SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

4/2025
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: 3

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

Strany/Pages: 52 – 67

Dátum vydania/Publication date: 15. október 2025/October 15, 2025

Informatívny článok/Information article

Roman PAVELKA
Štatistický úrad Slovenskej republiky

PARAMETRIZACE MODELU A MODELOVÁNÍ LINEÁRNÍCH HYPOTÉZ VYUŽITÍM SYSTÉMU SAS

A MODEL PARAMETRISATION AND MODELLING OF LINEAR HYPOTHESES USING SAS

ABSTRAKT
Cílem článku je přiblížit metody tvorby a interpretace lineárních hypotéz vycházejících z lineárních regresních modelů s využitím analytických možností vybraných procedur statistického systému SAS. Parametry sledovaných statistických modelů budou odhadovány především řešením normálních rovnic metodou nejmenších čtverců. Analytické procedury v systému SAS dokáží nejen odhadovat parametry statistických modelů, ale jsou vybaveny také funkcionalitou k odhadům lineárních funkcí parametrů statistického modelu. Odhady lineárních funkcí odhadovaných parametrů modelů tak vytvářejí podmínky pro testování obecných lineárních hypotéz, které mohou být konstruovány od jednoduchých až po komplexní porovnání. Modelování zkoumaných statistických hypotéz bude realizováno s pomocí různých způsobů parametrizace regresních efektů statistického modelu, které svým programovým vybavením umožňuje analytický systém SAS. Různá parametrizační pravidla, a tedy různý způsob konstrukce matice plánu statistického modelu, budou proto uplatňovány při tvorbě a modelování lineárních hypotéz. Pro realizaci modelování lineárních hypotéz budou v systému SAS využity analytické procedury REG, GLM a postodhadová procedura PLM.

ABSTRACT
The aim of this article is to present methods for constructing and interpreting linear hypotheses based on linear regression models using the analytical capabilities of selected procedures of the SAS statistical system. The parameters of the examined statistical models will be estimated primarily by solving normal equations using the least squares method. The analytical procedures in the SAS system can not only estimate the parameters of statistical models, but are also equipped with functionality for estimating linear functions of the parameters of the statistical model. Estimates of linear functions of the estimated model parameters thus provide a basis for testing general linear hypotheses, which can range from simple to complex comparisons. The modelling of the examined statistical hypotheses will be implemented using various methods of parameterization of the regression effects of the statistical model, as enabled by the SAS analytical system with its software. Various parameterization rules, and thus different methods of constructing the design matrix of the statistical model, will therefore be applied in the construction and modelling of linear hypotheses. For the implementation of linear hypothesis modelling in the SAS, the analytical procedures REG, GLM and the postestimation procedure PLM will be used.

KLÍČOVÁ SLOVA
analytický systém SAS, lineární hypotéza, odhadnutelná funkce, statistické testování

KEY WORDS
analytical system SAS, linear hypothesis, estimable function, statistical testing

1. ÚVOD

V mnoha studiích se datoví analytici zajímají o testování různých hypotéz. Výchozí výstupy z analytických procedur, který SAS generuje, je omezen na odhady parametrů modelu a porovnání specifických parametrů. V průběhu let byly v systému SAS vyvinuty procedury a příkazy pro odhad lineárních funkcí těchto parametrů a pro testování obecných lineárních hypotéz (v anglickém originále *General Linear Hypothesis*, ve zkratce dále jen jako GLH) pro lineární modely. GLH lze konstruovat jako jednoduchá nebo komplexní porovnání. Mnoho procedur pro statistické modelování, jako například GLM pro modelování obecných lineárních modelů, GENMOD pro modelování nespojitých náhodných proměnných s uživatelsky definovanou korelační strukturou, LOGISTIC pro logistickou regresi a odhady poměru šancí, MIXED pro lineární smíšené modely, GLIMMIX k modelování zobecněných smíšených lineárních modelů, jsou těmito funkcionalitami vybaveny. Konstrukce uvedených příkazů pro generování správného výstupu pro požadovaný odhad nebo hypotézu je pro mnohé analytiky ne zcela jasná. Předkládaný článek chce přiblížit, jak definovat lineární funkci na základě statistického modelu, ilustruje potřebnou syntaxi příkazů a ukazuje různé parametrizace a uspořádání na úrovni faktorů. Součástí tohoto přehledu jsou také uvedeny příklady, které ilustrují celý proces.

2. OBECNÝ LINEÁRNÍ MODEL A TESTOVÁNÍ LINEÁRNÍCH HYPOTÉZ

2.1. OBECNÝ LINEÁRNÍ MODEL (GLM) A ANAÝZA ROZPTYLU

Východiskem pro modelování lineárních hypotéz bude obecný lineární model, který je podle Rutherforda (2011) dán výrazem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

kde

- $\mathbf{Y} = (y_1, \dots, y_n)^T$ je $(n\ x\ 1)$ vektor n hodnot závislých proměnných,
- $\mathbf{X}$ je matice plánu (matice modelu) o rozměrech $n\ x(k+1)$ s n řádky, které odpovídají n pozorováním, a $(1+k)$ sloupci, kde sloupec jedniček představuje absolutní člen a k sloupců reprezentuje počet nezávislých proměnných, tj.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \tag{2.2}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^T$ je sloupcový vektor o rozměrech $(k+1)x1$ obsahující k neznámých regresních parametrů a konstantu β_0 reprezentující absolutní člen,
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ je $(n\ x\ 1)$ vektor n nepozorovatelných chyb modelu.

Pro ukázku možné analýzy v prostředí SAS je výhodné obecný lineární model (2.1) přeformulovat do tvaru

$$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk}, \tag{2.3}$$

kde

- y_{ijk} je *k*-té pozorování vysvětlované (závislé) na *i*-té úrovně faktoru *A* a *j*-té úrovni faktoru *B*,
- μ je absolutní člen (intercept),
- A_i je *i*-tá úroveň faktoru *A*,
- B_j je *j*-tá úroveň faktoru *B,*
- $(AB)_{ij}$ je interakce mezi hodnotou *i*-té úrovně faktoru *A* a *j*-té úrovně faktoru *B*,
- ε_{ijk} je *k*-tá reziduální hodnota *i*-té úrovně faktoru *A* na *j*-té úrovni faktoru *B*.

