# SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

## SLOVAK STATISTICS and DEMOGRAPHY

4/2023
ročník/volume 33


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

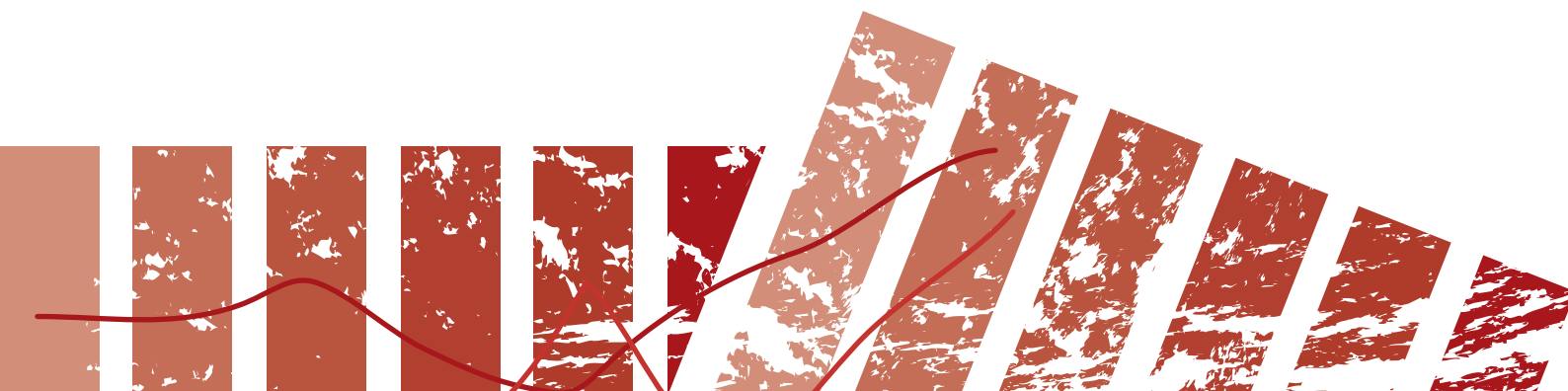
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.


Článok/Article: 3

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

Strany/Pages: 41 – 61

Dátum vydania/Publication date: 15. október 2023/October 15, 2023

Informatívny článok/Informative article

**Roman PAVELKA**
**Štatistický úrad Slovenskej republiky**

# ÚVOD DO STATISTICKÉHO MODELOVÁNÍ POMOCÍ ANALYTICKÉHO SYSTÉMU SAS

# INTRODUCTION TO STATISTICAL MODELLING USING THE SAS ANALYSIS SYSTEM

## ABSTRAKT
Mohutný nárůst výpočetního výkonu ve druhé polovině 20. století a vznik nových moderních technologií na počátku 21. století podstatně ovlivnil rozvoj metodologie statistického modelování. Od skromných začátků v zemědělství se aplikace statistického modelování, a to zejména lineární modely, staly zásadními v mnoha oblastech vědy a výzkumu. V průběhu několika desítek let se metodologie statistických modelů přetvořila do obecných analytických postupů datové vědy a tvoří jádro moderních statisticko-analytických metod. V popředí vývoje a implementace uvedených metod stojí programový systém SAS, který již ve svých základních modulech umožňuje tyto poznatky využívat v reálné praxi. I když je systém SAS vybaven mnohými funkcionalitami, které odpovídají současnému stavu statistické vědy, není v našich podmínkách příliš rozšířen, a to především z důvodu vyšší pořizovací ceny. Proto je cílem tohoto článku přiblížit některé možnosti tohoto softvéru z oblasti statistického modelování, které zůstávají mnohým analytikům skryté.

## ABSTRACT
The massive increase in computing power in the second half of the 20th century and the emergence of new modern technologies in the early 21st century have significantly influenced the development of statistical modelling methodology. From humble beginnings in agriculture, applications of statistical modelling, and in particular linear models, have become fundamental in many areas of science and research. Over the course of several decades, statistical model methodology has evolved into general data science analytical procedures and forms the core of modern statistical analytical methods. At the forefront of the development and implementation of these methods is the SAS software system, which enables these findings to be used in real practice already in its basic modules. Although the SAS system is equipped with many functionalities that correspond to the current state of statistical science, it is not very widespread in our conditions, mainly due to its higher purchase price. Therefore, the aim of this article is to present some of the capabilities of this software in the field of statistic modelling, which remain hidden to many analysts.

## KLÍČOVÉ SLOVA
analýza rozptylu, regresní analýza, SAS, statistické modelování, testování hypotéz

## KEY WORDS
analysis of variance, regression analysis, SAS, statistical modelling, hypothesis testing

## 1. ÚVOD

Programový systém SAS vznikl v roce 1976 pod označením SAS-76. Již od počátku existence systému SAS se vlajkovou lodí statistického modelování stala procedura GLM. Na svou dobu byla tato procedura velmi inovativní a upoutala pozornost statistiků a dalších osob zabývajících se analýzou dat v USA i mimo ně. Procedura GLM poskytovala komplexní platformu, která umožňovala práci s různými typy lineárních modelů. Uživatelům umožnila získat řešení pro většinu problémů spadajících do oblasti lineárních modelů, pro regresní analýzu, analýzu lineárních modelů a analýzu rozptylu i kovariance a analýzu vícerozměrných dat. Procedura GLM se stala vzorem pro vývoj další statistické procedury v systému SAS[1] a pro většinu statistiků a datových analytiků základním postupem při práci s daty [4].

V průběhu následujících let byla vytvořena statistická procedura REG, která rozšířila možnosti regresní analýzy o diagnostické nástroje. Nyní měl uživatel nejen možnost odhadovat inferenční statistiky v regresní analýze, ale mohl také získat statistiky, které mu pomohou rozhodnout, jaké proměnné do analýzy zahrnout, a identifikovat problematické data [1]. Postupující pokrok výpočetní techniky přinesl překotný rozvoj ve vývoji statistických procedur systému SAS. Možnosti statického modelování byly rozšířeny na nelineární modely, modely s pevnými i náhodnými efekty i na problémy spojené s analýzou korelovaných dat. V prvních verzích systému SAS byly značně omezené možnosti analýzy kategoriálních dat. Proto byl tento analytický softvér obohacen o statistické procedury, které inovovaly použití lineárních modelů pro analýzu kategoriálních dat. Toto vylepšení proto umožňovalo řešení problémů zobecněných lineárních modelů [6]. V současnosti systém SAS svým uživatelům nabízí v modulu SAS/STAT několik desítek procedur k analýze dat [8], která pokrývají většinu problémů moderní statistiky včetně vybraných oblastí strojového učení a umělé inteligence. Systém SAS je navíc průběžně aktualizován o nejnovější postupy datové vědy.

## 2. STATISTICKÉ MODELOVÁNÍ V PROSTŘEDÍ SYSTÉMU SAS

### 2.1. ZÁKLADNÍ POJMY A DEFINICE

Statistický model [9] je stochastický model, který obsahuje parametry (neznámé konstanty), které je třeba odhadnout na základě předpokladů o modelu a pozorovaných datech. Statistické modely mohou být jednoduché, ale také i velmi komplexní. Proto je nejvhodnější klasifikovat modely podle jednoduchých kritérií, jako je např. přítomnost náhodných efektů, přítomnost nelinearity, charakteristiky dat atd. Statistický model popisuje distribuční vlastnosti jedné nebo více vysvětlovaných proměnných. Tento popis má často jednoduchou podobu modelu s aditivní chybovou strukturou:

Vysvětlovaná proměnná = průměr + chyba

V matematickém zápisu má tato jednoduchá rovnice modelu tvar:

$$\boldsymbol{Y} = f\left(x_1, \cdots, x_k;\ \beta_1, \cdots, \beta_p\right) + \boldsymbol{\varepsilon} \tag{1}$$

[1] *Statistický systém SAS (z angl. Statistical Analysis System) je integrovaný systém softvérových produktů pro analýzu dat vyráběný americkou firmou SAS Institute, Inc.*

V rovnici (1) je proměnná *Y* nazývaná závislou, vysvětlovanou nebo výstupní proměnnou. Výrazy $x_1$, ..., $x_k$ označují hodnoty *k* regresorových proměnných, také označovaných jako kovariáty, nezávislé nebo vysvětlující proměnné. Neznámé konstanty modelu, které jsou předmětem statistického odhadu, jsou označeny jako $\beta_1$, ..., $\beta_p$. Člen rovnice (1) označený $\boldsymbol{\varepsilon}$ vyjadřuje náhodné poruchy modelu; nazývá se také reziduální nebo chybový člen modelu.

