

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103005/I/2023/421000214357

**ANALÝZA MARGINÁLNYCH STREDNÝCH
HODNÔT EKVIVALENTNÉHO DISPONIBILNÉHO
PRÍJMU ČLENOV SLOVENSKÝCH DOMÁCNOSTÍ
V PROGRAMOVACÍCH JAZYKOCH SAS A R**

Diplomová práca

2023

Bc. Sofia Bániková

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

**ANALÝZA MARGINÁLNYCH STREDNÝCH
HODNÔT EKVIVALENTNÉHO DISPONIBILNÉHO
PRÍJMU ČLENOV SLOVENSKÝCH DOMÁCNOSTÍ
V PROGRAMOVACÍCH JAZYKOCH SAS A R**

Diplomová práca

Študijný program: Data science v ekonómii

Študijný odbor: Ekonómia a manažment

Školiace pracovisko: Katedra štatistiky

Vedúci záverečnej práce: prof. Mgr. Erik Šoltés, PhD.

Bratislava 2023

Bc. Sofia Bániková

Zadanie

Čestné vyhlásenie

Čestne vyhlasujem, že som predkladanú diplomovú prácu vypracovala samostatne a že som uviedla všetky použité zdroje literatúry.

Dátum:

Podpis študenta:

Pod'akovanie

Touto cestou by som sa chcela pod'akovať vedúcemu záverečnej práce prof. Mgr. Erikovi Šoltésovi, PhD. za jeho ochotu, pomoc, trpezlivosť a taktiež cenné rady a pripomienky, poskytnuté počas písania tejto záverečnej práce.

Abstrakt

BÁNIKOVÁ, Sofia: *Analýza marginálnych stredných hodnôt ekvivalentného disponibilného príjmu členov slovenských domácností v programovacích jazykoch SAS a R.*

– Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra štatistiky.

– Vedúci záverečnej práce: prof. Mgr. Erik Šoltés, PhD. – Bratislava: FHI EU, 2023, 76 s.

Hlavným cieľom záverečnej práce je priblíženie a porovnanie možností analýz marginálnych stredných hodnôt ekvivalentného disponibilného príjmu domácností SR v programovacích jazykoch SAS a R. Práca je rozdelená na štyri kapitoly a obsahuje 33 výstupov, 9 obrázkov a 11 tabuliek. V prvej kapitole práce sa venujeme teoretickému vymedzeniu základných pojmov riešenej problematiky a taktiež súčasnému stavu doma a v zahraničí. Druhá kapitola bližšie definuje hlavný cieľ záverečnej práce. V tretej kapitole popisujeme metodiku práce a metódy skúmania, ktoré budú v ďalšej kapitole použité na analýzu a naplnenie cieľa práce. Posledná kapitola sa venuje výsledkom práce a teda samotnej analýze marginálnych stredných hodnôt v oboch programovacích jazykoch. Výsledky práce teda poskytujú porovnanie analýz marginálnych stredných hodnôt všeobecných lineárnych modelov v dvoch programovacích jazykoch a taktiež pohľad na vplyv rôznych faktorov na ekvivalentný disponibilný príjem domácností SR.

Kľúčové slová: ekvivalentný disponibilný príjem, všeobecný lineárny model, marginálne stredné hodnoty, EMMEANS

Abstract

BÁNIKOVÁ, Sofia: *Marginal means analysis of equivalised disposable income of Slovak households members in programming languages SAS and R.* – The University of Economics in Bratislava. Faculty of Economic Informatics; Department of Statistics. – Thesis supervisor: prof. Mgr. Erik Šoltés, PhD. – Bratislava: FHI EU, 2023, 76 p.

The main objective of the final thesis is to present and compare the possibilities of analyzing the marginal mean values of the equivalent disposable income of Slovak households using SAS and R programming languages. The thesis is divided into four chapters and includes 33 outputs, 9 figures and 11 tables. The first chapter focuses on the theoretical definition of the key concepts related to the problem being addressed, as well as the current situation both domestically and internationally. The second chapter provides a more detailed definition of the main objective of the final thesis. In the third chapter, the methodology of the study and the research methods to be used in the subsequent chapter are described, which will be employed to analyze and achieve the objective of the thesis. The final chapter is dedicated to presenting the results of the study, specifically the analysis of marginal mean values in both programming languages. Consequently, the results of the study offer a comparison of analyses of the marginal mean values of general linear models in the two programming languages, as well as insights into the impact of various factors on the equivalized disposable income of Slovak households.

Keywords: equivalized disposable income, general linear model, least squares means, EMMEANS

Obsah

Úvod	8
1 Súčasný stav riešenej problematiky doma a v zahraničí	10
1.1 Štatistické zisťovanie EU-SILC	10
1.2 Ekvivalentný disponibilný príjem	11
1.3 Vybrané indikátory chudoby a sociálneho vylúčenia	13
2 Cieľ práce	19
3 Metodika práce a metódy skúmania	20
3.1 Všeobecný lineárny model (GLM)	20
3.2 Overenie štatistickej významnosti modelu	21
3.3 Overenie štatistickej významnosti vplyvu premenných	23
3.4 Overenie štatistickej významnosti regresných koeficientov	23
3.5 SAS Enterprise Guide 7.1	24
3.6 Programovací jazyk R	27
4 Výsledky práce	28
4.1 Všeobecný lineárny model EDP členov domácností	29
4.2 Všeobecný lineárny model EDP členov domácností s upravenými premennými	44
4.3 Všeobecný lineárny model EDP členov domácností s interakciou	49
4.4 Programovací jazyk R	63
Záver	71
Zoznam použitej literatúry	74
Prílohy	76

Úvod

Marginálne stredné hodnoty predstavujú dôležitý ukazovateľ v rôznych oblastiach, ako napríklad ekonometria, medicína, sociológia či demografia. Sú používané na meranie a analýzu vplyvu zmien jednej premennej na druhú v rámci analýz, ktoré obsahujú viacero nezávislých premenných. Obzvlášť užitočné je ich skúmanie pri modelovaní vplyvu kategoriálnych premenných na spojitú závislú premennú, keďže umožňujú porovnať stredné hodnoty závislej premennej pre rôzne kategórie nezávislých premenných. V tejto práci sa zameriame na analýzu marginálnych stredných hodnôt ekvivalentného disponibilného príjmu domácností SR v závislosti od ukazovateľov ako napríklad najvyššie dosiahnuté vzdelanie, ekonomická aktivita či kraj SR, z ktorého členovia domácnosti pochádzajú. Na analýzu sa v oblasti ekonometrie a štatistiky využíva množstvo nástrojov na analýzu dát, pričom programovacie jazyky SAS a R, prostredníctvom ktorých budeme analýzu vykonávať, patria medzi najpoužívanejšie.

Cieľom práce je teda vysvetliť koncept marginálnych stredných hodnôt a ekvivalentného disponibilného príjmu, ako aj metódy ich analýzy pomocou programovacích jazykov SAS a R.

Práca je rozdelená na štyri kapitoly, pričom v prvej kapitole sa zameriame na súčasný stav riešenej problematiky a teda vymedzíme ekvivalentný disponibilný príjem a jeho význam pri hodnotení chudoby a sociálneho vylúčenia ako na Slovensku tak aj v EÚ. Taktiež priblížime podstatu zisťovania EU-SILC, ktoré je hlavným zdrojom údajov o ekvivalentnom disponibilnom príjme domácností.

V druhej kapitole detailne popíšeme nie len hlavný cieľ práce a taktiež všetky čiastkové ciele, ktoré nám napomôžu pri dosahovaní hlavného cieľa.

V tretej kapitole sa budeme venovať metodike práce a metódam skúmania. Popíšeme teda všeobecný lineárny model a princípy overovania štatistickej významnosti ako celého modelu, tak aj jednotlivých premenných či regresných koeficientov. Taktiež sa bližšie pozrieme na oba programovacie jazyky a ich konkrétne procedúry či balíčky, ktoré budú v ďalšej kapitole práce priamo využité na analýzu.

Vo štvrtej a zároveň poslednej kapitole pristúpime ku samotnej analýze marginálnych stredných hodnôt. V prvej časti tejto kapitoly budeme analyzovať ekvivalentný disponibilný príjem za využitia softvéru SAS, v druhej časti na analýzu využijeme programovací jazyk R. Postupy a získané výsledky následne interpretujeme a porovnáme.

Hlavným prínosom práce bude teda aplikácia teoretických vedomostí na reálne dáta a interpretácia výsledkov získaných pomocou programovacích jazykov SAS a R. Práca ukáže využitie štatistických nástrojov v praxi a pomôže lepšie pochopiť ekonomické ukazovatele ako marginálne stredné hodnoty a ekvivalentný disponibilný príjem. Taktiež prinesie porovnanie práce v oboch programovacích jazykoch, pričom poukáže na ich prednosti či konkrétne rozdiely.

1 Súčasný stav riešenej problematiky doma a v zahraničí

Chudoba je komplexný, viacrozmerný sociálny problém pri definovaní a meraní ktorého vždy vychádzame z konkrétneho konceptu. Neexistuje teda jediná správna definícia chudoby, ani jediný univerzálny spôsob jej merania. Každé meranie chudoby predstavuje meranie len určitého jej konceptu. (Vlačuha, 2019)

Rozoznávame viacero konceptov chudoby, ktoré sú kategorizované na základe rôznych kritérií, pričom sa tieto kategórie prekrývajú medzi sebou. My sa ďalej budeme zameriavať na analýzu medzinárodne definovaných a porovnateľných indikátorov chudoby a sociálneho vylúčenia, ktoré vychádzajú z ekvivalentného disponibilného príjmu domácností definovaného v prieskume EU-SILC. Tieto indikátory merajú nasledujúce koncepty chudoby:

- relatívny koncept – relatívna miera chudoby sa vypočítava porovnaním príjmu s určitou strednou hodnotou, v našom prípade s národnou hranicou chudoby,
- nepriamy koncept – meranie chudoby sa vykonáva prostredníctvom disponibilného príjmu domácností, pričom sa nezohľadňujú výdavky domácností,
- objektívny koncept – hranica chudoby sa stanovuje na základe nezávislých a objektívnych kritérií, nie subjektívne. (Vlačuha, 2019)

1.1 Štatistické zisťovanie EU-SILC

Zisťovanie o príjmoch a životných podmienkach domácností – EU-SILC (z angl. European Union Statistics on Income and Living Conditions), ktorého cieľom je zabezpečiť pravidelné, kvalitné a včasné informácie o príjmoch, chudobe a sociálnom vylúčení sa na Slovensku realizuje od roku 2005. Toto zisťovanie je súčasťou harmonizovaného zisťovania členských krajín Európskej únie, ktoré sa uskutočňuje na základe porovnateľnej medzinárodnej metodiky (Nariadenie (ES) č. 1177/2003 Európskeho parlamentu a Rady EÚ). Na Slovensku ho realizuje Štatistický úrad Slovenskej republiky. So zisťovaním sa začalo v roku 2003 na základe džentlmenskej dohody s Eurostatom v Belgicku, Dánsku, Írsku, Grécku, Luxembursku a Rakúsku s následným rozšírením do celej Európy. Do roku 2011 bolo do tohto zisťovania zahrnutých už všetkých 27 členských štátov EÚ a taktiež Island, Nórsko, Srbsko a Švajčiarsko. (Štatistika rozdelenia príjmov, 2015)

Výsledky tohto zisťovania sú dôležitým zdrojom informácií o životnej úrovni obyvateľstva a slúžia ako podklad pre plánovanie opatrení na zvyšovanie kvality života občanov. Jednými z hlavných výstupov tohto zisťovania sú indikátory chudoby a sociálneho

vylúčenia, na základe ktorých je na národnej úrovni možné sledovať chudobu a sociálne vylúčenie z rôznych hľadísk. V rámci krajín EÚ sa na analýzy využívajú medzinárodne definované a porovnateľné indikátory chudoby a sociálneho vylúčenia, vychádzajúce z ekvivalentného disponibilného príjmu domácností. (Vlačuha, 2019)

1.2 Ekvivalentný disponibilný príjem

Koncept ekvivalentného disponibilného príjmu domácností bol vyvinutý v 50. a 60. rokoch 20. storočia. V minulosti sa tomuto konceptu venovalo mnoho sociológov, ekonómov, či štatistikov. Medzi najvýznamnejších patrí britský sociálny teoretik a ekonóm Richard M. Tilmuss, ktorý sa vo svojej práci „Income Distribution and Social Change“ z roku 1962 zaoberal nerovnosťou v príjme a životnej úrovni rôznych skupín obyvateľstva. Toto dielo patriace medzi kľúčové diela v oblasti výskumu príjmovej nerovnosti predstavuje významný prínos k pochopeniu významu a dôsledkov tohto konceptu pre spoločnosť a ponúka nástroje na jeho meranie a zmiernenie.

Konceptu ekvivalentného disponibilného príjmu sa venovali aj ďalší ekonómovia a sociológovia ako napríklad Branko Milanović, ktorý sa v diele „Income, Inequality, and Poverty during the Transition from Planned to Market Economy“ zameriaval na vývoj príjmu a chudoby v krajinách strednej a východnej Európy po páde komunizmu. Autor tu využíval koncept ekvivalentného disponibilného príjmu na meranie životnej úrovne a porovnávanie vývoja príjmu a nerovnosti v rôznych krajinách. (Milanović, 1998)

Walter Garrison Runciman vo svojom diele „Relative Deprivation and Social Justice“ z roka 1966 predstavil koncept materiálnej deprivácie a zaoberal sa taktiež ekvivalentným disponibilným príjmom ako nástrojom na meranie chudoby a nerovnosti. Taktiež poukazoval na fakt, že chudoba a sociálna nerovnosť majú negatívny vplyv na spoločnosť ako celok.

Títo významní predstavitelia výskumu ekvivalentného disponibilného príjmu prispeli k rozvoju tohto konceptu a jeho využitiu v rôznych oblastiach výskumu a analýzy. V súčasnosti sa týmto konceptom zaoberajú vedci a analytici z rôznych oblastí, vrátane ekonómie, sociológie, politológie a demografie, a využívajú ho pri výskume sociálnej stratifikácie, nerovnosti príjmov a chudoby.

Ekvivalentný disponibilný príjem predstavuje jednotnú výšku príjmu všetkých členov domácnosti. Pre zohľadnenie veľkosti a zloženia domácnosti je teda disponibilný príjem domácnosti delený ekvivalentnou veľkosťou domácnosti, pričom v zisťovaní EU-

SILC sa na výpočet ekvivalentnej veľkosti domácnosti využíva tzv. modifikovaná OECD škála, ktorá každému členovi domácnosti priradzuje určitý koeficient. Každý prvý člen domácnosti má priradený koeficient 1, každý ďalší člen domácnosti vo veku 14 a viac rokov má priradený koeficient 0,5 a každý ďalší člen vo veku menej ako 14 rokov má priradený koeficient 0,3. Ekvivalentná veľkosť domácnosti je potom rovná súčtu týchto koeficientov priradených každému členovi domácnosti. (Eurostat, 2021) Napríklad pri domácnosti pozostávajúcej z dvoch dospelých osôb a dvoch detí mladších ako 14 rokov by bola ekvivalentná veľkosť domácnosti rovná súčtu $1 + 0,5 + 0,3 + 0,3$, teda 2,1. V prípade, že by bol celkový disponibilný príjem tejto domácnosti 18 900€, ekvivalentný disponibilný príjem každého jej člena by bol na úrovni 9 000€.

Celkový disponibilný príjem domácnosti predstavuje súčet zložiek hrubého osobného príjmu členov domácnosti a zložiek príjmu na úrovni domácnosti. Od tohto súčtu sú následne odčítané pravidelné dane z majetku, pravidelné platené transfery medzi domácnosťami (napr. výživné, pravidelná peňažná výpomoc), daň z príjmu a príspevky na sociálne poistenie. Do osobného príjmu členov domácností zaraďujeme pracovné príjmy (peňažné príjmy, nepeňažné príjmy a peňažné zisky (straty) z podnikania a inej samostatnej zárobkovej činnosti), príjmy z majetku, sociálne príjmy (starobné dávky, dávky v nezamestnanosti, dávky v invalidite, nemocenské dávky a iné) a ďalšie príjmové komponenty. Do príjmu na úrovni domácnosti zaraďujeme napr. príjem z prenájmu majetku alebo prijaté transfery od iných domácností. (Vlačuha, 2019)

Referenčné obdobie pre výpočet príjmu sa líši v závislosti od krajiny. Väčšina krajín používa 12-mesačné obdobie, zvyčajne predchádzajúci kalendárny alebo daňový rok. V prípade Írska, je zisťovanie príjmu priebežné a zahrňuje posledných 12 mesiacov pred zisťovaním. (Štatistika rozdelenia príjmov, 2015)

Jednotkami zisťovania sú súkromné hospodáriace domácnosti a ich aktuálni členovia, ktorí žijú na území danej krajiny v čase zberu údajov. Na Slovensku sa zber údajov realizuje prostredníctvom terénneho zberu dát v náhodne vybraných domácnostiach, kde opytovatelia priamo zisťujú požadované údaje od jej členov. Pri zisťovaní bol doteraz používaný dvojstupňový, stratifikovaný výber, pričom stratifikačnými kritériami boli stupeň urbanizácie (7 úrovní podľa veľkostnej skupiny obce) a geografická stratifikácia (8 úrovní podľa krajov zodpovedajúcich úrovni NUTS 3). Aplikovaním týchto stratifikačných kritérií vzniklo 48 strát. V rámci jednotlivých strát boli domácnosti vyberané jednoduchým náhodným výberom s proporcionálnym počtom domácností v každej strate. (Vlačuha, 2022)

Ako už bolo spomínané vyššie, množstvo medzinárodne definovaných a porovnateľných indikátorov kvality života, chudoby a sociálneho vylúčenia, vychádza práve z ekvivalentného disponibilného príjmu. Tento ukazovateľ však nie je významným len na medzinárodnej, ale taktiež aj na národnej úrovni, nakoľko sa o neho pri svojej činnosti opierajú mnohé štátne inštitúcie.

Inštitút pre výskum práce a rodiny napríklad vo výskumnej správe „*Analýza výdavkov domácností na bývanie*“ prináša analýzu vybraných aspektov bývania na Slovensku. Ekvivalentný disponibilný príjem je v tejto správe využívaný najmä na analýzu percentuálneho zastúpenia nákladov na bývanie na príjmoch domácností. Skúma sa taktiež miera zaťaženia výdavkami na bývanie v rôznych príjmových kvintilochoch. (Gerbery, 2019)

Rada pre rozpočtovú zodpovednosť sa o ekvivalentný disponibilný príjem opierala napríklad pri kvantifikácii vplyvu opatrení tzv. prorodinného balíčka, kde skúmala okrem iného aj zmeny disponibilného príjmu domácností po zavedení týchto opatrení. (Kancelária RRZ, 2022)

Úrad pre reguláciu sieťových odvetví poskytuje v strategickej koncepcii na ochranu odberateľov spĺňajúcich podmienky energetickej chudoby súbor odporúčaní na systémové opatrenia, ktorých implementácia môže zvýšiť mieru dlhodobej udržateľnej dostupnosti energií a pitnej vody pre sociálne zraniteľné domácnosti SR. V tejto koncepcii sa ekvivalentný disponibilný príjem využíva práve na odhad počtu domácností ohrozených energetickou chudobou.

1.3 Vybrané indikátory chudoby a sociálneho vylúčenia

Indikátor miery rizika chudoby alebo sociálneho vylúčenia AROPE (z angl. At risk of poverty or social exclusion) je základným medzinárodne definovaným a porovnateľným indikátorom založeným na ekvivalentnom disponibilnom príjme. Je to hlavný ukazovateľ na monitorovanie dosahovania cieľa EÚ 2030 v oblasti chudoby a sociálneho vylúčenia a bol hlavným ukazovateľom na monitorovanie pokroku pri napĺňaní cieľa stratégie EÚ 2020 v oblasti chudoby. V tomto indikátore sa dokopy zohľadňujú tri dimenzie chudoby a teda je tvorený kombináciou nasledovných čiastkových indikátorov:

- miera rizika príjmovej chudoby,
- miera závažnej materiálovej a sociálnej deprivácie,
- miera veľmi nízkej pracovnej intenzity.

