Využitie procedúr LOGISTIC a GENMOD v SAS-e pri analýze veľmi nízkej intenzity práce

Martina Košíková[1]

Abstrakt
Cieľom príspevku je analýza veľmi nízkej intenzity práce osôb slovenských domácností. Na údajoch zo štatistického zisťovania EU-SILC 2021 a s využitím procedúr LOGISTIC a GENMOD v rámci štatistického softvéru SAS Enterprise Guide aplikujeme metódy logistickej regresie a zovšeobecnených lineárnych modelov na kvantifikáciu vplyvu relevantných kategoriálnych faktorov na binárnu premennú veľmi nízka intenzita práce. Prostredníctvom analýzy marginálnych stredných hodnôt (príkaz LSMEANS) a kontrastnej analýzy (príkaz CONTRAST) identifikujeme skryté vzťahy medzi jednotlivými úrovňami faktora a príkazom ESTIMATE odhadneme pravdepodobnosť, že osoba bude čeliť riziku vylúčenia z trhu práce.

Kľúčové slová
veľmi nízka intenzita práce, marginálne stredné hodnoty, pomer šancí, logistická regresia

Abstract
The aim of the article is the analysis of the very low work intensity of persons in Slovak households. On the data from the statistical survey EU-SILC 2021 and using the LOGISTIC and GENMOD procedures within the statistical software SAS Enterprise Guide, we apply the methods of logistic regression and generalized linear models to quantify the effect of relevant categorical factors on the binary variable very low work intensity. Based on the analysis of least squares means (LSMEANS statement) and contrast analysis (CONTRAST statement), we identify hidden relationships between individual levels of the factor, and the ESTIMATE statement estimates the probability that a person will be at risk of being excluded from the labor market.

Key words
very low work intensity, least squares means, odds ratio, logistic regression

JEL classification
C12; C51; R29

1 Úvod

Veľmi nízka intenzita práce patrí k jednému z troch ukazovateľov monitorovania chudoby v rámci kontextu stratégie Európa 2030. Rôzne štúdie, napr. (Cantillon a Vandenbroucke, 2014), (Rastrigina a kol, 2015), (Johnston a McGauran, 2018), odhalili, že riziko chudoby závisí od rôznych faktorov, avšak v prípade, že by sa na tento problém pozeralo z hľadiska veľmi nízkej intenzity práce, tak ohrozenie domácnosti chudobou závisí predovšetkým od zloženia domácnosti, od počtu ekonomicky neaktívnych osôb, vzdelanostnej úrovne osôb a zároveň od osôb, ktoré sú spôsobilé vykonávať pracovnú činnosť, čo sa nazýva pracovná intenzita.

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

2 Logistická regresia

Logistická regresia je rozšírením klasickej lineárnej regresie, pretože kvantifikuje odhad asociácie jednej alebo viacerých nezávislých premenných s binárnou závislou premennou. Inými slovami, používa sa na odhad pravdepodobnosti sledovanej udalosti vzhľadom na hodnoty nezávislých premenných.

Na vyjadrenie tvaru modelu logistickej regresie je dôležité transformovať binárnu závislú premennú na spojitú premennú, t. j. vyjadríme logaritmus šancí:

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right)$$
$$odds = \frac{p_i}{1-p_i}$$

V takomto prípade, kde závislú premennú vyjadruje $logit(p_i)$, dostaneme lineárny vzťah medzi závislou premennou a vektorom nezávislých premenných. Výsledný tvar rovnice modelu logistickej regresie vyjadríme nasledovne:

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_k x_{ik}$$

kde β_j sú neznáme parametre modelu.

Významnosť modelu logistickej regresie sa overuje pomocou troch chí-kvadrát testov (test vierohodnostným pomerom, skóre test alebo Waldov test), pričom (Allison, 2012) uvádza, že v prípade dostatočne veľkých súborov nie je nutné uprednostňovať niektorý z uvedených testov. Na overenie signifikantnosti vplyvu nezávislých premenných na závislú premennú používame Waldov test:

$$Wald = \widehat{\boldsymbol{\beta}}^{\mathrm{T}}.\mathbf{S}_{\mathrm{b}}^{-1}.\widehat{\boldsymbol{\beta}}$$

kde $\widehat{\boldsymbol{\beta}}$ je vektor odhadov regresných koeficientov a $\mathbf{S}_{\mathrm{b}}^{-1}$ je variančno-kovariančná matica vektora $\widehat{\boldsymbol{\beta}}$.

Výhodou použitia logistickej regresie je to, že regresné koeficienty možno interpretovať ako pomer šancí. Pomer šancí vyjadruje, do akej miery sa mení pravdepodobnosť sledovanej udalosti v prípade jednotkového nárastu nezávislej premennej (kvantitatívne premenné) alebo oproti referenčnej kategórii (kvalitatívne premenné) pri zachovaní podmienky ceteris paribus.

$$OR = \frac{odds_1}{odds_2} = e^{\beta_j}$$

Ako uviedli autori (Schober a Vetter, 2021) podmienkou použitia logistickej regresie je splnenie nasledujúcich predpokladov:

- pokiaľ je nezávislá premenná kvantitatívna, musí mať lineárny vzťah s logitom (prirodzený logaritmus šancí),
- jednotlivé pozorovania musia byť nezávislé,
- model musí byť správne špecifikovaný – Hosmer-Lemeshow test dobrej zhody.

