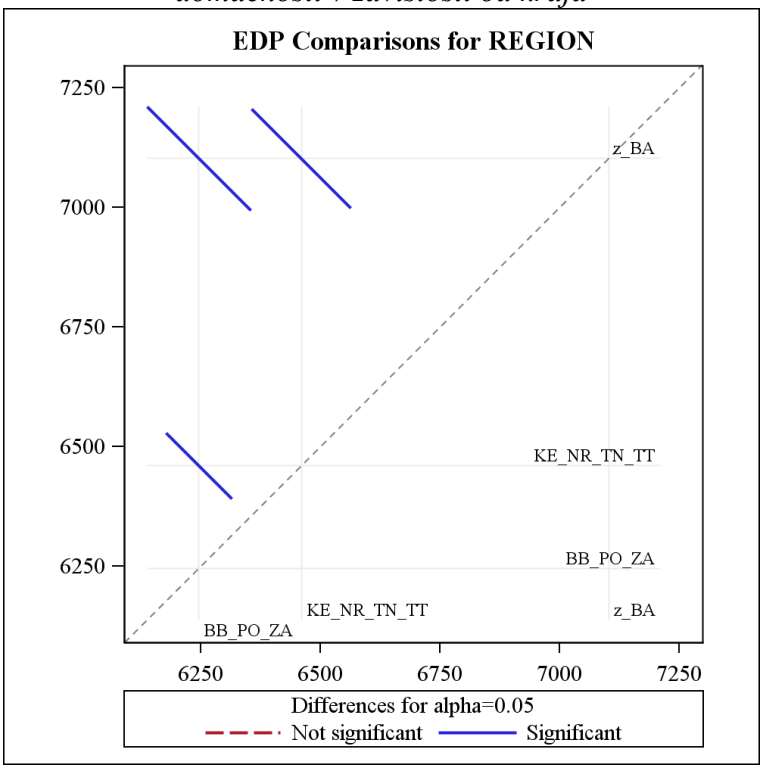
Obr. 1: Mapa regiónov SR - bodové odhady marginálnych stredných hodnôt EDP domácností v závislosti od kraja



Zdroj: EU-SILC 2018, vlastné spracovanie v QGIS 2.14.18

Pri premennej *REGION* v pôvodnom modeli bol medzi väčšinou dvojíc krajov štatisticky nevýznamný rozdiel v priemernom ekvivalentnom disponibilnom príjme domácností, čo dokumentujú *p*-hodnoty v matici v tab. 3. Opierajúc sa o výsledky simultánneho testovania realizovaného príkazom *CONTRAST* sme zlúčili najpodobnejšie kategórie z hľadiska priemerného *EDP* a vytvorili sme novú premennú s tromi kategóriami. Obr. 2 potvrdzuje, že medzi novovytvorenými kategóriami je štatisticky významný rozdiel v priemernom ekvivalentnom disponibilnom príjme.

Obr. 2: Intervalové odhady marginálnych stredných hodnôt EDP domácností v závislosti od kraja



Zdroj: EU-SILC 2018, vlastné spracovanie v SAS Enterprise Guide

5 Záver

Cieľom nášho príspevku bolo poukázať na možnosti využitia všeobecných lineárnych modelov na analýzu životnej úrovne domácností Slovenskej republiky. Prostredníctvom metód, ktoré boli pre analýzu všeobecných lineárnych modelov vyvinuté, sme analyzovali vplyv rôznych faktorov (status ekonomickej aktivity, typ domácnosti, vzdelanostná úroveň, zdravotný a rodinný stav, kraj a stupeň urbanizácie) na cieľovú premennú – ekvivalentný disponibilný príjem domácností.

Využitím príkazu *LSMEANS* sme zistili, že medzi niektorými kategóriami premennej *REGION* (kraj) nie je štatisticky významný rozdiel v priemernej výške ekvivalentného disponibilného príjmu, čo nás viedlo k predpokladu, že úroveň stredných hodnôt je pre viaceré kraje rovnaká. Na overenie predpokladu sme využili príkaz *CONTRAST*, ktorý bolo potrebné do procedúry *GLM* doplniť zásahom do programovacieho kódu. Podstatou využitia príkazu bolo porovnanie stredných hodnôt tých krajov, medzi ktorými bol najmenší a štatisticky nevýznamný rozdiel v priemernom ekvivalentnom disponibilnom príjme. Vzhľadom na to, že sa predpoklad o zhode stredných hodnôt potvrdil, rozhodli sme sa najpodobnejšie kraje zlúčiť do jednej kategórie a vytvoriť novú premennú, ktorá v ďalšej analýze nahradila pôvodnú premennú.

Do druhého modelu sme namiesto pôvodnej premennej *REGION* zaradili novovytvorenú premennú. Problém, ktorý v prvom modeli premenná *REGION* spôsobovala, konkrétne štatisticky nevýznamný rozdiel v prípade niektorých dvojíc kategórií, sa nám jej úpravou podarilo vyriešiť.