AROPE teda predstavuje počet osôb, ktoré sú buď ohrozené rizikom príjmovej chudoby a/alebo sú závažne materiálne a sociálne deprivované a/alebo žijú v domácnosti s veľmi nízkou pracovnou intenzitou, pričom osoby nachádzajúce sa vo viac ako jednej z vyššie uvedených situácií sú započítavané iba raz. Miera AROPE predstavuje podiel osôb v riziku chudoby alebo sociálneho vylúčenia ku celkovej populácii. (Vlačuha, 2022)

Čo sa týka čiastkových indikátorov miery rizika a chudoby alebo sociálneho vylúčenia, miera rizika príjmovej chudoby predstavuje percentuálny podiel osôb z celkovej populácie, ktorých ekvivalentný disponibilný príjem je pod hranicou rizika chudoby, teda pod 60 % mediánu národného ekvivalentného disponibilného príjmu.¹ Miera rizika chudoby je ovplyvnená množstvom faktorov, okrem iného sem patrí regionálne hľadisko, vek, typ domácnosti, pracovná intenzita domácnosti, najčastejší status ekonomickej aktivity a vlastnícky vzťah k obydlíu. V roku 2021 žilo pod hranicou príjmovej chudoby 12,3 % obyvateľov SR. (Vlačuha, 2022)

Miera závažnej materiálnej a sociálnej deprivácie vyjadruje percentuálny podiel osôb čeliacich vynútenému nedostatku v aspoň siedmych zo zoznamu trinástich depriváčnych položiek. Tento zoznam bol v priebehu roka 2021 upravený, z pôvodného zoznamu deviatich depriváčnych položiek bolo ponechaných šesť a pridaných sedem nových depriváčnych položiek. Táto úprava bola zavedená na národnej i medzinárodnej úrovni s cieľom lepšie zohľadniť súčasné životné štandardy a tým pádom lepšie identifikovať sociálne vylúčené osoby či domácnosti. V roku 2021 pocíťovalo na Slovensku závažnú materiálovú a sociálnu depriváciu 5,7 % obyvateľov, čo predstavuje nárast o 1,2 p. b. oproti predchádzajúcemu roku. (Vlačuha, 2022)

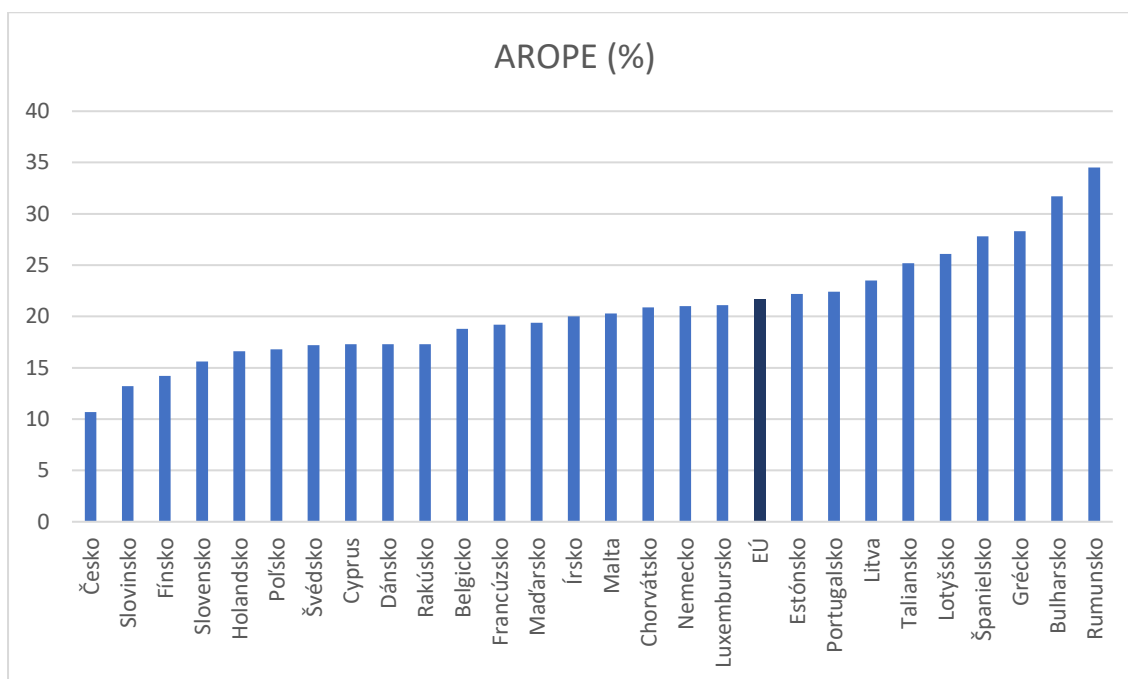
Miera veľmi nízkej pracovnej intenzity vyjadruje podiel osôb vo veku do 64 rokov žijúcich v domácnostiach s veľmi nízkou pracovnou intenzitou na celkovej populácii. Medzi domácnosti s veľmi nízkou pracovnou intenzitou zaraďujeme domácnosti, ktorých členovia v pracovnom veku odpracovali menej ako 20 % z celoročného pracovného potenciálu počas príjmového referenčného obdobia. Definícia pracovného veku sa vzťahuje na osoby vo veku od 18 rokov a nezahrňuje študentov vo veku od 18 do 24 rokov, prípadne osoby na predčasnom či invalidnom dôchodku. Miera veľmi nízkej pracovnej intenzity dosahovala na Slovensku v roku 2021 úroveň 3,9 %. (Vlačuha, 2022)

Obrázok 1 zobrazuje hodnoty indikátora miery rizika chudoby alebo sociálneho vylúčenia v krajinách EÚ v roku 2021. Na Slovensku dosahovala táto miera hodnotu 15,6 %,

¹ Medián ekvivalentného disponibilného príjmu je hodnota ekvivalentného disponibilného príjmu, ktorá rozdeľuje súbor podľa výšky príjmu na dve rovnako početné časti podľa počtu osôb.

čo predstavuje nárast o 1,8 p. b. oproti predchádzajúcemu roku, pričom národným cieľom je do roku 2030 dosiahnuť úroveň 13,3 %. Najvyššia hodnota miery rizika chudoby alebo sociálneho vylúčenia bola v roku 2021 dosahovaná v Rumunsku (34,5 %), naopak najnižšia miera bola zaznamenaná v Česku (10,7 %). V EÚ bola v roku 2021 priemerná hodnota tohto indikátora na úrovni 21,7 %, čo predstavuje 95,4 milióna ľudí. (Vlačuha, 2022)

Na základe EU-SILC 2021 bolo riziko chudoby alebo sociálneho vylúčenia zvýšené u žien (v porovnaní s mužmi), u mladých dospelých vo veku 18 – 24 rokov (v porovnaní s osobami stredného alebo vyššieho veku) a u ľudí s nízkou úrovňou dosiahnutého vzdelania (úrovne ISCED 0 – 2). Vyššiemu riziku chudoby alebo sociálneho vylúčenia čelili taktiež domácnosti so závislými deťmi, v roku 2021 bola až 1/5 obyvateľov EÚ žijúcich v domácnostiach so závislými deťmi ohrozená chudobou alebo sociálnym vylúčením. Čo sa týka pracovného statusu, ten predstavuje jednu z hlavných sociálno-ekonomických charakteristík ovplyvňujúcich riziko chudoby alebo sociálneho vylúčenia, pričom nezamestnané osoby čelia tomuto riziku najviac. V roku 2021 bolo až 64,5 % nezamestnaných osôb vo veku nad 17 rokov ohrozených chudobou alebo sociálnym vylúčením). Pre porovnanie, podiel zamestnaných osôb, ktoré boli ohrozené chudobou alebo sociálnym vylúčením bol 11,1 %. (Living conditions in Europe..., 2022)



Obrázok 1 Indikátor miery rizika chudoby alebo sociálneho vylúčenia v krajinách EÚ v roku 2021

Zdroj: EUROSTAT, vlastné spracovanie

Vzhľadom na to, že sa v práci budeme zaoberať výlučne skúmaním ekvivalentného disponibilného príjmu domácností SR a jeho porovnávaním v závislosti od rôznych vplyvov, priblížime si taktiež mieru rizika chudoby alebo sociálneho vylúčenia v členení podľa krajov SR. Je známe, že Slovensko má dlhodobo výrazné hospodárske a sociálne rozdiely medzi regiónmi, ktoré súvisia s nerovnomerným ekonomickým rozvojom jednotlivých krajov. Chudobnejšie regióny trpia nedostatočne rozvinutou infraštruktúrou, pomalým tempom budovania dopravnej siete, slabšie rozvíjajúcim sa podnikateľským prostredím, či nižšou úrovňou vzdelania niektorých obyvateľov. Tieto problémy vedú k prehlbovaniu nerovností medzi jednotlivými krajinami, čo potvrdzuje aj tabuľka 1 nižšie. (Vlačuha, 2022)

Môžeme vidieť, že v Banskobystrickom kraji bol v roku 2021 najvyšší podiel ľudí ohrozených chudobou alebo sociálnym vylúčením (24,9 %). Prešovský a Košický kraj mali tiež vysoký podiel, s hodnotami 21,3 % a 19,9 %. Naopak, v Bratislavskom kraji bolo chudobou alebo sociálnym vylúčením ohrozených iba 6,1 % ľudí.

Tabuľka 1 Miera rizika chudoby alebo sociálneho vylúčenia v členení podľa krajov SR v roku 2021

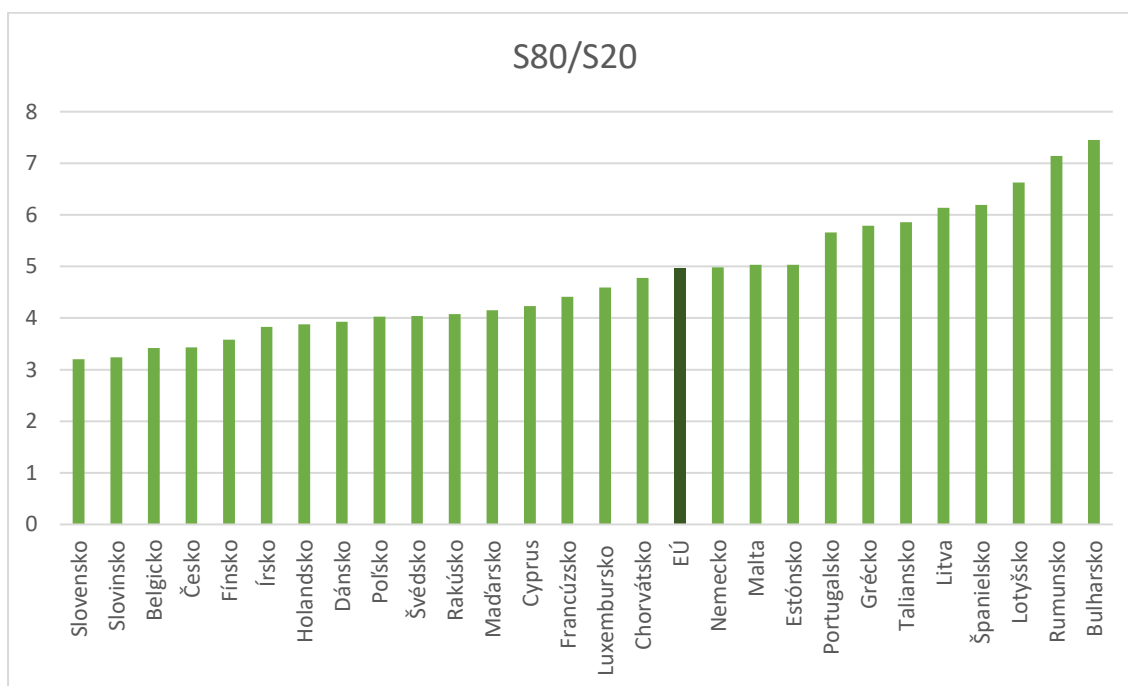
Kraj SR	AROPE (%)
Bratislavský kraj	6,1
Trnavský kraj	11,9
Trenčiansky kraj	11,7
Nitriansky kraj	12,1
Žilinský kraj	14,3
Banskobystrický kraj	24,9
Prešovský kraj	21,3
Košický kraj	19,9

Zdroj: EU-SILC 2021, vlastné spracovanie

Ako bolo uvedené, indikátor AROPE je hlavným ukazovateľom na monitorovanie dosahovania cieľa EÚ 2030 v oblasti chudoby a sociálneho vylúčenia. Európsky pilier sociálnych práv (EPSP) vytýčil v oblasti zamestnanosti, zručností a sociálnej ochrany na úrovni EÚ tri ciele. Za účelom znižovania chudoby a sociálneho vylúčenia bol stanovený nasledovný cieľ: znížiť počet ľudí ohrozených chudobou alebo sociálnym vylúčením aspoň o 15 miliónov, z čoho by aspoň 5 miliónov mali predstavovať deti. To zodpovedá hodnote indikátora miery rizika chudoby alebo sociálneho vylúčenia v EÚ na úrovni 17,3 %. (The European Pillar..., 2021)

Pomer príjmov horného a dolného kvintilu ($S80/S20$) je indikátor peňažnej chudoby, vyjadrujúci podiel sumy príjmov 20 % osôb z populácie s najvyššími príjmami (horný kvintil) k sume príjmov 20 % osôb z populácie s najnižšími príjmami (dolný kvintil). Tento koeficient nerovnosti príjmového rozdelenia môže nadobúdať hodnoty od 1 do nekonečna. V krajinách EÚ sa v praxi zväčša pohybuje v rozmedzí 2 až 12, pričom čím je hodnota tohto koeficienta vyššia, tým väčšia je nerovnomernosť príjmového rozdelenia v spoločnosti. (Vlačuha, 2019)

Obrázok 2 zobrazuje pomer príjmov horného a dolného kvintilu v krajinách EÚ v roku 2021. Medzi členskými štátmi EÚ dosahoval tento indikátor v roku 2021 najnižšiu hodnotu na Slovensku (3,2) a naopak najvyššia hodnota bola evidovaná v Bulharsku (7,45). V EÚ bola priemerná hodnota na úrovni 4,97.



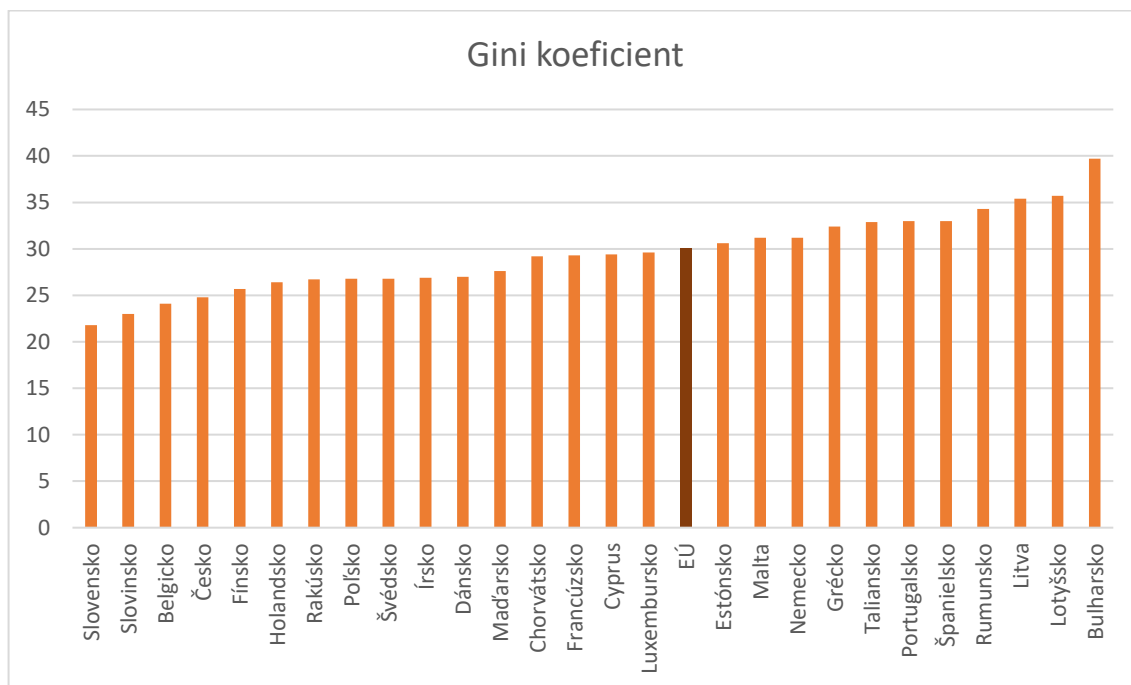
Obrázok 2 Pomer príjmov horného a dolného kvintilu ($S80/S20$) v krajinách EÚ v roku 2021

Zdroj: EUROSTAT, vlastné spracovanie

Giniho koeficient je známy ukazovateľ príjmovej nerovnosti, vyjadrujúci vzťah medzi kumulatívnymi podielmi populácie rozdelených podľa výšky ekvivalentných disponibilných príjmov a kumulatívnymi podielmi celkového ekvivalentného disponibilného príjmu. Ukazovateľ $S80/S20$ porovnáva len príjmy medzi hornými a dolnými vrstvami spoločnosti, zatiaľ čo Giniho koeficient berie do úvahy príjmy celej spoločnosti. Tento koeficient môže nadobúdať hodnoty od 0 do 1, resp. v od 0 do 100 percentuálnom

vyjadrení a taktiež vyjadruje príjmovú nerovnosť v spoločnosti. Ak by v spoločnosti existovala absolútna rovnosť v príjmoch, Giniho koeficient by bol rovný 0. Ak by všetky príjmy v spoločnosti patrili len jednej osobe, Giniho koeficient by nadobudol hodnotu 1. Čím je teda Giniho koeficient vyšší, tým je väčšia aj príjmová nerovnosť v spoločnosti. (Vlačuha, 2019)

Obrázok 3 zobrazuje hodnoty Giniho koeficienta v krajinách EÚ v roku 2021. V roku 2021 boli najväčšie rozdiely v príjmoch spomedzi členských štátov EÚ zaznamenané v Bulharsku, Lotyšsku a Litve. Naopak, najrovnomernejšie boli príjmy rozdelené na Slovensku a taktiež v Slovinsku, Belgicku a Česku. Priemerná hodnota tohto koeficienta v EÚ bola v roku 2021 na úrovni 30,1 %.



Obrázok 3 Giniho koeficient v krajinách EÚ v roku 2021

Zdroj: EUROSTAT, vlastné spracovanie

2 Cieľ práce

Hlavným cieľom práce je pomocou analýzy marginálnych stredných hodnôt a kontrastnej analýzy v programovacích jazykoch SAS a R kvantifikovať vplyv relevantných faktorov na ekvivalentný disponibilný príjem slovenských domácností. Konkrétne budeme na základe údajov pochádzajúcich zo zisťovania EU-SILC analyzovať vplyv ukazovateľov ako vzdelanie, typ domácnosti alebo kraj na výšku ekvivalentného disponibilného príjmu členov domácností SR za rok 2021.

V prvej časti práce bolo teda čiastkovým cieľom priblíženie konceptu ekvivalentného disponibilného príjmu a jeho významu a dôležitosti v kontexte rôznych ukazovateľov chudoby a príjmových nerovností. V oblasti týchto indikátorov sme taktiež mali za cieľ priblížiť súčasný stav na Slovensku a v zahraničí a jednotlivé krajiny medzi sebou porovnať. V súvislosti s ekvivalentným disponibilným príjmom bolo čiastkovým cieľom tejto kapitoly taktiež objasnenie významu štatistického zisťovania EU-SILC.

V ďalšej kapitole sme si za čiastkový cieľ práce stanovili vysvetlenie podstaty všeobecných lineárnych modelov a testov na vyhodnocovanie kvality týchto modelov. Taktiež popísanie prostredia štatistického softvéru SAS Enterprise Guide a procedúry GLM ako aj programovacieho jazyka R a balíčka EMMEANS.

V praktickej časti práce je čiastkovým cieľom práce skonštruovať všeobecný lineárny model vystihujúci vplyv vybraných faktorov na ekvivalentný disponibilný príjem, overiť štatistickú významnosť tohto modelu a pomocou analýzy marginálnych stredných hodnôt a kontrastnej analýzy v programovacích jazykoch SAS a R bližšie popísať vzťahy medzi vybranými faktormi a ekvivalentným disponibilným príjmom členov domácností.

V závere práce je cieľom zhodnotiť výsledky predošlých analýz a zhrnúť výhody a nevýhody používania programovacích jazykov SAS a R pri analýze marginálnych stredných hodnôt.

3 Metodika práce a metódy skúmania

V tejto kapitole záverečnej práce sa budeme venovať priblíženiu metód a postupov použitých pri analýze všeobecných lineárnych modelov v praktickej časti práce. Nakoľko správnosť výsledkov analýz závisí hlavne od správneho zostavenia a kvality modelu, je nutné si túto kvalitu overiť pomocou štatistických testov.

3.1 Všeobecný lineárny model (GLM)

Všeobecné lineárne modely (z angl. General Linear Models – GLM) využívame na modelovanie vzťahu medzi spojitou závislou premennou a jednou alebo viacerými nezávislými premennými, ktoré môžu byť ako spojité, tak aj kategoriálne. Všeobecný lineárny model je možné zapísať v tvare viacnásobného regresného modelu nasledovne:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i \quad (1)$$

kde:

y_i – i -te pozorovanie závislej (vysvetľovanej) premennej

x_{ik} – množina nezávislých (vysvetľujúcich) premenných

β_0 – lokujúca konštanta

β_k – množina k regresných koeficientov

ε_i – množina i náhodných chýb

Odhadnutá hodnota lokujúcej konštanty predstavuje hodnotu závislej premennej v prípade, že sú všetky nezávislé premenné rovné 0 (v prípade kategoriálnych premenných sú rovné referenčnej kategórii). Pre každú nezávislú premennú je ďalej odhadnutý regresný koeficient, ktorý vystihuje vzťah medzi touto nezávislou premennou x a závislou premennou y .

Viacnásobný regresný model môžeme rovnocenne vyjadriť aj maticovým zápisom

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

kde:

$\mathbf{y}$ – n -prvkový stĺpcový vektor pozorovaných hodnôt závislej premennej

$\mathbf{X}$ – matica typu $n \times (k + 1)$ pozorovaných hodnôt nezávislých premenných

$\boldsymbol{\beta}$ – $(k + 1)$ -prvkový stĺpcový vektor neznámych parametrov modelu

$\boldsymbol{\varepsilon}$ – n -prvkový stĺpcový vektor náhodných chýb

V prípade, že budeme v modeli uvažovať aj s interakciou niektorých nezávislých premenných, všeobecný lineárny model môžeme zapísať v tvare dvojfaktorového modelu nasledovne:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (3)$$

kde:

y_{ijk} – k -te pozorovanie závislej premennej na i -tej úrovni faktora A a súčasne j -tej úrovni faktora B

μ – stredná hodnota závislej premennej pre i -tu obmenu faktora A a j -tu obmenu faktora B

α_i – efekt faktora A , pričom i je počet obmien tohto faktora

β_j – efekt faktora B , pričom j je počet obmien tohto faktora

$(\alpha\beta)_{ij}$ – efekt interakcie medzi faktormi A a B

ε_{ijk} – množina náhodných chýb

3.2 Overenie štatistickej významnosti modelu

To, do akej miery odhadnutý model vystihuje variabilitu závislej premennej je možné posúdiť na základe analýzy rozptylu nezávislej premennej vychádzajúc z rovnosti

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \quad (4)$$

Úpravou rovnosti umocnením a sumáciou dostávame nasledovný tvar:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

kde rozdiel $y_i - \bar{y}$ predstavuje odchýlku i -tého pozorovania závislej premennej od priemeru a ľavá strana rovnice (5) predstavujúca súčet štvorcov týchto odchýlok vyjadruje celkovú variabilitu závislej premennej (SST). Súčet druhých mocnín rozdielov $(\hat{y}_i - \bar{y})$ vyjadruje variabilitu vyrovnaných hodnôt okolo priemeru $\bar{y}$, teda variabilitu závislej premennej vysvetlenú modelom (SSM). Súčet druhých mocnín rezíduí $(y_i - \hat{y}_i)$ predstavuje variabilitu nevysvetlenú modelom (SSR). Počet nezávislých sčítancov, z ktorých pozostáva súčet druhých mocnín determinuje počet stupňov voľnosti. Priemerný štvorec následne dostaneme ako podiel súčtu štvorcov odchýlok (SS) a zodpovedajúceho počtu stupňov voľnosti (DF). Podielom priemerných štvorcov MSM a MSR , teda podielom variability závislej premennej vysvetlenej modelom k variabilite závislej premennej nevysvetlenej modelom, dostávame testovaciu F štatistiku. Spojením týchto vzťahov do prehľadnej tabuľky dostávame tzv. tabuľku analýzy rozptylu zobrazenú na nasledujúcej strane.

