SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

2/2025
ročník/volume 35

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 4

Typ článku/Type of article: informatívny článok/informative article

Strany/Pages: 50 – 66

Dátum vydania/Publication date: 15. apríl 2025/April 15, 2025

Roman PAVELKA
Štatistický úrad Slovenskej republiky

MODELOVÁNÍ EFEKTŮ A JEJICH VISUALIZACE V PROSTŘEDÍ SYSTÉMU SAS

MODELLING OF EFFECTS AND THEIR VISUALIZATION IN THE SAS SYSTEM ENVIRONMENT

ABSTRAKT
Cílem článku je přiblížit metody interpretace statistické interakce prostřednictvím analýzy podmíněných efektů, které interakci tvoří, s využitím analytických možností vybraných procedur statistického systému SAS. Předkládaný článek se bude zejména zabývat odhady jednoduchých efektů, resp. sklonů (směrnic) obecných lineárních modelů a testování jejich významnosti. Podmíněný efekt se regresí odhadne přímo na referenční úrovni moderující proměnné. K odhadům efektů na ostatních úrovních jsou nutné další kroky. Testovat se budou i rozdíly mezi jednoduchými efekty, resp. sklony (směrnicemi) jako testy rozdílů efektů na referenční úrovni moderující proměnné a efekty na ostatních úrovních této proměnné. Vizuální znázornění interakce bude ilustrováno pomocí grafů jednoduchých efektů, resp. sklonů (směrnic), které obvykle poskytují nejjednodušší a nejrychlejší interpretaci. Ukázky modelování statistických efektů a jejich vizualizace bude uskutečněna pomocí procedury PLM. Funkcionalita této procedury však dává datovým analytikům široké možnosti v analýzách odhadnutého statistického modelu, testování statistických hypotéz o efektech a jejich vizualizaci, reparametrizaci modelu pro odhadování kontrastů i modelování hodnot predikce. Pro realizaci odhadů statistických modelů bude využita analytická procedura GLM.

ABSTRACT
The aim of the paper is to present the methods of interpretation of statistical interaction by analysing the conditional effects that constitute the interaction, using the analytical capabilities of selected procedures of the SAS statistical system. The presented paper will mainly deal with the estimation of simple effects or slopes of general linear models and testing their significance. The conditional effect is estimated by regression directly on the reference level of the moderating variable. Further steps are required to estimate effects at other levels. The differences between simple effects or slopes will also be tested as tests of differences in effects at the reference level of the moderating variable and effects at other levels of this variable. The visual representation of the interaction will be illustrated using graphs of simple effects or slopes, which usually provide the simplest and fastest interpretation. Examples of modelling of statistical effects and their visualization will be demonstrated using a PLM procedure. However, the functionality of this procedure gives data analysts a wide range of possibilities in analysing the estimated statistical model, testing statistical hypotheses about effects and their visualization, reparametrizing the model for estimating contrasts, and modelling prediction values. The GLM analytical procedure will be used for the estimation of statistical models.

KLÍČOVÁ SLOVA
analytický systém SAS, hlavní efekt, jednoduchý sklon, statistická interakce

KEY WORDS
analytical system SAS, main effect, simple slope, statistical interaction

1. ÚVOD

Statistické regresní modely odhadují účinky nezávislých proměnných (prediktorů) na závislé proměnné (výstupy). V rámci regresní analýzy se často modeluje také modifikace účinku jedné nezávislé proměnné na jinou nezávislou proměnnou, nazývanou také moderující proměnnou (Aguinis, 2004). Tato modifikace efektu je označována jako statistická interakce. Interakční proměnné jsou generovány vzájemným vynásobením nezávislé a moderující proměnné a tato výsledná součinová proměnná je pak vložena do regrese, typicky spolu s nezávislou a moderující proměnnou. Interakční regresní koeficient lze následně použít k analýze, zda účinek nezávislé proměnné závisí na jiné nezávislé proměnné (ve skutečnosti však regresní model nerozlišuje mezi nezávislou a moderující proměnnou, protože účinek moderující proměnné je také ovlivňován nezávislou proměnnou v modelu a je reprezentován stejným interakčním koeficientem). K úplnému pochopení a interpretaci interakce je potřebná znalost nejen velikosti působení jedné nezávislé proměnné na druhou, ale také i směru a významu efektu nezávislé proměnné na různých úrovních moderující proměnné.

2. REGRESNÍ MODELY A MODELY S KLASIFIKAČNÍMI EFEKTY

2.1. OBECNÝ LINEÁRNÍ MODEL (GLM)

Ukázky analýzy v předkládaném příspěvku vychází z obecného lineárního modelu. Obecný lineární regresní model je podle (Rutherford, 2011, s. 10) dán výrazem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

kde

- $\mathbf{Y} = (y_1, \dots, y_n)^T$ je $(n\ x\ 1)$ vektor n hodnot závisle proměnných,
- $\mathbf{X}$ je $n\ x(\ k+1)$ matice plánu (designová matice) s n řádky pozorování a k sloupců nezávislých proměnných a sloupcem 1, tj.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \tag{1.2}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^T$ je $1x(k+1\ )$ vektor k neznámých regresních parametrů a konstanty β_0 reprezentující absolutní člen modelu,
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ je $(n\ x\ 1)$ vektor n nepozorovatelných chyb modelu.

Pokud jsou všechny regresorové proměnné (efekty) v modelu spojité, regresní model je tvořen výlučně lineární kombinací regresorových proměnných **X**, přičemž každá spojitá proměnná (efekt) v modelu přispívá jedním sloupcem do matice plánu **X**, a tedy jedním parametrem do modelu jako celku.

Klasifikační efekt (pevný i náhodný) je naopak spojen s více než jedním sloupcem matice **X**. Klasifikace vzhledem k proměnné je proces, při kterém je každé pozorování přiřazeno jednomu z *k* úrovní; proces určení těchto *k* úrovní se označuje jako

tzv. levelizace proměnné (SAS Institute Inc., 2023a). Klasifikace proměnných se v regresních modelech používá k určení experimentálních podmínek, příslušnosti ke skupině, ošetření atd. a podobně. Skutečné hodnoty klasifikační proměnné nejsou důležité a proměnná může být číselná nebo číselně vyjádřená znaková proměnná. Důležitá je asociace diskrétních hodnot nebo úrovní klasifikační proměnné se skupinami pozorování. Zpravidla se klasifikační efekty v regresním modelu označují jako hlavní efekty.

2.2. PARAMETRIZACE REGRESNÍCH EFEKTŮ MODELU

Parametrizace regresních efektů modelu popisuje způsob konstrukce matice plánu. Tato parametrizační pravidla (SAS Institute Inc., 2023b, s. 401) jsou aplikována na regresní modely, modely s klasifikačními efekty i na všechny ostatní statistické modely.