Je-li v rovnici modelu (1) zdrojem náhodnosti pouze chybový člen $\boldsymbol{\varepsilon}$ a jestliže mají tyto náhodné chyby nulový průměr, potom funkce (1) se stává funkcí střední hodnoty daného statistického modelu ve tvaru

$$E[Y] = f\left(x_1, \cdots, x_k;\ \beta_1, \cdots, \beta_p\right) \tag{2}$$

kde E [.] označuje operátor střední hodnoty (očekávání).

V mnoha aplikacích je jednoduchá formulace modelu nedostatečná. Často je potřebné specifikovat nejen stochastické vlastnosti jednoduchého chybového členu, ale také to, jak se chyby modelu spojené s různými pozorováními navzájem ovlivňují. Pokud chyby nemají nulovou střední hodnotu nebo pokud rozptyl pozorování závisí na podmíněných středních hodnotách, k popisu mechanismu generujícího data jednoduchý aditivní chybový model je obvykle nedostatečný. Modely pro taková data obvykle vyžadují složitější formulace zahrnující různá rozdělení pravděpodobnosti.

## 2.2. TŘÍDY STATISTICKÝCH MODELŮ DOSTUPNÉ V SYSTÉMU SAS

### Lineární a nelineární modely

Problém statistického odhadu je nelineární, pokud odhadové rovnice – rovnice, jejichž řešení dávají odhady parametrů, závisejí na parametrech nelineárně. Takové problémy odhadů obvykle nemají řešení v uzavřené formě a musí se řešit iteračními numerickými technikami.

K rozlišení mezi lineárními a nelineárními modely se často používá nelinearita funkce střední hodnoty. Model má nelineární funkci střední hodnoty, jestliže derivace funkce střední hodnoty vzhledem k jejím parametrům závisí alespoň na jednom dalším parametru. Nelineární funkce střední hodnoty vede k nelineárním odhadům.

### Jednorozměrné a vícerozměrné modely

Vícerozměrný statistický model je model, ve kterém je více vysvětlovaných proměnných modelováno současně. Například pokud data obsahují hodnoty výšky ($h_i$) a váhy ($w_i$) dětí sbíraných za několik let, potom vícerozměrný statistický model je dán výrazem:

$$\boldsymbol{Y}_i = \begin{bmatrix} w_i \\ h_i \end{bmatrix} = \boldsymbol{X\beta} + \begin{bmatrix} \varepsilon_{wi} \\ \varepsilon_{hi} \end{bmatrix} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}_i \tag{3}$$

Vektory vysvětlovaných proměnných $\boldsymbol{Y}_i$ a náhodných chyb $\boldsymbol{\varepsilon}_i$ obsahují po 2 pozorování patřících *i*-tému dítěti. Chyby téhož dítěte mají tak korelaci:

$$CORR[\varepsilon_{wi}, \varepsilon_{hi}] = \frac{\sigma_{wh}}{\sqrt{\sigma_h^2 \sigma_w^2}} \qquad \boldsymbol{\varepsilon}_i = \left(\mathbf{0}, \begin{bmatrix} \sigma_w^2 & \sigma_{wh} \\ \sigma_{wh} & \sigma_h^2 \end{bmatrix}\right), \tag{4}$$

kde $\sigma_w^2$ označuje rozptyl hodnot váhy dětí, $\sigma_h^2$ je rozptyl hodnot výšky dětí a $\sigma_{wh}$ jejich kovariance.

**Regresní modely a modely s klasifikačními efekty**

Regresní model v užším slova smyslu v porovnání s klasifikačním modelem je modelem lineárním, pokud jsou všechny regresorové proměnné (efekty) v modelu spojité. Jinak řečeno, každý spojitý efekt v modelu přispívá jedním sloupcem do matice modelu $\boldsymbol{X}$, a tedy jedním parametrem do modelu jako celku.

Klasifikační efekt je naopak spojen s více než jedním sloupcem matice $\boldsymbol{X}$. 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 [9]. Klasifikace proměnných se v 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. Příkladem modelu s klasifikačním efektem je model, ve kterém je zahrnuta dvojúrovňová klasifikační proměnná GENDER. Hodnoty této klasifikační proměnné jsou kódovány jako '*F*' a '*M*'. Potom symbolické vyjádření modelu ve tvaru

$$weight = age + bmi + gender + error \tag{5}$$

se rozšiřuje do statistického modelu

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \tau_1 I(gender = 'F') + \tau_1 I(gender = 'M') + \varepsilon_i \tag{6}$$

kde *I(gender=*'F'*)* je indikátorová funkce, která vrací 1, pokud je hodnota proměnné pohlaví '*F*', a 0 v opačném případě. Parametry $\tau_1$ a $\tau_2$ jsou spojeny s efektem klasifikace proměnné GENDER. Tato forma zahrnutí klasifikační proměnné do modelu je pouze jedním z několika různých způsobů, jak do modelu zahrnout vliv úrovní klasifikační proměnné. Jedná se o tzv. singulární parametrizaci[2], která je obecně používaným přístupem a používá se v procedurách GLM, MIXED a GLIMMIX. Jiné tvary parametrizace klasifikačních efektů pomocí různých forem tzv. nesingulární parametrizace jsou k dispozici v procedurách GENMOD a LOGISTIC.

Modely, které obsahují pouze klasifikační efekty, se často ztotožňují s modely analýzy rozptylu (ANOVA), protože při jejich analýze se často používají metody ANOVA. To platí zejména pro experimentální data, kde modelové efekty zahrnují efekty ošetření a návrhu řízení chyb a pod. Nicméně klasifikační proměnné se objevují

[2] *Provedenou parametrizací se stanou sloupce matice X lineárně závislými, matice X'X je singulární a pro řešení normálních rovnic je nutné použít zobecněnou inverzní matici X'X podle (Ben-Israel, Greville, 2003).*

ve větší míře i v jiných modelech. Například mnoho smíšených modelů, kde se parametry odhadují pomocí maximální věrohodnosti, se skládá výhradně z klasifikačních efektů. Tyto modely ale neumožňují rozklad součtu čtverců typický pro techniky ANOVA. Mnoho modelů obsahuje jak spojité, tak klasifikační efekty. Například statistický model se spojitou proměnnou v jednotlivých úrovních klasifikačního efektu. Takové efekty jsou vhodné například k tomu, aby se v regresním modelu měnily sklony podle úrovní klasifikační proměnné.

**Modely s pevnými efekty, s náhodnými efekty a modely se smíšenými efekty**

Ve statistickém modelu představuje každá nezávislá proměnná pevný nebo náhodný efekt. Pevné efekty v modelu jsou neznámé konstanty (parametry). Naopak náhodné efekty konstantní nejsou a podléhají normálnímu rozdělení. Modely, ve kterých jsou všechny efekty pevné, se nazývají modely s pevnými efekty. Podobně modely, v nichž jsou všechny efekty náhodné – kromě případného absolutního členu (intercept) – se nazývají modely s náhodnými efekty. Smíšené modely jsou tedy takové modely, které obsahují pevné a náhodné efekty. V maticovém zápisu lineární model s efekty pevnými, s efekty náhodnými nebo lineární model se smíšenými efekty představují následující modelové rovnice:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (7)$$

$$\boldsymbol{Y} = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \quad (8)$$

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \quad (9)$$

V rovnicích (7) až (9) jsou $\boldsymbol{X}$, resp. $\boldsymbol{Z}$ regresní matice modelu spojené s pevnými, resp. náhodnými efekty. Vektor $\boldsymbol{\beta}$ je vektor parametrů pevných efektů a vektor $\boldsymbol{\gamma}$ reprezentuje náhodné efekty. Procedury pro modelování se smíšenými modely v softvéru SAS/STAT předpokládají, že náhodné efekty sledují normální rozdělení s rozptylovou a kovarianční maticí $\boldsymbol{G}$ a ve většině případů mají náhodné efekty nulovou střední hodnotu.

**Zobecněné lineární modely**

Třídou modelů, která v posledních desetiletích nabývá na významu, je třída zobecněných lineárních modelů. Teorie zobecněných lineárních modelů vychází z odborných statí [6, 10], následně byla zpopularizována v monografii [5].

