UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Lineární modely s kategoriálními vysvětlujícími proměnnými

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **doc. RNDr. Eva Fišerová, Ph.D.**
Vypracoval(a): **Ludmila Skákalová**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Ludmila Skákalová

Název práce: Lineární modely s kategoriálními vysvětlujícími proměnnými

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová, Ph.D.

Rok obhajoby práce: 2016

Abstrakt: Diplomová práce se zaměřuje na kategoriální vysvětlující proměnné v lineárních modelech. Zejména se věnuje různým a často ne příliš známým druhům kódování, které vzájemně srovnává na konkrétních datech. Dále se práce zabývá metodou umělých proměnných v lineárních modelech a jejím aplikacím. Vybrané aplikace metody umělých proměnných jsou provedeny na konkrétních datech.

Klíčová slova: kategoriální proměnná, umělá proměnná, regrese, anova, ancova, dummy

Počet stran: 89

Počet příloh: 1 CD

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Ludmila Skákalová

Title: Linears models with categorical explanatory variables

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Mathematical Applications

Supervisor: doc. RNDr. Eva Fišerová, Ph.D.

The year of presentation: 2016

Abstract: The Master's thesis focuses on categorical explanatory variables in linear models. It focuses mainly to various and often not too known types of coding that mutually compares on concrete data. This thesis also deals with a method of dummy variables in linear models and its applications. Selected applications of the method of dummy variables are performed on specific data.

Key words: categorical variable, dummy variable, regresion, anova, ancova, dummy

Number of pages: 89

Number of appendices: 1 CD

Language: Czech

Prohlášení

Prohlašuji, že jsem vytvořila tuto diplomovou práci samostatně za vedení doc. RNDr. Evy Fišerové, Ph.D. a že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 20. dubna 2016

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce doc. RNDr. Evy Fišerové, Ph.D. za obětavou spolupráci i za čas, který mi věnovala při konzultacích. Dále si zaslouží poděkování můj počítač, že vydržel moje pracovní tempo, a typografický systém TEX, kterým je práce vysázena.

Obsah

Úvod **7**

1 Lineární regresní model **9**

2 Druhy kódování kategoriálních proměnných **14**
2.1 Indikátory dummy 14
2.2 Indikátory simple 17
2.3 Indikátory effect 19
2.4 Helmertovy indikátory 21
2.5 Inverzní Helmertovy indikátory 23
2.6 Indikátory Profile 24
2.7 Inverzní indikátory Profile 26
2.8 Srovnání druhů kódování na konkrétních datech 27
2.9 Vlastní typ kódování kategoriální proměnné 36

3 Metoda umělých proměnných v regresní analýze **40**
3.1 Dichotomické (binární) umělé proměnné v modelu regresních přímek 40
3.1.1 Dichotomická umělá proměnná v regresním modelu s interakcemi 43
3.2 Vícekategoriální kvalitativní znaky 46
3.2.1 Interakce s vícekategoriálními kvalitativními proměnnými . 48
3.3 Modely s více kategoriálními proměnnými 48

4 Aplikace metody umělých proměnných **52**
4.1 Porovnávání separovaných regresních modelů 52
4.1.1 Modely s odlišným absolutním členem a rozdílnou směrnicí 52
4.1.2 Modely se stejnou směrnicí a rozdílnou úrovňovou konstantou 53
4.1.3 Modely s totožnou úrovňovou konstantou a rozdílnou směrnicí 54
4.2 Agregace kvantitativní proměnné 55

5 Umělé proměnné v ANOVĚ **58**

6 Analýza kovariance ANCOVA **60**

7 Umělé proměnné v ekonometrických modelech **63**
7.1 Posun lineárního modelu v prostoru 63
7.2 Posun lineárního modelu v čase 63
7.3 Změna regresních parametrů v prostoru i čase 64
7.4 Umělé proměnné a sezónnost 64

8 Po částech spojitý regresní model **66**

9 Příklady **69**
9.1 ANOVA s kódováním dummy a effect 69
9.2 ANCOVA 74
9.3 Regresní model s jednou 4-úrovňovou kategoriální proměnnou a jednou kvantitativní proměnnou 78