Nejjednodušší a nejobecnější parametrizací regresních modelů je tzv. **GLM parametrizace**, jejíž název byl odvozen podle definice uvedené v (1.1). V rámci GLM parametrizace je matice plánu (designová matice) vytvářena sloupcovým vektorem jedniček reprezentující absolutní člen regresní rovnice (v tabulce č. 1 sloupec označený I). Podobně tak i numerické proměnné nazývané regresními efekty mohou být zahrnuty do modelu jako sloupcové vektory. Má-li klasifikační proměnná k úrovní, potom GLM parametrizace generuje k sloupců (nazývaných tzv. umělými proměnnými s hodnotami 0/1 podle příslušnosti pozorování do jednotlivých úrovní klasifikační proměnné) pro hlavní efekty této klasifikační proměnné v matici plánu. GLM parametrizace pro klasifikační proměnné generuje obvykle pro tyto efekty více sloupců v matici plánu než je počet stupňů volnosti k jejich odhadu. Jinými slovy to znamená, že parametrizace metodou GLM je singulární. Do regresního modelu jsou mnohdy zahrnovány i tzv. křížové efekty (interakce) odrážející účinek změny jedné nezávislé proměnné na hodnoty druhé nezávislé proměnné. Tyto interakce mohou být pevné, náhodné, ale i smíšené, což záleží na typu vstupujícího efektu. Příklad matice plánu (designové matice) s výše uvedenými typy efektů je ilustrován v tabulce č. 1.

Tabulka č. 1: Ukázka GLM parametrizace a matice plánu

		Matice plánu (designová matice)							
Data		**I**	**X**	**M**			**X·M**		
X	**M**			**M=1**	**M=2**	**M=3**	**X(M=1)**	**X(M=2)**	**X(M=3)**
1.1	1	1	1.1	1	0	**0**	1.1	0	**0**
1.3	2	1	1.3	0	1	**0**	0	1.3	**0**
2.4	3	1	2.4	0	0	**1**	0	0	**2.4**
4.1	1	1	4.1	1	0	**0**	4.1	0	**0**
5.1	2	1	5.1	0	1	**0**	0	5.1	**0**
7.1	3	1	7.1	0	0	**1**	0	0	**7.1**

Zdroj: vlastní zpracování podle (SAS Institute Inc., 2023b)

Tabulka zobrazuje matici plánu vytvořenou z regresního efektu spojité proměnné X a moderující proměnné, M se 3 kategoriemi. Ačkoli by se mohli vytvořit 3 umělé proměnné pro reprezentaci 3 kategorií klasifikační proměnné M, do regrese lze vložit pouze 2 umělé proměnné, protože jednu umělou proměnnou lze vygenerovat lineární kombinací ostatních dvou. Vynechaná kategorie (v tabulce č. 1 vyznačena šrafováním) je zpravidla označována jako referenční kategorie. Interakce spojité proměnné X s kategorickou proměnnou M se dosáhne vynásobením spojité proměnné X každou

z umělých proměnných. I v tomto případě nebude do matice zahrnuta interakce proměnných vytvořená s vynechanou umělou proměnnou. Existují efekty i složitější, jako jsou efekty vložené, jiných typů, jejichž analýzy se však konceptuálně neliší.

Ve většině procedur SAS, když jsou v příkazu CLASS specifikovány kategorické proměnné, SAS automaticky vytvoří umělé (nula-jedničkové) proměnné pro každou jejich úroveň a vloží tyto proměnné (ne všechny, ale jednu) do regresní rovnice. Zároveň se vytvoří referenční kategorie implicitně reprezentovaná vynechanou umělou proměnnou (ve výchozím nastavení se vynechá poslední hodnota kategorické proměnné seřazená podle formátovaných hodnot). Aby se zajistilo, že se jedná o GLM parametrizaci, pro většinu analytických procedur lze do příkazu CLASS přidat volbu PARAM=GLM. Pokud analytická procedura neobsahuje funkcionalitu automatické parametrizace příkazem CLASS, parametrizace kategorické proměnné zajistí například dodatečným datovým krokem (týká se například analytické procedury REG).

V programovém prostředí SAS je možné také změnit způsob parametrizace. Například, pokud je analyzován efekt jednotlivých kategorií kategorické proměnné vzhledem k průměrnému efektu všech kategorií této proměnné, nastaví se parametr PARAM=EFFECT. Tím dojde ke změně parametrizace způsobem, že se v referenční kategorii namísto 0 použije -1. Zvolená metoda parametrizace má zásadní vliv na způsob testování hypotéz a na interpretaci odhadů parametrů (Pritchard & Pasta, 2004). Bližší informace lze najít v originální dokumentaci systému SAS (například v SAS Institute Inc., 2023a).

2.3. POSTODHADOVÁ ANALYTICKÁ PROCEDURA PLM

Následující analýzy a modelování efektů, vizualizace interakcí a testování uživatelsky vytvořených hypotéz budou realizovány procedurou SAS nazývanou PLM.

Analytická procedura PLM svou funkcionalitou zajišťuje v systému SAS různé analýzy a vykreslování funkcí poté, co je odhadnut počáteční regresní model jinou odhadovou procedurou včetně testování hypotéz o efektech modelu a kontrastech, výpočtu předpovídaných hodnot výsledku (skóre) a vykreslování těchto předpovědí. Na rozdíl od většiny ostatních analytických procedur SAS vstupem do procedury PLM není soubor dat. Vstupem do procedury PLM je soubor, který obsahuje informace o regresním modelu odhadnutém v jiné proceduře SAS. Soubor s informacemi o regresním modelu může být vytvořen mnoha běžně používanými regresními procedurami, jako jsou GLM pro modelování obecných lineárních modelů, GENMOD pro modelování nespojitých náhodných proměnných s uživatelsky definovanou korelační strukturou, LOGISTIC předurčené pro logistickou regresi, odhady poměru šancí, MIXED pro lineární smíšené modely, GLIMMIX k modelování zobecněných smíšených lineárních modelů a dalších. Prostřednictvím příkazu STORE se zadaným názvem úložiště se odhadnutý model uloží a následně se pomocí parametru RESTORE vloží jako vstup procedury PLM.

Při analýzách, modelování a vizualizaci budou využity zejména následující příkazy procedury PLM:

- příkaz SLICE: porovnává okrajové (marginální) střední hodnoty, které jsou využívány k odhadu jednoduchých efektů,
- příkaz LSMEANS: používá se k odhadu okrajových (marginálních) průměrů a výpočtu rozdílů mezi nimi,

- příkaz ESTIMATE: k odhadu průměrů, kontrastů a jednoduchých sklonů, podmíněných interakcí a rozdílů mezi nimi, a to zadáním lineárních kombinací koeficientů modelu; velmi flexibilní příkaz – všechny výpočty lze provádět pomocí tohoto příkazu, i když často vyžadují složitější reparametrizaci,
- příkaz LSESTIMATE: kombinuje funkcionalitu příkazu ESTIMATE a příkazu LSMEANS při odhadu jednoduchých efektů pomocí rozdílů mezi průměry,
- příkaz EFFECTPLOT: vykresluje predikované hodnoty výsledků v rozsahu hodnot jednoho a více prediktorů, vizualizuje jednoduché efekty, jednoduché sklony (směrnice) a podmíněné interakce. Pokud jsou v modelu další prediktory, lze v rámci EFFECTPLOT měnit jejich hodnoty, ve výchozím nastavení na střední hodnotu pro spojité prediktory a na referenční úroveň pro kategoriální prediktory.