Náhodné chyby ε_{ijk} jsou nezávislé a stejně rozdělené s rozdělením $N(0, \sigma_E^2)$.

Tvar modelu podle (2.3) popisuje obecný lineární model se 2 faktory (označené *A* a *B*) s jejich vzájemnou interakcí $(AB)_{ij}$. Pro účely analýzy rozptylu (porovnání průměrů na různých úrovních obou faktorů) je vhodné model (2.3) přepsat do tvaru

$$y_{ijk} = \mu + (\mu_{i\cdot} - \mu) + \left(\mu_{\cdot j} - \mu\right) + \left(\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu\right) + \varepsilon_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad (2.4)$$

kde

- y_{ijk} je *k*-té pozorování *i*-té úrovně faktoru A na *j*-té úrovni faktoru B,
- μ je absolutní člen (intercept),
- $(\mu_{i\cdot} - \mu)$ je efekt faktoru *A*, kde $\mu_{i\cdot}$ je marginální střední hodnota *i*-té úrovně faktoru *A*,
- $\left(\mu_{\cdot j} - \mu\right)$ je efekt faktoru *B*, kde $\mu_{\cdot j}$ je marginální střední hodnota *j*-té úrovně faktoru *B*,
- $\left(\mu_{\cdot ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu\right)$ je efekt interakce *i*-té úrovně faktoru *A* a *j*-té úrovně *B*,
- indexy obsahující symbol · označují průměrné hodnoty na dané úrovni.

Symbolický zápis vektoru parametrů **β** jako ukázka struktury efektů podle (2.3) je:

$$\boldsymbol{\beta} = (\mu\ \alpha_1\ \alpha_2\ \alpha_3\ \ \beta_1\ \ \beta_2\ (\alpha\beta)_{11}\ \ (\alpha\beta)_{12}\ (\alpha\beta)_{21}\ (\alpha\beta)_{22}\ (\alpha\beta)_{31}\ (\alpha\beta)_{32}\ ) \quad (2.5)$$

Programový systém SAS je k analýze rozptylu vybaven několika statistickými procedurami. Pro zpracování vyvážených dat (tj. dat se stejným počtem pozorování pro každou kombinaci klasifikačních faktorů) je určena především procedura ANOVA. Pokud každá kombinace úrovní klasifikačních faktorů neobsahuje stejný počet pozorování (tj. jedná se o nevyvážená data), musí se použít procedura GLM. Procedura GLM poskytuje analýzu vyvážených i nevyvážených dat s možností odhadu parametrů obecného lineárního modelu (SAS Institute Inc., 2025a).

Tabulka č. 1: Analýza rozptylu (ANOVA) obecného lineárního modelu se 2 faktory

Source	DF	Sum of Squares	Mean Square	F Value
Faktor A	$\mathrm{n}_A - 1$	$SS_A = n_B * n_I * \sum_{i=1}^{n_A} (\mu_{i\cdot} - \mu)^2$	$MS_A = \frac{\mathrm{SS}_A}{\mathrm{n}_A - 1}$	$F_A = \frac{\mathrm{MS}_A}{\mathrm{MS}_E}$
Faktor B	$\mathrm{n}_B - 1$	$SS_B = n_A * n_I * \sum_{j=1}^{n_B} (\mu_{\cdot j} - \mu)^2$	$MS_B = \frac{\mathrm{SS}_B}{\mathrm{n}_B - 1}$	$F_B = \frac{\mathrm{MS}_B}{\mathrm{MS}_E}$
Interakce A*B	$(\mathrm{n}_A - 1) * (\mathrm{n}_B - 1)$	$SS_{AB} = n_I * \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (\mu_{\cdot ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu)^2$	$MS_{AB} = \frac{\mathrm{SS}_{AB}}{(\mathrm{n}_A - 1) * (\mathrm{n}_B - 1)}$	$F_{AB} = \frac{\mathrm{MS}_{AB}}{\mathrm{MS}_E}$
Error	$\mathrm{n}_A * \mathrm{n}_B * (\mathrm{n}_I - 1)$	$SS_E = \sum_{k=1}^{n_I} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (y_{ijk} - \mu_{ij})^2$	$MS_E = \frac{\mathrm{SS}_E}{\mathrm{n}_A * \mathrm{n}_B * (\mathrm{n}_I - 1)}$	
Corrected Total	$\mathrm{n}_A * \mathrm{n}_B * \mathrm{n}_I - 1$	$SS_T = \sum_{k=1}^{n_I} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (y_{ijk} - \mu)^2$		

*Poznámka: n_A označuje počet pozorování pro faktor A; n_B označuje počet pozorování pro faktor B; n_I označuje počet pozorování interakce faktorů A*B; · označují průměrné hodnoty na dané úrovni.*

***Zdroj**: vlastní zpracování podle SAS Institute Inc. (2025a)*

Výsledkem analýzy rozptylu procedury GLM je tabulka analýzy rozptylu s relativními podíly rozptylu připadajícími na vliv jednotlivých faktorů a jejich interakce na rozptylu celkovém. Jednotlivé podíly představují statistiky *F* určené k testování významnosti faktorů a jejich interakce v modelu. Vyšší hodnota pravděpodobnosti (*p*-value > 0,05) nasvědčuje platnosti nulové hypotézy o nevýznamnosti vlivu faktoru anebo interakce. Pokud je toto číslo nižší než $0{,}05$[1], je obvykle vliv faktoru, resp. interakce považován za dostatečně významný. Matematické vztahy pro analýzu rozptylu dvoufaktorového modelu s interakcemi jsou zaznamenány v tabulce č. 1.

Regresorové proměnné v modelu (nazývané také jako tzv. efekty) mohou být spojité i nespojité, tzv. klasifikační efekty (zpravidla považované za hlavní efekty). Pokud jsou všechny regresorové proměnné (efekty) v modelu spojité, regresní model je tvořen výlučně lineární kombinací regresorových proměnných $\mathbf{X}$, přičemž každá spojitá proměnná (efekt) v modelu přispívá jedním sloupcem do matice plánu (matice modelu) $\mathbf{X}$, a tedy jedním parametrem do modelu jako celku. Klasifikační efekt (pevný i náhodný) je naopak spojen s více než jedním sloupcem matice $\mathbf{X}$.