Toto zobecnění vyžaduje podstatně složitější nastavení statistického modelu, než je tomu v případě lineárních modelů s normálně rozdělenými daty. Zobecněný lineární model se skládá ze:

- **systematické složky**, která v modelu vystupuje jako lineární prediktor podobně jako v lineárních modelech. Lineární prediktor $\boldsymbol{\eta} = \boldsymbol{x}'\boldsymbol{\beta}$ je lineární funkce v parametrech. Na rozdíl od lineárního modelu nepředstavuje tzv. funkci střední hodnoty podle (2).

- **spojovací funkce** $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$, což je funkce, která spojuje lineární prediktor s celkovým průměrem $\boldsymbol{\mu}$. Spojovací funkce je monotónní, inverzní funkce. Střední hodnotu lze tedy vyjádřit jako inverzní lineární prediktor, $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$. Například běžnou spojovací funkcí pro binární a binomická data je logitová spojovací funkce, $g(t) = ln\{t/(1-t)\}$.

- **náhodná složka** zobecněného lineárního modelu, která je reprezentována rozdělením pravděpodobnosti. Předpokládá se, že rozdělení pravděpodobnosti náleží do skupiny exponenciálních. Mezi diskrétní rozdělení této rodiny patří např. Bernoulliho (binární), binomické, Poissonovo, geometrické a záporné binomické rozdělení (pro danou hodnotu škály). Ke spojitým rozdělení pravděpodobnosti patří normální (Gaussovo), beta, gama, inverzní Gaussovo a exponenciální rozdělení.

Speciálním případem zobecněného lineárního modelu je klasický lineární model, kde spojovací funkce je funkcí identity a rozdělení pravděpodobnosti je normální.

Programový systém SAS pro odhad parametrů statistických modelů používá klasické odhadové metody, a to metodu nejmenších čtverců (včetně vážených, s iteračními výpočty) a metody založené na věrohodnostní funkci. Možnosti statistické inference nad daty ze statistických zjišťování poskytují procedury SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC a SURVEYPHREG, které při odhadech dovolují zohlednit také i mechanismus výběru do zjišťování.

## 3. ZÁKLADNÍ PRINCIPY REGRESNÍ ANALÝZY V SYSTÉMU SAS

Programový systém SAS dokáže analyzovat široké spektrum regresních modelů. Statistické procedury, kterými je SAS vybaven, podporují regresní modely:

- s diskrétní nebo spojitou vysvětlovanou proměnnou, jejíž rozdělení může, ale také nemusí podléhat normálnímu rozdělení pravděpodobnosti,
- lineární, nelineární anebo zobecněné lineární formy statistických modelů,
- modely obsahující klasifikační proměnné (nevstupují do modelu prostřednictvím svých hodnot, ale prostřednictvím svých úrovní),
- u modelů s klasifikačními efekty se pro zajištění jednoznačnosti odhadů využívá konceptu tzv. odhadnutelných funkcí,
- statistické modely s daty generovanými z experimentů, pozorování anebo statistických šetření,
- pro metody odhadů pomocí metod založených na funkci maximální věrohodnosti a minimalizaci součtu reziduálních čtverců.

### 3.1. ODHADY PARAMETRŮ LINEÁRNÍHO MODELU V SYSTÉMU SAS

Odhady parametrů modelu $\boldsymbol{\beta}$ metodou nejmenších čtverců (příp. vážených $\boldsymbol{W}$) jsou dány řešením normálních rovnic:

$$(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} \tag{10}$$

Jediným předpokladem, který je nutný k tomu, aby odhady metodou nejmenších čtverců byly nestranné, je nulová střední hodnota chyb modelu. Získané odhady

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} \tag{11}$$

mají minimální rozptyl ve třídě odhadů, které jsou nestranné a jsou lineárními funkcemi vysvětlované proměnné. Pokud je splněn další předpoklad normálně rozdělených chyb, pak platí následující [9]:

- vypočtené statistiky mají výběrová rozdělení vhodná pro testování hypotéz,
- odhady parametrů jsou normálně rozděleny,
- různé součty čtverců jsou alespoň za platnosti nulových hypotéz rozděleny úměrně chí-kvadrát rozdělení,

- poměry odhadů ke standardním chybám se za platnosti nulových hypotéz řídí Studentovým *t* rozdělením,
- poměry středních čtverců se řídí *F* rozdělením za platnosti nulových hypotéz.

Pokud jsou ke statistickému modelování použita data, která nesplňují výše uvedené předpoklady normality rozdělení chyb, výsledky analýzy je nutné interpretovat opatrně. Pravděpodobnosti významnosti jsou za těchto okolností nespolehlivé.

Pro realizaci odhadů lineárního modelu s funkcí střední hodnoty $E[\boldsymbol{Y}] = \boldsymbol{\beta X}$ metodou maximální věrohodnosti je potřebné specifikovat rozdělení pravděpodobnosti vysvětlované proměnné $\boldsymbol{Y}$. Pro odhady parametrů lineárního modelu lze podle [8] použít věrohodnostní funkci (resp. její logaritmus) ve tvaru

$$\boldsymbol{l(\beta, \sigma^2; y)} = -\frac{n}{2} ln\{\boldsymbol{2\pi}\} - \frac{n}{2} ln\{\boldsymbol{\sigma^2}\} - \frac{1}{2\boldsymbol{\sigma^2}} (\boldsymbol{y - X\beta})'(\boldsymbol{y - X\beta}) \qquad (12)$$

Funkce (12) je maximalizována, když je minimalizován součet čtverců $(\boldsymbol{y - X\beta})'(\boldsymbol{y - X\beta})$. Odhad parametrů $\boldsymbol{\beta}$ v lineárním modelu metodou maximální věrohodnosti je proto totožný s odhadem metodou nejmenších čtverců. Pokud matice $\boldsymbol{X'WX}$ nemá plnou hodnost, je nutné k řešení normálních rovnic pro minimalizaci součtu čtverců $(\boldsymbol{y - X\beta})'(\boldsymbol{y - X\beta})$ použít metodu zobecněné inverzní matice [2].

Pokud modelová matice $\boldsymbol{X}$ obsahuje sloupec, jehož nenulové hodnoty se nemění, obvykle sloupec jedniček, obsahuje lineární model absolutní člen (intercept). Možnost potlačit automatické přidání sloupce jedniček umožňují statistické procedury, které podporují parametr NOINT v příkazu MODEL. Obecně by modely bez absolutního členu (interceptu) měly být výjimkou, zejména pokud navrhovaný model neobsahuje klasifikační proměnné.

Vyrovnané hodnoty $\hat{y}_i$ pro $i = 1 \cdots n$ jsou reprezentovány rovnicí:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \qquad (13)$$

Po provedeném odhadu parametrů chybová složka pro *i*-tou hodnotu je dána jako

$$\hat{\varepsilon}_i = y_i - \hat{y}_i. \qquad (14)$$

***Tabulka č. 1: Analýza rozptylu (ANOVA) lineárního modelu***

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | k | $SS_M = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ | $MS_M = \frac{\mathrm{SS}_M}{k}$ | $F = \frac{\mathrm{MS}_M}{\mathrm{MS}_E}$ | Hodnota p-value< |
| **Error** | n - k – 1 | $SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ | $MS_E = \frac{\mathrm{SS}_E}{\mathrm{n - k - 1}}$ | | |
| **Corrected Total** | n – 1 | $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$ | | | |

*Poznámka: Symbol n vyjadřuje počet pozorování a k+1 vyjadřuje počet parametrů v modelu, z toho k je počet vysvětlujících proměnných.*
***Zdroj: vlastní zpracování autora podle [9]***

Současně s odhady parametrů regresní procedurou v systému SAS se analýzou rozptylu testuje významnost odhadnutého modelu (tzv. celkový *F*-test). Vyšší hodnota pravděpodobnosti (*p*-value > 0,05) nasvědčuje nevýznamnosti modelu. Analýza rozptylu s rozkladem sum čtverců a testové *F* statistiky jsou ilustrovány v tabulce č. 1.