3 Analýza veľmi nízkej intenzity práce osôb slovenských domácností

Prvotným a dôležitým krokom k získaniu spoľahlivých výsledkov jednotlivých analýz je relevantná údajová základňa. Keďže cieľom príspevku je analyzovať veľmi nízku intenzitu práce osôb žijúcich na Slovensku, tak vstupnú databázu tvoria údaje získané zo štatistického zisťovania EU-SILC 2021. Na základe skúseností a zároveň výsledkov z rôznych vedeckých prác (Šoltés a kol, 2018), (Glaser-Opitzová a Vojtková, 2020), (Ionescu, 2014), (Ward a Ozdemir, 2013) a pod. predpokladáme, že veľmi nízka intenzita práce môže byť ovplyvnená faktormi ako sú napríklad vzdelanie, ekonomická aktivita, rodinný alebo zdravotný stav, typ domácnosti, urbanizácia alebo kraj, v ktorom osoba žije. V rámci nasledujúcich analýz budeme pracovať so závislou premennou veľmi nízka intenzita práce (***VLWI*** – *very low work intensity*), ktorá je binárna s obmenami „***yes***“ (osoba ohrozená veľmi nízkou intenzitou práce) a „***no***“ (osoba, ktorá nie je ohrozená veľmi nízkou intenzitou práce). Nezávislé premenné vstupujúce do analýzy sú kategoriálne s niekoľkými obmenami:

EDUCATION (vzdelanie)
- ***Less_than_Secondary*** (nižšie ako sekundárne vzdelanie)
- ***Post_Secondary*** (postsekundárne vzdelanie)
- ***Tertiary_1*** (vysokoškolské vzdelanie 1. stupňa)
- ***Upper_Secondary*** (vyššie sekundárne vzdelanie)
- ***z_Tertiary_2_3*** (vysokoškolské vzdelanie 2. a 3. stupňa - **referenčná kategória)**

EA (ekonomická aktivita)
- ***Disabled_person*** (invalidná osoba)
- ***Inactive_person*** (iná neaktívna osoba)
- ***Person_in_household*** (osoba v domácnosti)
- ***Student*** (študent)
- ***Unemploy***ed (nezamestnaná osoba)
- ***z_at_Work*** (zamestnaná osoba – **referenčná kategória**)

HT (typ domácnosti)
- ***1A_at_least_1Ch*** (domácnosť 1 dospelej osoby aspoň s jedným závislým dieťaťom)
- ***1Adult*** (1 dospelá osoba)
- ***2A_1Ch*** (domácnosť 2 dospelých osôb s jedným závislým dieťaťom)
- ***2A_1R*** (domácnosť 2 dospelých osôb pričom aspoň jedna z nich je vo veku 65+)
- ***2A_at_least_3Ch*** (domácnosť 2 dospelých osôb aspoň s 3 závislými deťmi)
- ***2Adult*** (domácnosť 2 dospelých osôb)
- ***Other_0Ch*** (iná domácnosť bez závislých detí)
- ***Other_with_Ch*** (iná domácnosť so závislými deťmi)
- ***z_2A_2Ch*** (domácnosť 2 dospelých s 2 závislými deťmi – **referenčná kategória**)

MARITAL_STATUS (rodinný stav)
- ***Divorced*** (rozvedená osoba)
- ***Never_married*** (slobodná osoba)
- ***Widowed*** (ovdovelá osoba)
- ***z_Married*** (osoba v manželskom zväzku – **referenčná kategória**)

HEALTH (všeobecné zdravie)
- ***Bad*** (zlý zdravotný stav)
- ***Fair*** (priemerný zdravotný stav)
- ***z_Good*** (dobrý zdravotný stav – **referenčná kategória**)

URBANISATION (urbanizácia)

- ***Intermediate*** (územie s priemerne hustým osídlením)
- ***Sparse*** (územie s riedkym osídlením)
- ***z_Dense*** (územie s hustým osídlením – **referenčná kategória**)

REGION (kraj)

- ***BB*** (Banskobystrický kraj)
- ***KE*** (Košický kraj)
- ***NR*** (Nitriansky kraj)
- ***PO*** (Prešovský kraj)
- ***TN*** (Trenčiansky kraj)
- ***TT*** (Trnavský kraj)
- ***ZA*** (Žilinský kraj)
- ***z_BA*** (Bratislavský kraj – **referenčná kategória**)

3.1 Model veľmi nízkej intenzity práce osôb slovenských domácností

Prostredníctvom procedúry LOGISTIC v štatistickom programe SAS Enterprise Guide skonštruujeme model logistickej regresie. Na overenie vplyvu zaradených faktorov do modelu aplikujeme *Waldov test*, ktorým testujeme nulovú hypotézu, že nezávislá premenná nemá signifikantný vplyv na závislú premennú.

Tab. 1: Test štatistickej významnosti vplyvu faktorov na VLWI

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
EA	5	464.8397	<.0001
Marital_status	3	9.7072	0.0212
Education	4	92.7375	<.0001
HT	8	137.3158	<.0001
Health	2	8.3396	0.0155
Urbanisation	2	29.1665	<.0001
Region	7	44.9132	<.0001

Zdroj: vlastné spracovanie v SAS Enterprise Guide

Významnosť vplyvu sa potvrdila v prípade všetkých uvažovaných faktorov (tab. 1), pretože výsledné *p* - hodnoty boli dostatočne malé na to, aby nepresiahli bežne používanú hladinu významnosti 0,05. Veľkosť vplyvu môžeme posúdiť na základe výsledných hodnôt testovacej štatistiky, pričom najväčší vplyv má premenná EA, HT, Education alebo Region.

Tab. 2: Test štatistickej významnosti modelu logistickej regresie

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	1902.6024	31	<.0001
Score	2578.5303	31	<.0001
Wald	763.3943	31	<.0001

Zdroj: vlastné spracovanie v SAS Enterprise Guide

Významnosť modelu s týmito faktormi overíme prostredníctvom troch testov (tab. 2) – *Likelihood Ratio* (test vierohodnostným pomerom), *Score* (skóre test), *Wald* (Waldov test). Vzhľadom na dostatočne veľkú vzorku nie je dôvod ani jeden z uvedených testov uprednostňovať. Spomínanými testami overujeme platnosť nulovej hypotézy, že regresné koeficienty sú nulové, resp. model nie je štatisticky významný. Výsledné p - hodnoty uvedených testov nás jednoznačne informujú o tom, že nulovú hypotézu zamietame, pretože model je signifikantný na akejkoľvek používanej hladine významnosti.