Důvody, proč je výhodnější pro postodhadové analýzy používat proceduru PLM:

- Pokud je regresní model správně odhadnutý, není pro analýzu interakce nutný model při každém spuštění znovu upravovat. Model je uložen v úložišti položek, což může ušetřit spoustu času a vytváří méně nadbytečných výstupů z ostatních procedur.
- Příkazy procedury PLM obsahují často více funkcí (možností) než v jiných procedurách, např. modelování hodnot predikovaných použitým modelem apod.

3. MODELOVÁNÍ EFEKTŮ A VIZUALIZACE INTERAKCÍ

3.1. SOUBOR DAT POUŽITÝ K ANALÝZÁM

Datový soubor se skládá z údajů popisujících množství úbytku hmotnosti, kterého dosáhlo 900 účastníků roční studie 3 různých cvičebních programů, a to: běhání, plavání a čtení podle (Schwartz & Barrett, 2021). Jedná se o umělá data vytvořená Monte Carlo simulací. Jako referenční (kontrolní) aktivita slouží jako program čtení. Pro ilustraci předmětných metod a postupů byla autorem stanovena otázka: Jak týdenní počet hodin, které se subjekty rozhodly věnovat cvičení, předpovídá úbytek hmotnosti? Mezi proměnné použité při analýze souboru dat patří:

- LOSS: spojitá, normálně rozdělená proměnná popisující průměrný týdenní úbytek hmotnosti účastníků. Kladné skóre znamená úbytek hmotnosti, zatímco záporné skóre znamená přírůstek hmotnosti. Používá se jako závislá proměnná (výstup) ve většině analyzovaných regresních modelů,
- HOURS: spojitá proměnná popisující průměrný počet hodin cvičení týdně. Pro analýzy obvykle použita jako hlavní spojitý prediktor,
- EFFORT: spojitá proměnná v rozmezí od 0 do 50, která je průměrem týdenních skóre námahy uváděných subjektem, rovněž v rozmezí od 0 do 50, přičemž 0 znamená minimální fyzickou námahu a 50 maximální námahu,
- PROG: proměnná o třech úrovních, která podrobně popisuje, jakým cvičebním programem se subjekt řídil: běh = 1, plavání = 2, nebo čtení = 3 (čtení představuje kontrolní, resp. referenční úroveň),
- SEX: binární kategoriální proměnná označující pohlaví, muž = 0, žena = 1,
- SATISFIED: binární (0/1) proměnná označující spokojenost (spokojen = 1) nebo nespokojenost (spokojen = 0) subjektu s množstvím zhubnutých kilogramů.

Pro účely analýzy je v předkládaném příspěvku soubor nazván jako EXERCISE.

3.2. LINEÁRNÍ REGRESNÍ MODELY S INTERAKCÍ SPOJITÝCH PROMĚNNÝCH

Lineární regresní model obsahující 1 závislou proměnnou a 2 nezávislé spojité proměnné, tzv. prediktory a jejich interakci je definován rovnicí

$$Y = \beta_0 + \beta_x X + \beta_z Z + \beta_{xz} XZ + \varepsilon, \tag{3.1}$$

kde symboly X a Z představují nezávislé spojité proměnné, Y odpovídá vysvětlované proměnné, ε jsou nepozorovatelné chyby modelu pro n počet pozorování. V modelu (3.1) je interakční člen dán jako součin nezávislých proměnných X a Z.

Efekt spojité nezávislé proměnné X na proměnnou závislou Y se nazývá sklon (směrnice) nezávislé proměnné X (Rutherford, 2011, s. 256). Sklon (směrnice) nezávislé proměnné X představuje změnu závislé proměnné Y na jednotku změny nezávislé proměnné X (za jinak nezměněných podmínek). Pokud spojitá nezávislá proměnná X interaguje s další spojitou proměnnou, Z, účinek nezávislé proměnné na konkrétní úrovni $Z = z$ se opět nazývá sklon (směrnice). Podobně se účinek Z na konkrétní úrovni nezávislé proměnné X také nazývá sklon (směrnice). Při interpretaci interakce regresní model nerozlišuje role nezávislé proměnné X a Z.

Odhad sklonu proměnné X v libovolné hodnotě $Z = z$ vychází z rovnice modelu (3.1). Změnu závislé proměnné Y na jednotku změny nezávislé proměnné X pro $Z = z$ (podobně jako pro sklon nezávislé proměnné Z pro $X = x$) lze odvodit pomocí parciální derivace závislé proměnné Y vzhledem k nezávislé proměnné X pro $Z = z$, tj.

$$\begin{aligned} \frac{\delta}{\delta X} Y \quad &= \frac{\delta}{\delta X}[\beta_0 + \beta_x X + \beta_z Z + \beta_{xz} XZ + \varepsilon] \\ sklon_{x|Z=z} \quad &= \beta_x + \beta_{xz} Z \end{aligned} \tag{3.2}$$

V ekonometrii se (3.2) označuje jako tzv. marginální efekt (Greene, 2003, s. 674).

K modelování efektů průměrného počtu cvičení týdně (proměnná $HOURS$) a průměrného týdenního skóre úsilí (proměnná $EFFORT$) a jejich interakce se použije níže uvedený programový kód v programovacím jazyku SAS s využitím procedury GLM. Cílem je zjištění, jestli se vliv průměrného týdenního počtu hodin cvičení liší v závislosti na množství úsilí, které subjekt vynakládá.

```
proc glm data=exercise;
     model hours effort hours*effort / solution;
     store contcont;
run;
```

Programový kód zajišťuje:

- parametrem SOLUTION vytvoření tabulky regresních koeficientů (odhady parametrů) na výstupu, která je potřebná při konstrukci odhadů pro výpočet jednoduchých sklonů a efektů,
- příkazem STORE vytvoření úložiště položek s názvem contcont, které obsahuje informace o regresním modelu potřebné pro proceduru PLM.

Regresní koeficienty (odhady parametrů) modelu jsou zobrazeny v tabulce č. 2. Interakční člen $HOURS \cdot EFFORT$ je statisticky významný na úrovni 0.05 ($Pr = 0.0362$). Ostatní regresní koeficienty statisticky významné nejsou na hladině významnosti 0.05.

Tabulka č. 2: Odhad regresních koeficientů modelu včetně testů významnosti

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	7.798637340	11.60361526	0.67	0.5017
hours	-9.375681298	5.66392068	-1.66	0.0982
effort	-0.080276367	0.38464693	-0.21	0.8347
hours*effort	0.393346795	0.18750440	2.10	0.0362

Zdroj: vlastní zpracování v programovacím jazyku SAS

Výsledek odhadů procedurou GLM generuje následující regresní rovnici:

$$\widehat{LOSS}_i = 7.8 - 9.38 \cdot HOURS_i - 0.08 \cdot EFFORT_i + 0.39 \cdot HOURS_i \cdot EFFORT_i. \quad (3.3)$$

Možnou interpretací významné interakce hodin ($HOURS$) a úsilí ($EFFORT$) je, že vliv průměrného týdenního počtu hodin na úbytek hmotnosti závisí na úrovni vynaloženého úsilí; jinými slovy, sklon hodin závisí na úsilí.