Tabuľka 2 Analýza rozptylu pre lineárny regresný model

Variabilita závislej premennej	Súčet štvorcov odchýlok SS	Stupne voľnosti DF	Priemerný štvorec MS	Testovacia štatistika F
Variabilita vysvetlená modelom <i>M</i>	$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	k	$MSM = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{k}$	$F = \frac{MSM}{MSR}$
Variabilita nevysvetlená modelom <i>R</i>	$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - k - 1$	$MSR = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k - 1}$	
Celková variabilita <i>T</i>	$SST = \sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$		

Zdroj: vlastné spracovanie

Testovacia štatistika zobrazená v poslednom stĺpci tabuľky má Fisherovo rozdelenie pravdepodobnosti s počtom stupňov voľnosti k a $(n - k - 1)$ a slúži na overenie nulovej hypotézy

H_0 : model nie je štatisticky významný
oproti alternatívnej hypotéze

H_1 : model je štatisticky významný.

K zamietnutiu nulovej hypotézy na hladine významnosti α prichádza v prípade, že je hodnota testovacej štatistiky F väčšia ako hodnota kritickej oblasti $F_{1-\alpha}(k; n - k - 1)$. Ak daná nerovnosť neplatí, nulovú hypotézu nie je možné zamietnuť a teda môžeme konštatovať, že žiadna z premenných zahrnutých do daného modelu nemá štatisticky významný vplyv na závislú premennú. V opačnom prípade nulovú hypotézu zamietame a môžeme tvrdiť, že aspoň jedna z premenných zahrnutých do modelu má štatisticky významný vplyv na závislú premennú. To, ktoré premenné majú štatisticky významný vplyv na závislú premennú bude potom predmetom ďalšieho testovania.

Kvalitu odhadnutého modelu môžeme posúdiť na základe koeficienta determinácie, predstavujúceho podiel variability závislej premennej vysvetlenej modelom oproti celkovej variabilite:

$$r_{xy}^2 = \frac{SSM}{SST} \quad (6)$$

Z vyššie uvedeného je zrejmé, že tento koeficient nadobúda hodnoty z intervalu 0 a 1 a po prenasobení 100 ho môžeme interpretovať aj ako percento variability závislej premennej vysvetlené modelom.

3.3 Overenie štatistickej významnosti vplyvu premenných

Ak prišlo v teste štatistickej významnosti modelu k zamietnutiu nulovej hypotézy a teda vytvorený model považujeme za štatisticky významný, môžeme prejsť k testovaniu štatistickej významnosti vplyvu konkrétnych nezávislých premenných na závislú premennú. Testovať budeme teda platnosť nulovej hypotézy

H_0 : vplyv nezávislej premennej nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : vplyv nezávislej premennej je štatisticky významný.

porovnávaním testovacej štatistiky

$$F = \frac{\frac{\Delta var}{1}}{\frac{SSR^{(1)}}{n - k - 1}} = \frac{\Delta var}{MSR^{(1)}} \quad (7)$$

s kritickou hodnotou $F_{1-\alpha}(k; n - k - 1)$. Výraz Δvar (označovaný aj ako prínos k -tej nezávislej premennej) vo vzťahu testovacej štatistiky pritom predstavuje rozdiel variability závislej premennej vysvetlenej modelom a variability závislej premennej vysvetlenej redukovaným modelom, t. j. modelom, z ktorého vynecháme premennú, ktorej vplyv testujeme. Ak bude hodnota testovacej štatistiky väčšia ako kritická hodnota, zamietame nulovú hypotézu a vplyv tej-ktorej nezávislej premennej na závislú premennú môžeme považovať za štatisticky významný. V opačnom prípade by sme nezávislú premennú z modelu vylúčili.

3.4 Overenie štatistickej významnosti regresných koeficientov

V prípade, že sú do modelu zahrnuté aj kategóriálne premenné, v modeli je pre každú kategóriu tej-ktorej kategóriálnej premennej (s výnimkou referenčných kategórií) vytvorená umelá premenná. Na overenie štatistickej významnosti jednotlivých regresných koeficientov modelu budeme testovať nulovú hypotézu

H_0 : regresný koeficient nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : regresný koeficient je štatisticky významný

porovnávaním testovacej štatistiky

$$t = \frac{b_j}{s_{b_j}}, \quad (8)$$

kde b_j predstavuje hodnotu testovaného regresného koeficienta a s_{b_j} predstavuje jeho štandardnú odchýlku. Testovacia štatistika má Studentovo rozdelenie s počtom stupňov voľnosti $(n - k - 1)$. Vzhľadom na symetrické rozdelenie pravdepodobnosti testovacej štatistiky, s kritickou oblasťou budeme porovnávať testovaciu štatistiku v absolútnej hodnote. K zamietnutiu nulovej hypotézy na hladine významnosti α prichádza v prípade, ak je absolútna hodnota testovacej štatistiky t väčšia ako hodnota kritickej oblasti $t_{1-\frac{\alpha}{2}}(n - k - 1)$.

3.5 SAS Enterprise Guide 7.1

SAS (Statistical Analysis System) je softvérový produkt vyvinutý v 60. rokoch 20. storočia v Severnej Karolíne v USA, slúžiaci na spracovanie a analýzu dát. Obsahuje rôzne nástroje pre spracovanie dát a ich analýzu, vrátane nástrojov na vytváranie štatistických modelov, vizualizáciu dát, spracovanie textu, business intelligence a mnoho ďalších. Súčasťou tohto softvéru je aj prostredie SAS Enterprise Guide, ktoré budeme v ďalšej kapitole využívať na analýzu.

V softvéri SAS EG sa na odhad všeobecných lineárnych modelov pomocou metódy najmenších štvorcov využíva procedúra GLM. Medzi štatistické metódy dostupné v tejto procedúre patrí regresia, analýza rozptylu, analýza kovariancie, multivariačná analýza rozptylu a čiastočná korelácia.

V rámci procedúry GLM sú dostupné nasledovné funkcie:

PROC GLM <options>;

CLASS variable <(REF= option)> ...<variable <(REF= option)>> </ global-options>;

MODEL dependent-variables = independent-effects </ options>;

ABSORB variables;

BY variables;

CODE <options>;

FREQ variable;

ID variables;

WEIGHT variable;

CONTRAST 'label' effect values <...effect values> </ options>;

ESTIMATE 'label' effect values <...effect values> </ options>;
LSMEANS effects </ options>;
MANOVA <test-options> </ detail-options>;
MEANS effects </ options>;
OUTPUT <OUT=SAS-data-set> keyword=names <...keyword=names></ option>;
RANDOM effects </ options>;
REPEATED factor-specification </ options>;
STORE <OUT=>item-store-name </ LABEL='label'>;
TEST <H=effects> E=effect </ options>;

pričom funkcia MODEL je povinnou funkciou tejto procedúry, slúžiacou na zadefinovanie závislej a nezávislých premenných v odhadovanom modeli. V prípade, že sú do modelu zahrnuté aj kategoriálne premenné, spustenie procedúry je podmienené taktiež funkciou CLASS, ktorá slúži na deklaráciu kategoriálnych premenných použitých v modeli. Ostatné funkcie slúžia na dodatočnú úpravu modelu, doplnenie ďalších informácií či získanie výstupov potrebných pre želanú analýzu. (SAS Institute Inc., 2020) V našom prípade budeme na analýzu marginálnych stredných hodnôt využívať funkcie LSMEANS, CONTRAST a ESTIMATE.

V prípade, že sú do modelu zaradené kategoriálne premenné s viac ako dvoma obmenami, môžeme využiť funkciu LSMEANS a to konkrétne na výpočet marginálnych stredných hodnôt a taktiež posúdenie významnosti rozdielov medzi ich dvojicami. Na výpočet marginálnej strednej hodnoty pre určitú úroveň tej-ktorej premennej sa na základe postupu popísaného v publikácii Cary, NC: SAS Institute Inc., 2020 (s. 3809) konštruuje vektor lineárneho kontrastu **L**. Funkcia CONTRAST slúži na testovanie vlastných lineárnych hypotéz. Používame ju teda v prípade, že chceme posudzovať zhodu viacerých stredných hodnôt pre rozličné kategórie nezávislej premennej zahrnutej do modelu. Funkciu ESTIMATE používame na odhad strednej hodnoty závislej premennej pri ľubovoľnej kombinácii nezávislých premenných a taktiež na testovanie všeobecných lineárnych hypotéz. (SAS Institute Inc., 2020)

V prípade testovania rovnakých hypotéz dospejeme teda funkciou CONTRAST a ESTIMATE k rovnakému výsledku, pričom testovacia štatistika získaná funkciou CONTRAST je rovná druhej mocnine testovacej štatistiky získanej funkciou ESTIMATE. Štruktúra príkazov oboch funkcií je rovnaká, avšak na rozdiel od funkcie CONTRAST, funkciou ESTIMATE nie je možné testovať viaceré hypotézy simultánne. Konkrétna syntax týchto funkcií bude ďalej popísaná v praktickej časti práce.

V prípade testu zhody viacerých stredných hodnôt pre rozličné kategórie nezávislej premennej v rámci funkcie CONTRAST, testujeme teda nulovú hypotézu

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

oproti alternatívnej hypotéze

$$H_1: \mathbf{L}\boldsymbol{\beta} \neq 0.$$

O prijatí resp. zamietnutí nulovej hypotézy rozhodujeme na základe testovacej štatistiky:

$$F = \frac{\frac{(\mathbf{L}\mathbf{b}^\circ)^T \cdot [\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1} \cdot (\mathbf{L}\mathbf{b}^\circ)}{l}}{\frac{SSR}{n-p}}, \quad (9)$$

kde $\mathbf{b}^\circ$ predstavuje vektor parametrov modelu odhadnutého metódou zovšeobecnenej inverzie, teda $\mathbf{b}^\circ = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. l predstavuje počet stupňov voľnosti pre sumu štvorcov uvedenú v čitateli nad hlavnou zlomkovou čiarou a vyjadruje počet nezávislých riadkov matice $\mathbf{L}$. Menovateľ testovacej štatistiky predstavuje rozptyl rezíduí, kde p je hodnota matice $\mathbf{X}$. Testovacia štatistika má Fisherovo rozdelenie pravdepodobnosti s počtom stupňov voľnosti l a $(n-p)$. K zamietnutiu nulovej hypotézy príde na hladine významnosti α v prípade, ak bude hodnota testovacej štatistiky F väčšia ako hodnota kritickej oblasti $F_{1-\alpha}(l; n-p)$.

V prípade použitia funkcie ESTIMATE, o teste štatistickej významnosti regresného koeficienta pre novovytvorenú umelú premennú rozhodujeme na základe porovnania absolútnej hodnoty testovacej štatistiky

$$t = \frac{\mathbf{L}\mathbf{b}^\circ}{\sqrt{MSR \cdot [\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]}} \quad (10)$$

s kritickou oblasťou $t_{1-\frac{\alpha}{2}}(n-p)$. V prípade ak bude absolútna hodnota testovacej štatistiky väčšia ako hodnota kritickej oblasti, na hladine významnosti α zamietneme nulovú hypotézu a daný regresný koeficient budeme považovať za štatisticky významný.

3.6 Programovací jazyk R

R je programovací jazyk a prostredie určené predovšetkým na štatistickú analýzu a grafické spracovanie dát. Bol vyvinutý v roku 1993 v Austrálii štatistikom a výskumníkom Rossenom Ihakom a Robertom Gentlemanom. Je to open source jazyk, čo znamená, že jeho zdrojový kód je voľne dostupný a môže byť modifikovaný a distribuovaný komunitou používateľov. V súčasnosti sa často využíva v akademických kruhoch, výskumnej oblasti a priemysle na analýzu a vizualizáciu dát, modelovanie, strojové učenie a data mining. Disponuje množstvom rozšírení, knižníc, balíkov a nástrojov, ktoré umožňujú používateľom riešiť rôzne úlohy a problémy.

V tejto práci budeme na analýzu využívať prevažne balíček *emmeans*, ktorý slúži na odhad a vizualizáciu marginálnych stredných hodnôt a ich kontrastov pre rôzne druhy modelov, vrátane všeobecných lineárnych modelov. Okrem toho, tento balíček poskytuje možnosť vykonávať post hoc testy, teda testy na viacnásobné porovnanie marginálnych stredných hodnôt. Medzi základné a najviac využívané funkcie tohto balíčka patria napríklad:

1. `emmeans()` – funkcia umožňujúca výpočet marginálnych stredných hodnôt a kontrastov pre rôzne kombinácie faktorov v danom modeli
2. `pairs()` – funkcia slúžiaca na párové porovnanie marginálnych stredných hodnôt
3. `contrast()` – funkcia, ktorá umožňuje výpočet kontrastov medzi marginálnymi strednými hodnotami pre určité kombinácie faktorov. Taktiež umožňuje určiť intervaly spoľahlivosti pre tieto kontrasty.
4. `pwpm()` – táto funkcia umožňuje zobrazit' prehľadnú maticu poskytujúcu dôležité charakteristiky pre dvojice marginálnych stredných hodnôt. (Lenth, 2023)

4 Výsledky práce

V tejto kapitole práce sa budeme venovať aplikácii poznatkov a metodických postupov, ktoré boli vymedzené v predchádzajúcich častiach. Pomocou softvérov SAS Enterprise Guide 7.1 a RStudio 2022.07.2 analyzujeme marginálne stredné hodnoty všeobecných lineárnych modelov za účelom bližšieho popísania faktorov vplývajúcich na ekvivalentný disponibilný príjem členov domácností SR.

Vstupný dátový súbor s ktorým budeme pracovať pochádza zo štatistického zisťovania o chudobe – EU-SILC 2021. Z ukazovateľov sledovaných týmto zisťovaním bolo vybraných niekoľko faktorov, ktoré mali podľa nášho názoru významný vplyv na závislú premennú, vyššie spomínaný ekvivalentný disponibilný príjem. Nezávislými premennými, ktoré sme sa rozhodli do modelu zaradiť sú: ekonomická aktivita, vzdelanie, typ domácnosti, rodinný stav a kraj. Vzhľadom na to, že sú všetky nami vybrané nezávislé premenné kategóriálne, pred samotnou analýzou bolo nutné špecifikovať referenčnú, resp. porovnávajúcu kategóriu. Tá bola vybraná na základe logického uvažovania. Názvy kategórií boli následne prispôbené na analýzu softvérom SAS EG 7.1 a upravené tak, aby bola referenčná kategória podľa abecedy posledná. Tabuľka 3 nižšie už zobrazuje všetky premenné vstupujúce do modelu spolu s ich označením a jednotlivými obmenami, resp. kategóriami.

Tabuľka 3 Základné informácie o vstupných premenných

Označenie	Popis
EDP	Ekvivalentný disponibilný príjem
EA	Ekonomická aktivita
Disabled_person	osoba na invalidnom dôchodku
Inactive_person	iná neaktívna osoba
Person_in_household	osoba v domácnosti
Student	študent
Unemployed	nezamestnaný
z_at_Work (referenčná k.)	pracujúci
Education	Vzdelanie
ISCED_0_1_2	nižšie ako primárne, primárne a nižšie sekundárne
ISCED_3	vyššie sekundárne
ISCED_4	post-sekundárne, neterciárne
ISCED_5A	prvý stupeň terciárneho
z_ISCED_5B_6 (referenčná k.)	druhý a tretí stupeň terciárneho vzdelania

HT	Typ domácnosti
1A_at_least_1Ch	domácnosť 1 rodiča s 1+ závislým dieťaťom
1Adult	jednočlenná domácnosť
2A_1Ch	domácnosť 2 dospelých s 1 závislým dieťaťom
2A_1R	domácnosť 2 dospelých bez závislých detí – jeden dospelý vo veku 65r. a viac
2A_at_least_3Ch	domácnosť 2 dospelých s 3+ závislými deťmi
2Adult	domácnosť 2 dospelých bez závislých detí – obaja vo veku do 64r.
Other_Och	ostatné domácnosti bez závislých detí
Other_with_Ch	ostatné domácnosti so závislými deťmi
z_2A_2Ch (referenčná k.)	domácnosť 2 dospelých s 2 závislými deťmi
Marital_status	Rodinný stav
Divorced	rozvedený/á
Never_Married	slobodný/á
Widowsd	vdovec/vdova
z_Married (referenčná k.)	ženatý/vydatá
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 (referenčná k.)	Bratislavský kraj

Zdroj: EU-SILC 2021, vlastné spracovanie

Vstupný dátový súbor teda dokopy obsahuje 6 premenných, z toho jedna (závislá) premenná je kvantitatívna spojitá, ostatné (nezavislé) premenné sú kategóriálne. Nakoľko sme sa do analýzy rozhodli zahrnúť iba osoby v produktívnom veku (15 – 64 rokov), z dátového súboru bolo ešte nutné odstrániť pozorovania osôb predproduktívneho a poproduktívneho veku. Výberový súbor použitý na analýzu obsahoval po tejto úprave 7 356 pozorovaní.

4.1 Všeobecný lineárny model EDP členov domácností

V tejto a nasledujúcich kapitolách si priblížime analýzu marginálnych stredných hodnôt prostredníctvom procedúry GLM v softvéri SAS EG 7.1. Po úspešnom importovaní pripravených dát vyberieme v hlavnom menu *Tasks* → *ANOVA* → *Linear Models*.

V kontextovom okne následne v záložke *Data* priradíme závislej premennej rolu *Dependent variable*, nezávislým premenným priradíme rolu *Classification variables*. Následne prejdeme do záložky *Model*, kde pre nezávislé premenné nastavíme vplyv na závislú premennú. V tomto prípade pre všetky premenné vyberieme možnosť *Main*. Následne prejdeme do záložky *Post Hoc Tests*, kde pre všetky premenné nastavíme *Class effects to use* na možnosť *True*. Tlačidlom *Run* následne spustíme procedúru.

Pred samotnou analýzou marginálnych stredných hodnôt je nutné overiť, či nami odhadnutý model je štatisticky významný a teda, či vhodne opisuje vzťahy medzi závislou premennou a nezávislými premennými. Štatistickú významnosť lineárneho modelu posúdime na základe analýzy rozptylu závislej premennej, kde testujeme nulovú hypotézu

H_0 : model nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : model je štatisticky významný.

Záverečné rozhodnutie o prijatí resp. zamietnutí nulovej hypotézy vykonáme na základe porovnania vyčíslenej testovacej štatistiky s kritickou hodnotou, prípadne na základe p -hodnoty. Nulovú hypotézu na hladine významnosti α zamietneme v prípade, ak platí nerovnosť $F > F_{1-\alpha}(k; n - k - 1)$, teda ak je vyčíslená hodnota testovacej štatistiky väčšia ako kritická hodnota. V prípade rozhodovania o prijatí resp. zamietnutí nulovej hypotézy na základe p -hodnoty, nulovú hypotézu je možné zamietnuť, ak p -hodnota nepresiahne zvolenú hladinu významnosti α . p -hodnota je teda najnižšia hladina významnosti, pri ktorej možno zamietnuť nulovú hypotézu.

Tabuľka analýzy rozptylu zobrazená nižšie je súčasťou základného výstupu. Keďže táto tabuľka neobsahuje informáciu o kritickej hodnote, o teste štatistickej významnosti modelu rozhodneme na základe p -hodnoty uvedenej v stĺpci $Pr > F$. Na základe tejto hodnoty ($p < 0,0001$) môžeme na všetkých bežne používaných hladinách významnosti zamietnuť nulovú hypotézu a model považovať za štatisticky významný.

Výstup 1 Tabuľka analýzy rozptylu pre model EDP členov domácností

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	27	33914554059	1256094594.8	134.07	<.0001
Error	7328	68656663807	9369086.2182		
Corrected Total	7355	102571217866			

Zdroj: vlastné spracovanie v SAS EG

Kvalitu odhadnutého modelu môžeme posúdiť na základe koeficienta determinácie. Ten nájdeme taktiež v tabuľke, ktorá je súčasťou základného výstupu. Konkrétne ide o tabuľku zobrazenú vo výstupe 2. Koeficient determinácie sa nachádza v stĺpci R-Square a vyjadruje podiel variability závislej premennej vysvetlenej modelom k celkovej variabilite závislej premennej. V našom prípade tento koeficient dosahuje hodnotu 0,330644, čo znamená, že modelom vieme vysvetliť 33,06% variability ekvivalentného disponibilného príjmu členov domácností. Zvyšných 66,94% variability ekvivalentného disponibilného príjmu členov domácností ovplyvňujú činitele nezahrnuté do modelu. Vzhľadom na to, že príjem ovplyvňuje množstvo rôznych faktorov, ktoré v modeli neboli uvažované (ani zisťované v EU-SILC), nie je reálne očakávať vysokú hodnotu tohto koeficienta.

Výstup 2 Vybrané charakteristiky modelu EDP
členov domácností

R-Square	Coeff Var	Root MSE	EDP Mean
0.330644	34.06093	3060.896	8986.531

Zdroj: vlastné spracovanie v SAS EG

Zamietnutím nulovej hypotézy o štatistickej nevýznamnosti modelu sme vlastne prijali predpoklad, že aspoň jedna z nezávislých premenných zahrnutých do modelu má signifikantný vplyv na závislú premennú. Tento vplyv však nemusia mať všetky premenné zahrnuté do modelu. Ďalším krokom je teda otestovanie štatistickej významnosti vplyvu vybraných nezávislých premenných na závislú premennú. Budeme testovať nulovú hypotézu

H_0 : vplyv faktora na závislú premennú nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : vplyv faktora na závislú premennú je štatisticky významný.

Na základe výstupu 3 a stĺpca $Pr > F$ nulovú hypotézu zamietame pre každý uvažovaný faktor a konštatujeme, že všetky premenné zahrnuté do modelu majú štatisticky významný vplyv na ekvivalentný disponibilný príjem členov domácností na všetkých bežne používaných hladinách významnosti. Veľkosť vplyvu jednotlivých premenných na ekvivalentný disponibilný príjem samozrejme nie je rovnaká, dá sa orientačne zistiť na základe hodnoty vyčíslenej testovacej štatistiky v stĺpci F Value. Platí, že čím vyššia je hodnota testovacej štatistiky, tým väčší vplyv má daná premenná na závislú premennú. Na základe výstupu 3 teda môžeme konštatovať, že najväčší vplyv na výšku ekvivalentného

disponibilného príjmu členov domácností má premenná *HT* (typ domácnosti) a vysoký vplyv majú taktiež premenné *Education* (vzdelanie) a *EA* (ekonomická aktivita). Najnižší vplyv na ekvivalentný disponibilný príjem má premenná *Marital_status* (rodinný stav).