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., 2025a). 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

[1] Formálně lze p-hodnotu definovat i jako nejmenší hladinu významnosti testu, při níž na daných datech ještě zamítneme nulovou hypotézu. Platí tedy, že čím nižší p-hodnota testu je, tím menší nám tento test indikuje pravděpodobnost, že platí nulová hypotéza (Pavlík & Dušek, 2012).

skupinami pozorování. Zpravidla se klasifikační efekty v regresním modelu označují jako hlavní efekty. Způsob konstrukce matice plánu (matice modelu) **X** charakterizuje tzv. parametrizace.

2.2. PARAMETRIZACE REGRESNÍCH EFEKTŮ MODELU

Parametrizace regresních efektů modelu popisuje způsob konstrukce matice plánu (matice modelu). Tato parametrizační pravidla (SAS Institute Inc., 2025a) jsou aplikována na regresní modely, modely s klasifikačními efekty i na všechny ostatní statistické modely.

Nejjednodušší a nejobecnější parametrizací regresních modelů je tzv. parametrizace GLM, jejíž název byl odvozen podle definice uvedené v kapitole 2.1. V rámci parametrizace GLM je matice plánu vytvářena sloupcovým vektorem jedniček reprezentujícím absolutní člen regresní rovnice. Podobně tak i numerické proměnné nazývané regresními efekty mohou být zahrnuty do modelu jako sloupcové vektory. Má-li klasifikační proměnná k úrovní, potom parametrizace GLM 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. Parametrizace GLM 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 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 s klasifikační proměnnou (faktorem) A se 3 úrovněmi, A_1, A_2 a A_3, podle parametrizace GLM je ilustrován v tabulce č. 2.

Tabulka č. 2: Ukázka parametrizace GLM faktoru A se 3 úrovněmi (A_1, A_2 a A_3)

Faktor A	Indikátorové proměnné		
	D1	D2	D3
A_1	1	0	0
A_2	0	1	0
A_3	0	0	1

Zdroj: vlastní zpracování podle SAS Institute Inc. (2025a)

Referenční parametrizace (typu REF) pro kódování klasifikačních proměnných standardně používá jako referenční úroveň poslední úroveň, pokud není uvedeno jinak. Ukázka referenční parametrizace faktoru A se 3 hladinami je ilustrována v tabulce č. 3.

Tabulka č. 3: Ukázka REF parametrizace faktoru A se 3 úrovněmi (A_1, A_2 a A_3)

Faktor A	Umělé proměnné	
	D_1	D_2
A_1	1	0
A_2	0	1
A_3	0	0

Zdroj: vlastní zpracování podle SAS Institute Inc. (2025a)

Při referenční parametrizaci faktoru A do matice plánu vstupují 2 umělé proměnné a referenční kategorie A_3 (poslední řádek v tabulce č. 2) obsahuje 0.

Parametrizace faktoru A způsobem EFFECT v referenční kategorii A_3 namísto 0 použije -1. Podle Řehákové (2008) je také nazývána jako parametrizace typu DEVIATION. Uvedený způsob parametrizace je ilustrován v tabulce č. 4.

Tabulka č. 4: Ukázka EFFECT parametrizace faktoru A se 3 úrovněmi (A_1, A_2 a A_3)

Faktor A	Umělé proměnné	
	D_1	D_2
A_1	1	0
A_2	0	1
A_3	-1	-1

Zdroj: vlastní zpracování podle SAS Institute Inc. (2025a)

Většina analytických procedur ze systému SAS, které jsou v rámci své funkcionality vybaveny příkazem CLASS, uplatňuje jako výchozí (implicitní) parametrizaci klasifikačních proměnných a efektů parametrizaci GLM. Vybrané procedury navíc mají jako možnost parametrem PARAM= tuto výchozí parametrizaci změnit včetně výběru referenční kategorie pomocí parametru REF=. Volba parametrizace statistického modelu, tj. konstrukce matice plánu, má velký vliv na interpretaci odhadovaných parametrů statistického modelu. U většiny analytických procedur referenční úroveň klasifikační proměnné (faktoru) je možné specifikovat programovým příkazem podle požadavků uživatele.

Výše uvedené typy parametrizací patří k nejpoužívanějším pro odhady parametrů efektů klasifikační proměnné. V regresní analýze v systému SAS je možné nastavit pro proměnné i jiné typy parametrizace, které se však používají ve speciálních případech. Podle způsobu parametrizace jsou interpretovány odhadované parametry efektů.

2.3. ODHADNUTELNÉ FUNKCE A OBECNÉ LINEÁRNÍ HYPOTÉZY

Odhadnutelná funkce parametrů (Littell et al., 2002) je definována jako odhadovaná lineární kombinace $\mathbf{L\beta}$, ve které je vektor $\mathbf{L}$ konstruován podobně jako vektor parametrů modelu s vahami odpovídajícími každému efektu v modelu. Formálně lze vektor vah $\mathbf{L}$ vyjádřit

$$\mathbf{L} = \begin{pmatrix} Váha\ pro \\ Intercept \end{vmatrix} \begin{matrix} Váha\ pro \\ Faktor\ A \end{vmatrix} \begin{matrix} Váha\ pro \\ Faktor\ B \end{vmatrix} \begin{matrix} Váha\ pro \\ Interakci \end{matrix} \end{pmatrix} \tag{2.6}$$

Počet jedinečných symbolů ve vektoru představuje maximální počet lineárně nezávislých koeficientů odhadovaných modelem, který je roven hodnosti matice $\boldsymbol{X'X}$. Odhadnutelné funkce $\mathbf{L\beta}$ se vyznačují následujícími vlastnostmi:

- $\mathbf{L}\widehat{\boldsymbol{\beta}}$ a kovarianční matice $VAR(\mathbf{L}\widehat{\boldsymbol{\beta}})$ jsou jedinečné,
- $\mathbf{L}\widehat{\boldsymbol{\beta}}$ je jedinečným odhadem $\mathbf{L\beta}$, jejichž kovarianční matice je dána výrazem:

$$VAR(\mathbf{L}\widehat{\boldsymbol{\beta}}) = \sigma_\varepsilon^2[\mathbf{L}(\mathbf{X'WX})^{-}\mathbf{L'}], \tag{2.7}$$

kde

- $\mathbf{L}$ je vektor prvků reprezentující váhy odpovídající každému efektu,
- $\mathbf{X}$ je matice modelu,
- $\mathbf{W}$ je matice vah (v případě klasického lineárního modelu je jednotková),
- σ_ε^2 je rozptyl náhodné složky v modelu.