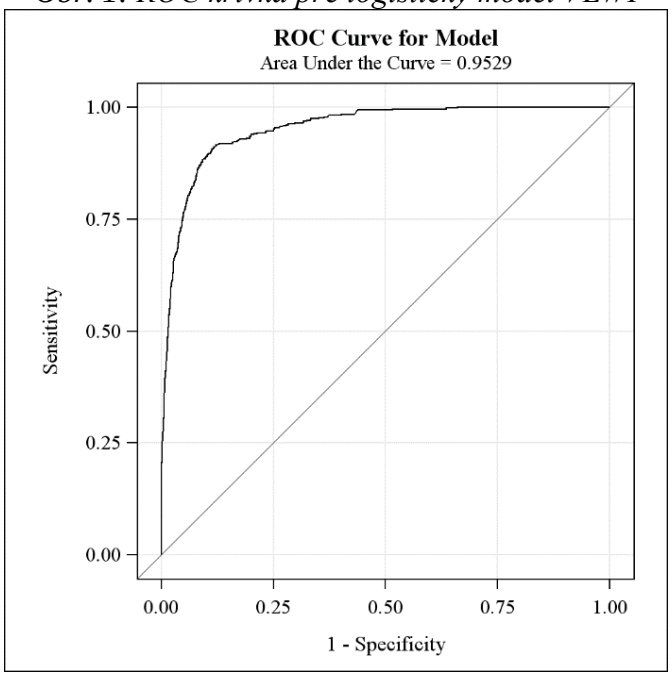
Adekvátnosť a úspešnosť modelu pri predikcii potvrdzujú miery asociácie z tab. 3 – Sommerovo D, Goodmanova-Kruskalova gamma a štatistika c, ktorých hodnoty sú dostatočne vysoké. Porovnanie konkordantných a diskordantných párov tiež výrazne svedčí o dostatočnej kvalite modelu. Štatistiku c môžeme aj graficky znázorniť pomocou ROC krivky (obr. 1), pričom jej hodnota predstavuje obsah plochy pod ROC krivkou (čím je väčšia plocha medzi krivkou a uhlopriečkou, tým je model kvalitnejší).

Tab. 3: Asociácia medzi odhadnutými pravdepodobnosťami získaných z logistického modelu VLWI a pozorovanými pravdepodobnosťami

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	95.3	**Somers' D**	0.906
Percent Discordant	4.7	**Gamma**	0.906
Percent Tied	0.0	**Tau-a**	0.116
Pairs	3453408	**c**	0.953

Zdroj: vlastné spracovanie v SAS Enterprise Guide

Obr. 1: ROC krivka pre logistický model VLWI



Zdroj: vlastné spracovanie v SAS Enterprise Guide

Vplyv nezávislej premennej na závislú premennú kvantifikujeme pomocou pomeru šancí, ktorý vyjadruje ako sa zmení šanca, že osoba bude čeliť veľmi nízkej intenzite práce oproti šanci, že osoba nebude ohrozená veľmi nízkou intenzitou práce (za podmienky ceteris paribus, t. j. ostatné faktory ostávajú fixované na referenčnej úrovni). Vzhľadom na rozsiahlosť výstupu uvedieme len niektoré výsledné hodnoty.

Veľmi nízka intenzita práce bola najviac determinovaná ekonomickou aktivitou (tab. 1). Najrizikovejšou kategóriou tejto premennej je Disabled_person pretože šanca, že takáto osoba bude ohrozená veľmi nízkou intenzitou práce je až 115,1 násobne vyššia oproti šanci, že ohrozená veľmi nízkou intenzitou práce bude osoba, ktorá je zamestnaná.

V prípade faktora HT je najkritickejšou kategóriou 2A_1R. Šanca, že osoba žijúca v domácnosti dvoch dospelých osôb (pričom aspoň jedna z nich je vo veku 65 rokov a viac) bude ohrozená veľmi nízkou intenzitou práce je 36,1 násobne vyššia, ako v prípade osoby žijúcej v domácnosti dvoch dospelých osôb s dvomi závislými deťmi.

Ďalším z najvplyvnejších faktorov bolo vzdelanie, pri ktorom na základe výsledných hodnôt pomerov šancí vyšla ako najkritickejšia kategória Less_than_Secondary. Pomer šancí ohrozenia veľmi nízkou intenzitou práce je v prípade osoby s nižším ako sekundárnym vzdelaním 8,6 násobne vyšší ako v referenčnej kategórii Tertiary_2_3.

V rámci premennej Region sú pomery šancí jednotlivých kategórií veľmi podobné, avšak najkritickejším krajom je na základe výsledkov Košický kraj (šanca 3,9 násobne vyššia ako v referenčnej kategórii).

Pomery šancí sa v prípade niektorých regresných koeficientov nedajú považovať za rozdielne, práve z toho dôvodu, že *p* - hodnoty testu o signifikantnosti jednotlivých kategórií presahujú úroveň hladiny významnosti. V takomto prípade je dôležité uvažovať, či nie je vplyv niektorých faktorov podmienený týmto problémom. Vzhľadom na tento fakt, budeme prostredníctvom ďalších analýz kvantifikovať hlbšie vzťahy medzi kategóriami faktora.

3.2 Analýza zhody marginálnych stredných hodnôt jednotlivých kategórií faktorov

Výsledné hodnoty predchádzajúcich analýz, konkrétne pri posudzovaní signifikantnosti jednotlivých kategórií faktorov, odhalili štatistickú nevýznamnosť niektorých regresných koeficientov. Využitím príkazu LSMEANS v procedúre GENMOD identifikujeme, či existuje zhoda marginálnych stredných hodnôt medzi niektorými dvojicami kategórií štyroch najvplyvnejších faktorov modelu.