Pro odhad předpokládaného úbytku hmotnosti ($LOSS$) u subjektu, který v průměru cvičí 2 hodiny týdně ($HOURS$) při intenzitě 30 ($EFFORT$), a testováním významnosti této předpovědi se využije příkaz ESTIMATE procedury PLM. Příkaz ESTIMATE se používá k odhadu lineárních kombinací (vážených součtů) regresních koeficientů, tj.

$$\widehat{LOSS}_i = 7.8 - 9.38 \cdot 2 - 0.08 \cdot 30 + 0.39 \cdot 2 \cdot 30. \quad (3.4)$$

Odhady pro předpověď této ztráty ($LOSS$) a testu významnosti jsou realizovány programovým kódem pro proceduru PLM s příkazem ESTIMATE uvedenými níže:

```
proc plm restore=contcont;
     estimate 'předpověď LOSS při HOURS=2 a EFFORT=30'
              intercept 1 HOURS 2 EFFORT 30 HOURS*EFFORT 60 / cl e;
run;
```

Programový kód zajišťuje:

- parametrem *e* se na výstupu procedury zobrazí koeficienty odhadované lineární kombinace odhadnutých parametrů statistického modelu,
- parametr *cl* umožňuje i odhady intervalu spolehlivosti (ve výchozím nastavení $95\ \%$),
- hodnotu testové statistiky *t* a hodnotu pravděpodobnosti $Pr > |t|$ potřebné pro testování, zda je odhadovaná hodnota statisticky významná.

Odhad předpovědi úbytku ($LOSS$) je významně odlišný od 0 (viz tabulka č. 3).

Tabulka č. 3: Předpověď úbytku hmotnosti (LOSS) včetně testu významnosti

Estimate Coefficients	
Effect	**Row1**
Intercept	1
hours	2
effort	30
hours*effort	60

Estimate					
Label	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
předpověď LOSS při HOURS=2 a EFFORT=30	10.2398	0.4531	896	22.60	<.0001

Zdroj: vlastní zpracování

Dalším krokem analýzy je odhad jednoduchého sklonu (směrnice) nezávislé proměnné $HOURS$ podle (3.2). Jelikož v získaných datech neexistuje nějaká důležitá hodnota proměnné $EFFORT$, je testování významnosti jednoduchého sklonu (směrnice) proměnné $HOURS$ realizováno pro průměrnou hodnotu proměnné $EFFORT$ plus nebo minus 1 směrodatná odchylka průměrné hodnoty proměnné $EFFORT$, tj. pro $EFFORT = 24.52,\ 29.66$ a 34.8. Testování významnosti odhadu směrnice proměnné $HOURS$ je realizováno procedurou PLM pomocí následujícího kódu:

```
proc plm restore=contcont;
     estimate
     'sklon HOURS, EFFORT=mean-sd' HOURS 1 HOURS*EFFORT 24.52,
     'sklon HOURS, EFFORT=mean' HOURS 1 HOURS*EFFORT 29.66,
     'sklon HOURS, EFFORT=mean+sd' HOURS 1 HOURS*EFFORT 34.8 / e;
run;
```

Výstupní odhady včetně testování významnosti jsou ilustrovány tabulkou č. 4:

Tabulka č. 4: Odhad sklonu (směrnice) HOURS včetně testů významnosti

Estimates					
Label	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
sklon HOURS, EFFORT=mean-sd	0.2692	1.3493	896	0.20	0.8419
sklon HOURS, EFFORT=mean	2.2910	0.9151	896	2.50	0.0125
sklon HOURS, EFFORT=mean+sd	4.3128	1.3084	896	3.30	0.0010

Zdroj: vlastní zpracování v programovacím jazyku SAS

Z odhadu vyplývá, že sklon proměnné $HOURS$ je signifikantní pouze u střední hodnoty a střední hodnoty plus jedna směrodatná odchylka proměnné $EFFORT$. Výsledky odhadu naznačují, že je třeba vynaložit určité úsilí, aby se počet hodin týdně projevil v očekávaném úbytku hmotnosti.

Aby se eliminoval vliv velikosti rozdílu proměnné $EFFORT$ při odhadech sklonu (směrnice) proměnné $HOURS$, je vhodné odhadovat rozdíl sklonů (směrnic) proměnné $HOURS$ v počáteční a koncové hodnotě velikosti tohoto rozdílu proměnné $EFFORT$.

Rozdíl jednoduchých sklonů (nebo jednoduchých efektů) se získá prostým odečtením lineární kombinace koeficientů pro jeden sklon proměnné $HOURS$ od druhého, tj.

$$\begin{aligned} & sklon_{HOURS|EFFORT=mean} - sklon_{HOURS|EFFORT=mean-sd} \\ & = \hat{\beta}_{HOURS} + \hat{\beta}_{HOURS \cdot EFFORT} \cdot 29.66 - \left(\hat{\beta}_{HOURS} + \hat{\beta}_{HOURS \cdot EFFORT} \cdot 24.52\right) \\ & = \hat{\beta}_{HOURS \cdot EFFORT} \cdot 5.14 \end{aligned} \qquad (3.5)$$

V rovnici (3.5) hodnota 29.66 je odhadem střední hodnoty proměnné $EFFORT$ (úsilí účastníků) a 24.52 představuje odhad střední hodnoty proměnné $EFFORT$ bez jeho směrodatné odchylky. Odhad rozdílu sklonů (směrnic) proměnné $HOURS$ se uskuteční opět využitím procedury PLM pomocí programového kódu:

```
proc plm restore=contcont;
     estimate
          'rozdíl sklonů proměnné HOURS, EFFORT=mean+sd - mean'
          HOURS*EFFORT 5.14;
run;
```

Odhad rozdílu sklonů (směrnic) proměnné $HOURS$ podle (3.5) je obsahem tabulky č. 5. Hodnota testové statistiky t a odpovídající pravděpodobnosti $Pr > |t|$ u rozdílu mezi sklony (směrnicemi) odpovídají hodnotám u odhadu koeficientu interakce $HOURS \cdot EFFORT$ z tabulky 2.