Například funkce střední hodnoty vycházející z rovnice (2.3) určená k odhadu střední hodnoty μ_1. úrovně 1 faktoru A, tj. A_1, se vyjádří ve tvaru:

$$\mu_{1\cdot} = \mu + \alpha_1 + \frac{\beta_1}{2} + \frac{\beta_2}{2} + \frac{(\alpha\beta)_{11}}{2} + \frac{(\alpha\beta)_{12}}{2}. \tag{2.8}$$

Vektor vah **L**, který odpovídá funkci odhadu podle (2.7), je definován jako:

$$\mathbf{L} = \left(1 \mid 1\ 0\ 0 \mid \frac{1}{2}\ \frac{1}{2} \middle| \frac{1}{2}\ \frac{1}{2}\ 0\ 0\ 0\ 0\right) \tag{2.9}$$

Váhy ve vektoru **L** musí být podle Rutherforda (2011) pro každý faktor rozděleny rovnoměrně a pro každý faktor musí být součet vah roven 1.

Obecná lineární hypotéza je založena na konceptu odhadnutelné funkce parametrů a je podle SAS Institute Inc. (2025a) definována jako

$$H\colon \mathbf{L}\boldsymbol{\beta} = \mathbf{0}. \tag{2.10}$$

Konkrétní tvar obecné lineární hypotézy (2.10) je vymezen lineární kombinací odhadovaných parametrů, a tedy vektorem vah **L**, tj. parametrizací modelu.

2.4. VYUŽÍVANÉ ANALYTICKÉ PROCEDURY STATISTICKÉHO SYSTÉMU SAS

Následující analýzy a modelování efektů, vizualizace interakcí a testování uživatelsky vytvořených hypotéz budou ve statistickém systému SAS realizovány analytickými procedurami REG, GLM a postodhadovou procedurou PLM.

Analytické procedury svou funkcionalitou zajišťuje v systému SAS různé analýzy a vykreslování funkcí včetně testování hypotéz o efektech modelu a kontrastech, výpočtu předpovídaných hodnot výsledku (skóre) a vykreslování těchto předpovědí.

3. UKÁZKY PARAMETRIZACE A MODELOVÁNÍ LINEÁRNÍCH HYPOTÉZ

3.1. SOUBOR DAT POUŽITÝ K ANALÝZÁM

Datový soubor, na kterém budou demonstrovány příklady konstrukce datových matic i ukázky tvorby uživatelských lineárních hypotéz, se skládá z údajů 200 studentů popisujících pohlaví (proměnná FEMALE), socioekonomický status (SES), typ navštěvované školy (SCHTYP) a etnický původ (RACE)[2]. Soubor obsahuje také řadu výsledků standardizovaných testů, včetně testů čtení (READ), psaní (WRITE), matematiky (MATH), vědy (SCIENCE) a společenských věd (SOCST) vystupujících v roli spojitých proměnných. Ukázky kódování budou prezentovány na kategorické proměnné RACE, jejímž obsahem jsou 4 úrovně (1 = Hispánec, 2 = Asiat, 3 = Afroameričan a 4 = běloch). Jako proměnná závislá je zvolena proměnná WRITE

[2] *High School and Beyond Longitudinal Study je dostupné na http://nces.ed.gov/surveys/hsb/.*

s výsledky testů ze psaní. Před vlastní analýzou je vhodné mít informaci o průměrných hodnotách závislé proměnné WRITE pro každou úroveň kategorické proměnné RACE. Znalost průměrných hodnot (viz tabulka č. 5) napomáhá interpretovat výsledky analýz.

Tabulka č. 5: Počet pozorování a průměry na úrovních kategorické proměnné RACE

Analysis Variable : write			
race	N Obs	Mean	N
1	24	46,4583333	24
2	11	58,0000000	11
3	20	48,2000000	20
4	145	54,0551724	145

Zdroj: vlastní zpracování podle Dostála (2021)

Celkový průměr ze všech 200 hodnot bez ohledu na rasu je $\overline{WRITE} = 52{,}775$ a průměr průměrů, tedy $(46{,}46 + 58{,}00 + 48{,}20 + 54{,}06)/4$, je $\overline{\overline{WRITE}} = 51{,}679$.

3.2. KÓDOVÁNÁNÍ NESPOJITÝCH REGRESORŮ A TESTOVÁNÍ HYPOTÉZ

Obecná lineární hypotéza (2.10) bude konstruována jako lineární kombinace odhadovaných parametrů pomocí vektoru vah (2.6) s využitím procedury:

- GLM, kde jsou klasifikační proměnné do regresního modelu zahrnuty programově, a to pomocí příkazu CLASS. Obecné lineární hypotézy jsou realizovány příkazem CONTRAST nebo ESTIMATE (provádí nejen statistické testování, ale i odhady kombinace parametrů), kde vektor vah pro odhadovatelnou funkci parametrů je vkládán analytikem na základě požadované statistické hypotézy,
- REG, kde klasifikační proměnné je nutno před vložením do regresního modelu překódovat do $k-1$ nových proměnných (kde k je počet úrovní kategoriální proměnné) v samostatném datovém kroku a použít tyto nové proměnné jako vysvětlované proměnné v regresním modelu.