## 3.2. ODHADY PARAMETRŮ OBECNÉHO LINEÁRNÍHO MODELU POMOCÍ SAS

Obecný lineární model lze podobně jako lineární model (regresní model v užším významu) vyjádřit v podobě modelu s aditivní chybovou složkou [7]. Na rozdíl od klasického lineárního modelu je koncepce obecného lineárního modelu založena na podstatně obecnějších základech. Původní chápání statistického modelu s aditivní strukturou chyby (vyrovnaná + reziduální složka) bylo nahrazeno koncepcí symbolicky vyjádřenou ve tvaru:

$$\boldsymbol{Data = Model + Error} \tag{15}$$

V rovnici (15) je model vyjádřením pochopení podstaty a mechanismu vzniku dat, resp. hypotézy o nich, které jsou aplikovány na analyzovaná data. Chybová složka označená *Error* je explicitním vyjádřením toho, že na data působí i jiné vlivy a okolnosti než ty zahrnuté v modelu. V rámci procesu obecného lineárního modelování se analytici pokoušejí porovnáním různých lineárních modelů určit *nejlepší* obecný lineární model pro data. Testované obecné lineární modely jsou posuzovány z hlediska relativního podílu rozptylu dat připisovaného modelu a chybových složek. U neměnného datového souboru je součet modelové a chybové složky variance (tj. rozptyl dat) konstantní, takže každé zvýšení variance, které je vysvětleno modelovou složkou, bude mít za následek ekvivalentní snížení variance chybové složky. Výraz *obecný* v modelu jednoduše označuje schopnost zohlednit rozdíly kvantitativních proměnných, které představují spojité míry (jako v regresi) a diskrétní rozlišení (kategoriální proměnné) reprezentující skupiny nebo experimentální podmínky. Podobně jako v analýze kovariance jsou v obecném lineárním modelu obsaženy proměnné reprezentující jak kvantitativní veličiny, tak kategoriální úrovně. Podobně jako v analýze rozptylu jsou porovnávány rozdíly mezi průměrnými hodnotami závislé (vysvětlované) proměnné na jednotlivých úrovních kategoriální proměnné (nazývané faktorem) a je testována významnost rozdílu uvedených průměrů. V obecném lineárním modelu může být zahrnut 1 nebo více různě uspořádaných faktorů (např. znáhodněné bloky, latinské čtverce atd.), opakovaná měření s vyváženými či nevyváženými návrhy. Proto regresní analýza (lineární modelování), analýza rozptylu a analýza kovariance jsou pouze speciálními případy obecných lineárních modelů.

Pro ukázku možné analýzy v prostředí SAS lze obecný lineární model formulovat ve tvaru

$$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk}, \tag{16}$$

kde $y_{ijk}$ je k-té pozorování vysvětlované (závislé) na *i*-té úrovně faktoru *A* a *j*-té úrovni faktoru *B*,
$\mu$ 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 úrovni faktoru *B* a
$\varepsilon_{ijk}$ je *k*-tá reziduální hodnota *i*-té úrovně faktoru *A* na *j*-té úrovni faktoru *B*.

Náhodné chyby $\varepsilon_{ijk}$ jsou nezávislé a stejně rozdělené s rozdělením $N(0, \sigma_E^2)$.[3]

Tvar modelu podle (16) 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 (16) přepsat do tvaru

$$y_{ijk} = \mu + \underbrace{\mu_{i\cdot} - \mu} + \underbrace{\mu_{\cdot j} - \mu} + \underbrace{\mu_{\cdot ij} - \mu_{i\cdot} - \mu_{\cdot j} - \mu} + \varepsilon_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \qquad (17)$$

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

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[4] poskytuje analýzu vyvážených i nevyvážených dat s možností odhadu parametrů obecného lineárního modelu [9].

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

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Faktor A | $n_A - 1$ | $SS_A = n_B * n_I * \sum_{i=1}^{n_A} (\mu_{i\cdot} - \mu)^2$ | $MS_A = \frac{SS_A}{n_A - 1}$ | $F_A = \frac{MS_A}{MS_E}$ | Hodnota p-value< |
| Faktor B | $n_B - 1$ | $SS_B = n_A * n_I * \sum_{j=1}^{n_B} (\mu_{\cdot j} - \mu)^2$ | $MS_B = \frac{SS_B}{n_B - 1}$ | $F_B = \frac{MS_B}{MS_E}$ | Hodnota p-value< |
| Interakce A*B | $(n_A - 1)*(n_B - 1)$ | $SS_{AB} = n_I * \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (\mu_{\cdot ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu)^2$ | $MS_{AB} = \frac{SS_{AB}}{(n_A - 1) * (n_B - 1)}$ | $F_{AB} = \frac{MS_{AB}}{MS_E}$ | Hodnota p-value< |
| Error | $n_A * n_B * (n_I - 1)$ | $SS_E = \sum_{k=1}^{n_I} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (y_{ijk} - \mu_{\cdot ij})^2$ | $MS_E = \frac{SS_E}{n_A * n_B * (n_I - 1)}$ | | |
| Corrected Total | $n_A * n_B * n_I - 1$ | $SS_T = \sum_{k=1}^{n_I} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (y_{ijk} - \mu)^2$ | | | |

*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 autora podle [9]***

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. Vysoká hodnota pravděpodobnosti nasvědčuje platnosti nulové hypotézy o nevýznamnosti modelu. Pokud je toto číslo nižší než 5 %

*[3] Normální rozdělení náhodných chyb (identicky a vzájemně nezávislé hodnoty) je hlavním rozdílem od zobecněných lineárních modelů, ve kterých tento předpoklad nemusí platit.*
*[4] GLM znamená General Linear Model tj. v překladu obecný lineární model.*

(0,05), je obvykle vliv faktoru, resp. interakce považován za dostatečně významný. Definice analýzy rozptylu pro dvoufaktorový model s interakcemi je v tabulce č. 2.

Odhad parametrů obecného lineárního modelu (16) je realizován procedurou GLM metodou nejmenších čtverců řešením normálních rovnic (10). Faktory jako kategorické proměnné se do matice modelu $\boldsymbol{X}$ zahrnou pomocí tzv. singulární parametrizace (6). Na základě toho je možné obecný lineární model přeformulovat do podoby vícenásobného regresního modelu, na jehož základě procedura GLM odhaduje parametry modelu:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{18}$$

Parametry regresního modelu (18) jsou odhadnuty v souladu s rovnicí (11). Odhadnuté parametry $\widehat{\boldsymbol{\beta}}$ v důsledku singulární parametrizace (6) však nepředstavují jedinečné řešení. Pro zajištění jednoznačnosti odhadu $\widehat{\boldsymbol{\beta}}$ se využívají tzv. odhadnutelné funkce $\boldsymbol{L\beta}$ [3]. Základem odhadnutelných funkcí $\boldsymbol{L\beta}$ je vektor prvků $\boldsymbol{L}$, které reprezentují váhy odpovídající každému efektu, případně interakci efektů ve funkci odhadů středních hodnot. Počet jedinečných symbolů ve vektoru představuje maximální počet lineárně nezávislých koeficientů odhadovaných modelem, který je roven hodnosti matice $\boldsymbol{X'X}$. Odhadnutelné funkce $\boldsymbol{L\beta}$ se vyznačují následujícími vlastnostmi:

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

$$VAR(\widehat{\boldsymbol{L\beta}}) = \sigma_\varepsilon^2[\boldsymbol{L}(\boldsymbol{X'WX})^{-1}\boldsymbol{L'}], \tag{19}$$

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

Na základě modelové složky dvoufaktorové rovnice obecného lineárního modelu jsou předpovídané hodnoty dány následujícím vztahem [7]:

$$\hat{y}_{ijk} = \mu + \underbrace{\mu_{i.} - \mu} + \underbrace{\mu_{.j} - \mu} + \underbrace{\mu_{.ij} - \mu_{i.} - \mu_{.j} - \mu}, \tag{20}$$

Rozdíly mezi hodnotami pozorovanými $y_{ijk}$ a vyrovnanými $\hat{y}_{ijk}$ modelovanými procedurou GLM jsou definovány výrazem

$$\hat{\varepsilon}_{ijk} = y_{ijk} - \hat{y}_{ijk}. \tag{21}$$

Pomocí příkazů procedury GLM lze odhadovat také marginální střední hodnoty jednotlivých faktorů (příkaz LSMEANS) a jejich interakce, průměry na jednotlivých úrovních faktorů a interakce (s využitím odhadnutelných funkcí $\boldsymbol{L}$) a další statistiky.

**Odhady lineárních funkcí parametrů lineárního modelu**

Procedura GLM jako jedna z nejstarších statistických procedur v systému SAS dokáže nejen odhadovat samotné parametry lineárních modelů, ale je vybavena také příkazy a parametry k odhadům lineárních funkcí parametrů modelu. Odhady

lineárních funkcí parametrů modelu vytváří podmínky pro testování lineárních modelů pomocí obecných lineárních hypotéz [4, 10], které tak mohou být konstruovány od jednoduchých až po komplexní porovnávání. Pro generování výstupů s lineárními kombinacemi odhadů parametrů modelu pro testování obecnými lineárními hypotézami je procedura GLM vybavena příkazy ESTIMATE a CONTRAST[5].