Výstup 3 Test štatistickej významnosti vplyvu vybraných premenných na závislú premennú

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EA	5	6245890336	1249178067	133.33	<.0001
Education	4	5202457600	1300614400	138.82	<.0001
HT	8	11083204885	1385400611	147.87	<.0001
Marital_status	3	188051365	62683788	6.69	0.0002
Region	7	3301207750	471601107	50.34	<.0001

Zdroj: vlastné spracovanie v SAS EG

Kvantifikovať prínos jednotlivých nezávislých premenných na vysvetlenie variability závislej premennej je možné na základe výstupu 4. Jednotlivé semi-parciálne koeficienty determinácie vypočítame predelením príslušnej hodnoty zo stĺpca *Type I SS* celkovou variabilitou modelu (*SST*) z výstupu 1. Súčet semi-parciálnych koeficientov determinácie pre všetky premenné zahrnuté do modelu potom musí byť rovný koeficientu determinácie (*R-square*) z výstupu 2. Semi-parciálny koeficient determinácie pre premennú *EA* má hodnotu 0,1330, teda môžeme konštatovať, že na základe ekonomickej aktivity je možné vysvetliť 13,30 % variability ekvivalentného disponibilného príjmu domácností.

Výstup 4 Vyhodnotenie prínosu nezávislých premenných na vysvetlenie variability závislej premennej

Source	DF	Type I SS	Mean Square	F Value	Pr > F
EA	5	13639659657	2727931931	291.16	<.0001
Education	4	5500040056	1375010014	146.76	<.0001
HT	8	11407739657	1425967457	152.20	<.0001
Marital_status	3	65906939	21968980	2.34	0.0709
Region	7	3301207750	471601107	50.34	<.0001

Zdroj: vlastné spracovanie v SAS EG

Konkrétne parametre nami odhadnutého regresného modelu sú uvedené vo výstupe 5. Tento výstup je taktiež súčasťou základného výstupu procedúry GLM. Keďže sú všetky nezávislé premenné zaradené do modelu kategoriálne, z každej kategórie jednotlivých premenných (s výnimkou referenčných kategórií) bola vytvorená umelá premenná, reprezentujúca tú-ktorú obmenu premennej. Na overenie štatistickej významnosti jednotlivých regresných koeficientov modelu budeme testovať nulovú hypotézu

H_0 : regresný koeficient nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : regresný koeficient je štatisticky významný.

Rozhodnutie o prijatí, resp. zamietnutí nulovej hypotézy vykonáme na základe porovnania vyčíslenej testovacej štatistiky Studentovho rozdelenia s kritickou hodnotou, prípadne na základe p -hodnoty. Vzhľadom na to, že výstup opäť neobsahuje informáciu o kritických hodnotách, rozhodnutie prijmeme na základe p -hodnoty. Väčšina hodnôt stĺpca $Pr > |t|$ neprekračuje bežne používané hladiny významnosti a pre tieto premenné môžeme teda zamietnuť nulovú hypotézu a regresné koeficienty považovať za štatisticky významné. Napríklad v prípade premennej *Region PO* môžeme na hladine významnosti 0,05 predpokladať, že ekvivalentný disponibilný príjem členov domácností v Prešovskom kraji je v priemere o 2 335,46€ nižší, ako príjem členov domácností v Bratislavskom kraji, ceteris paribus. V prípade umelej premennej *HT 1Adult* reprezentujúcej členov domácností pozostávajúcich z jedného dospelého človeka vidíme, že p -hodnota prekračuje hodnotu 0,05, na základe čoho považujeme tento regresný koeficient za štatisticky nevýznamný. Z toho vyplýva, že na hladine významnosti 0,05 môžeme predpokladať, že ekvivalentný disponibilný príjem členov domácností tohto typu je v priemere rovnaký ako ekvivalentný disponibilný príjem členov domácností referenčného typu (*2A_2Ch*), za ostatných nezmenených podmienok.

Výstup 5 Odhad parametrov modelu EDP členov domácností

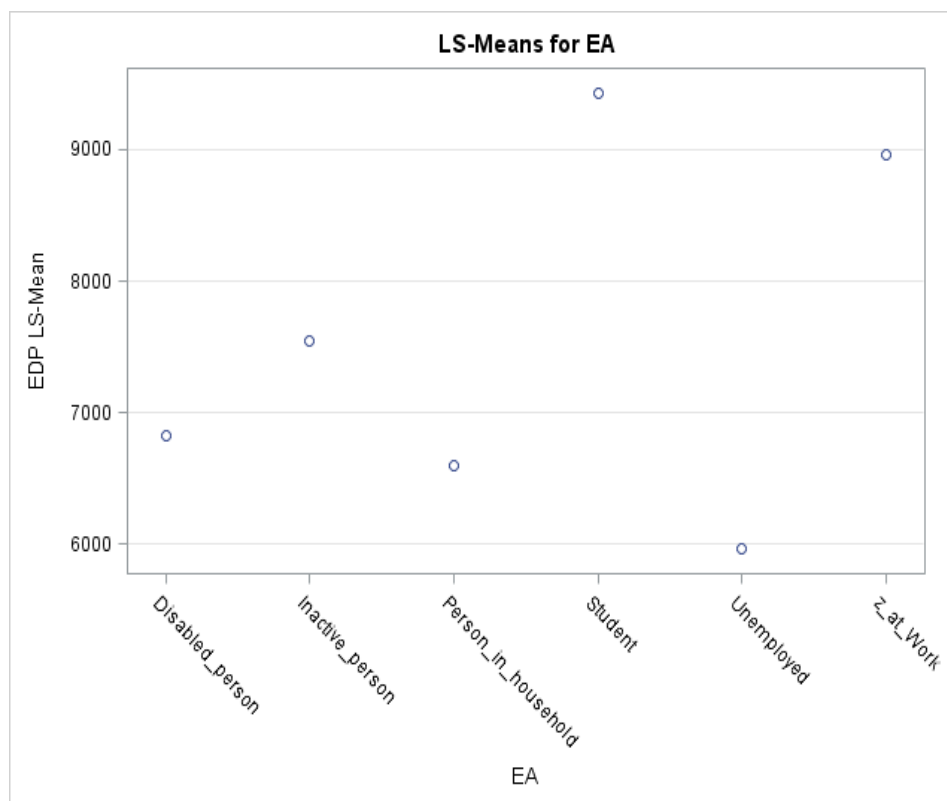
Parameter		Estimate		Standard Error	t Value	Pr > t
Intercept		10949.52074	B	144.6146362	75.72	<.0001
EA	Disabled_person	-2131.21781	B	155.8158762	-13.68	<.0001
EA	Inactive_person	-1416.51911	B	193.3174903	-7.33	<.0001
EA	Person_in_household	-2361.74873	B	368.7273343	-6.41	<.0001
EA	Student	473.65520	B	155.4683598	3.05	0.0023

	Parameter	Estimate		Standard Error	t Value	Pr > t
EA	Unemployed	-2995.60207	B	149.5176357	-20.04	<.0001
EA	z_at_Work	0.00000	B	.	.	.
Education	ISCED_0_1_2	-3296.47859	B	151.4246209	-21.77	<.0001
Education	ISCED_3	-1796.19411	B	95.6885963	-18.77	<.0001
Education	ISCED_4	-1534.57911	B	296.1346287	-5.18	<.0001
Education	ISCED_5A	-1183.60227	B	192.3229234	-6.15	<.0001
Education	z_ISCED_5B_6	0.00000	B	.	.	.
HT	1A_at_least_1Ch	-1314.15939	B	222.8329116	-5.90	<.0001
HT	1Adult	333.80547	B	186.7168043	1.79	0.0739
HT	2A_1Ch	565.81894	B	141.8160743	3.99	<.0001
HT	2A_1R	1863.28789	B	222.8034367	8.36	<.0001
HT	2A_at_least_3Ch	-1247.72791	B	197.5882556	-6.31	<.0001
HT	2Adult	2540.06143	B	132.6916068	19.14	<.0001
HT	Other_0Ch	3113.45715	B	136.5484851	22.80	<.0001
HT	Other_with_Ch	1303.52578	B	130.3274275	10.00	<.0001
HT	z_2A_2Ch	0.00000	B	.	.	.
Marital_status	Divorced	-336.90449	B	145.2410411	-2.32	0.0204
Marital_status	Never_married	-267.47103	B	90.7547800	-2.95	0.0032
Marital_status	Widowed	579.91105	B	253.5717577	2.29	0.0222
Marital_status	z_Married	0.00000	B	.	.	.
Region	BB	-1192.91352	B	136.0239270	-8.77	<.0001
Region	KE	-1752.01032	B	143.4104726	-12.22	<.0001
Region	NR	-877.85889	B	138.5634775	-6.34	<.0001
Region	PO	-2335.46412	B	133.0737095	-17.55	<.0001
Region	TN	-1113.41654	B	145.8932243	-7.63	<.0001
Region	TT	-977.46921	B	151.7771064	-6.44	<.0001
Region	ZA	-1510.90219	B	137.2902997	-11.01	<.0001
Region	z_BA	0.00000	B	.	.	.

Zdroj: vlastné spracovanie v SAS EG

Na zlepšenie kvality modelu ďalej otestujeme vhodnosť zlúčenia niektorých kategórií jednotlivých premenných do nových premenných, čím sa zníži počet regresných koeficientov a potenciálne zlepší kvalita modelu. Prejdeme teda k samotnej analýze

marginálnych stredných hodnôt. Súčasťou základného výstupu je tabuľka aj graf zobrazujúce bodové odhady marginálnych stredných hodnôt ekvivalentného disponibilného príjmu členov domácností v závislosti od jednotlivých kategoriálnych premenných.



Obrázok 4 Bodové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od statusu ekonomickej aktivity (EA)

Zdroj: vlastné spracovanie v SAS EG

Pre rozšírenie výstupu a zobrazenie intervalových odhadov ako aj *p*-hodnôt pre test štatistickej významnosti zhody marginálnych stredných hodnôt EDP členov domácností v závislosti od jednotlivých premenných je nutné upraviť kód procedúry GLM. Konkrétne je nutné upraviť príkaz LSMEANS a na jeho koniec za lomenu doplniť funkcie PDIF a CL. Funkciou PDIF vytvoríme tabuľku *p*-hodnôt pre test štatistickej významnosti zhody dvojíc marginálnych stredných hodnôt cieľovej premennej pre všetky kategórie a štatistickú významnosť párových diferencií pre všetky dvojice kategórií príslušného faktora. Funkciou CL pridáme intervalové odhady marginálnych stredných hodnôt. Hladina významnosti, resp. riziko odhadu môžu byť špecifikované pridaním možnosti ALPHA=option, no v našom prípade ponecháme automaticky prednastavenú hodnotu 0,05. Celý príkaz bude potom vyzeráť nasledovne:

LSMEANS EA Education HT Marital_status Region / PDIFF CL;

Výstup 6 a 7 a obrázok 5 nižšie zobrazujú tabuľky a graf pre premennú *EA*, ktoré pribudnú vo výstupe po uvedenej modifikácii kódu procedúry. Výstup 6 zobrazuje bodové a taktiež spomínané intervalové odhady marginálnych stredných hodnôt a taktiež poradia jednotlivých kategórií, ktoré budú použité ďalej pri teste štatistickej významnosti zhody marginálnych stredných hodnôt a taktiež vo výstupe 7.

Výstup 6 Bodové a intervalové odhady marginálnych stredných hodnôt
EDP členov domácností v závislosti od statusu ekonomickej aktivity (*EA*)

EA	EDP LSMEAN	95% Confidence Limits		LSMEAN Number
Disabled_person	6825.352683	6492.900397	7157.804970	1
Inactive_person	7540.051385	7133.711629	7946.391141	2
Person_in_household	6594.821763	5859.662976	7329.980551	3
Student	9430.225696	9112.724238	9747.727153	4
Unemployed	5960.968427	5636.129934	6285.806921	5
z_at_Work	8956.570496	8758.471665	9154.669326	6

Zdroj: vlastné spracovanie v SAS EG

Výstup 7 zobrazuje *p*-hodnoty testu štatistickej významnosti zhody marginálnych stredných hodnôt EDP členov domácností pre všetky dvojice kategórií premennej *EA*, pričom testujeme nulovú hypotézu

$$H_0: \mu_i = \mu_j$$

oproti alternatívnej hypotéze

$$H_1: \mu_i \neq \mu_j.$$

Opäť platí, že nulovú hypotézu je možné zamietnuť, ak *p*-hodnota nepresiahne zvolenú hladinu významnosti α . Z výstupu 6 vidíme, že *p*-hodnota prekračuje hladinu významnosti 0,05 v dvoch prípadoch a môžeme teda konštatovať, že na hladine významnosti 0,05 nie je štatisticky významný rozdiel v strednej hodnote ekvivalentného disponibilného príjmu osôb na invalidnom dôchodku (*Disabled_person*) a osôb v domácnosti (*Person_in_household*) a taktiež medzi osobami v domácnosti (*Person_in_household*) a nezamestnanými osobami (*Unemployed*).

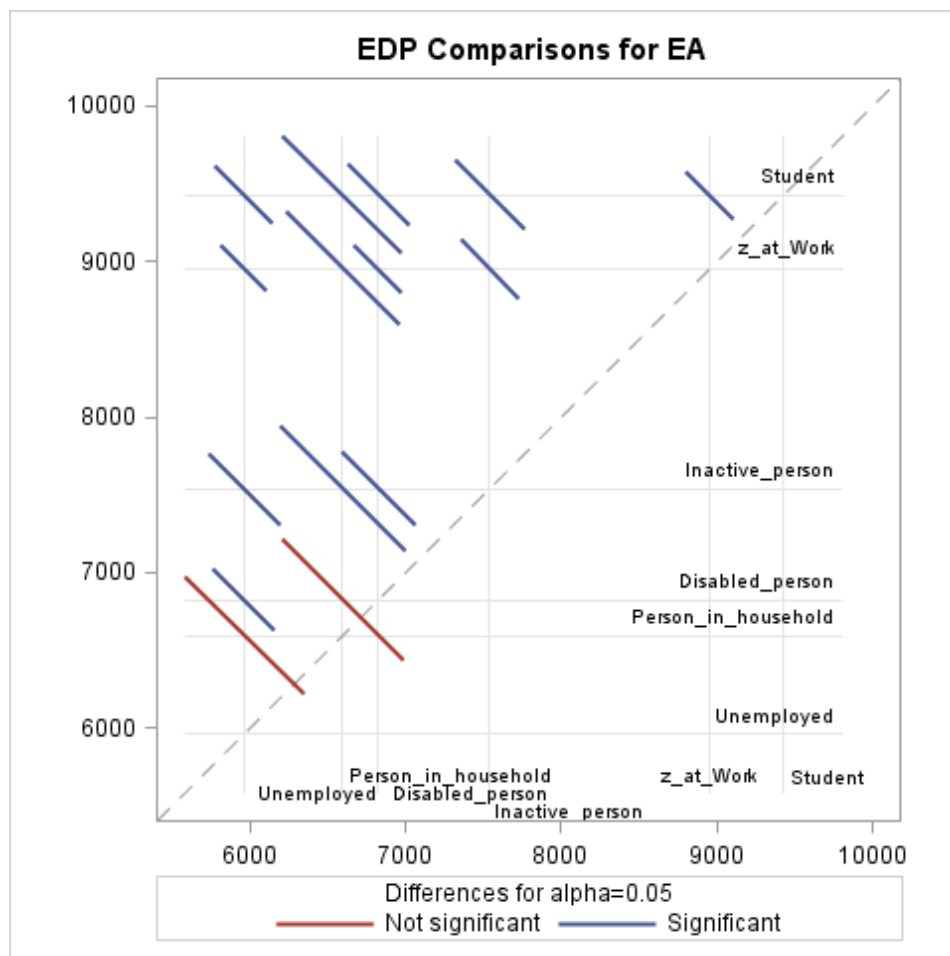
Výstup 7 Test štatistickej významnosti zhody marginálnych stredných hodnôt EDP členov domácností v závislosti od statusu ekonomickej aktivity (EA)

Least Squares Means for effect EA Pr > t for H0: LSMean(i)=LSMean(j)						
Dependent Variable: EDP						
i/j	1	2	3	4	5	6
1		0.0033	0.5586	<.0001	<.0001	<.0001
2	0.0033		0.0209	<.0001	<.0001	<.0001
3	0.5586	0.0209		<.0001	0.1035	<.0001
4	<.0001	<.0001	<.0001		<.0001	0.0023
5	<.0001	<.0001	0.1035	<.0001		<.0001
6	<.0001	<.0001	<.0001	0.0023	<.0001	

Zdroj: vlastné spracovanie v SAS EG

Vzhľadom na to, že medzi osobami na invalidnom dôchodku (*Disabled_person*) a nezamestnanými osobami (*Unemployed*) je štatisticky významný rozdiel v ich ekvivalentnom disponibilnom príjme, pristúpime k zlúčeniu dvojice kategórii *Disabled_person* a *Person_in_household* (nakoľko *p*-hodnota medzi týmito dvoma kategóriami je vyššia ako pri dvojici kategórii *Person_in_household* a *Unemployed*).

Obrázok 5 je grafickým zhrnutím výstupov 6 a 7 spomínaných vyššie. Horizontálne a vertikálne línie reprezentujú marginálne stredné hodnoty EDP členov domácností pre jednotlivé kategórie premennej *EA*. Na každej z týchto línii sú farebné úsečky reprezentujúce párové diferencie a ich významnosť. Diagonála je referenčnou liniou. V prípade, ak niektorá z úsečiek pretína referenčnú diagonálnu čiaru, potom sa marginálne stredné hodnoty vysvetľovanej premennej pre zodpovedajúce kategórie výrazne nelíšia (Wicklin, 2017). V tomto prípade môžeme vidieť dve červené úsečky naznačujúce štatisticky nevýznamný rozdiel v strednej hodnote EDP. Ako už bolo spomínané skôr, ide o kategórie *Disabled_person* a *Person_in_household* a taktiež *Person_in_household* a *Unemployed*.



Obrázok 5 Intervalové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od statusu ekonomickej aktivity (EA)

Zdroj: vlastné spracovanie v SAS EG

Rovnako, aj pri ostatných kategoriálnych premenných pristúpime k zlúčeniu ich kategórií a vytvoríme nové premenné, ktoré pôvodné premenné v modeli nahradia. Konkrétne pri premennej *Education* nie je na hladine významnosti 0,05 štatisticky významný rozdiel v ekvivalentnom disponibilnom príjme osôb so vzdelaním z kategórie *ISCED_3* a *ISCED_4* a taktiež medzi osobami so vzdelaním z kategórie *ISCED_4* a *ISCED_5A*. Nakoľko medzi osobami so vzdelaním z kategórie *ISCED_3* a *ISCED_5A* je štatisticky významný rozdiel v ich ekvivalentnom disponibilnom príjme, pristúpime k zlúčeniu dvojice kategórií *ISCED_3* a *ISCED_4* (vzhľadom na to, že *p*-hodnota medzi týmito dvoma kategóriami je vyššia ako pri dvojici kategórií *ISCED_4* a *ISCED_5A*). Pri premennej *HT* pristúpime k zlúčeniu kategórií *1A_at_least_1Ch* a *2A_at_least_3Ch* a taktiež kategórií *1Adult* a *2A_1Ch*. Čo sa týka premennej *Marital_status*, jediná dvojica kategórií, pre ktorú je možné na hladine významnosti 0,05 považovať ekvivalentný disponibilný príjem v priemere za rovnaký, sú osoby patriace do kategórie *Divorced* a

Never_married. Pristúpime teda k zlúčeniu týchto dvoch kategórii. Všetky výstupy, na základe ktorých sme dospeli k zlučovaniu konkrétnych kategórií sú vzhľadom na ich rozsah dostupné v prílohe (obrázok 1-6 a výstup 1-6).

Čo sa týka premennej *Region*, tam zlučovanie kategórií nie je také jednoznačné a jednoduché ako pri predchádzajúcich analyzovaných premenných. Na základe výstupov 8, 9 a obrázka 6 môžeme konštatovať, že ekvivalentný disponibilný príjem členov domácností je medzi viacerými krajinami SR podobný. Na hladine významnosti 0,05 nie je štatisticky významný rozdiel v ekvivalentnom disponibilnom príjme medzi členmi domácností Žilinského (ZA) a Košického kraja (KE). Štatistický nevýznamný rozdiel môžeme pozorovať taktiež medzi dvojicami týchto krajov: Nitriansky (NR), Trenčiansky (TN), Trnavský (TT) a prípadne Banskobystrický (BB). Opodstatnenosť zlučovania viacerých kategórií (umelých premenných) do jednej je však potrebné podložiť ďalšími testami. Popradský (PO) a Bratislavský kraj (BA), ktorý je zároveň aj referenčnou kategóriou tejto premennej ponecháme nezmenené, nakoľko sa ekvivalentný disponibilný príjem členov domácností v týchto krajinách signifikantne odlišuje od ostatných.