Výchozí volbou pro vizualizaci interakce spojitých proměnných $HOURS$ a $EFFORT$ je obrysový graf. Obě interagující proměnné mohou být reprezentovány spojitě v grafu, přičemž závislá proměnná je vykreslena jako obrysy. Obrysový graf lze vytvářet parametrem CONTOUR v příkazu EFFECTPLOT procedury PLM:

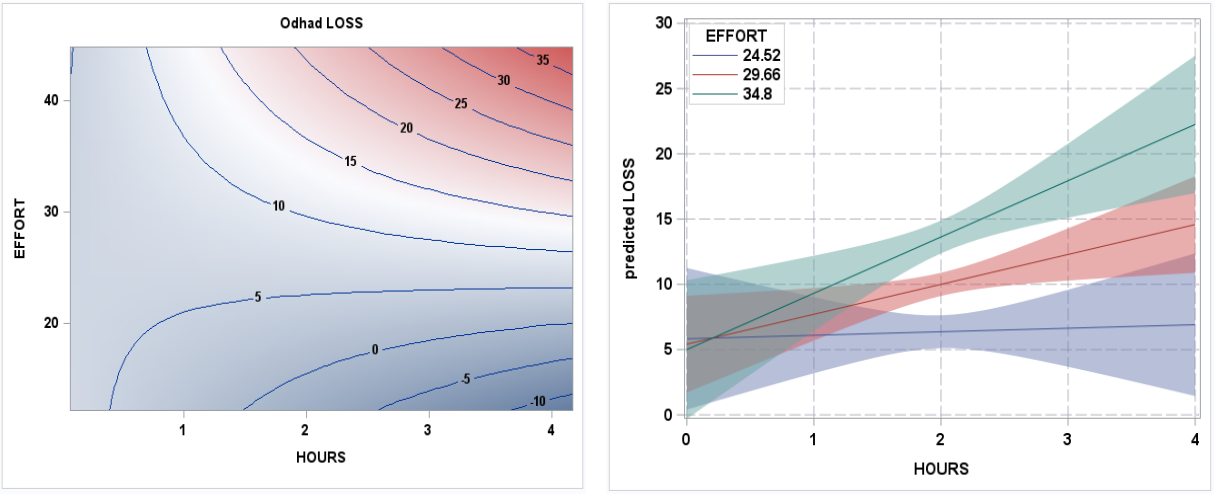
```
proc plm source=contcont;
     effectplot contour (x=HOURS y=EFFORT);
run;
```

Asi nejčastěji volenou možností, jak vykreslit interakci 2 spojitých proměnných, je vykreslit sklon jedné spojité proměnné jako přímky na vybraných úrovních interagující spojité proměnné. K vytvoření uvedených přímkových grafů se využije parametr FIT příkazu EFFECTPLOT procedury PLM. Predikovaná ztráta $LOSS$ v rozsahu proměnné $HOURS$ při vybraných hodnotách proměnné $EFFORT$ se vykreslí kódem:

```
proc plm source=contcont;
     effectplot fit (x=hours) / at(effort = 24.52 29.66 34.8);
run;
```

Vizualizace spojitých proměnných $HOURS$ a $EFFORT$ v podobě grafu obrysového a grafu vyrovnaných hodnot ilustruje obrázek č. 1.

Obrázek č. 1: Vizualizace interakce spojitých proměnných HOURS a EFFORT



Zdroj: vlastní zpracování

3.3. LINEÁRNÍ REGRESNÍ MODELY S INTERAKCÍ SPOJITÉ A KATEGORICKÉ PROMĚNNÉ

V případě, že se do statistického modelu (3.1) zahrne nezávislá proměnná spojitá X a kategorická nezávislá proměnná, označená jako M, lze jako interakční členy modelu odhadovat jednoduché sklony (směrnice) spojité proměnné X pro jednotlivé úrovně kategorické proměnné M. V tom případě je kategorická proměnná považovaná za moderující proměnnou. Pokud v roli moderující proměnné vystupuje nezávislá spojitá proměnná X, za interakční členy lze být považován jednoduchý efekt kategorické proměnné M napříč spojitou proměnnou X.

V regresních modelech se kategorické proměnné obvykle zadávají do jedné nebo více umělých (indikátorových) proměnných. Důležitou otázkou v analýze modelu je překódování kategorických proměnných jako umělých proměnných (parametrizace).

Rovnice lineárního regresního modelu spojité nezávislé proměnné, X, interagující s nezávislou kategorickou proměnnou M se 3 kategoriemi, je:

$$\begin{aligned} Y \quad &= \beta_0 \\ &+ \beta_x X \\ &+ \beta_{m1} M_{M=1} + \beta_{m2} M_{M=2} \\ &+ \beta_{xm1} X M_{M=1} + \beta_{xm2} X M_{M=2} \\ &+ \varepsilon \end{aligned} \quad . \tag{3.6}$$

Žádná z vynechaných umělých proměnných pro $M_{M=3}$ a $XM_{M=3}$ se neobjevuje v regresní rovnici. Odhadované parametry mají následující interpretaci:

- β_0: absolutní člen (konstanta), tj. odhad proměnné Y, když $X = 0$ a $M = 3$,
- β_x: sklon nezávislé proměnné X, když $M = 3$, sklon referenční kategorie,
- β_{m1}: efekt nezávislé proměnné $M = 1$ vs. $M = 3$, když $X = 0$,
- β_{m2}: efekt nezávislé proměnné $M = 2$ vs. $M = 3$, když $X = 0$,
- β_{xm1}: diference ve směrnici (sklonu) X, když $M = 1$ vs. $M = 3$, nebo-li efekt interakce nezávislých proměnných pro $M = 1$ vs. $M = 3$ pro každé dodatečné zvýšení X o jednotku,

- β_{xm2}: diference ve směrnici (sklonu) X, když $M = 2$ vs. $M = 3$, nebo-li efekt interakce nezávislých proměnných pro $M = 2$ vs. $M = 3$ pro každé dodatečné zvýšení X o jednotku.

Jednotlivé členy β_x, β_{m1}, a β_{m2} reprezentují jednoduché sklony a efekty na referenční úrovni interagující proměnné. Interakční členy β_{xm1}, a β_{xm2}, reprezentují změny v jednoduchých sklonech a efektech. Výše uvedená ukázka parametrizace umělých proměnných představuje příklad GLM parametrizace, která je schematicky uvedena v tabulce č. 1. V případě použití jiného způsobu kódování umělých proměnných – například parametrizace efektů – se efekty jednotlivých kategorií vztahují k celkovému průměrnému efektu všech kategorií (blíže např. Řeháková, 2008).

Vztahy pro jednoduché sklony (směrnice) se získají pomocí parciálních derivací závislé proměnné Y vzhledem ke spojité nezávislé proměnné X:

$$\begin{aligned} \frac{\delta Y}{\delta X} &= \beta_x + \beta_{xm1|M=1} + \beta_{xm2|M=2} \\ sklon_x &= \beta_x + \beta_{xm1|M=1} + \beta_{xm2|M=1} \end{aligned} \tag{3.7}$$

Rovnice pro odhad jednoduchého sklonu nezávislé spojité proměnné X pro každou kategorii proměnné M lze konstruovat tak, že se do rovnice sklonu nezávislé spojité proměnné X (3.7) dosadí jedničky a nuly v souladu s parametrizací modelu (viz podkapitola 2.2):

$$\begin{aligned} sklon_{x|M=1} &= \beta_x + \beta_{xm1|M=1} \\ sklon_{x|M=2} &= \beta_x + \beta_{xm2|M=2} \\ sklon_{x|M=3} &= \beta_x \end{aligned} \tag{3.8}$$