Závěr **84**

Úvod

Tato diplomová práce je věnována užití kategoriálních (kvalitativních) proměnných v regresní analýze, tedy znaků, u nichž nemůžeme zjistit měřitelné hodnoty, ale určujeme pouze rovnost či různost. V regresní analýze se ale často setkáváme se situacemi, kdy do modelu vstupují takovéto proměnné a pro zkoumání modelované situace mohou mít velký vliv. Kategoriální proměnné zavádíme do modelu především proto, aby se zabránilo neobjektivnímu hodnocení vlivu vysvětlujících proměnných na vysvětlovanou proměnnou v důsledku vynechání významné proměnné. Snažíme se vytvořit model s co nejmenší možnou chybou a komplexnějším výsledkem. Kategoriální proměnné budeme zavádět do regresních modelů pomocí tzv. umělých proměnných, které mají v regresní analýze velké využití. Setkáváme se s nimi v Ekonometrii, časových řadách při popisu sezónnosti, v analýze rozptylu a dalších oblastech.

Cílem této diplomové práce je prozkoumat možnosti použití lineárních statistických modelů v případě, kdy mezi vysvětlujícími proměnnými vystupují kategoriální proměnné. Práce se zabývá podrobnější analýzou umělých proměnných při různých typech kódování a jejich vybranými aplikacemi. Použití modelů a různých typů kódování je demonstrováno na řešení konkrétních úloh. Všechny výpočty a grafy jsou spočteny a vykresleny pomocí statistického softwaru `R`.

První část práce je věnována regresním modelům obecně a slouží spíše pro připomenutí některých základních vlastností regresního modelu a odhadu parametrů metodou nejmenších čtverců. Druhá část práce se již zabývá otázkou, jakým způsobem zavést kvalitativní proměnnou do regresního modelu. Jejím hlavním cílem je seznámit čtenáře s různými druhy kódování kategoriálního znaku. V této kapitole je zpracováno 7 typů kódování, ke každému typu kódování je odvozena designová matice. Různé druhy kódování se od sebe liší, mimo jiné interpretacemi odhadů parametrů. Pro lepší názornost jsou tyto interpretace odhadů demonstrovány na konkrétním příkladě. Třetí část práce popisuje metodu umělých proměnných. Tato část je rozdělena do tří hlavních oddílů a to na zavádění jedné

dichotomické (binární) umělé proměnné, jedné vícekategoriální umělé proměnné a nakonec kombinace těchto proměnných do modelu. V kapitolách je popsán postup vytvoření jednoduchých regresních modelů pomocí zmíněných typů umělých proměnných. Modely jsou následně znázorněny pomocí regresních přímek na nichž jsou demonstrovány interpretace odhadů parametrů regresního modelu. Poslední část práce je zaměřena na různé aplikace metody umělých proměnných. U vybraných aplikací je metoda předvedena na konkrétních datech.

1 Lineární regresní model

Než se dostaneme k vysvětlení pojmu kategoriální proměnná a ukázkám, jak s ní pracovat, začneme osvěžením pojmu lineární regresní model. Tato kapitola má za úkol připomenout základní vlastnosti lineárního regresního modelu, společně s cíly a vlastnostmi regresní analýzy. Na tyto poznatky se budeme v dalších kapitolách hojně odkazovat.

Lineární regresní model budeme uvažovat ve tvaru

$$\mathbf{Y}_{(n\times 1)} = \mathbf{X}_{(n\times k)}\boldsymbol{\beta}_{(k\times 1)} + \boldsymbol{e}_{(n\times 1)}, \tag{1}$$

kde $\mathbf{Y}_{(n\times 1)} = (Y_1, Y_2, \ldots, Y_n)'$ je náhodný vektor celkem n pozorování vysvětlované (závislé) proměnné Y. Nenáhodný vektor $\boldsymbol{\beta} = (\beta_1 \ldots, \beta_k)'$ je vektor k neznámých regresních parametrů, které vyjadřují vliv jednotlivých vysvětlujících proměnných na modelovanou veličinu a jejichž hodnoty hledáme. Vektor $\boldsymbol{e} = (e_1, \ldots, e_n)'$ je vektor náhodný chyb. Matice $\mathbf{X}_{(n\times k)}$ je tzv. **designová matice**, jejíž řádky odpovídají jednotlivým měřením a sloupce tvoří jednotlivé vysvětlující proměnné (znaky). Aby byl model jednoznačně definován, musí mít matice $\mathbf{X}$ plnou sloupcovou hodnost (v našem případě, kdy $k < n$, bude $rank(\mathbf{X}) = k$). Dále budeme předpokládat:

- Střední hodnota náhodného vektoru $\boldsymbol{e}$ je nulová. $E\boldsymbol{e} = \mathbf{0}$.
- Rozptyl složek náhodného vektoru $\boldsymbol{e}$ je konstantní: $var(e_i) = \sigma^2$, tzv. **Homoskedasticita**.
- Kovariance složek náhodného vektoru $\boldsymbol{e}$ je nulová $cov(e_i, e_j) = 0$, pro $i \neq j$, $\forall i, j = 1, \ldots, n$.
- Náhodný vektor $\boldsymbol{e}$ má normální rozdělení
- Regresní parametry mohou nabývat libovolných hodnot

Pokud platí podmínky modelu, pak pro náhodnou proměnnou $\mathbf{Y}$ platí:

- Střední hodnota náhodného vektoru $\mathbf{Y}$ je rovna $\mathbf{X}\boldsymbol{\beta}$: $E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$.

- Rozptyl složek náhodného vektoru $\mathbf{Y}$ je konstantní. $var(Y_i) = \sigma^2$ pro $\forall\ i\ =\ 1, \ldots, n$.
- Složky náhodného vektoru $\mathbf{Y}$ jsou nekorelované. $cov(Y_i, Y_j) = 0$ pro $i \neq j$, $\forall i, j = 1, \ldots, n$.
- Náhodný vektor $\mathbf{Y}$ má normální rozdělení o parametrech $\mathbf{X}\boldsymbol{\beta}$ a $\sigma^2\mathbf{I}_n$. $\mathbf{Y} \sim \mathbb{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.

Odhady parametrů vektoru $\boldsymbol{\beta}$ se v drtivé většině případů provádí metodou nejmenších čtverců, která dává nejlepší odhady s nejmenší chybou. Pro takto nadefinovaný regresní model je odhad MNČ tvaru

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{Y}, \tag{2}$$

kde součin matic $\mathbf{C}_{k\times n} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ nazýváme **matice kontrastů**.

Z podmínek modelu a metody nejmenších čtverců, pak plynou vlastnosti odhadů parametrů $\boldsymbol{\beta}$:

- Střední hodnota odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ je rovna parametru $\boldsymbol{\beta}$. $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$,
- Varianční matice odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ je rovna hodnotě $var\hat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$,
- Odhad $\hat{\boldsymbol{\beta}}$ má normální rozdělení o parametrech $\boldsymbol{\beta} = (\beta_1 \ldots, \beta_k)'$ a $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. $\hat{\boldsymbol{\beta}} \sim \mathbb{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

Připomeňme pár nejdůležitějších pojmů a vzorců:
Vyrovnané hodnoty $\hat{\mathbf{Y}}$ náhodné proměnné $\mathbf{Y}$ se vypočítají pomocí odhadů $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ vztahem $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. **Rezidui** lineárního regresního modelu $\hat{\mathbf{e}}$ se myslí odhadnutá chyba $\mathbf{e}$ reziduálního modelu, která se vypočte jako rozdíl původních napozorovaných hodnot náhodné veličiny $\mathbf{Y}$ a vyrovnaných hodnot $\hat{\mathbf{Y}}$, $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. Pro původní model napozorovaných náhodných veličin $\mathbf{Y}$ tedy platí vztah $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\mathbf{e}}$. **Reziduální součet čtverců RSČ** se pak vypočte jako součin vektorů reziduí $\hat{\mathbf{e}}$. RSČ= $\hat{\mathbf{e}}'\hat{\mathbf{e}}$. Reziduální součet udává modelem nevysvětlenou

část z původního **celkového součtu čtverců** $S_T^2 = (\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})$, kde $\bar{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ značí aritmetický průměr náhodné veličiny $\mathbf{Y}$. Reziduální součet čtverců se po dosazení za rezidua dá také vyjádřit vztahem:

$$\text{RSČ} = \hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$$

Reziduální součet čtverců se využívá k výpočtu jedné z nejdůležitějších charakteristik regresního modelu a to **indexu determinace** R^2:

$$R^2 = 1 - \frac{\text{RSČ}}{S_T^2}. \tag{3}$$

Index determinace udává jaký podíl rozptylu v pozorování náhodné veličiny $\mathbf{Y}$ se podařilo regresí vysvětlit. Jinými slovy čím větší je hodnota indexu determinace R^2, tím větší byla úspěšnost regrese. Jelikož je $R^2 \in < 0, 1 >$, pak nejlepší hodnotu $R^2 = 1$ dostaneme pro RSČ$= 0$. **Směrodatná odchylka (chyba) odhadu** $s(\hat{\beta}_j)$ regresních parametrů $\hat{\beta}_i$ je míra rozptýlení pozorovaných hodnot okolo regresní přímky.

$$s(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_{jj}}, \tag{4}$$

kde $\hat{\sigma}^2$ značí nestranný odhad rozptylu náhodné složky $\hat{\sigma}^2 = \frac{1}{n-k}$RSČ, v_{jj} jsou složky matice $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$.