Výstup 8 Bodové a intervalové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od premennej *Region*

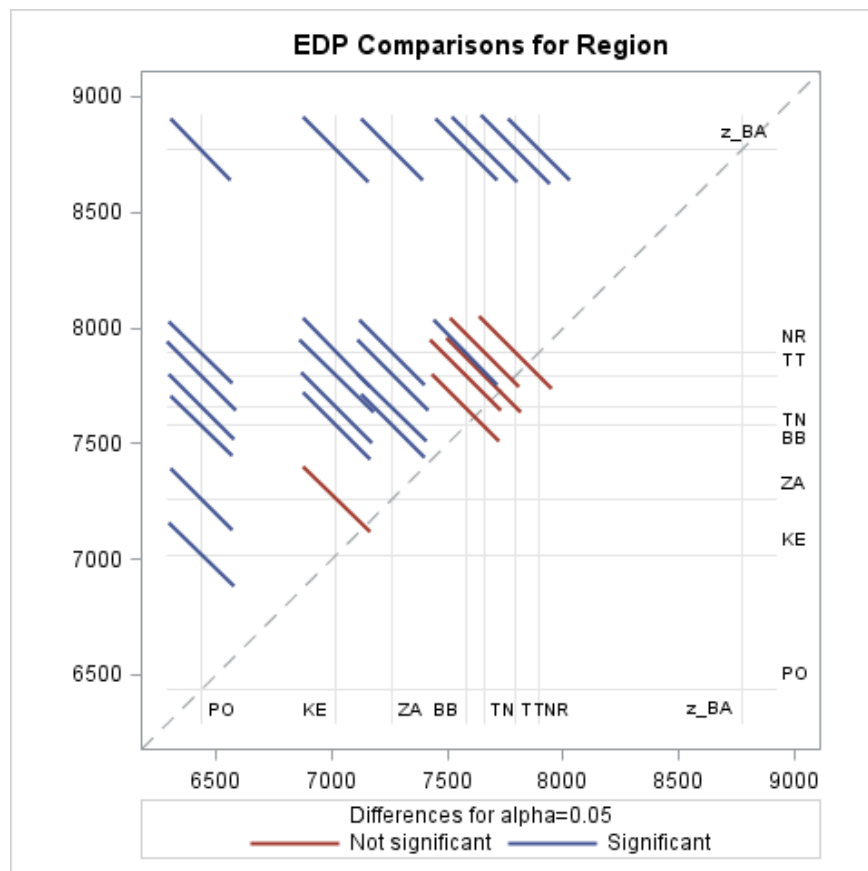
Region	EDP LSMEAN	95% Confidence Limits		LSMEAN Number
BB	7578.422566	7288.286892	7868.558240	1
KE	7019.325772	6718.713243	7319.938301	2
NR	7893.477201	7592.450502	8194.503900	3
PO	6435.871973	6153.062783	6718.681164	4
TN	7657.919547	7345.494376	7970.344717	5
TT	7793.866882	7472.704706	8115.029058	6
ZA	7260.433902	6962.703060	7558.164744	7
z_BA	8771.336091	8482.650967	9060.021214	8

Zdroj: vlastné spracovanie v SAS EG

Výstup 9 Test štatistickej významnosti zhody marginálnych stredných hodnôt EDP členov domácností v závislosti od premennej *Region*

Least Squares Means for effect Region Pr > t for H0: LSMean(i)=LSMean(j)								
Dependent Variable: EDP								
i/j	1	2	3	4	5	6	7	8
1		0.0001	0.0270	<.0001	0.5945	0.1652	0.0229	<.0001
2	0.0001		<.0001	<.0001	<.0001	<.0001	0.1017	<.0001
3	0.0270	<.0001		<.0001	0.1216	0.5285	<.0001	<.0001
4	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
5	0.5945	<.0001	0.1216	<.0001		0.4080	0.0079	<.0001
6	0.1652	<.0001	0.5285	<.0001	0.4080		0.0006	<.0001
7	0.0229	0.1017	<.0001	<.0001	0.0079	0.0006		<.0001
8	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	

Zdroj: vlastné spracovanie v SAS EG



Obrázok 6 Intervalové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od premennej *Region*

Zdroj: vlastné spracovanie v SAS EG

Ako sme uviedli v predchádzajúcej časti, rozličné kategórie nezávislých premenných, ktoré sú v modeli reprezentované umelými premennými je pre zjednodušenie modelu možné zlúčiť na základe testu zhody marginálnych stredných hodnôt. V prípade, ak ide o dvojice kategórií, záver je možné prijať na základe výstupu z príkazu `LSMEANS`. V prípade premennej *Region* sa však významnosť diferencií nepotvrdila pri viacerých kategóriách, ktoré by bolo vhodné spojiť do jednej umelej premennej. Overenie rovnosti viacerých stredných hodnôt budeme realizovať prostredníctvom príkazu `CONTRAST`, ktorého syntax je nasledovná:

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

kde `label` identifikuje názov konkrétneho porovnávania, z dôvodu čoho musí byť vždy uvedený v úvodzovkách, `effect` identifikuje premenné modelu, resp. kategórie premennej, pre ktoré testujeme zhodu stredných hodnôt závislej premennej. `Values` chápeme ako koeficienty kategórií premennej, ktoré vstupujú do analyzovaného faktora, pričom ich poradie závisí od poradia kategórii konkrétnej kategoriálnej premennej. Ak je špecifikovaných viac koeficientov ako je počet kategórií konkrétnej premennej, softvér nadbytočné koeficienty ignoruje. Naopak, ak je špecifikovaných menej koeficientov ako je počet kategórií premennej, zvyšným kategóriám je automaticky priradený koeficient 0. Súčet tých koeficientov, ktoré vstupujú do príkazu `CONTRAST` musí byť nulový, v opačnom prípade je príkaz zle zadáný a nezbehne. Ďalej nasledujú nepovinné časti príkazu, uvedené v ostrých zátvorkách. Príkazom je teda po zadaní viacerých premenných, resp. kategórií premennej a ich koeficientov možné testovať viacero nulových hypotéz súčasne. Za lomenom môžu byť v príkaze špecifikované ďalšie rôznorodé možnosti (SAS Institute Inc., 2015.)

Správnemu zostaveniu príkazu predchádza formulácia nulovej hypotézy a jej následná úprava na požadovaný tvar. V našom prípade budeme testovať nulovú hypotézu

$$H_0: \mu_{NR} = \mu_{TN} = \mu_{TT} = \mu_{BB}$$

oproti alternatívnej hypotéze

$$H_1: \text{aspoň jedna rovnosť neplatí,}$$

pričom nulovú hypotézu je nutné rozdeliť na 3 parciálne hypotézy, ktoré budú testované simultánne. Na overenie vyššie uvedenej nulovej hypotézy môžeme využiť simultánny test napr. týchto troch parciálnych nulových hypotéz:

$$H_0: \mu_{NR} = \mu_{TN} \wedge H_0: \mu_{TT} = \frac{1}{2}(\mu_{NR} + \mu_{TN}) \wedge H_0: \mu_{BB} = \frac{1}{3}(\mu_{NR} + \mu_{TN} + \mu_{TT}).$$

Po úprave hypotéz do tvarov lineárnych kombinácií je ich finálna forma nasledovná:

$$H_0: \mu_{NR} - \mu_{TN} = 0$$

$$H_0: \mu_{TT} - 0,5 \cdot \mu_{NR} - 0,5 \cdot \mu_{TN} = 0$$

$$H_0: 3 \cdot \mu_{BB} - \mu_{NR} - \mu_{TN} - \mu_{TT} = 0$$

Konkrétny zápis príkazu pre simultánný test troch vyššie uvedených nulových hypotéz musí reflektovať poradie kategórií premennej *Region* a v našom prípade použijeme syntax:

```
CONTRAST "BB=TT=NR=TN" REGION 0 0 1 0 -1, REGION 0 0 -0.5 0
-0.5 1, REGION 3 0 -1 0 -1 -1;
```

Výstup 10, ktorý dostaneme po úspešnom spustení a zbehnutí príkazu, obsahuje v stĺpci DF počet stupňov voľnosti. Tie závisia od počtu simultánne testovaných nulových hypotéz konkrétnym príkazom. Vzhľadom na to, že v našom prípade boli súčasne testované tri nulové hypotézy, počet stupňov voľnosti je taktiež rovný trom. Ako zvyčajne, výstup obsahuje v stĺpci F Value hodnotu testovacej štatistiky (viď vzťah (9) v 3. kapitole) a v stĺpci Pr > F *p*-hodnotu daného testu. Vzhľadom na to, že *p*-hodnota (*p* = 0,1338) prekračuje hladinu významnosti 0,05, nemáme dostatok dôkazov na zamietnutie nulovej hypotézy a môžeme teda konštatovať, že na hladine významnosti 0,05 nie je štatisticky významný rozdiel v strednej hodnote ekvivalentného disponibilného príjmu v Banskobystrickom, Trnavskom, Nitrianskom a Trenčianskom kraji. Tieto štyri kategórie premennej *Region* teda pre zjednodušenie modelu zlúčime do novej premennej, ktorá pôvodné premenné v modeli nahradí. Ako bolo už spomínané, k zlúčeniu príde taktiež v prípade Košického a Žilinského kraja a premenná *Region* bude po úprave obsahovať 4 obmeny, a to konkrétne *PO*, *KE_ZA*, *BB_NR_TN_TT* a *z_BA*, ktorá je zároveň referenčnou kategóriou.

Výstup 10 Výstup príkazu CONTRAST

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
BB=TT=NR=TN	3	52313515.04	17437838.35	1.86	0.1338

Zdroj: vlastné spracovanie v SAS EG

Pred samotným zaradením novovytvorených premenných do modelu zabezpečíme pomocou príkazu *ESTIMATE* konkrétne bodové odhady priemerného ekvivalentného disponibilného príjmu členov domácností charakterizovaných týmito novými premennými. Syntax príkazu *ESTIMATE* je veľmi podobná príkazu *CONTRAST* a vyzerá nasledovne:

```
ESTIMATE 'label' effect values <...effect values> </ options>;
```

Jedným z rozdielov príkazov *CONTRAST* a *ESTIMATE* je, že príkazom *ESTIMATE* je možné testovať iba jednu hypotézu, zatiaľ čo v prípade príkazu *CONTRAST* bolo možné

testovať viacero nulových hypotéz súčasne. Ďalším rozdielom je, že v príkaze ESTIMATE je nutné uviesť aj konštantu (*intercept*), ktorá predstavuje súčet koeficientov pri kategóriách analyzovaného faktora (v príkaze CONTRAST bol súčet koeficientov nulový a túto konštantu teda nebolo nutné uvádzať). Za lomenom taktiež využijeme funkciu DIVISOR=number, ktorá špecifikuje číslo, akým sú delené všetky koeficienty daného príkazu. Počet CONTRAST a ESTIMATE príkazov nie je limitovaný, avšak všetky tieto príkazy musia nasledovať až za príkazom MODEL v procedúre GLM (SAS Institute Inc., 2015)

V našom prípade budeme vytvárať päť nových premenných, teda aj počet ESTIMATE príkazov na odhad priemerného EDP členov domácností bude päť. Konkrétny zápis príkazov bude potom vyzeráť nasledovne:

```
ESTIMATE 'EA Disabled_person_Person_in_household' intercept 2
EA 1 0 1 / divisor=2;
```

```
ESTIMATE 'HT 1A_at_least_1Ch_2A_at_least_3Ch' intercept 2 HT
1 0 0 0 1 / divisor=2;
```

```
ESTIMATE 'HT 1Adult_2A_1Ch' intercept 2 HT 0 1 1 / divisor=2;
```

```
ESTIMATE 'Education ISCED_3_ISCED_4' intercept 2 Education 0
1 1 / divisor=2;
```

```
ESTIMATE 'Marital_status Divorced_Never_married' intercept 2
Marital_status 1 1 / divisor=2;
```

```
ESTIMATE 'Region ZA_KE' intercept 2 Region 0 1 0 0 0 0 1 /
divisor=2;
```

```
ESTIMATE 'Region BB_NR_TN_TT' intercept 4 Region 1 0 1 0 1 1
0 / divisor=4;
```

Ako už bolo spomínané vyššie, príkazy sme formulovali do tvaru, ktorý vyžaduje aj využitie funkcie DIVISOR=number, pričom hodnota tohto deliteľa závisí od počtu nenulových koeficientov v tom-ktorom príkaze. V prípade, že by sme funkciu DIVISOR=number do príkazu nechceli zakomponovať, bolo by nutné upraviť hodnoty koeficientov tak, aby mali rovnakú hodnotu a v súčte dávali hodnotu 1. Prvý príkaz by teda alternatívne mohol vyzeráť nasledovne:

```
ESTIMATE 'EA Disabled_person_Person_in_household' intercept 1
EA 0.5 0 0.5;
```

Výstup príkazu obsahuje v stĺpci *Estimate* bodové odhady priemerného ekvivalentného disponibilného príjmu členov domácností charakterizovaných novovytvorenými premennými. V stĺpci *Standard Error* nájdeme hodnoty štandardných chýb odhadu, na základe ktorých je ďalej možné konštruovať intervalové odhady, či

overovať ďalšie hypotézy. Vo výstupe sa samozrejme nachádza aj testovacia štatistika a príslušná p -hodnota testu.

Z výstupu 11 zistujeme, že všetky p -hodnoty nadobúdajú rovnakú hodnotu ($p < 0,001$), ktorá neprekračuje bežne používané hladiny významnosti. Vo všetkých 7 skupinách osôb uvedených vo výstupe 10 je na hladine významnosti 0,05 stredná hodnota EDP štatisticky významne odlišná od 0. V skupine zdravotne postihnutých osôb a osôb v domácnosti (*Disabled_person_Person_in_household*) odhadujeme ekvivalentný disponibilný príjem v priemere 6 710,09 € ročne.

Výstup 11 Výstup príkazu ESTIMATE

Parameter	Estimate	Standard Error	t Value	Pr > t
EA Disabled_person_Person_in_household	6710.08722	214.161801	31.33	<.0001
HT 1A_at_least_1Ch_2A_at_least_3Ch	5475.04705	164.351753	33.31	<.0001
HT 1Adult_2A_1Ch	7205.80291	137.091509	52.56	<.0001
Education ISCED_3_4	7448.11595	172.880498	43.08	<.0001
Marital_status Divorced_Never_married	7255.26010	117.669561	61.66	<.0001
Region ZA_KE	7139.87984	133.669583	53.41	<.0001
Region BB_NR_TN_TT	7730.92155	124.780198	61.96	<.0001

Zdroj: vlastné spracovanie v SAS EG

4.2 Všeobecný lineárny model EDP členov domácností s upravenými premennými

Na základe výsledkov testov štatistickej významnosti zhody marginálnych stredných hodnôt ekvivalentného disponibilného príjmu členov domácností v závislosti od rôznych kategoriálnych premenných sme v predchádzajúcej časti vyhodnotili, že by bolo vhodné kategórie niektorých premenných zlúčiť a vytvoriť nový model. Konkrétne sme zlučovali kategórie premenných *EA*, *HT*, *Education*, *Marital_status* a *Region*, pričom referenčné kategórie všetkých premenných ostali nezmenené.

Štatistickú významnosť modelu overíme na základe analýzy rozptylu závislej premennej, testovaním nulovej hypotézy

H_0 : model nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : model je štatisticky významný.

Na základe výstupu 12 a p -hodnoty v poslednom stĺpci tabuľky môžeme zamietnuť nulovú hypotézu a model považovať za štatisticky významný. Hodnota prvého stĺpca tabuľky vo výstupe 13 nám poskytuje informáciu o koeficiente determinácie, na základe ktorého môžeme tvrdiť, že model vysvetľuje 33,04% variability ekvivalentného disponibilného príjmu členov domácností.

Výstup 12 Tabuľka analýzy rozptylu pre model EDP členov domácností

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	33809356544	1878297585.8	200.42	<.0001
Error	7337	68761861323	9371931.4873		
Corrected Total	7355	102571217866			

Zdroj: vlastné spracovanie v SAS EG

Výstup 13 Vybrané charakteristiky modelu EDP členov domácností

R-Square	Coeff Var	Root MSE	EDP Mean
0.329618	34.06611	3061.361	8986.531

Zdroj: vlastné spracovanie v SAS EG

Ďalším krokom je test štatistickej významnosti vplyvu jednotlivých premenných na závislú premennú kde budeme testovať nulovú hypotézu

H_0 : vplyv faktora na závislú premennú nie je štatisticky významný
oproti alternatívnej hypotéze

H_1 : vplyv faktora na závislú premennú je štatisticky významný.

Na základe výstupu 14 a p -hodnôt testu už môžeme vidieť, že pre všetky premenné je možné zamietnuť nulovú hypotézu a teda všetky premenné zahrnuté do modelu majú na závislú premennú štatisticky významný vplyv. Poradie premenných z hľadiska veľkosti ich vplyvu na závislú premennú sa nezmenilo, stále platí, že najväčší vplyv na výšku ekvivalentného disponibilného príjmu členov domácností má premenná *HT* a teda typ domácnosti.

Výstup 14 Test štatistickej významnosti vplyvu vybraných premenných na závislú premennú

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EA	4	6405193051	1601298263	170.86	<.0001
Education	3	5216421848	1738807283	185.53	<.0001
HT	6	11625988879	1937664813	206.75	<.0001
Marital_status	2	223312518	111656259	11.91	<.0001
Region	3	3239179754	1079726585	115.21	<.0001

Zdroj: vlastné spracovanie v SAS EG

Štatistickú významnosť jednotlivých regresných koeficientov modelu overujeme testovaním nulovej hypotézy

H_0 : regresný koeficient nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : regresný koeficient je štatisticky významný.

Na základe výstupu 15 a p -hodnôt v poslednom stĺpci tabuľky môžeme konštatovať, že pri všetkých umelých premenných je možné na hladine významnosti 0,05 zamietnuť nulovú hypotézu. V porovnaní s prvým modelom teda prišlo k zlepšeniu, nakoľko v tomto prípade sú všetky regresné koeficienty na hladine významnosti 0,05 štatisticky významné. Vo všetkých prípadoch je teda štatisticky významný rozdiel medzi referenčnou kategóriou a ostatnými kategóriami daných premenných.

Ak sa pozrieme na odhadnuté hodnoty regresných koeficientov, pri premennej *EA* charakterizujúcej status ekonomickej aktivity členov domácnosti vidíme, že najvyšší ročný ekvivalentný disponibilný príjem dosahujú študenti. Konkrétne je príjem týchto členov v priemere o 543,78 € vyšší ako príjem, ktorý dosahuje referenčná kategória tejto premennej, teda zamestnané osoby. Ekvivalentný disponibilný príjem nezamestnaných osôb je naopak najnižší spomedzi všetkých kategórií tejto premennej. V priemere je tento ročný príjem o 3 009,56 € nižší ako pri zamestnaných osobách.

Čo sa týka premennej *Education* charakterizujúcej najvyššie dosiahnuté vzdelanie členov domácností, môžeme vidieť, že so zvyšujúcim sa stupňom vzdelania členov domácností rastie aj ich ekvivalentný disponibilný príjem. Ročný ekvivalentný disponibilný príjem členov domácností, ktorý majú vzdelanie patriace do kategórie *ISCED_0_1_2* je v priemere o 3 302,50 € nižší v porovnaní s referenčnou kategóriou, teda členmi domácností, ktorí majú vzdelanie patriace do kategórie *ISCED_5B_6*.

Ak sa pozrieme na premennú *HT*, teda typ domácnosti, môžeme si všimnúť, že pri všetkých typoch domácností s výnimkou jednej, je ekvivalentný disponibilný príjem ich členov vyšší ako dosahuje referenčná kategória domácností, teda domácnosť pozostávajúca z dvoch dospelých a dvoch detí (*z_2A_2Ch*). Spomínanou výnimkou sú domácnosti v zložení jeden dospelý a aspoň jedno dieťa alebo dvaja dospelí a aspoň tri deti (*1A_at_least_1Ch_2A_at_least_3Ch*), ktorých ročný ekvivalentný disponibilný príjem je oproti referenčnej kategórii v priemere nižší o 1 286,19 €. To, že členovia týchto domácností dosahujú najnižší ekvivalentný disponibilný príjem logicky potvrdzuje aj metodika výpočtu EDP, keďže sa príjem zárobkovo činných členov rozpočítava na viacerých členov domácnosti. Naopak, najvyšší príjem môžeme pozorovať u členov domácností patriacich do kategórie *Other_0Ch*, čiže domácnosti bez závislých detí, ktorých ročný ekvivalentný disponibilný príjem je oproti referenčnej kategórii v priemere vyšší o 3 154,04 €.

Premenná *Marital_status* obsahuje po zlúčení kategórií iba 3 kategórie, pričom referenčnou kategóriou sú ľudia v manželskom zväzku (*z_Married*). Ročný ekvivalentný disponibilný príjem rozvedených alebo slobodných ľudí bol v roku 2021 oproti referenčnej kategórii v priemere nižší o 328,42€. Naopak, ovdovení ľudia dosahovali v porovnaní s referenčnou kategóriou vyšší ročný ekvivalentný disponibilný príjem a to v priemere o 522,79 €.

Poslednou kategoriálnou premennou zahrnutou do modelu je premenná *Region*. Referenčnou kategóriou je v tomto prípade Bratislavský kraj, v ktorom členovia domácností dosahovali najvyšší ročný ekvivalentný disponibilný príjem spomedzi všetkých krajov SR. Naopak, ročný ekvivalentný disponibilný príjem členov domácností Prešovského kraja je najnižší, pričom v porovnaní s referenčnou kategóriou bol tento príjem v roku 2021 nižší o 2 326,44 €.

Lokujúca konštanta (*intercept*) je reprezentantom členov domácností spadajúcich do všetkých referenčných kategórií súčasne. Môžeme teda tvrdiť, že priemerný ročný ekvivalentný disponibilný príjem zamestnanej osoby s vysokoškolským vzdelaním 2. alebo 3. stupňa, osoby patriacej do domácnosti Bratislavského kraja, ktorú tvoria dvaja dospelí a dve závislé deti, bol v roku 2021 na úrovni 10 935,53 €. Taktiež je nutné podotknúť, že všetky interpretácie regresných koeficientov modelu uvedené vyššie platia iba v prípade dodržania podmienky *ceteris paribus*.