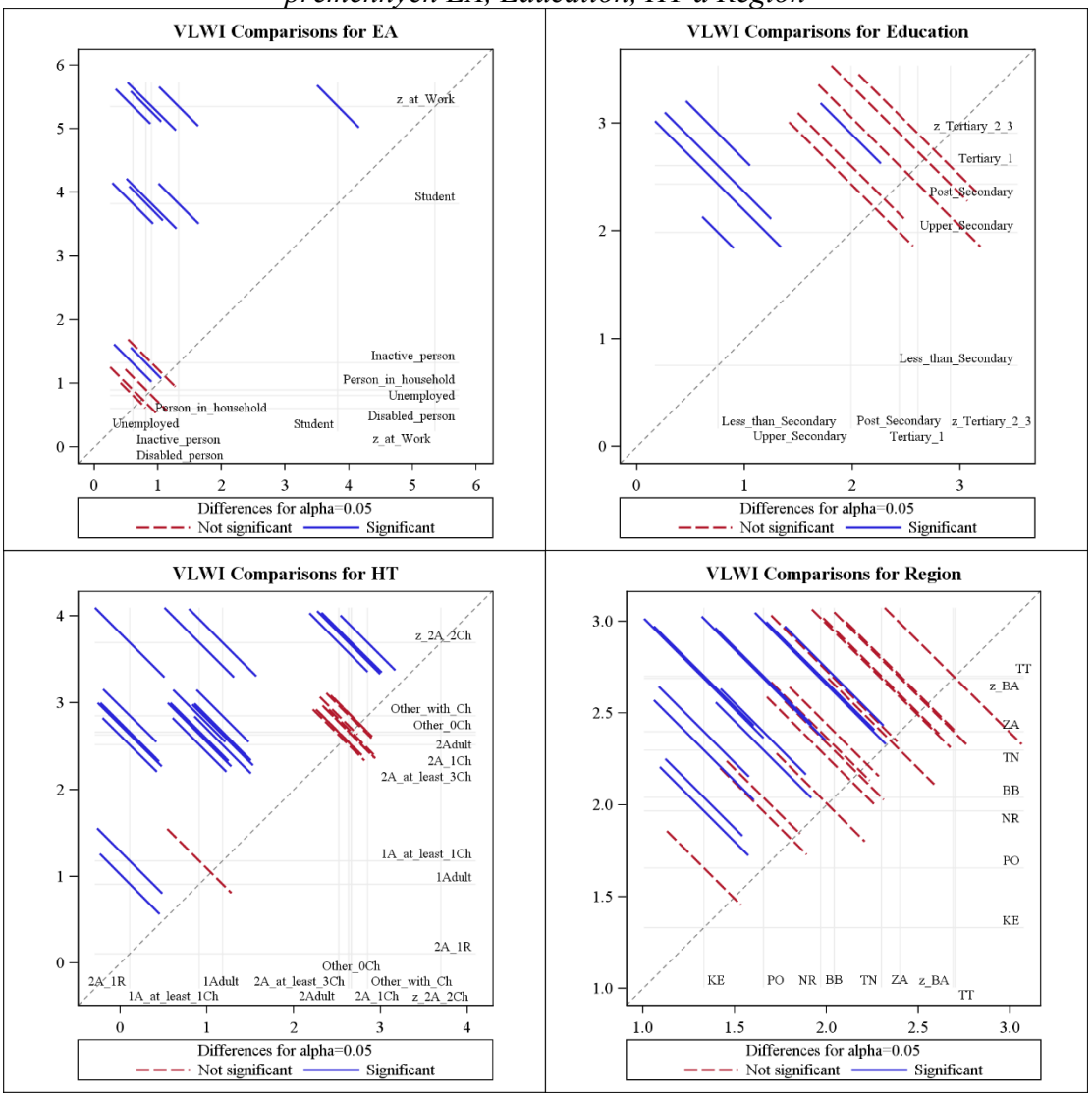
Štatisticky nevýznamný rozdiel stredných hodnôt medzi dvojicami kategórií faktora znázorňuje červená úsečka (obr. 2), presahujúca diagonálu. V prípade premennej EA môžeme vidieť štatisticky nevýznamný rozdiel medzi dvojicami kategórií Disabled_person a Person_in_household (*p = 0,4210*), Disabled_person a Unemployed (*p = 0,3134*) alebo medzi Person_in_household a Unemployed (*p = 0,7889*).

Signifikantnosť rozdielu stredných hodnôt sa nepotvrdila v prípade faktora Education pri kategóriách Post_Secondaty a Tertiary_1 (*p = 0,8234*), Post_Secondary a z_Tertiary_2_3 (*p = 0,4614*) a taktiež pri dvojici Tertiary_1 a z_Tertiary_2_3 (*p = 0,5895*).

Pri faktore HT je najväčšia podobnosť medzi dvojicou 2A_1Ch a Other_0Ch, kde *p* - hodnota je až 0,9945. Ďalej môžeme vidieť nesignifikantné rozdiely stredných hodnôt medzi 2Adult a 2A_1Ch (*p = 0,5790*), 2A_1Ch a 2A_at_least_3Ch (*p = 0,9187*), 2A_1Ch a Other_with_Ch (*p = 0,4392*), 2A_at_least_3Ch a 2Adult (*p = 0,6960*), 2A_at_least_3Ch a Other_0Ch (*p = 0,9194*), 2A_at_least_3Ch a Other_with_Ch (*p = 0,4050*), 2Adult a Other_0Ch (*p = 0,5061*), 2Adult a Other_with_Ch (*p = 0,1250*), Other_0Ch a Other_with_Ch (*p = 0,3544*) a medzi dvojicou 1A_at_least_1Ch a 1Adult (*p = 0,4758*).

Niekoľko štatisticky nevýznamných rozdielov vidíme na obr. 2 aj pri faktore Region, napríklad medzi dvojicou BB a NR (*p = 0,7576*), KE a PO (*p = 0,1154*), ďalej medzi TT a z_BA (*p = 0,9732*), TT a ZA (*p = 0,3131*), ZA a z_BA (*p = 0,4298*), TN a ZA (*p = 0,7281*), TN a TT (*p = 0,2076*) a taktiež aj medzi kategóriami TN a z_BA (*p = 0,3079*).

Obr. 2: Intervalové odhady marginálnych stredných hodnôt logitu šance VLWI v závislosti od premenných EA, Education, HT a Region



Zdroj: vlastné spracovanie v SAS Enterprise Guide

Štatistická nevýznamnosť stredných hodnôt logitu šance medzi niektorými kategóriami nás doviedla k predpokladu o ich zhode. Tento predpoklad je však dôležité overiť, a to využitím príkazu CONTRAST v rámci procedúry LOGISTIC. Dôležitým krokom je však zadefinovať nulové hypotézy, pretože koeficienty, ktoré ich úpravou získame, sú potrebné pre skonštruovanie samotného príkazu.