Váhy vkládané do příkazu ESTIMATE (případně CONTRAST) procedury GLM jsou podle (Rutherford, 2011) považovány za kontrasty (viz např. 2.9). Vytvoření obecné lineární hypotézy v proceduře REG je podstatně složitější, protože vyžaduje tzv. reparametrizaci[3]. Původní matice plánu $\mathbf{X}$ modelu v proceduře REG musí změnit na upravenou matici modelu $\boldsymbol{X}^*$, která zajistí požadovanou lineární hypotézu. Na základě konceptu odhadnutelné funkce parametrů (Littell et al., 2002) lze reparametrizovanou matici plánu pro odhady a testování v proceduře REG vyjádřit ve tvaru:

$$\mathbf{X}^* = \mathbf{L}(\mathbf{L}'\mathbf{L})^{-\mathbf{1}}, \tag{3.1}$$

kde

- $\mathbf{L}$ je vektor vah ESTIMATE (resp. CONTRAST) procedury GLM,
- $\mathbf{X}$ je původní matice modelu pro proceduru REG,
- $\mathbf{X}^*$ je reparametrizovaná matice modelu podle (3.1) pro proceduru REG.

[3] *Reparametrizace znamená změnu konstrukce matice plánu (matice modelu) – poznámka autora.*

Pro účely toho příspěvku se budou váhy pro proceduru GLM nazývat jako váhy kontrastů, váhy v modelu pro proceduru REG váhy regresní. Pro účely ukázky byly vybrány systémy kódování pomocí fiktivních (tzv. dummy) proměnných a nejpoužívanější systémy kódování efektů.

Princip zahrnutí nespojitých regresorů do lineárního regresního modelu byl již naznačen v kapitole 2.2. Ve většině případů pro statistickou analýzu dat nebude použití parametrizace GLM postačovat. Ve skutečnosti však existuje celá řada jiných kódování uplatnitelných ve statistické analýze dat. Uživatelská kódování dávají analytikům možnosti k testování různých, i komplikovanějších hypotéz při srovnávání uživatelsky definovaných skupin. Každý způsob kódování, který bude v následujícím textu zmiňován, bude aplikován na matici plánu s vektorem vah kontrastů a vektorem regresních vah po provedených odhadech.

Podle Řehákové (2008) pro obecné lineární hypotézy (kontrasty) existují 2 základní typy kódovacích systémů: kódování pomocí fiktivních proměnných a kódování efektů.

Kódování využitím fiktivních proměnných (parametrizace GLM)

Kódování využitím fiktivních proměnných představuje nejjednodušší a nejběžnější systém kódování. Při uvedeném kódování se na základě kategoriální proměnné vytvoří série dichotomických proměnných (nabývají pouze hodnot 0 a 1). Pro všechny úrovně kategoriální proměnné kromě jedné bude vytvořena nová proměnná, která má pro každé pozorování na dané úrovni hodnotu 1 a pro všechny ostatní 0. V příkladu s proměnnou RACE bude mít první nová proměnná (X_1) hodnotu 1 pro každé pozorování, ve kterém je osoba Hispánec, a 0 pro všechna ostatní pozorování. Podobně se vytvoří proměnná X_2 s hodnotu 1, pokud je osoba Asiat, a 0 jinak, a X_3 s hodnotou 1, pokud je osoba Afroameričan, a 0 jinak. Úroveň kategoriální proměnné, která je ve všech nových proměnných kódována jako nula, je referenční úroveň neboli úroveň, se kterou jsou porovnávány všechny ostatní úrovně. V našem příkladu je referenční úrovní úroveň 4 (běloch). Jako referenční úroveň je možné vybrat libovolnou úroveň kategoriální proměnné.

Po vytvoření nových proměnných X_1, X_2 a X_3 se tyto zadají do rovnice regrese (původní proměnná se nezadává). Výstup regrese bude obsahovat koeficienty pro každou z těchto nových proměnných. Koeficient pro X_1 představuje průměr závislé proměnné pro skupinu 1 mínus průměr závislé proměnné pro referenční (vynechanou) skupinu. Toto platí podobně pro koeficienty proměnné X_2 a X_3. Váhy kontrastů i váhy regresní jsou totožné a znázorněné v tabulce č. 6

Tabulka č. 6: Ukázka vektorů vah – kódování pomocí fiktivních proměnných

	ANOVA model (GLM)			Regresní model (REG)		
	L_1	L_2	L_3	X_1	X_2	X_3
1 (Hispánec)	1	0	0	1	0	0
2 (Asiat)	0	1	0	0	1	0
3 (Afroameričan)	0	0	1	0	0	1
4 (běloch)	0	0	0	0	0	0

Zdroj: vlastní zpracování podle UCLA (n.d.)

Kódování efektů

Tyto kódovací systémy na rozdíl od kódování s fiktivními proměnnými umožňují provádět i jiné typy porovnání, protože používají více hodnot než jen nulu a jednu. Kódování efektů umožňuje přiřadit různé váhy různým úrovním kategoriální proměnné. Zatímco pro fiktivní proměnné platí, že platné jsou pouze hodnoty 0 a 1; v kódování efektů platí, že součet všech kódových hodnot v jakékoli nové proměnné musí být roven 0. Není příliš důležité, které úrovni je přiřazena kladná nebo záporná hodnota, protože 0 1 -1 0 je totéž jako 0 -1 1 0 v tom, že obě tato kódování porovnávají druhou a třetí úroveň proměnné, i když znaménko koeficientu se změnilo. Ačkoliv se dá použít fiktivní proměnné s jakýmkoli typem kategoriální proměnné, některé formy kódování efektů dávají větší smysl s ordinálními kategoriálními proměnnými než s nominálními kategoriálními proměnnými. Tabulka č. 7 představuje nejdůležitější systémy kódování, které náleží do obecné skupiny kódování efektů.

Tabulka č. 7: Hlavní způsoby kódování efektů kategorických proměnných

Typ kódování efektů	Realizované porovnání
Simple	Porovnává každou úroveň proměnné s první úrovní (nebo s jakoukoli zadanou úrovní)
Deviation	Porovnává odchylky průměrů na jednotlivých úrovních od celkového průměru
Difference	Porovnává úrovně proměnné s průměrem předchozích úrovní proměnné
Helmert	Porovnává úrovně proměnné s průměrem následujících úrovní proměnné
Repeated	Porovnává sousední úrovně proměnné
Speciální	Uživatelsky definovaný kontrast

Zdroj: vlastní zpracování podle SAS Institute Inc. (2025b)

Kódování efektů (typu Simple)

Tento typ kódování je vhodný tehdy, když se sleduje porovnání efektů různých kategorií s efektem jedné úrovně, která se nazývá referenční. Lze ho použít jak pro nominální, tak pro ordinální proměnné, i když pro ordinální proměnné jsou vhodnější jiné typy. Podstata jednoduchého kódování spočívá v tom, že efekt na každé úrovni je porovnáván s efektem na referenční úrovni. V níže uvedeném příkladu je referenční úrovní úroveň 4 a první porovnání posuzuje úroveň 1 s úrovní 4, druhé porovnání srovnává úroveň 2 s úrovní 4 a třetí porovnání srovnává úroveň 3 vůči úrovni 4.