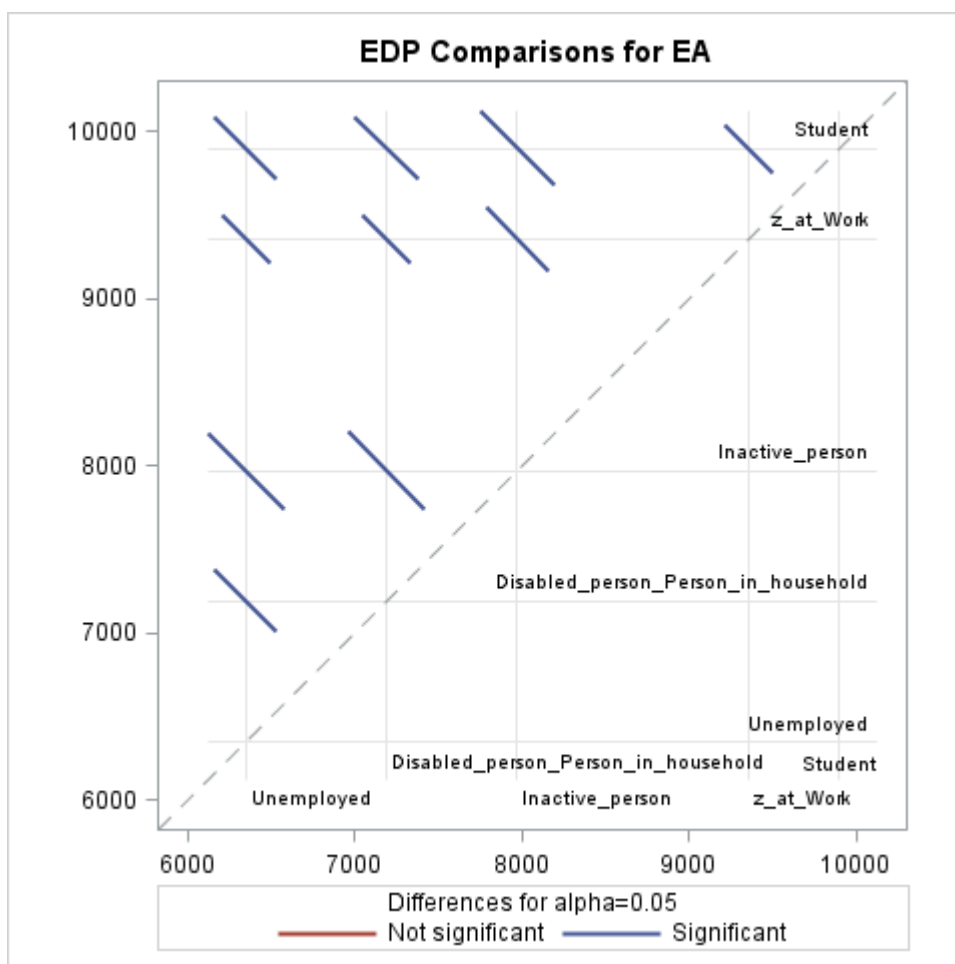
Výstup 15 Odhad parametrov modelu EDP členov domácností

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	10935.52679	144.2276025	75.82	<.0001
EA Disabled_person_Person_in_household	-2165.59545	145.4985683	-14.88	<.0001
EA Inactive_person	-1384.56034	192.0057840	-7.21	<.0001
EA Student	543.77925	148.6076560	3.66	0.0003
EA Unemployed	-3009.55531	148.7212894	-20.24	<.0001
EA z_at_Work	0.00000	.	.	.
Education ISCED_0_1_2	-3302.49578	150.8539976	-21.89	<.0001
Education ISCED_3_4	-1778.23720	95.2363459	-18.67	<.0001
Education ISCED_5A	-1176.79690	192.2499068	-6.12	<.0001
Education z_ISCED_5B_6	0.00000	.	.	.
HT 1A_at_least_1Ch_2A_at_least_3Ch	-1286.18576	161.8353803	-7.95	<.0001
HT 1Adult_2A_1Ch	495.90342	130.9911594	3.79	0.0002
HT 2A_1R	1901.71252	221.6369519	8.58	<.0001
HT 2Adult	2557.65766	132.1662456	19.35	<.0001
HT Other_0Ch	3154.04109	135.0564865	23.35	<.0001
HT Other_with_Ch	1311.58865	129.9933875	10.09	<.0001
HT z_2A_2Ch	0.00000	.	.	.
Marital_status Divorced_Never_married	-328.42494	81.5242583	-4.03	<.0001
Marital_status Widowed	522.79436	247.7012161	2.11	0.0348
Marital_status z_Married	0.00000	.	.	.
Region BB_NR_TN_TT	-1033.19762	107.0295597	-9.65	<.0001
Region KE_ZA	-1609.25832	118.7961157	-13.55	<.0001
Region PO	-2326.44360	132.5038574	-17.56	<.0001
Region z_BA	0.00000	.	.	.

Zdroj: vlastné spracovanie v SAS EG

Pri všetkých premenných zaradených do modelu prišlo na základe analýz v predchádzajúcej časti k zlúčeniu kategórií, ktoré si boli najviac podobné z hľadiska

ekvivalentného disponibilného príjmu. Z obrázka 7 a taktiež obrázkov 7-10, ktoré sú pre ich rozsah dostupné v prílohe vidíme, že v novovytvorenom modeli vyššie spomínaný problém nepretrváva a teda medzi kategóriami jednotlivých premenných je štatisticky významný rozdiel v ekvivalentnom disponibilnom príjme členov domácností.



Obrázok 7 Intervalové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od statusu ekonomickej aktivity (EA)

Zdroj: vlastné spracovanie v SAS EG

4.3 Všeobecný lineárny model EDP členov domácností s interakciou

V tejto časti práce budeme analyzovať model EDP členov domácností s interakciou, teda budeme skúmať vzájomný vplyv premenných. Konkrétne sme sa do modelu rozhodli zaradiť tri interakcie s premennou *Region*. Premenné sme vyberali na základe veľkosti ich vplyvu na výšku ekvivalentného disponibilného príjmu a teda zvolili sme premenné *HT*, *EA* a *Education*. Štatistickú významnosť tohto modelu sme opäť overovali pomocou analýzy rozptylu závislej premennej testovaním nulovej hypotézy

H_0 : model nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : model je štatisticky významný.

Vzhľadom na výšku p -hodnoty v poslednom stĺpci tabuľky vo výstupe 16 môžeme na všetkých bežne používaných hladinách významnosti zamietnuť nulovú hypotézu a model považovať za štatisticky významný. Na základe hodnoty koeficienta determinácie z výstupu 17 môžeme tvrdiť, že model vysvetľuje 33,83 % variability ekvivalentného disponibilného príjmu členov domácností a došlo k nárastu vysvetlenej variability o 0,87 p. b. oproti modelu, ktorý sme analyzovali v časti 1.2.

Výstup 16 Tabuľka analýzy rozptylu pre model EDP s interakciou

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	57	34699002885	608754436.57	65.46	<.0001
Error	7298	67872214982	9300111.672		
Corrected Total	7355	102571217866			

Zdroj: vlastné spracovanie v SAS EG

Výstup 17 Vybrané charakteristiky modelu EDP
členov domácností

R-Square	Coeff Var	Root MSE	EDP Mean
0.338292	33.93533	3049.608	8986.531

Zdroj: vlastné spracovanie v SAS EG

Štatistickú významnosť vplyvu jednotlivých premenných na závislú premennú overíme testovaním nulovej hypotézy

H_0 : vplyv faktora, resp. interakcie na závislú premennú nie je štatisticky významný

oproti alternatívnej hypotéze

H_1 : vplyv faktora, resp. interakcie na závislú premennú je štatisticky významný.

Na základe výstupu 18 a p -hodnôt v poslednom stĺpci tabuľky môžeme pre všetky premenné zamietnuť nulovú hypotézu a tvrdiť, že všetky premenné zahrnuté do modelu majú na závislú premennú štatisticky významný vplyv. Veľkosť vplyvu jednotlivých premenných sa s pridaním interakcie do modelu výrazne nezmenila, najvyšší vplyv na výšku ekvivalentného

disponibilného príjmu členov domácností má aj v tomto modeli premenná reprezentujúca typ domácnosti (*HT*).

Čo sa týka interakcií premennej *Region* s ostatnými premennými, vzhľadom na rozsah výstupov a veľké množstvo novovzniknutých umelých premenných, analyzovať a porovnávať marginálne stredné hodnoty ekvivalentného disponibilného príjmu členov domácností budeme iba pre nami zvolené kategórie jednotlivých premenných.

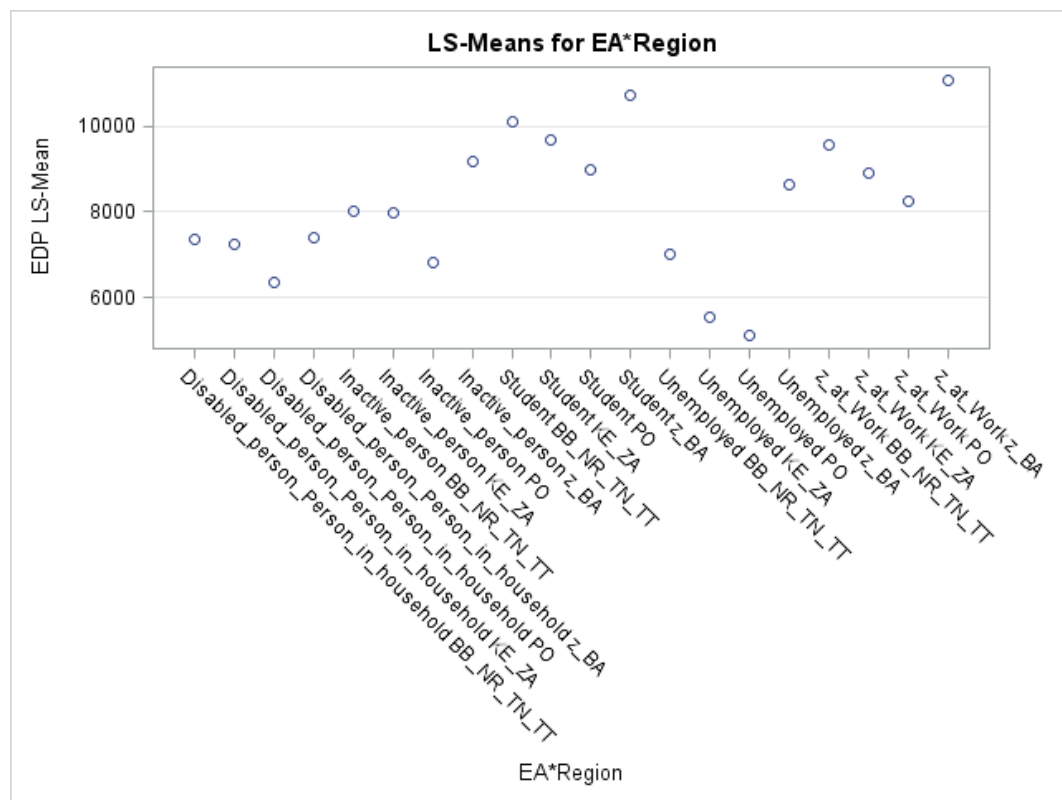
Výstup 18 Test štatistickej významnosti vplyvu vybraných premenných na závislú premennú

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EA	4	3185590445	796397611	85.63	<.0001
Education	3	3369663922	1123221307	120.78	<.0001
HT	6	9522406513	1587067752	170.65	<.0001
Marital_status	2	220876962	110438481	11.87	<.0001
Region	3	580671178	193557059	20.81	<.0001
EA*Region	12	250372633	20864386	2.24	0.0080
Education*Region	9	309390694	34376744	3.70	0.0001
HT*Region	18	353251698	19625094	2.11	0.0040

Zdroj: vlastné spracovanie v SAS EG

Na základe obrázka 8, zobrazujúceho bodové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od interakcie premenných *EA* a *Region*, môžeme predpokladať, že existuje viacero dvojíc, pre ktoré bude štatisticky nevýznamný rozdiel vo výške ich ekvivalentného disponibilného príjmu. Na základe obrázka 8 môžeme taktiež konštatovať, že najnižší ekvivalentný disponibilný príjem dosahovali v roku 2021 nezamestnaní členovia domácností Prešovského kraja. Ich priemerný ročný ekvivalentný disponibilný príjem bol na úrovni 5 130,75 €. Naopak, najlepšie na tom boli zamestnané osoby žijúce v Bratislavskom kraji. Ich priemerný ekvivalentný disponibilný príjem bol v roku 2021 na úrovni 11 074,87 € ročne. Tieto výsledky sa zhodujú s výsledkami z predchádzajúcich kapitol, kde sme ukázali, že pri premennej *Region* dosahovali najnižší ekvivalentný disponibilný príjem práve osoby z Prešovského kraja a naopak, najvyšší príjem dosahovali osoby z Bratislavského kraja. Rovnako to platí aj pri premennej *EA*, kde najnižší

príjem dosahovali v roku 2021 nezamestnané osoby (*Unemployed*) a naopak, najvyšší príjem mali zamestnané osoby (*z_at_Work*).



Obrázok 8 Bodové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od *EA* a *Region*

Zdroj: vlastné spracovanie v SAS EG

Výstup 19 Bodové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od interakcie premenných *EA* a *Region* pre zamestnané osoby

EA	Region	EDP LSMEAN	LSMEAN Number
z_at_Work	BB_NR_TN_TT	9568.2400	17
z_at_Work	KE_ZA	8899.5431	18
z_at_Work	PO	8231.0276	19
z_at_Work	z_BA	11074.8672	20

Zdroj: vlastné spracovanie v SAS EG

Pre detailnejšiu analýzu sme vybrali referenčnú a zároveň najpočetnejšiu kategóriu premennej *EA*, teda zamestnané osoby v interakcii s premennou *Region*. Na základe výstupu 19 vyššie môžeme konštatovať, že ako už bolo spomenuté, najvyšší ročný ekvivalentný disponibilný príjem dosahujú členovia domácností v Bratislavskom kraji. Naopak, najhoršie

sú na tom členovia domácností Prešovského kraja, ktorých priemerný ročný ekvivalentný disponibilný príjem bol v roku 2021 8 231,03 €.

Významnosť zhody marginálnych stredných hodnôt EDP členov domácností v závislosti od interakcie premenných *EA* a *Region* posúdime na základe výstupu 19, ktorý je čiastkovým výstupom zobrazujúcim iba potrebné *p*-hodnoty testu, pričom testujeme nulovú hypotézu

$$H_0: \mu_i = \mu_j$$

oproti alternatívnej hypotéze

$$H_1: \mu_i \neq \mu_j.$$

Opäť platí, že nulovú hypotézu je možné zamietnuť, ak *p*-hodnota nepresiahne zvolenú hladinu významnosti α . Z výstupu 20 vidíme, že žiadna z *p*-hodnôt neprekračuje hladinu významnosti 0,05. Môžeme teda konštatovať, že aj samostatne u zamestnaných osôb má kraj, z ktorého členovia domácností pochádzajú štatisticky významný vplyv na výšku ich ekvivalentného disponibilného príjmu.

Výstup 20 Test štatistickej významnosti zhody marginálnych stredných hodnôt EDP členov domácností v závislosti od interakcie premenných *EA* a *Region* pre zamestnané osoby

Least Squares Means for effect EA*Region Pr > t for H0: LSMean(i)=LSMean(j)				
Dependent Variable: EDP				
i/j	17	18	19	20
17		0.0002	<.0001	<.0001
18	0.0002		0.0043	<.0001
19	<.0001	0.0043		<.0001
20	<.0001	<.0001	<.0001	

Zdroj: vlastné spracovanie v SAS EG

(pozn. skupiny 17 až 20 sú definované vo výstupe 19)

Pri interakcii premenných *Education* a *Region* sme pre analýzu zhody marginálnych stredných hodnôt vybrali najpočetnejšiu kategóriu premennej *Education* – *ISCED_3_4* (vyššie sekundárne alebo post-sekndárne, neterciálne vzdelanie) a taktiež referenčnú kategóriu premennej, ktorou je *z_ISCED_5B_6* (2. a 3. stupeň terciárneho vzdelania).

Výstup 21 poskytuje konkrétne bodové odhady marginálnych stredných hodnôt EDP členov domácností, pričom vzhľadom na rozsah pôvodného výstupu uvádzame len časť relevantnú pre túto analýzu. Pri oboch vybraných stupňoch dosiahnutého vzdelania členov domácností je poradie výšky ich ekvivalentného disponibilného príjmu v závislosti od krajov SR rovnaké. Opäť platí, že najvyšší príjem dosahujú členovia domácností Bratislavského kraja, druhé miesto obsadzujú domácnosti Banskobystrického, Nitrianskeho, Trenčianskeho a Trnavského kraja. Ďalej nasledujú domácnosti Košického a Žilinského kraja. Najnižší príjem bol opäť dosahovaný členmi domácností Prešovského kraja. Nakoľko je ekvivalentný disponibilný príjem členov domácností pri tejto interakcii podobný a rozdiely príjmov medzi členmi domácností jednotlivých krajov nepôsobia markantne, ďalším krokom bude test štatistickej významnosti týchto diferencií a prípadne následné zlúčenie podobných kategórií.

Výstup 21 Bodové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od interakcie premenných *Education* a *Region* pre osoby so vzdelaním ISCED 3-4 a ISCED 5B-6

Education	Region	EDP LSMEAN	LSMEAN Number
ISCED_3_4	BB_NR_TN_TT	8146.2748	5
ISCED_3_4	KE_ZA	7596.2574	6
ISCED_3_4	PO	7182.9114	7
ISCED_3_4	z_BA	8880.0995	8
z_ISCED_5B_6	BB_NR_TN_TT	9947.8801	13
z_ISCED_5B_6	KE_ZA	9884.0303	14
z_ISCED_5B_6	PO	8646.8111	15
z_ISCED_5B_6	z_BA	10171.9604	16

Zdroj: vlastné spracovanie v SAS EG

Vo výstupe 22 vidíme, že p -hodnota testu v jednom prípade mierne prekračuje zvolenú hodnotu 0,05 a teda na hladine významnosti 0,05 môžeme konštatovať, že medzi osobami, ktoré dosiahli vzdelanie kategórie ISCED_3_4 a sú z Košického, Žilinského (KE_ZA) a Prešovského kraja (PO) nie je štatisticky významný rozdiel v ich ekvivalentnom disponibilnom príjme. Tieto kategórie budú teda pre zjednodušenie modelu následne zlúčené do jednej.

Výstup 22 Test štatistickej významnosti zhody
marginálnych stredných hodnôt EDP členov domácností
v závislosti od interakcie premenných *Education* a *Region*
pre osoby so vzdelaním ISCED 3-4

Least Squares Means for effect Education*Region Pr > t for H0: LSMean(i)=LSMean(j)				
Dependent Variable: EDP				
i/j	5	6	7	8
5		0.0020	<.0001	0.0079
6	0.0020		0.0805	<.0001
7	<.0001	0.0805		<.0001
8	0.0079	<.0001	<.0001	

Zdroj: vlastné spracovanie v SAS EG

Čo sa týka členov domácností, ktorí dosiahli vzdelanie referenčnej kategórie, teda najvyššie možné vzdelanie, na základe výstupu 23 môžeme konštatovať, že ekvivalentný disponibilný príjem týchto členov domácností je medzi viacerými dvojicami kategórií premennej *Region* veľmi podobný.

Výstup 23 Test štatistickej významnosti zhody
marginálnych stredných hodnôt EDP členov domácností
v závislosti od interakcie premenných *Education* a *Region*
pre osoby so vzdelaním ISCED 5B-6

Least Squares Means for effect Education*Region Pr > t for H0: LSMean(i)=LSMean(j)				
Dependent Variable: EDP				
i/j	13	14	15	16
13		0.8085	<.0001	0.5093
14	0.8085		0.0004	0.4292
15	<.0001	0.0004		0.0002
16	0.5093	0.4292	0.0002	

Zdroj: vlastné spracovanie v SAS EG

Konkrétne na hladine významnosti 0,05 môžeme pozorovať štatisticky nevýznamný rozdiel v ekvivalentnom disponibilnom príjme medzi dvojicami členov domácností kategórie *BB_NR_TN_TT* (skupina 13), *KE_ZA* (skupina 14) a *z_BA* (skupina 16) premennej *Region*.

Opodstatnenosť zlúčenia všetkých troch skupín do jednej je však potrebné podložiť ďalším testom.

Overenie rovnosti viacerých stredných hodnôt budeme opäť realizovať prostredníctvom príkazu CONTRAST. Testovať budeme teda nulovú hypotézu

$$H_0: \mu_{13} = \mu_{14} = \mu_{16}$$

oproti alternatívnej hypotéze

H_1 : aspoň jedna rovnosť neplatí,

pričom nulovú hypotézu je nutné rozdeliť na 2 samostatné hypotézy, ktoré budeme následne testovať simultánne. Overovať budeme teda súčasnú platnosť nasledovných dvoch nulových hypotéz:

$$H_0: \mu_{13} = \mu_{14} \wedge H_0: \mu_{16} = \frac{1}{2}(\mu_{13} + \mu_{14}).$$

Po úprave hypotéz do tvarov lineárnych kombinácií je ich finálna forma nasledovná:

$$H_0: \mu_{13} - \mu_{14} = 0$$

$$H_0: \mu_{16} - 0,5 \cdot \mu_{13} - 0,5 \cdot \mu_{14} = 0$$

Vzhľadom na to, že testujeme rovnosť stredných hodnôt EDP členov domácností so vzdelaním z kategórie *z_ISCED_5B_6* pre rôzne kraje SR, koeficienty jednotlivých faktorov zapíšeme pre lepšiu prehľadnosť do kontingenčných tabuliek. Koeficienty pre prvú nulovú hypotézu zapíšeme do tabuľky 4. Do tabuľky 5 zapíšeme koeficienty pre druhú nulovú hypotézu. V súčtovom riadku jednotlivých tabuliek potom nájdeme koeficienty pre premennú *Region*, v súčtovom stĺpci sa nachádzajú koeficienty pre premennú *Education*. Celkový súčet koeficientov predstavuje koeficient pre lokujúcu konštantu. Koeficienty pre interakciu premenných sú uvedené v poli tabuľky (biele pole), pričom hodnoty týchto koeficientov sú určované práve na základe testovaných nulových hypotéz.

Tabuľka 4 Koeficienty pre príkaz CONTRAST na overenie hypotézy $H_0: \mu_{13} - \mu_{14} = 0$

Education	Region				Spolu
	BB_NR_TN_TT	KE_ZA	PO	z_BA	
ISCED_0_1_2	0	0	0	0	0
ISCED_3_4	0	0	0	0	0
ISCED_5A	0	0	0	0	0
z_ISCED_5B_6	1	-1	0	0	0
Spolu	1	-1	0	0	0

Zdroj: vlastné spracovanie

Tabuľka 5 Koeficienty pre príkaz CONTRAST na overenie
hypotézy $H_0: \mu_{16} - 0,5 \cdot \mu_{13} - 0,5 \cdot \mu_{14} = 0$

Education	Region				Spolu
	BB_NR_TN_TT	KE_ZA	PO	z_BA	
ISCED_0_1_2	0	0	0	0	0
ISCED_3_4	0	0	0	0	0
ISCED_5A	0	0	0	0	0
z_ISCED_5B_6	-0,5	-0,5	0	1	0
Spolu	-0,5	-0,5	0	1	0

Zdroj: vlastné spracovanie

Zápis príkazu pre simultánny test vyššie uvedených nulových hypotéz bude teda na základe tabuliek 2 a 3 vyzerat' takto:

```
CONTRAST '13=14=16' Region 1 -1 Education*Region 0 0 0 0 0 0
0 0 0 0 0 0 1 -1, Region -0.5 -0.5 0 1 Education*Region 0 0 0
0 0 0 0 0 0 0 0 -0.5 -0.5 0 1;
```

Vzhľadom na to, že hodnota lokujúcej konštanty a všetkých koeficientov premennej *Education* sú nulové, nie je potrebné ich v príkaze uvádzať. Rovnako nie je potrebné uvádzať zvyšné hodnoty koeficientov premenných v prípade, že sú nulové.