K modelování předpovědi úbytku hmotnosti (závislá proměnná $LOSS$) v závislosti na spojité proměnné reprezentující průměrný počet hodin cvičení týdně (proměnná $HOURS$) a kategorické proměnné $PROG$ o 3 úrovních cvičebního programu a jejich vzájemné interakci podle modelu (3.6) se k počátečnímu odhadu využije procedura GLM. V rámci analýzy se budou odhadovat jednoduché sklony průměrného počtu hodin cvičení (proměnná $HOURS$) v jednotlivých programech cvičení (kategorická proměnná $PROG$) a jednoduché efekty jednotlivých cvičebních programů (proměnná $PROG$) na různých úrovních průměrného počtu hodin cvičení (proměnná $HOURS$). Pro analýzy se použije následující programový kód:

```
proc glm data=exercise order=internal;
     class prog;
     model loss = prog hours prog*hours / solution;
     store catcont;
run;
```

Programový kód zajišťuje:

- příkazem class $PROG$ se deklaruje cvičební programy $PROG$ jako kategorická proměnná. SAS vytvoří umělé (binární) proměnné pro každou kategorii $PROG$ a všechny je zadá do regrese (viz podkapitola 2.2 o parametrizaci kategorických (klasifikačních) prediktorů v SAS),

- příkazem order=internal se při použití formátů na proměnnou SAS ve výchozím nastavení změní pořadí úrovní proměnné podle abecedního pořadí formátů. Pro kategorickou (klasifikační) proměnnou $PROG$ byly použita pomocí procedury FORMAT následující formáty: 1 = jogging (běhání), 2 = reading (čtení) a 3 = swimming (plavání). Bez dalších instrukcí systém SAS nastaví v rámci analýz běhání na první úroveň, čtení na druhou a plavání na třetí. Pokud se však mají porovnat 2 programy cvičení vůči programu čtení, pomocí volby order=internal systém SAS nastaví pořadí kategorií původní řazení úrovní.

Tabulka č. 5: Tabulka regresních koeficientů modelu včetně testů významnosti

Source	DF	Type III SS	Mean Square	F Value	Pr > F
prog	2	2901.898322	1450.949161	34.32	<.0001
hours	1	3116.205464	3116.205464	73.71	<.0001
hours*prog	2	5319.048143	2659.524072	62.91	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		2.21637209	B	1.48612022	1.49	0.1362
prog	jogging	-8.99703193	B	2.21602813	-4.06	<.0001
prog	swimming	9.93260912	B	2.17711476	4.56	<.0001
prog	reading	0.00000000	B			
hours		-2.95616163	B	0.70795538	-4.18	<.0001
hours*prog jogging		10.40891093	B	1.07225125	9.71	<.0001
hours*prog swimming		9.83021575	B	1.05144861	9.35	<.0001
hours*prog reading		0.00000000	B			

Zdroj: vlastní zpracování programovacím jazykem SAS

Tabulka č. 5 ilustruje odhady pořízené procedurou GLM. V horní části odhadů je znázorněn *F* celkový test interakce $HOURS$ a $PROG$. Jedná se o společný test odhadů obou interakčních koeficientů, $\beta_{xm1|M=1}$, a $\beta_{xm2|M=2}$, proti 0. Významný test obvykle znamená, že alespoň jeden z koeficientů interakce není roven 0. Sledovaná interakce $HOURS$ a $PROG$ je statisticky významná ($Pr < 0.001$), což vytváří předpoklad tuto interakci hlouběji analyzovat. Obsahem dolní části jsou odhady parametrů modelu včetně testů jejich významnosti. Symbol B v tabulce odhadů informuje, že v důsledku použité parametrizace modelu je matice plánu singulární a k řešení normálních rovnic (metoda nejmenších čtverců) je nutná zobecněná inverze této matice.

Na základě tabulky odhadů na výstupu procedury GLM lze tedy zkonstruovat následující regresní rovnici (bez proměnných s nulovými odhady):

$$\begin{aligned}\widehat{LOSS} &= 2.22 \\ &-2.96 \cdot HOURS \\ &-9.0(PROG = 1) + 9.83(PROG = 2) \\ &+10.41 \cdot HOURS \cdot (PROG = 1) + 9.83 \cdot HOURS \cdot (PROG = 2)\end{aligned} \tag{3.9}$$

Rovnice pro jednoduché sklony nezávislé spojité proměnné $HOURS$ pro jednotlivé úrovně kategorické proměnné $PROG$ lze na základě (3.9) odvodit ve tvaru:

$$
\begin{aligned}
sklon_{HOURS|PROG=1} &= -2.96 + 10.41 = 7.45 \\
sklon_{HOURS|PROG=2} &= -2.96 + 9.83 = 6.87 \\
sklon_{HOURS|PROG=3} &= -2.96 + 0 = -2.96
\end{aligned}
\tag{3.10}
$$

Odhady sklonů spojité proměnné $HOURS$ dle (3.10) lze realizovat pomocí procedury PLM s využitím modelu catcont odhadnutého procedurou GLM. I když koeficient pro β_{xm3} je podle (3.8) omezen na 0, v kódu pro odhad sklonů (směrnic) je potřebné použít hodnotu 1 a pro rovnice odhadu sklonu (směrnice) platí:

$$
\begin{aligned}
sklon_{x|m=1} &= 1 \cdot \beta_x + 1 \cdot \beta_{xm1|M=1} + 0 \cdot \beta_{xm2|M=2} + 0 \cdot \beta_{xm3|M=3} \\
sklon_{x|m=2} &= 1 \cdot \beta_x + 0 \cdot \beta_{xm1|M=1} + 1 \cdot \beta_{xm2|M=2} + 0 \cdot \beta_{xm3|M=3} \\
sklon_{x|m=3} &= 1 \cdot \beta_x + 0 \cdot \beta_{xm1|M=1} + 0 \cdot \beta_{xm2|M=2} + 1 \cdot \beta_{xm3|M=3}
\end{aligned}
\tag{3.11}
$$

K tomuto účelu je nutná příkazová syntaxe, tj.

```
proc plm restore = catcont;
     estimate 'sklon HOURS, PROG=1 běhání' HOURS 1 HOURS*PROG 1 0 0,
          'sklon HOURS, PROG=2 plavání' HOURS 1 HOURS*PROG 0 1 0,
          'sklon HOURS, PROG=3 čtení' HOURS 1 HOURS*PROG 0 0 1 / e;
run;
```

Odhady sklonu (směrnice) proměnné $HOURS$ na jednotlivých úrovních proměnné $PROG$ jsou uvedeny v tabulce č. 6. Parametr *e* v příkazu pro odhady sklonů způsobí, že se ve výstupu procedury PLM zobrazí koeficienty odhadované lineární kombinace (odpovídá rovnici 3.11) odhadnutých parametrů statistického modelu. Parametr *e* může tedy sloužit na kontrolu správnosti odhadované lineární kombinace.

Tabulka č. 6: Odhad sklonu (směrnice) HOURS včetně testů významnosti

Estimates					
Label	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
sklon HOURS, PROG=1 běhání	7.4527	0.8053	894	9.25	<.0001
sklon HOURS, PROG=2 plavání	6.8741	0.7774	894	8.84	<.0001
sklon HOURS, PROG=3 čtení	-2.9562	0.7080	894	-4.18	<.0001

Estimate Coefficients				
Effect	**prog**	**Row1**	**Row2**	**Row3**
Intercept				
prog	jogging			
prog	swimming			
prog	reading			
hours		1	1	1
hours*prog	jogging	1		
hours*prog	swimming		1	
hours*prog	reading			1

Zdroj: vlastní zpracování v programovacím jazyku SAS

Z odhadů sklonu (směrnice) proměnné $HOURS$ na úrovních $PROG$ vyplývá:

- odhadované sklony proměnné $HOURS$ v jednotlivých kategoriích proměnné $PROG$ odpovídají výpočtům v souladu s (3.10) a jsou signifikantní pro všechny úrovně,
- k výpočtu jednoduchého sklonu v referenční skupině není třeba používat příkazy pro odhad; ten je již odhadnut v samotné regresi.