Vzhľadom na to, že sme pri premennej EA nepotvrdili signifikantnosť rozdielu stredných hodnôt medzi kategóriami Disabled_person (1 kategória faktora EA), Person_in_household (3 kategória faktora EA) a Unemployed (5 kategória), zadefinujeme si dve nulové hypotézy:

$$H_0\ (1): \mu_1 - \mu_3 = 0$$
$$H_0\ (2): 0{,}5 * \mu_1 + 0{,}5 * \mu_3 - \mu_5 = 0$$

a získané koeficienty budú vstupom do nasledujúceho príkazu:

```
CONTRAST 'D-PH-UN' EA 1 0 -1, EA 0.5 0 0.5 0 -1/ESTIMATE=ALL
ALPHA=0.05;
```

Tab. 4: Výstup príkazu CONTRAST pre kategórie faktora EA

Contrast Test Results			
Contrast	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
D-PH-UN	2	1.1811	0.5540

Zdroj: vlastné spracovanie v SAS Enterprise Guide

Keďže cieľom zadefinovania vyššie uvedených nulových hypotéz a zároveň zostavením príkazu CONTRAST bolo overiť zhodu stredných hodnôt logitu šance medzi tromi kategóriami faktora EA, môžeme na základe tab. 4 skonštatovať, že *p* - hodnota prekročila bežne používanú hladinu významnosti, z čoho vyplýva, že osoby, ktoré sú invalidné, nezamestnané alebo sú v domácnosti nemajú signifikantne odlišnú šancu, že budú čeliť riziku veľmi nízkej intenzity práce.

Rovnakým spôsobom si skonštruujeme príkazy aj pre kategórie ostatných troch faktorov. Pri faktore Education uvažujeme o zhode stredných hodnôt logitu šance medzi kategóriami Post_Secondary (2 kategória), Tertiary_1 (3 kategória) a z_Tertiary_2_3 (5 kategória):

```
CONTRAST 'PS-T' Education 0 1 -1, Education 0 0.5 0.5 0 -
1/ESTIMATE=ALL ALPHA=0.05;
```

Tab. 5: Výstup príkazu CONTRAST pre kategórie faktora Education

Contrast Test Results			
Contrast	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
PS-T	2	0.7086	0.7017

Zdroj: vlastné spracovanie v SAS Enterprise Guide

Výsledná *p* - hodnota v tab. 5 potvrdila predpokladané tvrdenie, t. j. osoby s post-sekundárnym vzdelaním, s vysokoškolským vzdelaním 1., 2. alebo 3. stupňa nemajú signifikantne odlišnú šancu, že budú čeliť riziku veľmi nízkej intenzity práce.

Príkazom LSMEANS sme odhalili niekoľko štatistických rozdielov stredných hodnôt logitu šance aj pri premennej HT. Zhodu medzi 1A_1Ch a 1Adult nemusíme príkazom CONTRAST overovať, pretože *p* - hodnota pri príkaze LSMEANS ich zhodu jednoznačne potvrdila. Avšak zhodu medzi kategóriami 2A_1Ch (3 kategória), 2A_at_least_3Ch (5 kategória), 2Adult (6 kategória), Other_0Ch (7 kategória) a Other_with_Ch (8 kategória) overíme opäť zadefinovaním príkazu:

```
CONTRAST 'HT5' HT 0 0 1 0 -1, HT 0 0 0.5 0 0.5 -1, HT 0 0 0.3333
0 0.3333 0.3333 -1 HT 0 0 0.25 0 0.25 0.25 0.25 -1/ESTIMATE=ALL
ALPHA=0.05;
```

Tab. 6: Výstup príkazu CONTRAST pre kategórie faktora HT

Contrast Test Results			
Contrast	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
HT5	3	2.5724	0.4623

Zdroj: vlastné spracovanie v SAS Enterprise Guide

Výsledná *p* - hodnota (tab. 6) nám jednoznačne potvrdila, že medzi strednými hodnotami logitu šance uvedených piatich kategórií nie je signifikantný rozdiel.

Posledným z najvplyvnejších faktorov je Region, pri ktorom sme prostredníctvom príkazu LSMEANS odhalili štatisticky nevýznamný rozdiel medzi dvojicou BB a NR a medzi KE a PO. Príkazom CONTRAST overíme, či medzi kategóriami TN, TT, ZA a z_BA existuje štatisticky nevýznamný rozdiel stredných hodnôt logitu šance:

```
CONTRAST 'TN_TT_ZA_BA' Region 0 0 0 0 1 -1, Region 0 0 0 0 0.5
0.5 -1, Region 0 0 0 0 0.3333 0.3333 0.3333 -1/ESTIMATE=ALL
ALPHA=0.05;
```

Tab. 7: Výstup príkazu CONTRAST pre kategórie faktora Region

Contrast Test Results			
Contrast	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
TN_TT_ZA_BA	3	2.1593	0.5400

Zdroj: vlastné spracovanie v SAS Enterprise Guide

P - hodnota v tab. 7 prevyšuje akúkoľvek bežne používanú hladinu významnosti, z čoho vyplýva, že osoby žijúce v Trenčianskom, Trnavskom, Žilinskom alebo v Bratislavskom kraji nemajú štatisticky významne odlišnú šancu, že budú čeliť riziku veľmi nízkej intenzity práce.