Výstup 24 zobrazuje výsledok simultánneho testu zhody marginálnych stredných hodnôt EDP členov domácností so vzdelaním *ISCED_5B_6* vo všetkých krajoch naprieč SR s výnimkou Prešovského. Vzhľadom na to, že *p*-hodnota ($p = 0,7255$) prekračuje všetky bežne využívané hladiny významnosti, nemáme dostatok dôkazov na zamietnutie nulovej hypotézy a môžeme teda konštatovať, že na hladine významnosti 0,05 nie je štatisticky významný rozdiel v ekvivalentnom disponibilnom príjme u vyššie spomínaných členov domácností. Všetky tieto kategórie interakcie medzi premennými *Education* a *Region* môžu byť teda zlúčené do jednej skupiny. Ako bolo už spomínané, k zlúčeniu by taktiež mohlo prísť v prípade členov domácností so vzdelaním kategórie *ISCED_3_4* v Košickom, Žilinskom a Prešovskom kraji.

Výstup 24 Výstup príkazu CONTRAST

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
13=14=16	2	5968492.115	2984246.057	0.32	0.7255

Zdroj: vlastné spracovanie v SAS EG

Pri interakcii premenných *HT* a *Region* sme pre analýzu zhody marginálnych stredných hodnôt vybrali referenčnú kategóriu premennej *HT* – z_2A_2Ch, teda členov

domácností pozostávajúcich z dvoch dospelých a dvoch detí. Výstup 25 potvrdil, že tak ako vo všeobecnosti, aj v prípade osôb z domácnosti typu 2A_2Ch mali v roku 2021 najvyšší priemerný ekvivalentný disponibilný príjem osoby z Bratislavského kraja a najnižší osoby z Prešovského kraja. V Bratislavskom kraj sme odhadli priemerný EDP vo výške 7 933,24 € ročne a v Prešovskom kraji 5 704,86 €.

Výstup 25 Bodové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od interakcie premenných *HT* a *Region* pre domácnosti 2 dospelých s 2 závislými deťmi

HT	Region	EDP LSMEAN	LSMEAN Number
z_2A_2Ch	BB_NR_TN_TT	7426.1216	25
z_2A_2Ch	KE_ZA	6665.4243	26
z_2A_2Ch	PO	5704.8605	27
z_2A_2Ch	z_BA	7933.2419	28

Zdroj: vlastné spracovanie v SAS EG

Výstup 26 Test štatistickej významnosti zhody marginálnych stredných hodnôt EDP členov domácností v závislosti od interakcie premenných *HT* a *Region* pre domácnosti 2 dospelých s 2 závislými deťmi

Least Squares Means for effect HT*Region Pr > t for H0: LSMean(i)=LSMean(j)				
Dependent Variable: EDP				
i/j	25	26	27	28
25		0.0073	<.0001	0.1620
26	0.0073		0.0088	0.0013
27	<.0001	0.0088		<.0001
28	0.1620	0.0013	<.0001	

Zdroj: vlastné spracovanie v SAS EG

Na základe výstupu 26 môžeme vzhľadom na výšky *p*-hodnôt na hladine významnosti 0,05 konštatovať, že medzi sledovanými členmi domácností existuje jedna dvojica kategórií premennej *Region*, pre ktoré sa diferencie ekvivalentného disponibilného príjmu v priemere štatisticky významne nelíšia. Konkrétne ide o kategórie *BB_NR_TN_TT* a *z_BA*. Tieto

kategórie teda zlúčime a následne odhadneme novú výšku priemerného EDP členov domácností charakterizovaných touto novou premennou.

Konkrétne bodové odhady ekvivalentného disponibilného príjmu členov domácností charakterizovaných novými premennými zabezpečíme, rovnako ako v predošlej časti práce, pomocou príkazu `ESTIMATE`. Pre lepšiu prehľadnosť a orientáciu si koeficienty opäť zapíšeme do kontingenčných tabuliek. Tabuľka 6 obsahuje koeficienty pre odhad marginálnej strednej hodnoty EDP členov domácností so vzdelaním *ISCED_3_4* v Košickom, Žilinskom a Prešovskom kraji. Tabuľka 7 obsahuje koeficienty pre odhad marginálnej strednej hodnoty EDP členov domácností so vzdelaním *ISCED_5B_6* naprieč všetkými kraji SR s výnimkou Prešovského. V tabuľke 8 nájdeme koeficienty pre odhad marginálnej strednej hodnoty EDP členov domácností pozostávajúcich z dvoch dospelých a dvoch detí súhrnne v Banskobystrickom, Nitrianskom, Trenčianskom, Trnavskom a Bratislavskom kraji.

Tabuľka 6 Koeficienty pre príkaz `ESTIMATE` na odhad marginálnej strednej hodnoty $\mu(\mu_6, \mu_7)$

Education	Region				Spolu
	BB_NR_TN_TT	KE_ZA	PO	z_BA	
ISCED_0_1_2	0	0	0	0	0
ISCED_3_4	0	1	1	0	2
ISCED_5A	0	0	0	0	0
z_ISCED_5B_6	0	0	0	0	0
Spolu	0	1	1	0	2

Zdroj: vlastné spracovanie

Tabuľka 7 Koeficienty pre príkaz `ESTIMATE` na odhad marginálnej strednej hodnoty $\mu(\mu_{13}, \mu_{14}, \mu_{16})$

Education	Region				Spolu
	BB_NR_TN_TT	KE_ZA	PO	z_BA	
ISCED_0_1_2	0	0	0	0	0
ISCED_3_4	0	0	0	0	0
ISCED_5A	0	0	0	0	0
z_ISCED_5B_6	1	1	0	1	3
Spolu	1	1	0	1	3

Zdroj: vlastné spracovanie

Tabuľka 8 Koeficienty pre príkaz ESTIMATE na odhad marginálnej strednej hodnoty $\mu(\mu_{25}, \mu_{28})$

Education	Region				Spolu
	BB_NR_TN_TT	KE_ZA	PO	z_BA	
1A_at_least_1Ch_2A_at_least_3Ch	0	0	0	0	0
1Adult_2A_1Ch	0	0	0	0	0
2A_1R	0	0	0	0	0
2Adult	0	0	0	0	0
Other_0Ch	0	0	0	0	0
Other_with_Ch	0	0	0	0	0
z_2A_2Ch	1	0	0	1	2
Spolu	1	0	0	1	2

Zdroj: vlastné spracovanie

Na základe tabuliek 4 až 6 sú zápisy príkazu ESTIMATE na odhad priemerného EDP pre uvedené skupiny osôb takéto:

```
ESTIMATE '6_7' intercept 2 Education 0 2 Region 0 1 1
Education*Region 0 0 0 0 0 1 1 / divisor=2;
```

```
ESTIMATE '13_14_16' intercept 3 Education 0 0 0 3 Region 1 1
0 1 Education*Region 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 /
divisor=3;
```

```
ESTIMATE '25_28' intercept 2 HT 0 0 0 0 0 0 2 Region 1 0 0 1
HT*Region 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
0 1 / divisor=2;
```

Podľa výstupu 27 na hladine významnosti 0,05 sú priemerné ekvivalentné disponibilné príjmy vo všetkých 3 novovytvorených skupinách signifikantne odlišné od 0. V skupine so vzdelaním *ISCED_3_4* v Košickom, Žilinskom a Prešovskom kraji odhadujeme ekvivalentný disponibilný príjem v priemere 7 389,58 € ročne. V skupine so vzdelaním *ISCED_5B_6* vo všetkých krajoch naprieč SR s výnimkou Prešovského odhadujeme ekvivalentný disponibilný príjem v priemere 10 001,29 € ročne. Čo sa týka skupiny osôb z domácností pozostávajúcich v dvoch dospelých a dvoch detí v Banskobystrickom, Nitrianskom, Trenčianskom, Trnavskom a Bratislavskom kraji, ich ekvivalentný disponibilný príjem odhadujeme v priemere na 7 679,68 € ročne, ceteris paribus.

Výstup 27 Výstup príkazu ESTIMATE

Parameter	Estimate	Standard Error	t Value	Pr > t
6_7	7389.5844	140.472001	52.61	<.0001
13_14_16	10001.2903	155.700344	64.23	<.0001
25_28	7679.6818	198.713852	38.65	<.0001

Zdroj: vlastné spracovanie v SAS EG

Vzhľadom na to, že zastúpenie členov domácností v jednotlivých kategóriách premenných vstupujúcich do odhadu nie je rovnaké, pristúpime taktiež k odhadu marginálnych stredných hodnôt EDP členov domácností so zohľadnením ich počtu. Vyššie uvedené kontingenčné tabuľky teda upravíme tak, že jednotlivé koeficienty nahradíme počtom členov domácností tej-ktorej kategórie. Upravené tabuľky budú potom vyzerat' nasledovne:

Tabuľka 9 Koeficienty pre príkaz ESTIMATE na odhad marginálnej strednej hodnoty $\mu(\mu_6, \mu_7)$

Education	Region				Spolu
	BB_NR_TN_TT	KE_ZA	PO	z_BA	
ISCED_0_1_2	0	0	0	0	0
ISCED_3_4	0	1 138	693	0	1 831
ISCED_5A	0	0	0	0	0
z_ISCED_5B_6	0	0	0	0	0
Spolu	0	1 138	693	0	1 831

Zdroj: vlastné spracovanie

Tabuľka 10 Koeficienty pre príkaz ESTIMATE na odhad marginálnej strednej hodnoty $\mu(\mu_{13}, \mu_{14}, \mu_{16})$

Education	Region				Spolu
	BB_NR_TN_TT	KE_ZA	PO	z_BA	
ISCED_0_1_2	0	0	0	0	0
ISCED_3_4	0	0	0	0	0
ISCED_5A	0	0	0	0	0
z_ISCED_5B_6	542	332	0	333	1 207
Spolu	542	332	0	333	1 207

Zdroj: vlastné spracovanie

Tabuľka 11 Koefficienty pre príkaz ESTIMATE na odhad marginálnej strednej hodnoty $\mu(\mu_{25}, \mu_{28})$

Education	Region				Spolu
	BB_NR_TN_TT	KE_ZA	PO	z_BA	
1A_at_least_1Ch_2A_at_least_3Ch	0	0	0	0	0
1Adult_2A_1Ch	0	0	0	0	0
2A_1R	0	0	0	0	0
2Adult	0	0	0	0	0
Other_0Ch	0	0	0	0	0
Other_with_Ch	0	0	0	0	0
z_2A_2Ch	442	0	0	181	623
Spolu	442	0	0	181	623

Zdroj: vlastné spracovanie

Príkaz ESTIMATE bude mať v prípade odhadu marginálnej strednej hodnoty EDP pre predmetné 3 skupiny osôb po zohľadnení váh uvedených v tabuľke 9 až 11 takéto zápisy:

```
ESTIMATE '6_7_w' intercept 1831 Education 0 1831 Region 0
1138 693 Education*Region 0 0 0 0 0 1138 693 / divisor=1831;
```

```
ESTIMATE '13_14_16_w' intercept 1207 Education 0 0 0 1207
Region 542 332 0 333 Education*Region 0 0 0 0 0 0 0 0 0 0 0 0
542 332 0 333 / divisor=1207;
```

```
ESTIMATE '25_28_w' intercept 623 HT 0 0 0 0 0 0 623 Region
442 0 0 181 HT*Region 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 442 0 0 181 / divisor=623;
```

Porovnaním výstupu 28 a 27 zisťujeme, že použitie váh nemalo výrazný vplyv na odhad marginálnych stredných hodnôt EDP členov domácností. Najväčší rozdiel v odhade marginálnej strednej hodnoty v prípade rovnakých a rôznych váh je u osôb, ktoré sú členmi domácností 2A_2Ch a sú z Banskobystrického, Nitrianskeho, Trenčianskeho, Trnavského alebo Bratislavského kraja, kde je rozdiel približne 106 €, čo v prepočte na 1 mesiac predstavuje necelých 9 €. V prípade použitia váh sme však získali menšie štandardné chyby odhadov, čo by viedlo aj k užším intervalovým odhadom.

Z výstupu 28 zisťujeme, že nakoľko všetky p -hodnoty nadobúdajú rovnakú hodnotu ($p < 0,001$), ktorá neprekračuje bežne používané hladiny významnosti, vo všetkých 3 skupinách osôb uvedených je na hladine významnosti 0,05 stredná hodnota EDP štatisticky významne odlišná od 0. V skupine so vzdelaním *ISCED_3_4* v Košickom, Žilinskom a Prešovskom kraji odhadujeme ekvivalentný disponibilný príjem v priemere 7 439,81 € ročne. V skupine so vzdelaním *ISCED_5B_6* vo všetkých krajoch naprieč SR s výnimkou Prešovského odhadujeme ekvivalentný disponibilný príjem v priemere 9 992,14 € ročne. Čo

sa týka skupiny osôb z domácností pozostávajúcich v dvoch dospelých a dvoch detí v Banskobystrickom, Nitrianskom, Trenčianskom, Trnavskom a Bratislavskom kraji, ich ekvivalentný disponibilný príjem odhadujeme v priemere na 7 573,46 € ročne, ceteris paribus.

Výstup 28 Výstup príkazu ESTIMATE

Parameter	Estimate	Standard Error	t Value	Pr > t
6_7_w	7439.8135	136.500226	54.50	<.0001
13_14_16_w	9992.1391	147.889081	67.57	<.0001
25_28_w	7573.4551	172.136187	44.00	<.0001

Zdroj: vlastné spracovanie v SAS EG

4.4 Programovací jazyk R

Na analýzu marginálnych stredných hodnôt sa v programovacom jazyku R využíva balíček *emmeans* alebo taktiež *lsmeans*. Tieto balíčky poskytujú nástroje na výpočet, porovnávanie a vizualizáciu odhadnutých marginálnych priemerov pre rôzne typy modelov, vrátane zovšeobecnených lineárnych modelov. Pred samotnou analýzou je nutné si importovať vybraný balíček príkazom `library(emmeans)` a taktiež balíček potrebný na otvorenie dátového súboru. Jeden z balíčkov využívaných na importovanie dátových súborov vo formáte .XLS importujeme príkazom `library(readxl)`. Predpokladom na importovanie týchto balíčkov je samozrejme ich predchádzajúce stiahnutie a nainštalovanie príkazom `install.packages("packagename")`. Ďalším krokom je načítanie konkrétneho dátového súboru a jeho uloženie do premennej *df* príkazom: `df <- read_excel("filename.xls")`. Po importovaní sú teda všetky dáta uložené v jednej premennej vo forme tabuľky, ktorá obsahuje rozličné dátové typy. Konkrétne stĺpec *EDP*, predstavujúci závislú premennú, je vo formáte *numeric*, má teda číselné hodnoty. Ostatné stĺpce tabuľky, ktoré budú zahrnuté do modelu ako nezávislé premenné, sú typu *character*, teda majú formu znakového reťazca. Tento formát nie je použiteľný pre analýzu a preto je nutné jednotlivé stĺpce pretransformovať na dátový typ umožňujúci priradiť jednotlivým kategóriám premenných zmysluplné hodnoty. Stĺpce budú konkrétne pretransformované na faktory, dátový typ, slúžiaci na reprezentáciu kategoriálnych premenných, obsahujúcich určité kategórie, ktoré nie je možné interpretovať

ako numerické hodnoty. Parameter *levels* potom určuje konkrétne obmeny, ktoré tá ktorá premenná nadobúda. Každý obmene premennej je následne priradená číselná hodnota, pričom číslo 1 je priradené prvej obmene podľa abecedy, ak nie je zadefinované inak. V programovacom jazyku R je ako referenčná kategória volená kategória s hodnotou *level* = 1. Vzhľadom na túto skutočnosť je v našom prípade, pre zachovanie rovnakých referenčných kategórií, nutné zmeniť predvolené abecedné poradie. Zmenu dátového typu stĺpcov a zároveň zmenu predvoleného poradia kategórii vykonáme príkazmi:

```
df$EA <- factor(df$EA, levels = c("z_at_Work",
"Disabled_person", "Inactive_person", "Person_in_household",
"Student", "Unemployed"))

df$Education <- factor(df$Education, levels =
c("z_ISCED_5B_6", "ISCED_0_1_2", "ISCED_3", "ISCED_4",
"ISCED_5A"))

df$HT <- factor(df$HT, levels = c("z_2A_2Ch",
"1A_at_least_1Ch", "1Adult", "2A_1Ch", "2A_1R",
"2A_at_least_3Ch", "2Adult_0Ch", "Other_0Ch",
"Other_with_Ch"))

df$Marital_status <- factor(df$Marital_status, levels =
c("z_Married", "Divorced", "Never_married", "Widowed"))

df$Region <- factor(df$Region, levels = c("z_BA", "BB", "KE",
"NR", "PO", "TN", "TT", "ZA"))
```

Samotnú analýzu budeme realizovať na základe lineárneho modelu, ktorý vytvoríme príkazom `model <- lm(formula = EDP ~ EA + Education + HT + Marital_status + Region, data = df)`. Následne príkazom `summary(model)` zobrazíme súhrnný výstup modelu, súčasťou ktorého sú mnohé štatistické ukazovatele opisujúce kvalitu modelu a štatistickú významnosť jeho koeficientov. V úvode výstupu nájdeme zápis modelu, na základe ktorého bol samotný model vytvorený. Ďalej vo výstupe nájdeme informácie o rezíduách, vrátane ich kvartilov, mediánu a taktiež minima a maxima. Nasleduje tabuľka zobrazujúca odhady regresných koeficientov modelu, spolu s ich štandardnými chybami, testovacími štatistikami a *p*-hodnotami. Môžeme vidieť, že na rozdiel od výstupu zo SAS EG, v tomto prípade nie sú súčasťou výstupu referenčné kategórie premenných. Štatistickú významnosť jednotlivých koeficientov je samozrejme

možné overiť na základe p -hodnôt, prípadne vyčíslených testovacích štatistík Studentovho rozdelenia, ak máme k dispozícii aj kritickú oblasť. Vo výstupe taktiež nájdeme vybrané charakteristiky modelu ako odhad štandardnej odchýlky rezíduí či koeficient determinácie. V poslednom riadku výstupu nájdeme F-testovaciu štatistiku pre test štatistickej významnosti modelu s príslušnými stupňami voľnosti a samozrejme príslušnú p -hodnotu testu.

Výstup 29 Súhrnné charakteristiky modelu EDP domácností

```
Call:
lm(formula = EDP ~ EA + Education + HT + Marital_status + Region,
    data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-11490.6  -1849.7   -193.5   1688.6  27121.7

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      10949.52    144.61   75.715 < 2e-16 ***
EADisabled_person -2131.22    155.82  -13.678 < 2e-16 ***
EAIinactive_person -1416.52    193.32   -7.327 2.60e-13 ***
EAPerson_in_household -2361.75    368.73   -6.405 1.60e-10 ***
EAStudent         473.66    155.47    3.047 0.00232 **
EAUnemployed     -2995.60    149.52  -20.035 < 2e-16 ***
EducationISCED_0_1_2 -3296.48    151.42  -21.770 < 2e-16 ***
EducationISCED_3    -1796.19     95.69  -18.771 < 2e-16 ***
EducationISCED_4    -1534.58    296.13   -5.182 2.25e-07 ***
EducationISCED_5A   -1183.60    192.32   -6.154 7.94e-10 ***
HT1A_at_least_1Ch  -1314.16    222.83   -5.898 3.85e-09 ***
HT1Adult          333.81    186.72    1.788 0.07386 .
HT2A_1Ch          565.82    141.82    3.990 6.68e-05 ***
HT2A_1R          1863.29    222.80    8.363 < 2e-16 ***
HT2A_at_least_3Ch  -1247.73    197.59   -6.315 2.86e-10 ***
HT2Adult         2540.06    132.69   19.143 < 2e-16 ***
HTOther_0Ch       3113.46    136.55   22.801 < 2e-16 ***
HTOther_with_Ch   1303.53    130.33   10.002 < 2e-16 ***
Marital_statusDivorced -336.90    145.24   -2.320 0.02039 *
Marital_statusNever_married -267.47     90.75   -2.947 0.00322 **
Marital_statuswidowed  579.91    253.57    2.287 0.02223 *
RegionBB          -1192.91    136.02   -8.770 < 2e-16 ***
RegionKE          -1752.01    143.41  -12.217 < 2e-16 ***
RegionNR           -877.86    138.56   -6.335 2.51e-10 ***
RegionPO          -2335.46    133.07  -17.550 < 2e-16 ***
RegionTN          -1113.42    145.89   -7.632 2.61e-14 ***
RegionTT           -977.47    151.78   -6.440 1.27e-10 ***
RegionZA          -1510.90    137.29  -11.005 < 2e-16 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3061 on 7328 degrees of freedom
Multiple R-squared:  0.3306,    Adjusted R-squared:  0.3282
F-statistic: 134.1 on 27 and 7328 DF,  p-value: < 2.2e-16
```

Zdroj: vlastné spracovanie v RStudio

Pred analýzou vplyvu jednotlivých faktorov na závislú premennú je nutné importovať jeden z najpopulárnejších balíčkov na analýzu dát, balíček s názvom *car*. Tento balíček obsahuje množstvo funkcií a nástrojov na manipuláciu s dátami, lineárnu a nelineárnu regresiu, analýzu rozptylu, analýzu časových radov, testovanie hypotéz, vizualizáciu dát a mnohé iné. Balíček importujeme príkazom `library(car)` a vplyv jednotlivých faktorov na závislú premennú následne otestujeme príkazom: `Anova(model, type = "III")`. Na základe výstupu 30 nižšie môžeme konštatovať, že na hladine významnosti 0,05 má každá zo zvolených nezávislých premenných štatisticky významný vplyv na závislú premennú *EDP*. Z výstupu môžeme taktiež na základe testovacej štatistiky každej premennej posúdiť veľkosť jej vplyvu na závislú premennú. Výstup 30 poskytuje rovnaké výsledky ako výstup 3 získaný aplikáciou v SAS Enterprise Guide.