Jedním z úloh regresní analýzy je zjištění, zda se sklony nezávislé proměnné na jednotlivých úrovních kategorické proměnné významně liší. Toto lze zjistit odhady rozdílů sklonů (směrnic) a jejich testováním vůči 0. Odhady rozdílů mezi jednoduchými sklony (směrnicemi) lze dosáhnout pomocí rozdílů hodnot sklonů definovaných podle (3.11), tj.

$$\begin{aligned} sklon_{x|m=2} - sklon_{x|m=1} &= -1 \cdot \beta_{xm1|M=1} + 1 \cdot \beta_{xm2|M=2} + 0 \cdot \beta_{xm3|M=3} \\ sklon_{x|m=3} - sklon_{x|m=1} &= -1 \cdot \beta_{xm1|M=1} + 0 \cdot \beta_{xm2|M=2} + 1 \cdot \beta_{xm3|M=3} \\ sklon_{x|m=2} - sklon_{x|m=3} &= +0 \cdot \beta_{xm1|M=1} - 1 \cdot \beta_{xm2|M=2} + 1 \cdot \beta_{xm3|M=3} \end{aligned} \qquad (3.12)$$

K odhadům rozdílů ve sklonech a testováním jejich statistické významnosti se použije procedura PLM s příkazem ESTIMATE:

```
proc plm restore = catcont;
     estimate
     'rozdíl sklonů HOURS, PROG=1 vs PROG=2' HOURS*PROG -1 1 0,
     'rozdíl sklonů HOURS, PROG=1 vs PROG=3' HOURS*PROG -1 0 1,
     'rozdíl sklonů HOURS, PROG=2 vs PROG=3' HOURS*PROG 0 -1 1 / e;
run;
```

Odhady rozdílů sklonu (směrnice) nezávislé spojité proměnné $HOURS$ mezi různými úrovněmi kategorické proměnné $PROG$ jsou ilustrovány v tabulce č. 7.

Tabulka č. 7: Odhad rozdílů sklonu (směrnice) HOURS včetně testů významnosti mezi různými úrovněmi kategorické proměnné PROG

Estimates					
Label	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
rozdíl sklonů HOURS, PROG=1 vs PROG=2	-0.5787	1.1193	894	-0.52	0.6053
rozdíl sklonů HOURS, PROG=1 vs PROG=3	-10.4089	1.0723	894	-9.71	<.0001
rozdíl sklonů HOURS, PROG=2 vs PROG=3	-9.8302	1.0514	894	-9.35	<.0001

Zdroj: vlastní zpracování v programovacím jazyku SAS

Hodnoty odhadnutých rozdílů sklonů (směrnic) proměnné $HOURS$ pro různé úrovně proměnné $PROG$ nasvědčují tomu, že:

- rozdíl sklonů (směrnic) mezi úrovněmi 1 a 2 proměnné $PROG$ je statisticky nevýznamný, o čemž vypovídá hodnota pravděpodobnosti chyby $(Pr = 0.6053)$,
- rozdíl sklonů (směrnic) úrovně 1 a 2 kategorické proměnné $PROG$ vůči referenční úrovni $(PROG = 3)$ odpovídá hodnotám odhadů interakčních koeficientů uvedených v tabulce č. 5.

Za účelem názornější prezentace jednoduchých sklonů (směrnic) spojité vysvětlující proměnné $HOURS$ lze opět využít post-odhadovou proceduru PLM. Podobně jako v předešlých případech lze k vizualizaci jednoduchých sklonů (směrnic) spojité vysvětlující proměnné $HOURS$ na každé úrovni kategorické proměnné $PROG$ použít proceduru PLM. Graf efektů, resp. jednoduchých sklonů (směrnic) spojité proměnné na jednotlivých úrovních kategorické proměnné je vytvářen příkazem EFFECTPLOT se zadaným parametrem $SLICEBY = PROG$. Jednotlivé směrnice spojité proměnné lze zobrazit včetně pásů spolehlivosti (nejčastěji se používá výchozí spolehlivost odhadu na úrovni 95%, kterou je ale možné také uživatelsky měnit). Sekvence příkazů procedury PLM k vykreslení grafu jednoduchých sklonů (směrnic) napříč všemi úrovněmi kategorické proměnné je

```
proc plm restore=catcont;
     effectplot slicefit (x=HOURS sliceby=PROG) / clm;
run;
```

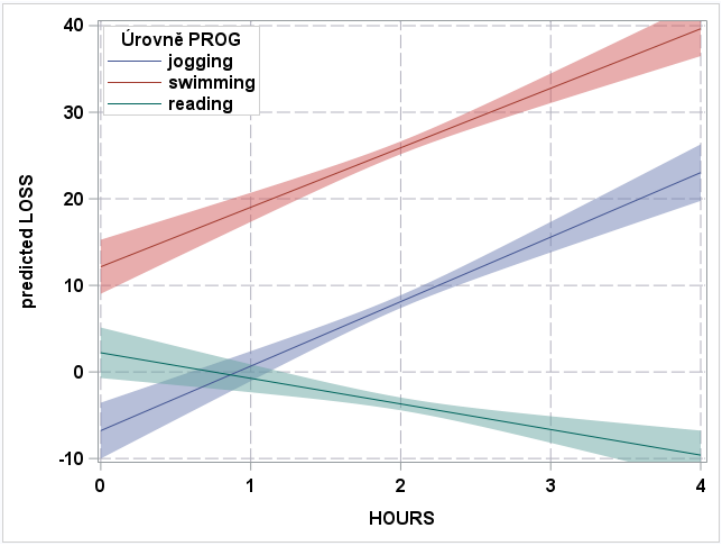
Výše uvedený programový kód procedury PLM zabezpečuje:

- pomocí příkazu EFFECTPLOT zobrazení hodnot spojité nezávislé proměnné $HOURS$ na ose x, přičemž při použití parametru $SLICEBY = PROG$ jsou jednotlivé sklony (směrnice) odhadnuty na každé úrovni kategorické proměnné $PROG$,
- využitím parametru CLM v příkazu EFFECTPLOT se odhadují pásy spolehlivosti pro sklony (směrnice) odhadnuté na každé úrovni kategorické proměnné $PROG$.

Ze sestrojeného grafu je zřetelně vidět, že se jednoduché sklony (směrnice) proměnné $HOURS$ mezi programy běhu (jogging) a plavání (swimming) až na konstantu neodlišují. Naopak sklon (směrnice) proměnné $HOURS$ je v programu čtení (reading) viditelně rozdílný od 2 předešlých na ostatních úrovní proměnné $PROG$.