Hodnoty vektorů vah pro model ANOVA (procedura GLM) i pro klasickou regresi (procedurou REG) pro jednotlivé úrovně klasifikační proměnné jsou uvedeny v tabulce č. 8. Oba vektory spolu souvisejí a vzájemný vztah vektoru vah modelu pro proceduru GLM a vektoru vah pro proceduru REG (reparametrizace statistického modelu) je dán rovnicí (3.1).

Tabulka č. 8: Ukázka vektorů vah – jednoduché kódování

	model ANOVA (GLM)			Regresní model (REG)		
	L_1	L_2	L_3	X_1	X_2	X_3
1 (Hispánec)	1	0	0	$\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
2 (Asiat)	0	1	0	$-\frac{1}{4}$	$\frac{3}{4}$	$-\frac{1}{4}$
3 (Afroameričan)	0	0	1	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{3}{4}$
4 (běloch)	-1	-1	-1	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$

Zdroj: vlastní zpracování podle UCLA (n.d.)

Při porovnání hodnot závislé proměnné na jednotlivých úrovních kategorické proměnné RACE při jednoduchém kódování efektů se pomocí procedury GLM pro každý kontrast použije samostatný příkaz ESTIMATE. Srovnání hodnot (včetně statistických hypotéz) závisle proměnné na jednotlivých úrovních (příp. i kombinací úrovní) kategorické proměnné RACE ilustruje tabulka č. 9.

Tabulka č. 9: Odhady lineárních kombinací parametrů modelu vč. hypotéz

Parameter	Estimate	Standard Error	t Value	Pr > |t|
1 (Hispánec) vs. 4 (běloch)	-7,60	1,99	-3,82	0,0002
2 (Asiat) vs. 4 (běloch)	3,94	2,82	1,40	0,1638
3 (Afroameričan) vs. 4 (běloch)	-5,86	2,15	-2,72	0,0071

Zdroj: vlastní zpracování podle SAS Institute Inc. (2025a)

Ukázka příkazu procedury GLM pro odhady lineární kombinace parametrů včetně obecných lineárních hypotéz s kódováním efektů (typu Simple) kategoriální proměnné RACE je uvedena ve tvaru:

```
proc glm data = work.HSB2;
      class race;
      model write = race;
      estimate '1 (Hispánec) vs. 4 (běloch)' race 1 0 0 -1;
      estimate '2 (Asiat) vs. 4 (běloch)' race 0 1 0 -1;
      estimate '3 (Afroameričan) vs. 4 (běloch)' race 0 0 1 -1;
run;

quit;
```

Kódování efektů (typu Deviation)

Používá se v případě, pokud je nutné sledovat efekty jednotlivých kategorií vzhledem k průměrnému efektu všech kategorií. Kódovací systém porovnává průměr závislé proměnné WRITE (průměrný efekt) pro danou úroveň kategorické proměnné RACE s celkovým průměrem této závislé proměnné. První srovnání porovnává průměr na úrovni 1 s celkovým průměrem proměnné WRITE, druhé porovnání posuzuje průměr na úrovni 2 vůči celkovému průměru a třetí srovnání porovnává úroveň 3 vůči celkovému průměru proměnné WRITE.

V regresním vektoru vah je hodnota 1 pro novou proměnnou X_1 na úrovni 1 proměnné RACE, úroveň 2 a úroveň 3 je kódována v proměnné X_1 hodnotou 0. Nová proměnná X_2 nabývá hodnoty 1 pro úroveň 2, nová proměnná X_3 má hodnotu 1 pro úroveň 3. Ve všech 3 porovnáních je úrovni 4 kategoriální proměnné RACE přiřazena hodnota -1. Ilustrace vektoru vah pro regresi je v tabulce č. 10.

U vektoru vah kontrastů první srovnání, které porovnává úroveň 1 s úrovněmi 2, 3 a 4, přiřadí úrovni 1 hodnotu 3/4 (0,75) a úrovním 2, 3 a 4 hodnotu -1/4 (0,25). Podobně druhé srovnání, které porovnává úroveň 2 s úrovněmi 1, 3 a 4, přiřadí úrovni 2 hodnotu 3/4 (0,75) a úrovním 1, 3 a 4 hodnotu -1/4 (0,25) atd. pro třetí porovnání (viz tabulka č. 10).

Tabulka č. 10: Ukázka vektorů vah – kódování efektů (typu Deviation)

	ANOVA model (GLM)			Regresní model (REG)		
	L_1	L_2	L_3	X_1	X_2	X_3
1 (Hispánec)	$\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	1	0	0
2 (Asiat)	$-\frac{1}{4}$	$\frac{3}{4}$	$-\frac{1}{4}$	0	1	0
3 (Afroameričan)	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{3}{4}$	0	0	1
4 (běloch)	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	-1	-1	-1

Zdroj: vlastní zpracování podle UCLA (n.d.)

Odhady lineárních kombinací parametrů spolu se statistickými testy uvedených odhadů jsou uvedeny v tabulce č. 11.