Podobně jako hypotéza o nevýznamnosti odhadnutého modelu (celkový *F*-test podle tabulky č. 1) nebo analýza rozptylu obecného lineárního modelu (viz tabulka č. 2) je obecná lineární hypotéza [4] založena na vhodném rozkladu součtu čtverců a testování relativního podílu rozptylu připadajícího na odhadovanou funkci parametrů modelu (*F* statistika) [7].

Základní syntaxe příkazu ESTIMATE je:

```
estimate 'label' effect-name effect-coefficients;,
```

kde koeficienty efektů jsou reprezentovány lineární kombinací parametrů modelu a '*label*' představuje znakový řetězec identifikující výsledek ve výstupu procedury GLM.

Nechť má obecný lineární model (16) 2 pevné faktory, a to faktor A se 3 úrovněmi a B se 2 úrovněmi. Potom lze obecný lineární model podle (16) vyjádřit rovnicí jako:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \tag{24}$$

kde $i = 1, 2, 3$, $j = 1, 2$ a $k = 1, \cdots, n_{ij}$.

Vektor odhadovaných parametrů $\boldsymbol{\beta}$ odpovídajících modelu podle vztahu (24) je:

$$\boldsymbol{\beta} = (\mu\ \alpha_1\ \alpha_2\ \alpha_3\ \ \beta_1\ \ \beta_2\ (\alpha\beta)_{11}\ \ (\alpha\beta)_{12}\ (\alpha\beta)_{21}\ (\alpha\beta)_{22}\ (\alpha\beta)_{31}\ (\alpha\beta)_{32}) \tag{25}$$

Vektor parametrů (25) se z pohledu jednotlivých faktorů dá zjednodušeně vyjádřit:

$$\boldsymbol{\beta} = (Intercept|Faktor\ A|Faktor\ B|Interakce) \tag{26}$$

Odhadovaná lineární kombinace parametrů se definuje jako odhadnutelná funkce $\boldsymbol{L\beta}$ [3], ve které je vektor $\boldsymbol{L}$ konstruován podobně jako vektor parametrů modelu s vahami odpovídajícími každému efektu v modelu. Formálně lze vektor vah $\boldsymbol{L}$ vyjádřit

$$\boldsymbol{L} = \begin{pmatrix} Váha\ pro \\ Intercept \end{vmatrix} \begin{matrix} Váha\ pro \\ Faktor\ A \end{vmatrix} \begin{matrix} Váha\ pro \\ Faktor\ B \end{vmatrix} \begin{matrix} Váha\ pro \\ Interakci \end{matrix} \end{pmatrix} \tag{27}$$

Například funkce střední hodnoty vycházející z rovnice (24) určená k odhadu střední hodnoty $\mu_{1\cdot}$ úrovně 1 faktoru A, tj. $A_1$, se vyjádří ve tvaru:

$$\mu_{1\cdot} = \mu + \alpha_1 + \frac{\beta_1}{2} + \frac{\beta_2}{2} + \frac{(\alpha\beta)_{11}}{2} + \frac{(\alpha\beta)_{12}}{2}. \tag{28}$$

*[5] Příkazy ESTIMATE a CONTRAST disponují také další procedury lineárního modelování jako je procedura GENMOD, GLIMMIX, LOGISTIC, MIXED, SURVEYREG a SURVEYLOGISTIC.*

Vektor vah ***L***, který odpovídá funkci odhadu podle (28), je definován jako:

$$\boldsymbol{L} = (1|\ 1\ 0\ 0|\ 0.5\ 0.5|\ 0.5\ \ 0.5\ 0\ 0\ 0\ 0\ ) \tag{29}$$

Váhy ve vektoru ***L*** musí být podle [7] pro každý faktor rozděleny rovnoměrně a pro každý faktor musí být součet vah roven 1. Na základě toho syntaxe příkazu ESTIMATE pro obecnou lineární hypotézu odhadu střední hodnoty $\mu_{1.}$ úrovně 1 faktoru A podle rovnice (28) dostane tvar:

```
estimate 'L:mean a1' intercept 1 a 1 0 0 b 0.5 0.5 a*b 0.5 0.5 0 0 0 0;
```

V případě, že obecný lineární model podle (24) se 2 faktory obsahuje také i 1 vysvětlující proměnnou (kovariátu) $x_{ijk}$, rovnice modelu (24) se změní na tvar:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma x_{ijk} + \varepsilon_{ijk}, \tag{30}$$

kde $i = 1, 2, 3$, $j = 1, 2$ a $k = 1, \cdots, n_{ij}$. Parametr $\gamma$ odpovídá směrnici regresní přímky.

Vektor odhadovaných parametrů $\boldsymbol{\beta}$ modelu (30) s 1 kovariátou se změní na:

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

K odhadu střední hodnoty $\mu_{1.}$ úrovně 1 faktoru A, tj. $A_1$, se funkce střední hodnoty vycházející z rovnice modelu se 2 faktory a 1 vysvětlující proměnnou vyjádří ve tvaru:

$$\mu_{1.} = \mu + \alpha_1 + \frac{\beta_1}{2} + \frac{\beta_2}{2} + \frac{(\alpha\beta)_{11}}{2} + \frac{(\alpha\beta)_{12}}{2} + \bar{x}_{ijk.} \tag{32}$$

Vektor vah ***L***, který odpovídá funkci odhadu podle (32), je definován rovnicí:

$$\boldsymbol{L} = \left(1|\ 1\ 0\ 0|\ 0.5\ 0.5|\ 0.5\ \ 0.5\ 0\ 0\ 0\ 0\ \middle|\bar{x}_{ijk}\ \right), \tag{33}$$

kde $\bar{x}_{ijk}$ je průměr vysvětlující proměnné $x_{ijk}$, $i = 1, 2, 3$, $j = 1, 2$ a $k = 1, \cdots, n_{ij}$.

Příkaz CONTRAST má syntax velmi podobnou jako příkaz ESTIMATE . V případě, že jsou váhy vektoru ***L*** součástí příkazu CONTRAST, musí jejich součet ve vektoru vah být roven 0. Podobným způsobem je možné odhadovat jakékoliv kombinace lineárních funkcí parametrů lineárního modelu s testováním významnosti takto získaných odhadů pomocí obecných lineárních hypotéz.

## 4. UKÁZKA ODHADU PARAMETRŮ LINEÁRNÍHO MODELU V SYSTÉMU SAS

Programový systém SAS je vybaven několika desítkami statistických procedur, které jsou předurčeny pro různé typy regresní analýzy. Modely pro regresi mají svůj původ v charakteristikách sledované proměnné (diskrétní nebo spojitá, normálně nebo nenormálně rozdělená), v předpokladech o tvaru modelu (lineární, nelineární nebo zobecněný lineární), o mechanismu generování dat (data pocházející z náhodného výběru, pozorování nebo experimentální data) a v metodě odhadu. Pro analýzu lineárních modelů se převážně využívá procedura REG a procedura GLM. Obě procedury umožňují odhady parametrů lineárních modelů.

Pro ukázku modelování lineárního modelu a analýzy rozptylu byl zvolen soubor dat CLASS ze systémové knihovny SASHELP. Datový soubor SASHELP.CLASS obsahuje informace o malé fiktivní třídě studentů. Proměnné souboru zahrnují 2 znakové a 4 numerické proměnné a struktura souboru je ilustrována obrázkem č. 1:

***Obrázek č. 1: Zobrazení struktury datového souboru SASHELP.CLASS***

Variables in Creation Order

| # | Variable | Type | Len |
|---|---|---|---|
| 1 | Name | Char | 8 |
| 2 | Sex | Char | 1 |
| 3 | Age | Num | 8 |
| 4 | Height | Num | 8 |
| 5 | Weight | Num | 8 |

***Zdroj: vlastní zpracování autora podle SAS Institute Inc. 2023b***

Pro účely ukázky použití regresní procedury REG bude použit lineární model ve tvaru:

$$Weight_i = Height_i + \varepsilon_i. \tag{34}$$

Modelování obecného lineárního modelu bude realizování pomocí procedury GLM s využitím singulární parametrizace podle (6). Obecný lineární model tak bude tvořen výrazem:

$$Weight_i = Sex + Height_i + Height_i * Sex + \varepsilon_i. \tag{35}$$

Znaková proměnná Sex bude v modelu (35) sehrávat úlohu klasifikačního faktoru. Tato datová sada se často používá v dokumentaci SAS k ilustraci kódování SAS.