Graf jednoduchých sklonů (směrnic) spojité proměnné $HOURS$ na jednotlivých úrovních kategorické proměnné $PROG$ s pásy spolehlivosti ilustruje obrázek č. 2.

Obrázek č. 2: Vizualizace sklonů (směrnic) spojité proměnné HOURS



Zdroj: vlastní zpracování

4. ZÁVĚR

Lineární regresní modely představují statistické modely, které se používají v mnoha oblastech přírodních a společenských věd. Vysvětlování pozorovaných hodnot (modelování) výsledkové (závislé) proměnné prostřednictvím jednoho nebo více prediktorů (spojitých nebo kategorických nezávisle proměnných, vysvětlujících proměnných) již po mnoho let patří ke standardním postupům statistické analýzy. I když tyto modely musí splňovat celkem přísné předpoklady pro jejich adekvátní použití (homogenní v rozptylu, nesystematičnost a nekorelovanost odchylek od lineárního vztahu), mnoho výzkumníků s nimi pracuje ve svých analýzách. Po specifikaci pravděpodobnostního rozdělení odchylek od očekávaných hodnot (predikce) jednotlivých pozorování (normální rozdělení s nulovou střední hodnotou a konstantním rozptylem) navíc lineární regresní modely umožňují předvídat rozdělení výsledků, konstruovat intervaly spolehlivosti a testovat statistické hypotézy.

Článek chce přiblížit korektní analýzu uvedených lineárních modelů pomocí analytického systému SAS. Modelování efektů a vizualizace statistických interakcí náleží k nejdůležitějším problémům regresní analýzy. Zjištění, která nezávislá proměnná (prediktor, vysvětlující proměnná) má největší vliv na výsledkovou proměnou, je cílem převážné většiny analytiků. A k tomuto cíli je analytický systém SAS velmi vhodný. Tento softvér je vybaven mnoha funkcionalitami, které je však nutné správně pochopit a v souladu se statistickou teorií aplikovat. Základní otázkou definice statistického modelu v programovém systému SAS je parametrizace, tj. způsob konstrukce matice plánu pro různé typy proměnných (spojité i kategoriální).

I když se ukázky modelování efektů a vizualizace interakce týkají nejjednodušších statistických modelů, koncepce analýzy složitějších modelů je stejná. Článek představuje úvod do statistického modelování v programovém systému SAS.

LITERATURA

Aguinis, H. (2004). *Regression Analysis for Categorical Moderators.* Guilford.

Greene, W. H. (2003). *Econometric Analysis.* 4th ed. Pearson Education.

Pritchard, M. L. & Pasta, D. J. (2004). Head of the CLASS: Impress your colleagues with a superior understanding of the CLASS statement in PROC LOGISTIC [Příspěvek na konferenci]. In *SUGI 29 Proceedings* (pp. 194-29). SAS Institute.

Řeháková, B. (2008) Kontrasty v logistické regresi. *Sociologický časopis/Czech Sociological Review*, *44*(4), 745-765.

Rutherford, A. (2011). *ANOVA and ANCOVA: a GLM Approach*. 2th ed. John Wiley & Sons.

SAS Institute Inc. (2023a). *SAS/STAT® 15.3 User's Guide. Chapter 4: Introduction to Statistical Modelling with SAS/STAT Software.* SAS Institute Inc.

SAS Institute Inc. (2023b). *SAS/STAT® 15.3 User's Guide. Chapter 20: Shared Concepts and Topics.* SAS Institute Inc.

Schwartz, S. & Barrett, T. (2021). The Weight Loss Study. *Encyclopedia of Quantitative Methods in R*, vol. 4: Multiple Linear Regression. https://cehs-research.github.io/eBook_regression/interactions-example.html

RESUMÉ

Modelování efektů, jejich vzájemných interakcí, mnohonásobné porovnávání včetně vhodné vizualizace náleží k nejdůležitějším otázkám analýzy dat. Například sledování účinnosti a interakce léků, navrhování a vyhodnocování experimentů, realizace

a vyhodnocení klinických studií a další představují některé z nejdůležitějších oblastí, ve kterých je modelování efektů a vizualizace interakcí využíváno. Proto je důležité mít vhodné analytické nástroje, jako je například systém SAS, kterými je možné tyto efekty a interakce správně vyhodnocovat.
Programový systém SAS je vybaven mnoha funkcionalitami, které umožňují analytikům správně analyzovat a modelovat hodnocená data. Při analýze dat, modelování efektů a vizualizaci interakcí je však nutné rozlišovat, jestli proměnné vstupující do statistického modelu jsou spojité anebo nespojité, a podle toho zvolit ty správné metody analýzy dat. Základní otázkou analýzy dat a statistického modelování v systému SAS je volba parametrizace modelu. Výběr způsobu konstrukce matice plánu modelu by neměl být samoúčelný, ale měl by odpovídat tomu, co chceme zjistit, resp. jakou hypotézu chceme ověřit. Pro různé typy parametrizace budou odpovídat různé hodnoty parametrů a jejich různá interpretace.

RESUME
Modelling of effects, their mutual interactions, multiple comparisons including appropriate visualisation are among the most important issues of data analysis. For example, monitoring the effectiveness and interaction of drugs, designing and evaluating experiments, conducting and evaluating clinical trials, and others are some of the most important areas in which modelling of effects and visualization of interaction are used. Therefore, it is important to have appropriate analytical tools-such as the SAS system by means of which these effects and interactions can be properly evaluated.
The SAS software system is equipped with many functionalities that allow analysts to correctly analyse and model the evaluated data. However, when analysing data, modelling effects and visualising interactions, it is necessary to distinguish whether the variables entering the statistical model are continuous or discontinuous and to choose the right data analysis methods accordingly. A fundamental issue in data analysis and statistical modelling in the SAS system is the choice of model parameterisation. The choice of the method of construction of the model plan matrix should not be an end in itself, but should correspond to what we want to find out or what hypothesis we want to test. Different parameter values and their different interpretations will correspond to different types of parameterisation.

PROFESIJNÝ ŽIVOTOPIS
Ing. Roman Pavelka, PhD.*, v rokoch 1995 – 2010 pracoval v poradenskej spoločnosti Trexima, s. r. o. Na pozícii štatistik – analytik sa zaoberal najmä analýzami mzdových a personálnych údajov. Podieľal sa na tvorbe pravidelných štatistických prehľadov a správ. Spolupracoval s akademickými pracoviskami, agentúrami i súkromnými subjektmi na realizácii a vyhodnocovaní ad hoc štatistických výskumov. Oblasť jeho vedeckého záujmu predstavujú výberové zisťovania, odhady a štatistické modely. V rokoch 2012 až 2013 sa zúčastnil na zahraničnej stáži v Spojenom kráľovstve. Od roku 2013 pôsobil v Národnom ústave certifikovaných meraní vzdelávania (NÚCEM), kde zaisťoval štatistické vyhodnocovanie výsledkov testovania žiakov a študentov. Od roku 2015 pracuje v odbore metód štatistických zisťovaní Štatistického úradu SR.*

KONTAKT
roman.pavelka@statistics.sk