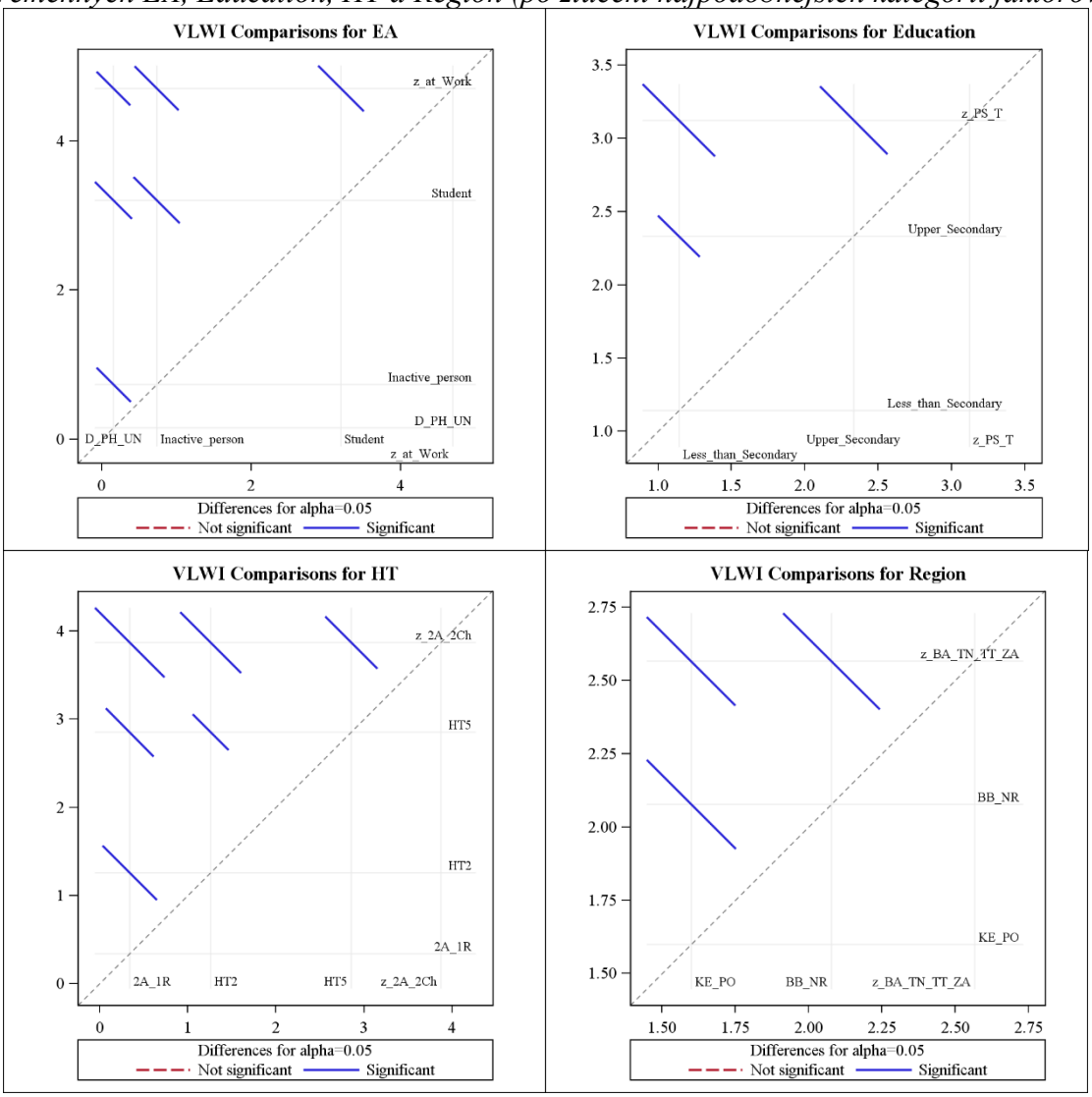
Na základe získaných výsledkov analýz sme usúdili, že kategórie, medzi ktorými sa nepotvrdil štatisticky významný rozdiel stredných hodnôt logitu šance, môžeme zlúčiť a vytvoriť tak novú kategóriu premennej.

Odhad modelu po modifikácií premenných potvrdil opodstatnenosť zlúčenia jednotlivých kategórií. Napriek tomu, že sa významnosť modelu ani vplyvu faktorov na závisle premennú nezmenila, zmenili sa rozdiely medzi marginálnymi strednými hodnotami, pretože vďaka zlúčeniu kategórií do jednej novej kategórie sú už vidieť iba štatisticky významné rozdiely (modré úsečky) medzi kategóriami štyroch najvplyvnejších faktorov (obr. 3).

Pomery šancí odhalili, že najrizikovejšie sú nasledujúce kategórie:

- pri premennej EA je to kategória D_PH_UN – šanca veľmi nízkej intenzity práce v prípade osoby, ktorá je invalidná/v domácnosti/nezamestnaná je 93,5-násobne vyššia ako u zamestnanej osoby,
- pri premennej Marital_status kategória Never_married – šanca veľmi nízkej intenzity práce slobodnej osoby je 1,6-násobne vyššia ako u osoby v manželskom zväzku,
- pri premennej Education kategória Less_than_Secondary – šanca veľmi nízkej intenzity práce u osoby s nižším ako sekundárnym vzdelaním je 7,1-násobne vyššia ako u osoby s post-sekundárnym/vysokoškolským vzdelaním 1., 2. alebo 3. stupňa,
- pri premennej HT kategória 2A_1R - šanca veľmi nízkej intenzity práce v prípade osoby žijúcej v domácnosti 2A_ 1R je 34,2-násobne vyššia ako v prípade osoby žijúcej v domácnosti dvoch dospelých osôb s dvomi závislými deťmi,
- pri premennej Health kategória Bad – šanca veľmi nízkej intenzity práce u osoby so zlým zdravotným stavom je 1,9-násobne vyššia ako u osoby s dobrým zdravotným stavom,
- pri premennej Urbanisation kategória Intermediate – šanca veľmi nízkej intenzity práce u osoby žijúcej na území so stredne hustým osídlením je 1,5-násobne vyššia ako u osoby žijúcej na území s hustým osídlením,
- a pri premennej Region kategória KE_PO – šanca veľmi nízkej intenzity práce u osoby žijúcej v KE alebo PO je 2,6-násobne vyššia ako u osoby žijúcej v TN, TT, ZA alebo v BA.

Obr. 3: Intervalové odhady marginálnych stredných hodnôt logitu šance VLWI v závislosti od premenných EA, Education, HT a Region (po zlúčení najpodobnejších kategórií faktorov)



Zdroj: vlastné spracovanie v SAS Enterprise Guide

Pomocou príkazu ESTIMATE v procedúre GENMOD odhadneme v nasledujúcej časti pravdepodobnosť, že osoba bude čeliť veľmi nízkej intenzite práce v závislosti od kraja a ekonomickej aktivity, pričom ostatné faktory ostatnú fixované na referenčnej úrovni. Pôvodné referenčné úrovne jednotlivých faktorov sme nahradili novými kategóriami, konkrétne tými, ktoré nám v rámci predchádzajúcich analýz vyšli ako najkritickejšie.