Tabulka č. 11: Odhady lineárních kombinací parametrů modelu vč. hypotéz

Parameter	Estimate	Standard Error	t Value	Pr > |t|
1 (Hispánec) vs. 2 (Asiat), 3 (Afroameričan) a 4 (běloch)	-5,22	1,63	-3,20	0,0016
2 (Asiat) vs. 1 (Hispánec), 3 (Afroameričan) a 4 (běloch)	6,32	2,16	2,93	0,0038
3 (Afroameričan) vs. 1 (Hispánec), 2 (Asiat) a 4 (běloch)	-3,48	1,73	-2,01	0,0460

Zdroj: vlastní zpracování podle SAS Institute Inc. (2025a)

Kódování efektů (typu Helmert)

Kódování efektů typu Helmert porovnává efekt každé úrovně kategoriální proměnné s průměrným efektem následujících úrovní. První porovnání tedy porovnává průměr závislé proměnné pro úroveň 1 proměnné RACE s průměrem všech následujících úrovní proměnné RACE (úrovně 2, 3 a 4), druhé porovnání srovnává průměr závislé proměnné pro úroveň 2 proměnné RACE s průměrem všech následujících úrovní proměnné RACE (úrovně 3 a 4) a třetí porovnání porovnává průměr závislé proměnné pro úroveň 3 proměnné RACE s průměrem všech následujících úrovní kategoriální proměnné RACE (úroveň 4). I když tento typ kódovacího systému nedává velký smysl u nominální proměnné, jako je kategoriální proměnná RACE, je užitečný v situacích, kdy jsou úrovně kategoriální proměnné seřazeny například od nejnižší po nejvyšší nebo od nejmenší po největší, atd.

Pro první novou proměnnou X_1 regresní koeficienty nabývají stejných hodnot jako u typu Simple. U proměnných X_2 a X_3 hodnoty regresních koeficientů se mění v závislosti na počtu porovnávaných úrovní. Podobně se chová vektor vah kontrastů. Konkrétní hodnoty Helmertovy typu kódování jsou uvedeny v tabulce č. 12.

Pro porovnání efektů v kontextu Helmertova kódování lze opět využít příkaz ESTIMATE z analytické procedury. GLM. Odhady lineárních kombinací parametrů spolu se statistickými testy uvedených odhadů jsou zaznamenány v tabulce č. 13.

Tabulka č. 12: Ukázka vektorů vah – kódování efektů (typu Helmert)

	model ANOVA (GLM)			Regresní model (REG)		
	L_1	L_2	L_3	X_1	X_2	X_3
1 (Hispánec)	1	0	0	$\frac{3}{4}$	0	0
2 (Asiat)	$-\frac{1}{3}$	1	0	$-\frac{1}{4}$	$\frac{2}{3}$	0
3 (Afroameričan)	$-\frac{1}{3}$	$-\frac{1}{2}$	1	$-\frac{1}{4}$	$-\frac{1}{3}$	$\frac{1}{2}$
4 (běloch)	$-\frac{1}{3}$	$-\frac{1}{2}$	-1	$-\frac{1}{4}$	$-\frac{1}{3}$	$-\frac{1}{2}$

Zdroj: vlastní zpracování podle UCLA (n.d.)

Tabulka č. 13: Odhady lineárních kombinací parametrů modelu vč. hypotéz

Parameter	Estimate	Standard Error	t Value	Pr > |t|
1 (Hispánec) vs. 2 (Asiat), 3 (Afroameričan) a 4 (běloch)	-6,96	2,18	-3,20	0,0016
2 (Asiat) vs. 3 (Afroameričan) a 4 (běloch)	6,87	2,93	2,35	0,0198
3 (Afroameričan) vs. 4 (běloch)	-5,86	2,15	-2,72	0,0071

Zdroj: vlastní zpracování podle SAS Institute Inc. (2025a)

Kódování efektů (typu uživatelsky definované kódování – příklad)

Statistický systém SAS dovoluje použít pro jakékoliv porovnávání jakékoliv obecné kódovací schéma. Porovnávat se mohou efekty na samostatných úrovních kategorických proměnných, průměrné efekty pro skupiny úrovní i efekty v rámci celého sledovaného souboru dat.

V následujícím příkladu se pro další analýzy vytvoří speciální systém kódování proměnných, které umožňuje porovnat plánované cíle studie. Jedná se o porovnání:

- úroveň 1 (Hispánec) s úrovní 3 (Afroameričan),
- úroveň 2 (Asiat) s úrovněmi 1 (Hispánec) a 4 (běloch),
- úrovně 1 (Hispánec) a 2 (Asiat) s úrovněmi 3 (Afroameričan) a 4 (běloch).

U prvního kontrastu se má porovnat efekt na úrovni 1 s efektem na úrovni 3. Při realizaci porovnávání se průměr závislé proměnné vynásobí vektorem koeficientů kontrastu. Proto se musí koeficienty kontrastu nastavit na 1 0 -1 0. Znamená to, že srovnávání efektů se realizuje na úrovních s jednotkovými hodnotami koeficientů kontrastu, ale se znaménkem opačným. Úrovně 2 a 4 se tedy do porovnání nezapojují, protože se vynásobí nulou a do dalšího porovnávání nebudou zařazeny. Součet koeficientů kontrastu musí být vždy nulový. Pokud se součet koeficientů kontrastu nerovná nule, kontrast nelze odhadnout a systém SAS vypíše chybovou zprávu. Současně není příliš důležité, které kategoriální úrovni je přiřazena hodnota kladná nebo záporná. Bude li vektor koeficientů kontrastu obsahovat 1 0 -1 0 anebo -1 0 1 0, vždy se bude porovnávat efekt z úrovně 1 kategorické proměnné s efektem na úrovni 3 u proměnné RACE. Při změně znaménka ve vektoru koeficientů kontrastu dojde také i ke změně znaménka u příslušného koeficientu ve vektoru regresních koeficientů.

Druhý případ uživatelsky definovaného vektoru kontrastů slouží k porovnání efektu úrovně 2 proměnné RACE s průměrným efektem úrovní 1 a 4 u této kategorické proměnné. V případě, že ke srovnávání je vybrána více než 1 úroveň (jako je tomu v případě průměrného efektu úrovní 1 a 4 u kategorické proměnné RACE), musí se

jednotkový koeficient rovnoměrně rozdělit. Aby se zajistilo splnění podmínky, že suma vah kontrastů je jednotková, jednotkový koeficient se rovnoměrně rozdělí na všechny ty úrovně, které byly zahrnuty do 1 kombinované skupiny porovnání. To znamená, že každý efekt na úrovni 1 a 4 proměnné RACE má při požadovaném porovnávání koeficient kontrastu o hodnotě 1/2. Celkový efekt kombinace úrovní 1 a 4 má proto koeficient kontrastu roven 1. Výsledný vektor koeficientů kontrastu pro splnění požadovaného zadání je -1/2 1 0 -1/2. Odpovídající vektor regresních koeficientů je roven -1 1 -1 1.