V článku je prostredníctvom odhadov regresných koeficientov všeobecného lineárneho modelu kvantifikovaný vplyv uvažovaných faktorov na *EDP* slovenských domácností. Ukázali sme ako možno analýzu marginálnych stredných hodnôt realizovanú prostredníctvom príkazov *LSMEANS* a *CONTRAST* využiť pri posúdení významnosti rozdielov v strednej hodnote cieľovej premennej na rôznych úrovniach toho-ktorého faktora. Kontrastnú analýzu sme

v našom modeli uskutočnili pre faktor *REGION* a zistili sme, že domácnosti z Banskobystrického, Prešovského a Žilinského kraja v roku 2018 nemali štatisticky významne odlišný priemerný *EDP* a ten bol pri fixovaní ostatných faktorov použitých v modeli *GLM* na úrovni 6 254 EUR/rok. V týchto troch krajoch evidujeme najnižšiu priemernú úroveň *EDP*. Nevýznamné rozdiely v priemernom *EDP* sme odhalili aj medzi Košickým, Nitrianskym, Trenčianskym a Trnavským krajom, kde sme odhadli priemerný *EDP* o 214,21 EUR/rok vyšší ako v prvom zhluku krajov. Podľa očakávania sme štatisticky významne vyšší priemerný *EDP* ako v ostatných krajoch identifikovali v Bratislavskom kraji, a to na úrovni približne 7 100 EUR/rok. Aj keď článok poskytuje ukážku redukcie úrovní len jedného faktora (*REGION*), vo všeobecnosti sa môže takáto redukcia týkať viacerých faktorov. Opodstatnená redukcia počtu úrovní viackategoriálnych faktorov založená na kontrastnej analýze má niekoľko výhod:

- výsledky analýz sa sprehľadňujú,
- eliminuje sa riziko vzniku prázdnych kategórií, ktoré vzniká pri uvažovaní interakcií medzi viacerými faktormi,
- zlúčené kategórie sú početnejšie ako pôvodné kategórie, čo väčšinou vedie k menšej štandardnej chybe odhadu marginálnej strednej hodnoty, čo má samozrejme dopad na induktívne úsudky o LS-means.

Príspevok bol spracovaný v rámci riešenia grantovej úlohy KEGA 007EU-4/2020 ***Interaktívna a interdisciplinárna výučba predmetov Služby a Inovácie v cestovnom ruchu s využitím informačných technológií.***

Literatúra

[1] Anyaegbu, G. (2010). Using the OECD equivalence scale in taxes and benefits analysis [online]. Dostupné na: https://link.springer.com/content/pdf/10.1057/_elmr.2010.9.pdf [cit. 2020-04-03].

[2] Atkinson, A. B., Rainwater, L. & Smeeding, T. M. (1995). *Income Distribution in OECD Countries* [online]. Paris, France, [cit. 2020-04-03].

[3] Bellu, L. G. (2005). *Equivalence Scales: General Aspects* [online]. Urbino, Italy, Dostupné na: http://www.fao.org/docs/up/easypol/325/equiv_scales_general_032en.pdf [cit. 2020-04-02].

[4] Bowley, S. R. (2013). *Constructing SAS Contrast/Estimate Statements* [online]. University of Guelph, Canada, Dostupné na: https://www.plant.uoguelph.ca/sites/plant.uoguelph.ca/files/forages/Constructing%20SAS%20EstimateContrast%20Statements-2013.pdf [cit. 2020-02-18].

[5] Howell, D. C. (2010). *Statistical Methods for Psychology* [online]. Belmont, CA, USA: Cengage Wadsworth, [cit. 2020-03-26].

[6] Chen, H. (2008). Using ESTIMATE and CONTRAST Statements for Customized Hypothesis Tests. *SAS Institute Inc. Paper SP09-2008*.

[7] Kováčová, Y. & Vlačuha, R. (2019). *Indikátory chudoby a sociálneho vylúčenia* [online]. Bratislava, [cit. 2020-02-12].

[8] Kováčová, Y. & Vlačuha, R. (2019). *Zisťovanie o príjmoch a životných podmienkach domácností v SR* [online]. Bratislava, [cit. 2020-02-12].

[9] Lenth, R., V. (2016). Least-squares means: the R package lsmeans. *Journal of Statistical Software*. *69*(1), 1-33.

[10] Littell, R. C., Stroup, W. W., & Freund, R. J. (2010). *SAS for Linear Models.* 4th ed. Cary, NC: SAS Institude Inc.

[11] Mangiafico, S. S. (2015). *An R Companion for the Handbook of Biological Statistics* [online]. New Brunswick, NJ, Dostupné na: https://rcompanion.org/documents/RCompanionBioStatistics.pdf [cit. 2020-04-02].

[12] OECD, (2008). *Income Distribution and Poverty in OECD Countries* [online]. Paris, France, [cit. 2020-04-05].
[13] Pituch, K. A., & Stevens, J. P. (2015). *Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS*. Routledge.
[14] Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: a GLM Approach*. Sage.
[15] Savarese, P. T. & Patetta. J. (2010). *An Overview of the CLASS, CONTRAST, and HAZARDRATIO Statements in the SAS® 9.2 PHREG Procedure* [online]. SAS Institute Inc., Cary, NC, Dostupné na: https://www.lexjansen.com/pharmasug/ 2010/SAS/SAS-SP-SAS01.pdf [cit. 2020-02-22].
[16] Šoltés, E., Hurbánková, Ľ., Kotlebová, E., Šoltésová, T. & Vojtková, M. (2018). *Chudoba a sociálne vylúčenie v EÚ a v SR: v kontexte stratégie Európa 2020*. Pardubice: Univerzita Pardubice Fakulta ekonomicko-správní.
[17] ŠÚ SR, (2019). Tlačová správa. Bratislava, Dostupné na: https://www7.statistics.sk/wps/wcm/connect/c3c666ca-41da-47669d284b97599ec1e9/TS_chudoba_EU_SILC_018.pdf [cit. 2020-09-01].
[18] UNECE, (2011). *Handbook on Household Income Statistics* [online]. UN, Dostupné na: https://www.unece.org/fileadmin/DAM/stats/publications/Canberra_Group_Handbook_2nd_edition.pdf [cit. 2020-04-05].