Zadefinovaním prvého príkazu odhadneme pravdepodobnosť, že invalidná/v domácnosti/ nezamestnaná osoba (D_PH_UN - μ_1) žijúca v Košickom/Prešovskom kraji (KE_PO -μ_2) bude čeliť veľmi nízkej intenzite práce:

```
estimate 'KE_PO and D_PH_UN' intercept 1 EA 1 Region 0 1 / cl
alpha=0.05 exp;
```

Tab. 8: Výstup príkazu ESTIMATE pre kategórie D_PH_UN a KE_PO

Estimate							
Label	**Estimate**	**St. Error**	**z**	**Pr > |z|**	**Lower**	**Upper**	**Exp**
KE_PO and D_PH_UN	4.0571	0.3383	11.99	<.0001	3.3940	4.7201	57.8041

Zdroj: vlastné spracovanie v SAS Enterprise Guide

Výsledkom príkazu ESTIMATE je tab. 8, ktorá poskytuje bodový odhad logaritmu šance (Estimate), bodový odhad šance (Exp), avšak neposkytuje priamo bodový odhad pravdepodobnosti, preto je nutné pomocou nasledujúcich vzťahov odhad pravdepodobnosti, že invalidná/v domácnosti/ nezamestnaná osoba žijúca v KE alebo v PO kraji, kvantifikovať:

$$\hat{\pi}_i = \frac{Exponentiated}{1 + Exponentiated}$$

$$\hat{\pi}_i = \frac{57{,}8041}{1 + 57{,}8041} = 0{,}9830$$

Bodový odhad poukazuje na veľmi vysokú pravdepodobnosť rizika ohrozenia veľmi nízkou intenzitou práce. V prípade osoby, ktorá je invalidná/v domácnosti/nezamestnaná žijúca v Košickom alebo Prešovskom kraji je 98,3 % pravdepodobnosť, že bude čeliť veľmi nízkej intenzite práce, a to za podmienky, že ostatné faktory sú fixované na referenčnej (najkritickejšej) úrovni. Výsledok potvrdzuje aj obrázok 4.

Prostredníctvom druhého príkazu odhadneme pravdepodobnosť, že pracujúca osoba (z_at_Work - μ_4) žijúca v BA/TN/TT/ZA kraji (z_BA_TN_TT_ZA - μ_3) bude čeliť veľmi nízkej intenzite práce:

```
estimate 'BA_TN_TT_ZA and at_Work' intercept 1 EA 0 0 0 1 Region
0 0 1 / cl alpha=0.05 exp;
```

Tab. 9: Výstup príkazu ESTIMATE pre kategórie at_Work a BA_TN_TT_ZA

Estimate							
Label	**Estimate**	**St. Error**	**z**	**Pr > |z|**	**Lower**	**Upper**	**Exp**
BA_TN_TT_ZA and at_Work	-1.4484	0.3623	-4.00	<.0001	-2.1585	-0.7383	0.2349

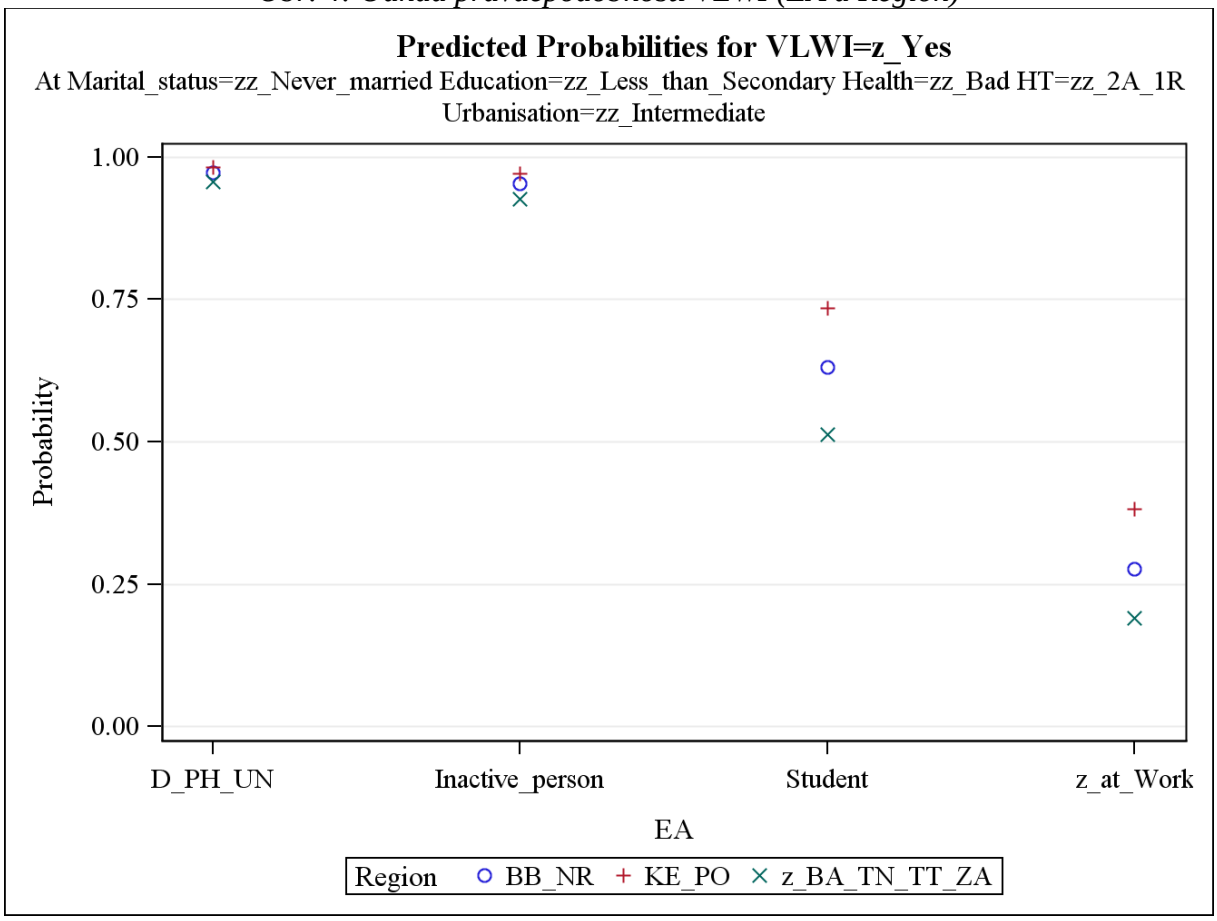
Zdroj: vlastné spracovanie v SAS Enterprise Guide

Bodový odhad pravdepodobnosti, že pracujúca osoba žijúca v Bratislavskom, Trenčianskom, Trnavskom alebo v Žilinskom kraji bude ohrozená veľmi nízkou intenzitou práce, kvantifikujeme nasledovne:

$$\hat{\pi}_i = \frac{0{,}2349}{1 + 0{,}2349} = 0{,}1902$$

Zamestnaná osoba žijúca v BA, TN, TT alebo ZA kraji má 19,02 % pravdepodobnosť, že bude čeliť veľmi nízkej intenzite práce, a to za podmienky, že ostatné faktory sú fixované na referenčnej (najkritickejšej) úrovni. Výsledok opäť potvrdzuje aj obrázok 4.