## 4.1. ODHADY PARAMETRŮ LINEÁRNÍHO MODELU PROCEDUROU REG

Pro odhady parametrů pomocí regresních procedur systému SAS se požadovaný model ve tvaru podle (34) v proceduře REG nastaví příkazem MODEL. Graf modelované závislosti, diagnostické grafy a jiné grafy zajistí příkaz PLOTS. Syntaxe příkazu pro modelování lineární závislosti (pro data CLASS ze systémové knihovny SASHELP) procedurou REG je následující:

```
title 'Simple Linear Regression';
ods graphics on;

proc reg data=sashelp.class plots(unpack);
     model Weight = Height;
run;

quit;
```

Odhadovaný lineární regresní model (34) obsahuje absolutní člen (intercept) a 1 nezávislou proměnnou HEIGHT. Analýzou rozptylu byla zjištěna hodnota *F* statistiky v testu významnosti modelu ve výši 57,08, na základě které lze

s pravděpodobností <0,0001 zamítnout hypotézu o statistické nevýznamnosti odhadnutého modelu. Odhadnutý regresní model vysvětluje velkou část variability vysvětlované proměnné HEIGHT, o čem vypovídá dosažená hodnota koeficientu determinace $R^2$. Zjištěné hodnoty *t* statistik dovolují zamítnout nulovou hypotézu o nevýznamnosti koeficientu beta u vysvětlující proměnné WEIGHT. Lze zamítnout také i nulovou hypotézu o nevýznamnosti absolutního členu (interceptu) v modelu. Uvedená interpretace statistických testů je oprávněná jen za předpokladu normálně rozdělených chyb v odhadovaném modelu. Informace po provedeném odhadu ze statistické regresní procedury REG jsou ilustrovány na obrázku č. 2.

***Obrázek č. 2: Výstupy regresních procedur při odhadech parametrů lineárního modelu***

**Simple Linear Regression**

**Model: MODEL1**
**Dependent Variable: Weight**

| | |
|---|---|
| Number of Observations Read | 19 |
| Number of Observations Used | 19 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 7193.24912 | 7193.24912 | 57.08 | <.0001 |
| Error | 17 | 2142.48772 | 126.02869 | | |
| Corrected Total | 18 | 9335.73684 | | | |

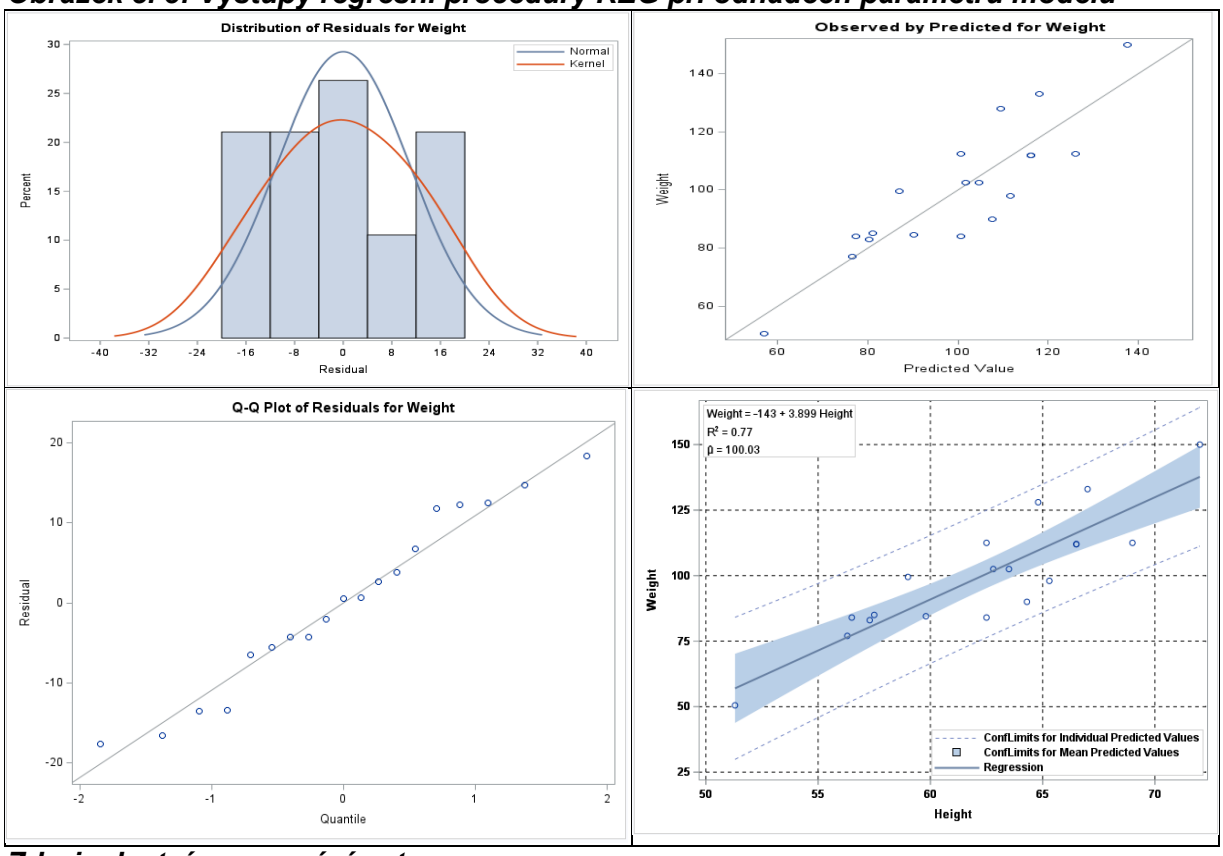
| | | | |
|---|---|---|---|
| Root MSE | 11.22625 | R-Square | 0.7705 |
| Dependent Mean | 100.02632 | Adj R-Sq | 0.7570 |
| Coeff Var | 11.22330 | | |

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -143.02692 | 32.27459 | -4.43 | 0.0004 |
| Height | 1 | 3.89903 | 0.51609 | 7.55 | <.0001 |

***Zdroj: vlastní zpracování autora***

***Obrázek č. 3: Výstupy regresní procedury REG při odhadech parametrů modelu***




***Zdroj: vlastní zpracování autora***

Grafické zobrazení závislosti naměřených pozorování a odhadnutého modelu včetně rovnice a vybraných statistik – koeficientu determinace a průměru závisle proměnné je znázorněno na obrázku č. 3. Na obrázku č. 3 jsou vybrány také některé diagnostické grafy, jejichž pomocí se testuje splnění předpokladů, například normalita a nezávislost reziduí.

Použití procedury REG pro analýzu lineárního modelu je jednoduché, intuitivní a uživatelsky dostatečně přívětivé. Statistická procedura REG poskytuje také i velké množství diagnostických nástrojů, které pomáhají v mnoha otázkách regresní analýzy. Podrobnější popis příkazů a parametrů procedury REG a přesné matematické definice jednotlivých dostupných statistik na výstupu regresní procedury REG lze dohledat v originální dokumentaci programového systému SAS [9].

## 4.2. ODHADY PARAMETRŮ LINEÁRNÍHO MODELU PROCEDUROU GLM

### Odhad parametrů obecného lineárního modelu bez klasifikačního faktoru

V případě, že se do obecného lineárního modelu (35) nezahrne klasifikační faktor, lze nahlížet na obecný lineární model jako na lineární model podle vztahu (34). Syntaxe příkazů procedury GLM se stává shodnou se syntaxí příkazů procedury REG. Jediný rozdíl v použití procedury GLM je vložení parametru SOLUTION v příkazu MODEL:

```
title 'Simple Linear Regression';
ods graphics on;
```

```
proc glm data=sashelp.class plots(unpack);
     model Weight = Height / solution;
run;

quit;
ods graphics off;
```

Odhadnuté parametry i ostatní přidružené statistiky mají stejné hodnoty i stejnou interpretaci jako u vykonaného odhadu parametrů procedurou REG.