Závěrečný případ uživatelsky definovaného porovnání je analogickým případem předešlého zadání. Průměrný efekt kombinace 2 úrovní se bude porovnávat s průměrným efektem kombinace 2 jiných úrovní u kategorické proměnné RACE. Této situaci musí odpovídat také i hodnoty vektoru koeficientů kontrastu a hodnoty vektoru regresních koeficientů. Hodnoty koeficientů vektoru kontrastu ve výši -1/2 a 1/2 nasvědčují skutečnosti, že efekty se porovnávají v kombinaci 2 úrovní. Výsledné požadované uživatelsky definované vektory koeficientů kontrastů a regresních koeficientů jsou uvedeny v tabulce č. 14.

Tabulka č. 14: Ukázka uživatelského porovnávání a uživatelské parametrizace

	model ANOVA (GLM)			Regresní model (REG)		
	L_1	L_2	L_3	X_1	X_2	X_3
1 (Hispánec)	1	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	-1	$1\frac{1}{2}$
2 (Asiat)	0	1	$\frac{1}{2}$	$\frac{1}{2}$	1	$-\frac{1}{2}$
3 (Afroameričan)	-1	0	$-\frac{1}{2}$	$-1\frac{1}{2}$	-1	$1\frac{1}{2}$
4 (běloch)	0	$-\frac{1}{2}$	$-\frac{1}{2}$	$1\frac{1}{2}$	1	$-2\frac{1}{2}$

Zdroj: vlastní zpracování podle UCLA (n.d.)

I při uživatelsky definovaných porovnáních lze pro odhady lineárních kombinací parametrů odhadu využít příkaz ESTIMATE analytické procedury GLM. Odhady lineárních kombinací parametrů se statistickými testy uvedených uživatelsky definovaných porovnání a uživatelské parametrizace jsou vyobrazeny v tabulce č. 15.

Tabulka č. 15: Odhady lineárních kombinací parametrů modelu vč. hypotéz

Parameter	Estimate	Standard Error	t Value	Pr > |t|
1 (Hispánec) vs. 2 (Asiat), 3 (Afroameričan) a 4 (běloch)	-1,74	2,73	-0,64	0,5246
2 (Asiat) vs. 3 (Afroameričan) a 4 (běloch)	7,74	2,90	2,67	0,0082
3 (Afroameričan) vs. 4 (běloch)	1,10	1,96	0,56	0,5756

Zdroj: vlastní zpracování podle SAS Institute Inc. (2025a)

4. ZÁVĚR

Již po mnoho let patří ke standardním postupům statistické analýzy vysvětlování pozorovaných hodnot (modelování) zájmové (závislé) proměnné prostřednictvím jednoho nebo více prediktorů (spojitých nebo kategorických vysvětlujících proměnných). Mezi nejpoužívanější statistické modely v mnoha oblastech přírodních a společenských věd patří obecné lineární modely. I když tyto modely musí splňovat celkem přísné předpoklady pro svoje adekvátní použití (homogenní v rozptylu, nesystematičnost a nekorelovanost odchylek od lineárního vztahu), jejich popularita

mezi výzkumníky neustále roste. Navíc při dodržení předpokladů na rozdělení odchylek od očekávaných hodnot (predikce) jednotlivých pozorování (normální rozdělení s nulovou střední hodnotou a konstantním rozptylem) lze lineární regresní modely využít k předvídání rozdělení výsledků, ke konstrukci intervalů spolehlivosti odhadů a k testování statistických hypotéz, a to od těch nejjednodušších až po velmi komplexní.

Základní otázkou definice statistického modelu je v programovém systému SAS parametrizace, tj. způsob konstrukce matice plánu pro různé typy proměnných (spojité i kategoriální). Volba způsobu parametrizace při tvorbě statistického modelu i v systému SAS patří k nejdůležitějším otázkám, kterým musí většina analytiků čelit. Správné pochopení konstrukce matice plánu pro různé typy proměnných umožňuje vytvářet statistické modely, které odpovídají sledovaným závislostem a také poskytují možnost jejich adekvátnímu předvídání. Způsob tvorby matice plánu, a tedy parametrizace, je také určující i pro interpretaci a význam odhadovaných parametrů a jejich statistickému testování pomocí velmi pružně definovaných statistických hypotéz.

Ačkoliv se ukázky parametrizace a statistických hypotéz týkají nejjednodušších statistických modelů, jsou principy a použití u složitějších modelů stejné. Článek představuje základní principy pro statistické modelování v analytickém systému SAS.

LITERATURA

Dostál, D. (2021). *Lineární statistické modely v psychologii.* Univerzita Palackého v Olomouci

Littell, R. C., Stroup, W. W. & Freund, R. J. (2002). *SAS for Linear Models*. SAS Press.

Pavlík, T. & Dušek, L. (2012). *Biostatistika.* Akademické nakladatelství CERM.

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

Řeháková, B. (2008). Kontrasty v logistické regresi [Contrasts in Logistic Regression]. *Sociologický časopis [Czech Sociological Review], 44*(4), 745 – 765. https://doi.org/10.13060/00380288.2008.44.4.07

SAS Institute Inc. (2025a). *SAS/STAT® 15.3 user's guide: Chapter 3. Introduction to statistical modeling with SAS/STAT software* (pp. 55 – 66). SAS Institute Inc.

SAS Institute Inc. (2025b). *SAS/STAT® 15.3 user's guide: Chapter 20. Shared concepts and topics* (pp. 399 – 401). SAS Institute Inc.

UCLA Office of Advanced Research Computing. (n.d.). *Introduction to the features of SAS / SAS Learning Modules.* UCLA Statistical Methods and Data Analytics. https://stats.oarc.ucla.edu/sas/modules/introduction-to-the-features-of-sas/

RESUMÉ

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

The SAS software system is equipped with many functionalities that allow analysts to properly 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 appropriate data analysis methods accordingly. A fundamental question in data analysis and statistical modelling in the SAS system is the choice of model parameterisation. The selection of the method for constructing the design matrix of the model 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 types of parametrization can lead to different parameter values and their corresponding interpretations.

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