Obr. 4: Odhad pravdepodobnosti VLWI (EA a Region)



Zdroj: vlastné spracovanie v SAS Enterprise Guide

4 Záver

Cieľom príspevku bolo poukázať na možnosti využitia procedúr LOGISTIC a GENMOD pri kvantifikácii vplyvu relevantných faktorov na veľmi nízku intenzitu práce osôb slovenských domácností.

Pomocou PROC LOGISTIC sme odhadli model logistickej regresie a na základe testov, ktoré nám procedúra poskytla, sme mohli posúdiť, ktorý zo signifikantných faktorov má najväčší vplyv na binárnu závislú premennú VLWI. Metódou maximálnej vierohodnosti boli odhadnuté regresné koeficienty a im prislúchajúce pomery šancí, prostredníctvom ktorých sme mohli kvantifikovať vplyv faktora na závislú premennú. Testy štatistickej významnosti regresných koeficientov odhalili, že pomery šancí sa v prípade niektorých regresných koeficientov nedajú považovať za rozdielne, čo nás viedlo k hlbšej analýze vzťahov medzi kategóriami faktorov.

Pomocou procedúry GENMOD a aplikáciou príkazu LSMEANS sme zistili, že pri štyroch najvplyvnejších faktoroch sa v prípade niektorých dvojíc kategórií nepotvrdil signifikantný rozdiel stredných hodnôt logitu šance. Vzhľadom na to, že analýza marginálnych stredných hodnôt (LSMEANS) odhaľuje signifikantnosť rozdielu len medzi dvojicami kategórií, pri úvahe o zhode niekoľkých stredných hodnôt sme museli rozšíriť procedúru LOGISTIC o príkaz CONTRAST. Pomocou tohto príkazu sme overovali platnosť nulovej hypotézy, že stredné hodnoty logitu šance u jednotlivých kategórií faktora nie sú signifikantne rozdielne. Výsledok príkazu nám jednoznačne potvrdil uvažovaný predpoklad, preto sme najpodobnejšie kategórie zlúčili do jednej novovytvorenej kategórie faktora, a tým pádom

nahradili štyri najvplyvnejšie nezávisle premenné (EA, HT, Education a Region) novými modifikovanými faktormi.

Rozšírením procedúry LOGISTIC o príkaz ESTIMATE sme kvantifikovali odhad pravdepodobnosti, že osoba, ktorá je invalidná/v domácnosti/ nezamestnaná žijúca v KE alebo v PO kraji, bude čeliť veľmi nízkej intenzite práce a rovnako sme odhadli aj pravdepodobnosť, že zamestnaná osoba žijúca v BA, TN, TT alebo v ZA kraji bude ohrozená veľmi nízkou intenzitou práce, pričom ostatné faktory ostali fixované na najkritickejšej úrovni.

Aj keď v príspevku uvádzame redukciu kategórií len v prípade štyroch najvplyvnejších faktorov, vo všeobecnosti je možné redukovať aj kategórie ostatných faktorov. Zahrnutie príkazu CONTRAST do analýz má výhodu v odhalení skrytých a hlbších vzťahov medzi kategóriami, zlúčením napríklad menej početných kategórií do jednej novej kategórie má za následok zmenšenie štandardnej chyby odhadu marginálnych stredných hodnôt a zároveň sú výsledky analýz prehľadnejšie.

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] Allison, P. D. (2012). *Logistic Regression using SAS. Theory and Application.* Cary, NC: SAS Institute Inc., second edition.

[2] Cantillon, B., Vandenbroucke, F. (2014). Reconciling Work and Poverty Reduction: *How Successful are European Welfare States?* Oxford University Press.

[3] Glaser-Opitzová, H., Vojtková, M. (2020). *The influence of selected factors on the at-risk-of-poverty rate of Slovak households* [online]. Dostupné na: https://www.itema-conference.com/wp-content/uploads/2021/04/ITEMA.S.P.2020.107.pdf [cit. 2022-9-11].

[4] Ionescu, R. V. (2014). *Education, Employment and Poverty in the context of the Europe 2020 Strategy* [online]. Dostupné na: https://www.researchgate.net/publication/318306809_EDUCATION_EMPLOYMENT_AND_POVERTY_IN_THE_CONTEXT_OF_THE_EUROPE_2020_STRATEGY [cit. 2022-9-20].

[5] Johnston, H., McGauran, A. M. (2018). *Low Work Intensity Households and the Quality of Supportive Services: Detailed Research Report* [online]. Dostupné na: http://www.tara.tcd.ie/bitstream/handle/2262/100570/Research_Series_Paper_12_Low_Work_Intensity_Households.pdf?sequence=1 [cit. 2022-9-13].

[6] Rastrigina, O., Leventi, Ch., Sutherland, H. (2015). *Nowcasting risk of poverty and low work intensity in Europe* [online]. Dostupné na: https://www.iser.essex.ac.uk/research/publications/working-papers/euromod/em9-15 [cit. 2022-9-20].

[7] Schober, P., Vetter, T. (2021). *Logistic Regression in Medical Research* [online]. Dostupné na: https://journals.lww.com/anesthesia-analgesia/fulltext/2021/02000/logistic_regression_in_medical_research.12.aspx [cit. 2022-9-11].

[8] 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.

[9] Ward, T., Ozdemir, E. (2013). *Measuring low work intensity – an analysis of the indicator* [online]. Dostupné na: https://ideas.repec.org/p/hdl/improv/1309.html [cit. 2022-9-13].