**Odhad parametrů obecného lineárního modelu s klasifikačním faktorem**

V případě, že je v modelu klasifikační faktor zahrnut, jedná se o obecný lineární model (35). Proto k odhadování parametrů modelu lze použít výlučně proceduru GLM opět s parametrem SOLUTION v příkazu MODEL:

```
title 'One-Factor ANOVA';
ods graphics on;

proc glm data=sashelp.class plots(unpack);
     class Sex (ref='F');
     model Weight = Sex Height Sex*Height / solution;
run;
ods graphics off;
```

Ve srovnání s příkazy procedury REG je v příkazech procedury GLM příkaz CLASS, kterým se do modelu vkládá klasifikační faktor SEX. Singulární parametrizace v modelu (35) je vyjádřena explicitně výrazem (ref=‘ *F*‘), a proto úroveň *F* klasifikačního faktoru je úrovní referenční. Jedná se tedy o obecný lineární model v aditivním tvaru s 1 klasifikačním faktorem SEX a jeho interakcí s vysvětlovanou proměnnou HEIGHT.

Výstupem procedury GLM po realizované analýze rozptylu a odhadu parametrů modelu podle (35) jsou výstupy ilustrované na obrázku č. 4. Informace, které procedura GLM ve svém implicitní nastavení představuje, lze rozdělit na několik relativně samostatných částí. Podobně jako u ostatních regresních procedur (i u procedury REG popsané v předešlé podkapitole) jsou nejprve uvedeny informace o počtech pozorování vstupujících do zpracování procedurou GLM. Následují informace o provedené analýze rozptylu vycházející z definičních vztahů tabulky č. 2. Přehled výsledků uzavírá výstup s hodnotami odhadnutých parametrů obecného lineárního modelu. Podobně jako u procedury REG informace na výstupu slouží k testování hypotézy o významnosti parametrů, faktorů i modelu jako celku. Výstupy procedury GLM po analýze obecného lineárního modelu a odhadu jeho parametrů představuje obrázek č. 4.

***Obrázek č. 4: Výstupy procedury GLM po analýze rozptylu obecného lineárního modelu***

**One-Factor ANOVA**

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| Sex | 2 | M F |

| | |
|---|---|
| Number of Observations Read | 19 |
| Number of Observations Used | 19 |

Page Break

**One-Factor ANOVA**

**Dependent Variable: Weight**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 7402.992420 | 2467.664140 | 19.15 | <.0001 |
| Error | 15 | 1932.744422 | 128.849628 | | |
| Corrected Total | 18 | 9335.736842 | | | |

| R-Square | Coeff Var | Root MSE | Weight Mean |
|---|---|---|---|
| 0.792974 | 11.34821 | 11.35120 | 100.0263 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Sex | 1 | 1681.122953 | 1681.122953 | 13.05 | 0.0026 |
| Height | 1 | 5696.840666 | 5696.840666 | 44.21 | <.0001 |
| Height*Sex | 1 | 25.028801 | 25.028801 | 0.19 | 0.6657 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Sex | 1 | 15.202417 | 15.202417 | 0.12 | 0.7360 |
| Height | 1 | 5654.260964 | 5654.260964 | 43.88 | <.0001 |
| Height*Sex | 1 | 25.028801 | 25.028801 | 0.19 | 0.6657 |

| Parameter | | Estimate | | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | | -117.3697952 | B | 48.60161448 | -2.41 | 0.0290 |
| Sex | M | -23.7312215 | B | 69.08843486 | -0.34 | 0.7360 |
| Sex | F | 0.0000000 | B | . | . | . |
| Height | | 3.4244052 | B | 0.79971932 | 4.28 | 0.0007 |
| Height*Sex | M | 0.4881440 | B | 1.10756565 | 0.44 | 0.6657 |
| Height*Sex | F | 0.0000000 | B | . | . | . |

*Poznámka: Výrazy, za jejichž odhady následuje písmeno "B", nejsou jednoznačně odhadnutelné. Matice X'X je singulární, a k řešení normálních rovnic byla použita zobecněná inverze podle [2].*

***Zdroj: vlastní zpracování autora***

Možná interpretace informací z výstupu analýzy obecného lineárního modelu a odhadu parametrů je následující:

V testu významnosti modelu (tzv. nulová hypotéza pro odhadovaný model) byla zjištěna hodnota statistiky *F* ve výši 19,15, která umožňuje hypotézu o nevýznamnosti odhadovaného modelu zamítnout. Byl zaznamenán vysoký podíl variability vysvětlované proměnné HEIGHT modelem podle (35), čemuž nasvědčuje koeficient determinace na úrovni 0,792974. Při testování hypotézy o významnosti klasifikačního faktoru SEX a interakci tohoto faktoru s numerickou proměnnou WEIGHT nabízí procedura GLM v základních nastaveních k posouzení 2 typy součtů čtverců. Jedná

se o rozklady součtů čtverců, které jsou podle [9] ve výstupech procedury GLM označeny jako Type I SS a Type III SS.

Oba typy rozkladů součtů čtverců vycházejí ze stejných definičních vztahů (podle tabulky č. 2). Rozdíl mezi rozklady sum čtverců typu I a rozklady sum čtverců typu III spočívá ve způsobu zahrnutí faktoru do analyzovaného modelu při testování jeho významnosti. Zatímco u rozkladů podle typu I jsou hodnoty součtu čtverců jednotlivých faktorů závislé na hodnotě součtu čtverců faktoru již zahrnutého v modelu, rozklady součtů čtverců podle typu III na pořadí zahrnutí faktoru do modelu závislé nejsou. V obou typech rozkladů sum čtverců má zdroj variability (faktor nebo interakce), který je do rozkladů typu I vybrán jako poslední, stejnou hodnotu u rozkladů podle typu III. Porovnání součtů čtverců obou typů u jednotlivých faktorů umožňuje lépe kvantifikovat vliv faktorů, resp. jejich interakcí, a tedy lépe posoudit významnost konkrétního faktoru pro daný model[6].

V tabulce rozkladu čtverců typu I jednotlivé řádky postupně shora dolů udávají, o kolik se přidáním daného členu zmenší reziduální součet čtverců. Obecně tedy závisí na pořadí, v jakém se jednotlivé členy (faktory, interakce) objevují. V každém řádku tedy statistika *F* (prostřednictvím příslušné dosažené hladiny testu *p*) vypovídá o významnosti té části variability závisle proměnné, kterou nelze vysvětlit pomocí všech výše uvedených členů a kterou daný faktor či interakce vysvětluje. Rozklad čtverců typu III hodnotí přínos daného faktoru (interakce faktorů) po adjustaci vůči všem ostatním členům bez ohledu na jejich pořadí. Pro tento rozklad je obtížné hledat interpretaci, protože hodnotí vzrůst reziduálního součtu čtverců způsobený vyloučením daného faktoru či interakce, když v modelu zůstanou (je provedena adjustace vůči nim) všechny ostatní faktory či interakce včetně případných interakcí, v nichž je člen (faktor či interakce) obsažen.

Na základě testování významnosti jednotlivých zdrojů variability (faktorů a interakcí) byla u proměnné HEIGHT hypotéza o nevýznamnosti faktoru zamítnuta – a to i podle rozkladů typu I a typu III. U ostatního faktoru, resp. interakce se nedá na základě zjištěných hodnot *F* statistik nevýznamnost vyloučit. S podobným výsledkem je možné interpretovat i odhady parametrů, kdy se za statisticky významné dají považovat odhadnuté parametry absolutního členu (intercept) a parametru přidruženého k proměnné HEIGHT. Parametrizací modelu z důvodu zahrnutí klasifikační proměnné se matice $\boldsymbol{X'WX}$ stala singulární a k řešení normálních rovnic musela být použita její zobecněná inverzní matice. Podobně jako u procedury REG jsou součástí odhadu procedurou GLM diagnostické grafy pro kontrolu splnění předpokladů lineární regrese.

**Porovnání hodnot úrovní klasifikačního faktoru**

Důležitým úkolem při analýze dat s klasifikačními efekty je odhadnout typickou hodnotu vysvětlované proměnné pro každou úroveň daného efektu; často je potřebné tyto odhady také porovnat, které úrovně jsou z pohledu vysvětlované proměnné rovnocenné. Procedura GLM se s úkolem porovnávání hodnot na jednotlivých úrovních klasifikačního faktoru vypořádává 2 způsoby, a to:

- pomocí výběrových aritmetických průměrů,
- tzv. průměry nejmenších čtverců (*LS*-průměry).