Výstup 30 Test štatistickej významnosti vplyvu premenných na závislú premennú

Anova Table (Type III tests)

Response: EDP

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	5.3711e+10	1	5732.7859	< 2.2e-16 ***
EA	6.2459e+09	5	133.3298	< 2.2e-16 ***
Education	5.2025e+09	4	138.8198	< 2.2e-16 ***
HT	1.1083e+10	8	147.8693	< 2.2e-16 ***
Marital_status	1.8805e+08	3	6.6905	0.0001661 ***
Region	3.3012e+09	7	50.3359	< 2.2e-16 ***
Residuals	6.8657e+10	7328		

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Zdroj: vlastné spracovanie v RStudio

Na detailnejšiu analýzu vplyvu jednotlivých faktorov na závislú premennú pristúpime k využitiu niektorých funkcií balíčka *emmeans*. Najskôr je však nutné vytvoriť tzv. referenčnú mriežku modelu, potrebnú pre pokročilú analýzu efektov kategoriálnych premenných vrátane rozdielov či kontrastov medzi jednotlivými kategóriami.

Príkazom `model.rg <- ref_grid(model)` vytvoríme novú premennú *model.rg*, v ktorej bude uložená táto referenčná mriežka obsahujúca všetky informácie o kategoriálnych premenných modelu potrebné na ďalšiu analýzu. Následne je možné funkciou *emmeans* zobrazit' bodové a intervalové odhady marginálnych stredných hodnôt ekvivalentného disponibilného príjmu členov domácností v závislosti od jednotlivých

kategoriálnych premenných. Hodnoty si pre neskoršie použitie taktiež uložíme do nových premenných. Konkrétna syntax tejto funkcie pre jednotlivé premenné bude nasledovná:

```
model.rg <- ref_grid(model)
EA.emm <- emmeans(model.rg, "EA")
eduaction.emm <- emmeans(model.rg, "Education")
HT.emm <- emmeans(model.rg, "HT")
marital_status.emm <- emmeans(model.rg, "Marital_status")
region.emm <- emmeans(model.rg, "Region").
```

Po spustení kódu dostaneme napríklad pre premennú *EA* výstup 31 zobrazený nižšie.

Výstup 31 Bodové a intervalové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od *EA*

EA	emmean	SE	df	lower.CL	upper.CL
z_at_Work	8957	101	7328	8758	9155
Disabled_person	6825	170	7328	6493	7158
Inactive_person	7540	207	7328	7134	7946
Person_in_household	6595	375	7328	5860	7330
Student	9430	162	7328	9113	9748
Unemployed	5961	166	7328	5636	6286

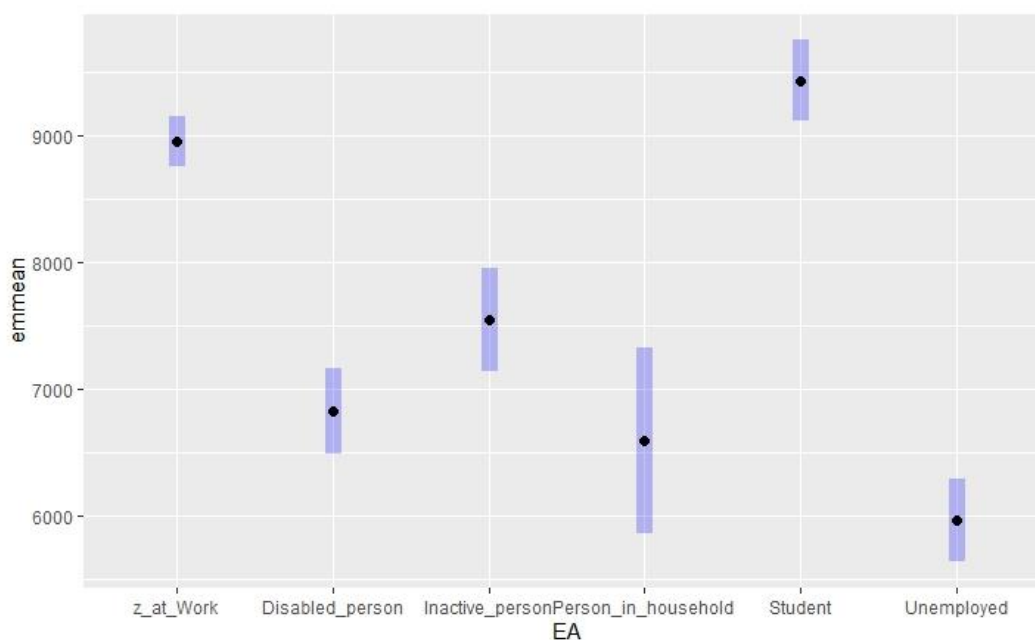
Results are averaged over the levels of: Education, HT, Marital_status, Region
Confidence level used: 0.95

Zdroj: vlastné spracovanie v RStudio

Grafické zobrazenie bodových a intervalových odhadov marginálnych stredných hodnôt premenných zabezpečíme príkazmi:

```
plot(EA.emm)+coord_flip()
plot(eduaction.emm)+coord_flip()
plot(HT.emm)+coord_flip()
plot(marital_status.emm)+coord_flip()
plot(region.emm)+coord_flip().
```

Nakoľko tento konkrétny typ grafu má prednastavené hodnoty osí *x* a *y* opačne v porovnaní so softvérom SAS EG, pre lepšiu vzájomnú porovnateľnosť týchto grafov sme použili funkciou `coord_flip()` na výmenu osí. Výsledný graf pre premennú *EA* zobrazuje obrázok 9.



Obrázok 9 Bodové a intervalové odhady marginálnych stredných hodnôt EDP členov domácností v závislosti od EA

Zdroj: vlastné spracovanie v RStudio

Maticu p -hodnôt pre test štatistickej významnosti zhody marginálnych stredných hodnôt je možné zobraziť napríklad funkciou `pwpm`, ktorá slúži na porovnávanie dvojíc marginálnych stredných hodnôt. Syntax pre použitie tejto funkcie bude nasledovná:

```
pwpm(EA.emm, adjust="none")
pwpm(eduaction.emm, adjust="none")
pwpm(HT.emm, adjust="none")
pwpm(marital_status.emm, adjust="none")
pwpm(region.emm, adjust="none").
```

V prípade, že by sme metódu úpravy p -hodnôt nenastavili na možnosť *none*, p -hodnoty by boli upravené Tukeyho metódou, ktorá je pre túto funkciu predvolená. Výstupom tejto funkcie bude matica obsahujúca 3 rôzne informácie o dvojiciach marginálnych stredných hodnôt. V hornom trojuholníku matice budú zobrazené spomínané p -hodnoty pre test štatistickej významnosti zhody marginálnych stredných hodnôt, v dolnom trojuholníku matice budú zobrazené jednotlivé párové diferencie. Na diagonále nájdeme konkrétne bodové odhady marginálnych stredných hodnôt pre jednotlivé kategórie tej-ktorej premennej. Výstup 32 nižšie je výstupom príkazu pre premennú *Region*.

Výstup 32 Diferencie a *p*-hodnoty pre test štatistickej významnosti zhody marginálnych stredných hodnôt EDP domácností v závislosti od premennej *Region*

	z_BA	BB	KE	NR	PO	TN	TT	ZA
z_BA	[8771]	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
BB	1192.9	[7578]	0.0001	0.0270	<.0001	0.5945	0.1652	0.0229
KE	1752.0	559.1	[7019]	<.0001	<.0001	<.0001	<.0001	0.1017
NR	877.9	-315.1	-874.2	[7893]	<.0001	0.1216	0.5285	<.0001
PO	2335.5	1142.6	583.5	1457.6	[6436]	<.0001	<.0001	<.0001
TN	1113.4	-79.5	-638.6	235.6	-1222.0	[7658]	0.4080	0.0079
TT	977.5	-215.4	-774.5	99.6	-1358.0	-135.9	[7794]	0.0006
ZA	1510.9	318.0	-241.1	633.0	-824.6	397.5	533.4	[7260]

Row and column labels: Region

Upper triangle: P values

Diagonal: [Estimates] (emmean)

Lower triangle: Comparisons (estimate) earlier vs. later

Zdroj: vlastné spracovanie v RStudio

Výstup 32 nad hlavnou diagonálou poskytuje informácie, ktoré sme získali v SAS Enterprise Guide prostredníctvom výstupu 9, na hlavnej diagonále sú bodové odhady marginálnych stredných hodnôt, ktoré korešpondujú s výsledkami vo výstupe 8 (SAS Enterprise Guide) a pod hlavnou diagonálou sú odhady diferencií marginálnych stredných hodnôt, ktoré sme SAS Enterprise Guide neodhadovali.

Testy zhody marginálnych stredných hodnôt pre viaceré kategórie premennej *Region* budeme rovnako ako v SAS EG realizovať cez funkciu `contrast`. Najskôr si do premennej *contrasts* vložíme vektory koeficientov kategórií premennej, ktoré budú vstupovať do analyzovaného faktora. Následne do premennej *Test* vložíme funkciu kontrastu pre vybranú kategoriálnu premennú a na záver príkazom `test` otestujeme zhodu marginálnych stredných hodnôt pre premennú *Test*. Zápis príkazov bude nasledovný:

```
contrasts=list(NR_TN = c(0,0,0,1,0,-1,0,0), TT_NR_TN =
c(0,0,0,-0.5,0,-0.5,1,0), BB_NR_TN_TT = c(0,3,0,-1,0,-1,
-1,0))
Test = contrast(region.emm, contrasts)
test(Test, joint=TRUE)
```

Na základe výstupu 33 *p*-hodnoty môžeme konštatovať, že na hladine významnosti 0,05 nezamietame nulovú hypotézu a teda nie je štatisticky významný rozdiel v strednej hodnote ekvivalentného disponibilného príjmu v Banskobystrickom, Trnavskom, Nitrianskom a Trenčianskom kraji.

Výstup 33 Výstup príkazu *contrast*

df1	df2	F.ratio	p.value
3	7328	1.861	0.1338

Zdroj: vlastné spracovanie v RStudio

Výstup 33 poskytuje identický výsledok, ako výstup 10, ktorý sme získali aplikáciou príkazu CONTRAST v programovacom jazyku SAS.

K príkazu ESTIMATE využívaného v programovacom jazyku SAS neexistuje ekvivalentný príkaz v balíčku EMMEANS. Odhad strednej hodnoty závislej premennej pri ľubovoľnej kombinácii nezávislých premenných teda za použitia balíčka EMMEANS nie je možný. Alternatívou spomínaného príkazu by v programovacom jazyku R mohla byť napríklad funkcia ESTM z balíčka sasLM, ktorý je však iba implementáciou niektorých SAS procedúr do programovacieho jazyka R.

V tejto časti práce sme ukázali, že analýzu všeobecných lineárnych modelov, analýzu marginálnych stredných hodnôt a kontrastnú analýzu je možné rovnocenne realizovať, či už v profesionálnom softvéri (v našom prípade SAS) alebo v open-source prostredí (v našom prípade R). Ide len o dostupnosť softvéru a preferenciu dátového analytika, ktorý nástroj použije. Navyše, na aplikácii príkazu CONTRAST, či už v programovacom jazyku SAS alebo R sme demonštrovali, že postupy kontrastnej analýzy podrobne vysvetlené na programovacom jazyku SAS v častiach 4.1 a 4.2 sa analogicky používajú aj v programovacom jazyku R.

Záver

Cieľom práce bolo ukázať a porovnať možnosti analýzy marginálnych stredných hodnôt ekvivalentného disponibilného príjmu domácností SR v programovacích jazykoch SAS a R. Prostredníctvom všeobecných lineárnych modelov sme teda analyzovali vplyv rozličných ukazovateľov na túto závislú premennú. Postupne sme skonštruovali tri modely, pričom každý ďalší model bol modifikáciou predošlého, čím sme dosiahli postupné zvyšovanie kvality a percenta variability závislej premennej vysvetleného modelom. Na dosiahnutie cieľa sme v softvéri SAS využili procedúru GLM a predovšetkým príkazy CONTRAST a ESTIMATE. V programovacom jazyku R bol na tieto účely vytvorený a teda aj využitý balíček EMMEANS.

V prvom a druhom modeli sme analyzovali vplyv piatich nezávislých premenných, pričom úprava v prípade druhého modelu spočívala v zlúčení kategórií premenných, pre ktoré sme na základe predošlej analýzy usúdili, že ich diferencie sa štatisticky významne nelíšia. Na overenie tohto predpokladu sme v programovacom jazyku SAS využili funkcie CONTRAST a ESTIMATE a dospeli sme k záveru, že každá z premenných zaradených do prvého modelu obsahuje kategórie, pre ktoré nie je štatisticky významný rozdiel vo výške ekvivalentného disponibilného príjmu. V prípade premennej *EA* sme zlučovali kategórie *Disabled_person*, teda osoba na invalidnom dôchodku a *Person_in_household*, teda osoba v domácnosti. V prípade premennej *HT* sme zlučovali kategórie *1A_at_least_1Ch* (domácnosť 1 rodiča s 1+ závislým dieťaťom) a *2A_at_least_3Ch* (domácnosť 2 dospelých s 3+ závislými deťmi) a taktiež *1Adult* (jednočlenná domácnosť) a *2A_1Ch* (domácnosť 2 dospelých s 1 závislým dieťaťom). Výška ekvivalentného disponibilného príjmu sa taktiež dala považovať za rovnakú pre ľudí so vzdelaním kategórie *ISCED_3* (vyššie sekundárne) a *ISCED_4* (post-sekundárne, neterciárne vzdelanie). Pri premennej *Marital_status* prišlo k zlúčeniu kategórií *Divorced* (rozvedený/á) a *Never_married* (slobodný/á). Čo sa týka krajov SR, ekvivalentný disponibilný príjem sme považovali za rovnaký v Žilinskom a Košickom kraji a taktiež v Banskobystrickom, Nitrianskom, Trnavskom a Trenčianskom kraji. Príkazom ESTIMATE sme následne odhadli konkrétne bodové odhady priemerného ekvivalentného disponibilného príjmu členov domácností charakterizovaných týmito novými kategóriami premenných.

Do tretieho modelu sme zaradili tri ďalšie premenné a to konkrétne interakcie premenných *EA*, *Education* a *HT* s premennou *Region*. Na základe analýzy marginálnych stredných hodnôt sme opäť dospeli k záveru, že ekvivalentný disponibilný príjem členov

domácností je pre viaceré kategórie premenných možné považovať za rovnaký. Toto tvrdenie sme opäť podložili testami vykonanými prostredníctvom príkazu CONTRAST. Vzhľadom na rozsah práce sme však testovali štatistickú významnosť zhody marginálnych stredných hodnôt EDP iba pre vybrané kategórie osôb. Následne sme opäť príkazom ESTIMATE odhadli konkrétne bodové odhady priemerného ekvivalentného disponibilného príjmu pre kategórie premenných, ktoré by bolo možné zlúčiť do jednej. Vzhľadom na to, že zastúpenie členov domácností v jednotlivých kategóriách premenných nebolo rovnaké, pristúpili sme taktiež k zohľadneniu ich počtu a teda zavedeniu váh, čo však nemalo výrazný vplyv na odhad marginálnych stredných hodnôt EDP členov domácností. Vo všetkých troch modeloch mala najvyšší vplyv na výšku ekvivalentného disponibilného príjmu členov domácností premenná *HT*, teda typ domácnosti.

Na základe vykonaných analýz môžeme vo všeobecnosti konštatovať, že najnižší ekvivalentný disponibilný príjem dosahovali v roku 2021 nezamestnané, rozvedené, prípadne slobodné osoby s nižším ako primárnym, primárnym alebo nižším sekundárnym vzdelaním, patriace do domácností Prešovského kraja, pozostávajúcich z 1 rodiča s 1+ závislým dieťaťom, prípadne z 2 dospelých s 3+ závislými deťmi. Tieto osoby teda patria medzi najohrozenejšie aj z pohľadu rizika sociálneho vylúčenia či rizika chudoby.

Prvý model a jeho marginálne stredné hodnoty sme následne názorne analyzovali aj v programovacom jazyku R, pričom v tejto časti práce sme našu pozornosť sústredili skôr na vysvetlenie konštrukcie príkazov ako na konkrétnu štatistickú analýzu, ktorá už bola dôsledne vykonaná v predchádzajúcich kapitolách práce.

Na základe vykonaných analýz za hlavný rozdiel medzi programovacími jazykmi SAS a R považujeme užívateľské rozhranie a znalosti programovania potrebné na prácu s jednotlivými programovacími jazykmi. Zatiaľ čo v softvéri SAS EG nie je potrebná pokročilá znalosť programovania, nakoľko základný model a výstupy je možné získať bez písania jediného riadka kódu, v programovacom jazyku R je analýza postavená výlučne na zadanom kóde. Analýza v programovacom jazyku R je preto najmä pre neskúsených programátorov časovo náročnejšia. Výhodou programovacieho jazyka R v porovnaní so SAS môže byť jeho dostupnosť, nakoľko prostredie RStudio je voľne stiahnuteľné. Vzhľadom na to, že v programovacom jazyku R je nutné celý kód programu manuálne písať, analýzy je taktiež možné viac prispôbiť vlastným potrebám. Napríklad je možné zobrazit marginálne stredné hodnoty pre kombinácie faktorov, ktoré v modeli nie sú zahrnuté vo forme interakcie premenných. Takáto možnosť v programovacom jazyku SAS nie je dostupná a teda marginálne stredné hodnoty je možné vyčíslit iba pre kombinácie faktorov

zaradených do modelu. Naopak, grafické zobrazenie marginálnych stredných hodnôt považujeme za prehľadnejšie a detailnejšie v softvéri SAS EG.

Diplomová práca teda poskytla náhľad na analýzu marginálnych stredných hodnôt ekvivalentného disponibilného príjmu členov domácností SR v roku 2021 pomocou programovacích jazykov SAS a R. Za prínosné teda považujeme priblíženie využitia ako procedúry GLM a príkazov CONTRAST a ESTIMATE, tak aj balíčka EMMEANS a taktiež poskytnutie výsledkov analýz ekvivalentného disponibilného príjmu slovenských domácností, ktorý predstavuje dôležitý sociálno-ekonomický ukazovateľ.

Zoznam použitej literatúry

EUROSTAT: *Glossary: Equivalised disposable income*. [online]. 2021. [cit. 2023-03-29].

Dostupné na: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Equivalised_disposable_income

EUROSTAT: *The European Pillar of Social Rights Action Plan (EU 2030 targets)*.

[online]. 2021. [cit. 2023-04-03]. Dostupné na: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:The_European_Pillar_of_Social_Rights_Action_Plan_%28EU_2030_targets%29

EUROSTAT: *Living conditions in Europe - poverty and social exclusion*. [online]. 2022.

[cit. 2023-04-01]. Dostupné na: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Living_conditions_in_Europe_-_poverty_and_social_exclusion#Key_findings

EUROSTAT: *Štatistika rozdelenia príjmov*. [online]. 2015. [cit. 2023-04-01]. Dostupné na:

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Income_poverty_statistics/sk&oldid=267180

GERBERY, D. *Analýza výdavkov domácností na bývanie*. [online]. 2019. [cit. 2023-04-20].

Dostupné na: <https://ivpr.gov.sk/wp-content/uploads/2020/07/Anal%C3%BDza-v%C3%BDdavkov-na-b%C3%BDvanie-fin%C3%A1lna-verzia.pdf>

Kancelária RRZ. *KVANTIFIKÁCIA VPLYVU OPATRENÍ, Zákon o financovaní voľného času dieťaťa (tzv. prarodinný balíček)*. [online]. 2022. [cit. 2023-04-21]. Dostupné na:

https://www.rrz.sk/wp-content/uploads/2022/06/KO_202202_Pomoc_rodinam.pdf

LENTH, R. V. *Package 'emmeans'* [online]. 2023. [cit. 2023-04-04]. Dostupné na:

<https://cran.r-project.org/web/packages/emmeans/index.html>

LENTH, R. V. *Least-Squares Means: The R Package lsmeans*. [online]. 2016. [cit. 2023-04-04]. Dostupné na:

<https://www.jstatsoft.org/index.php/jss/article/view/v069i01/974>

MILANOVIĆ, B. *Income, Inequality, and Poverty during the Transition from Planned to Market Economy*. Washington, D.C.: The Woeld Bank, 1998, 237 s. ISBN 0-8213-3994-X

SAS Institute Inc. *SAS/STAT® 14.1 User's Guide*. [online]. Cary, NC: SAS Institute Inc. 2015. [cit. 2023-02-17]. Dostupné na:

<https://support.sas.com/documentation/onlinedoc/stat/141/glm.pdf>

SAS Institute Inc. *SAS/STAT® 15.2 User's Guide*. [online]. Cary, NC: SAS Institute Inc..

2020. [cit. 2023-04-05]. Dostupné na:

http://documentation.sas.com/api/docsets/statug/15.2/content/glm.pdf?locale=sk#nameddest=statug_glm_toc

SAS Institute Inc. *SAS® Technical Report R-103, Least-Squares Means in the Fixed-Effects General Linear Models*. Cary, NC: SAS Institute Inc., 1997, 9 s. ISBN 1-55544-967-7

ŠOLTÉS, E. *Regresná a korelačná analýza s aplikáciami*. Bratislava: Wolters Kluwer (Iura Edition), 2008. 287 s. ISBN 978-80-8078-163-7

ŠOLTÉS, E. *Využitie všeobecných lineárnych modelov na analýzu výšky poistných škôd*. Bratislava, 2020. 170 s. ISBN 978-80-899-64-8

VLAČUHA, R. – KOVÁČOVÁ, Y. *Indikátory chudoby a sociálneho vylúčenia*. Bratislava: Štatistický úrad SR, 2019, 57 s. ISBN 978-80-8121-695-4

VLAČUHA, R. – KOVÁČOVÁ, Y. – KUBALA, M. *Indikátory chudoby a sociálneho vylúčenia*. Bratislava: Štatistický úrad SR, 2022, 81 s. ISBN 978-80-8121-867-5

WICKLIN, R. *The diffogram and other graphs for multiple comparisons of means*.

[online]. 2017. [cit. 2023-03-28]. Dostupné na:

<https://blogs.sas.com/content/iml/2017/10/18/diffogram-multiple-comparisons-sas.html>