*[6] Rozklady čtverců typu I jsou nazvány sekvenčními rozklady součtu čtverců a součty čtverců typu III parciálními.*

Tyto statistiky (výběrové aritmetické průměry a *LS*-průměry) obecně nejsou shodné a svým rozdílem odrážejí míru nevyváženosti dat pro analýzu rozptylu, resp. vliv faktorů anebo interakcí. *LS*-průměry jsou podle [3] považovány za odhadované marginální populační střední hodnoty. Porovnání skupin se uskutečňuje na základě příkazu MEANS a LSMENS procedury GLM. Příkladem může být syntaxe:

```
ods graphics on;
proc glm data=sashelp.class plots(unpack)=all;
     class Sex (ref='F') / ;
     model Weight = sex Height sex*height / solution;
     means Sex;
     lsmeans Sex;
     lsmeans Sex / at means;
     lsmeans Sex / at Height=80;
run;

quit;

ods graphics off;
```

Na základě výše uvedené sekvence příkazů procedura GLM příkazem MEANS počítá průměry pro všechny úrovně faktoru SEX. Příkaz LSMEANS odhaduje střední hodnoty, a to marginální pro úrovně faktoru SEX. Obsahuje-li příkaz LSMEANS také parametr AT, procedura GLM provádí odhady v příslušné hodnotě, resp. úrovni faktoru, např. příkaz LSMEANS Sex / AT HEIGHT = 80 odhaduje marginální průměr pro HEIGHT = 80.

## 5. ZÁVĚR

Cílem článku bylo přiblížit význam a možnosti programového systému SAS z pohledu statistického modelování. Statistické modelování prošlo v posledních desetiletích bouřlivým vývojem, který byl urychlován neustálým rozvojem moderních technologií a rozšiřováním neustále se zkvalitňující výpočetní techniky. Z celé široké oblasti modelování ve statistice byl ve článku položen důraz na klasické a obecné lineární modely, protože patří v oblasti statistického modelování k těm nejdůležitějším.

Článek zavádí nezbytné základní pojmy a definice, a tak poskytuje teoretický úvod do sledované tématiky. Na několika vybraných příkladech byly demonstrovány ukázky modelovací techniky procedurami REG a GLM, které tvoří pilíře regresních technik programového systému SAS. Výstupy z regresní analýzy obou procedur byly dostatečně interpretovány tak, aby poskytly všem výzkumníkům a datovým analytikům návod k dalšímu zkoumání.

## LITERATURA

[1] BELSLEY, D.A. – KUH, E. – WELSCH, R. E.: Regression diagnostics: Identifying Influential Data and Sources of Collinearity.. New York: John Wiley and Sons. 1980. 321 s. ISBN ISBN 0-47 1-69 1 17-8.
[2] BEN-ISRAEL, A. – GREVILLE, T. N. E.: Generalized inverses: theory and applications. Springer-Verlag New York, 2003. Inc.:New York. 436.s. ISBN 0-387-00293-6.

[3] LITTELL, C. L. – STROUP, W. W. – FREUND, R. J.: SAS for Linear Models, 2002. 4th ed. Cary, NC: SAS Institute Inc. xv + 478 s. ISBN 978-1-59047-023-7.
[4] LITTELL, R. C.: The Evolution of Linear Models in SAS: A Personal Perspective. SAS Global Forum 2011. [cit. 02. 07. 2023]. Dostupné na: http://support.sas.com/resources/papers/proceedings11/325-2011.pdf
[5] McCULLAGH, P. – NELDER, J. A.: Generalized Linear Models. 2nd ed. London: Chapman & Hall, 1989. 526 s.ISBN 0-412-31760-5.
[6] NELDER, J.A. – WEDDERBURN, R.W.M.: Generalized Linear Models. In: Journal of the Royal Statistical Society. Series A (General). Vol. 135, No. 3 1972, pp. 370 – 384.
[7] RUTHERFORD, A.: ANOVA and ANCOVA: a GLM approach, 2nd Edition. John Wiley & Sons, Inc.: Hoboken, New Jersey. 2011. xvi + 344 s. ISBN 978-0-470-38555-5.
[8] SAS Institute Inc. 2023a. SAS/STAT® 15.3 User's Guide. Chapter 1 Procedures by Category. Cary, NC: SAS Institute Inc. [cit. 02. 07. 2023]. Dostupné na: http://support.sas.com/documentation/onlinedoc/stat/153/statug.pdf.
[9] SAS Institute Inc. 2023b. SAS/STAT® 15.3 User's Guide. Chapter 4 Introduction to Statistical Modelling with SAS/STAT Software. Cary, NC: SAS Institute Inc. s. 26. [cit. 03. 07. 2023].
Dostupné na: http://support.sas.com/documentation/onlinedoc/stat/153/statug.pdf.
[10] TOBIAS, R. – KIERNAN, K. – TAO, J. – GIBBS, P.: CONTRAST and ESTIMATE Statements Made Easy: The LSMESTIMATE Statement. SAS Global Forum 2011. [cit. 22. 08. 2023]. Dostupné na:
https://support.sas.com/resources/papers/proceedings11/351-2011.pdf.

**RESUMÉ**
Význam statistického modelování v analýze dat se v posledních několika desetiletích rapidně zvýšil. Při zkoumání přírodních, či ekonomických nebo i jiných dat již nestačí pouhý popis těchto dat nebo klasická statistická inference o datech. V současnosti se do popředí zájmu datových analýz dostávají takové metody sofistikovaného zpracování dat, které umožňují pochopit mechanismus vzniku sledovaných dat, hledat a nacházet v datech nové poznatky a souvislosti mezi nimi a s rozumnou pravděpodobností dovolují předvídat vývoj pozorovaného jevu do určitého předem stanoveného horizontu. V průběhu let se navíc změnila i samotná sledovaná data. V důsledku rychlého vývoje výpočetní techniky a obecné digitalizace života se objem zkoumaných dat neustále zvyšuje, data jsou rozmanitější a jsou získávána vyšší rychlostí. Ze zkoumaných dat se tak zpravidla stávají větší a komplexnější soubory údajů, které pocházejí z nových zdrojů.
Prostředí, které je schopné čelit výše uvedeným výzvám v této oblasti, je právě programový systém SAS. SAS je analytický modulární systém, který již ve svém základním sestavení nabízí svým uživatelům velmi silné a výkonné funkcionality k analýzám a zpracování zkoumaných dat. Například pro statistické modelování tento analytický systém zahrnuje několik desítek statistických procedur, přičemž každá z nich se specializuje pro určité třídy statistického modelování. Příkladem takových funkcionalit, ve kterých jsou implementovány nejdůležitější poznatky moderní regresní vědy, jsou procedury REG a GLM.

**RESUME**
The importance of statistical modelling in data analysis has increased rapidly over the last few decades. When examining natural or economic or even other data, it is no

longer sufficient to simply describe the data or to make classical statistical inference about the data. Nowadays, data analysis is becoming more interested in sophisticated data processing methods that allow to understand the mechanism of the observed data, to search for and find new insights and connections between data and to predict with reasonable probability the development of the observed phenomenon within a certain predetermined horizon. In addition, the observed data itself has changed over the years. As a result of the rapid development of computer technology and the general digitisation of life, the volume of data examined is constantly increasing, the data are more varied and are being acquired faster. As a result, the examined data are generally larger and more complex data sets, arising from new sources.
An environment that is capable of meeting the above challenges in this area is the SAS software system. SAS is an analytical modular system that, already in its basic setup, offering its users very powerful functionalities for analysing and processing the data under study. For example, for statistical modelling, this analytical system includes several dozens of statistical procedures, each specialized for a particular class of statistical modelling. The examples of such functionalities, which implement the most important findings of the modern regression science, are the REG and the GLM procedures.

***PROFESIJNÝ ŽIVOTOPIS***

***Ing. Roman Pavelka, PhD****., v letech 1995 – 2010 pracoval v poradenské společnosti Trexima, s. r. o. Na pozici statistik – analytik se zabýval analýzami zejména mzdových a personálních dat. Podílel se na tvorbě pravidelných statistických přehledů a reportů. Spolupracoval s akademickými pracovišti, agenturami i soukromými subjekty na realizaci a vyhodnocování ad hoc statistických výzkumů. Oblast jeho vědeckého zájmu představují výběrová šetření, odhady a statistické modely. V letech 2012 až 2013 se zúčastnil zahraniční stáže ve Velké Británii. Od roku 2013 působil v Národnom ústave certifikovaných meraní vzdelávania (NÚCEM), kde zajišťoval statistické vyhodnocování výsledků testování žáků a studentů. Od roku 2015 pracuje v odboru metod statistických zjišťování Štatistického úradu SR.*

## KONTAKT

roman.pavelka@statistics.sk