Jedním z hlavních cílů regresní analýzy je ověření závislosti vysvětlované proměnné Y na vysvětlující proměnné x. K ověření této závosloti používáme především testování různých hypotéz, pomocí kterých testujeme, zda jsou vysvětlující proměnné x_i pro regresní model významné či nikoli. Nejvíce využívané testy jsou v tomto ohledu tzv. **dílčí t-testy** a **F-test**. Podívejme se proto na tyto testy z blízka.

Dílčí t-testy jsou dílčí testy významnosti jednotlivých parametrů β_i, $i = 1, \ldots, k$. Testujeme postupně nulovou hypotézu ve tvaru $H_0 : \beta_i = 0$ proti alternativní hypotéze $H_A : \beta_i \neq 0$. V případě zavedení absolutního členu do modelu pak testujeme také hypotézu tvaru: $H_0 : \mu = 0$ proti alternativě $H_A : \mu \neq 0$. Testová statistika má Studentovo t-rozdělení s $n - p$ stupni volnosti, kde $p = k$

pro model bez absolutního členu a $p = k + 1$ pro model s absolutním členem. Testová statistika dílčího t-testu (t-statistika) je dána vztahem [5]:

$$T_i = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 v_{ii}}}. \tag{5}$$

Hodnota $\hat{\sigma}^2$ značí nestranný odhad rozptylu náhodné složky, v_{ii} jsou složky matice $V = (X'X)^{-1}$. Hypotézy dílčích t-testů se vyhodnotí na základě p-hodnoty. P-hodnota představuje maximální možnou hladinu testu α, pro kterou hypotézu H_0 ještě nezamítneme. Pokud tedy hladina testu α bude větší než vypočtená p-hodnota, zamítneme H_0. V opačném případě nulovou hypotézu H_0 nelze zamítnout. V práci je volena hladina testu $\alpha = 0,05$. Pokud se ukáže, že pro konkrétní i nelze zamítnout nulovou hypotézu H_0, je třeba zvážit setrvání příslušné vysvětlující proměnné x_i v modelu. Pokud by se totiž parametr u příslušné proměnné neodlišoval významně od nuly, pak taková proměnná do modelu nic nového nepřináší a je na zvážení takovou proměnnou do modelu již nezahrnovat.

F-testy, oproti dílčím t-testům testují nulovost parametrů β komplexně. Testujeme hypotézu $H_0 : \mu, \beta_i = 0$ pro $\forall i = 1, \ldots, k$ oproti alternativě $H_A : \mu \neq 0$ nebo alespoň jedno $\beta_i \neq 0$. Testová statistika má F-rozdělení s $k-l$ a $n-k$ stupni volnosti, kde n je celkový počet pozorování náhodné veličiny $\mathbf{Y}$, k je počet parametrů β a l je počet parametrů v modelu za platnosti nulové hypotézy. Testová statistika F-testu (F-statistika) je tvaru [5] :

$$F = \frac{(n-k)[\text{RSČ}(RM) - \text{RSČ}(FM)]}{(k-l)\text{RSČ}(FM)} \sim F_{k-l,n-k},$$

Značkou RSČ značíme tzv, **reziduální součet čtverců**. Jako FM bereme celkový model se všemi vysvětlujícími veličinami X. Podmodel (RM) tohoto modelu je pro případ platnosti nulové hypotézy například, kde $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. Jestliže uvažujeme model s absolutním členem a nulová hypotéza předpokládá, že $\beta_i = 0$ pro $\forall i = 1, \ldots, k$, pak $l = 1$. F-statistiku lze v tomto případě vyjádřit také vztahem:

$$F = \frac{(n-k)}{(k-1)} \frac{R^2}{1-R^2} \sim F_{k-1,n-k}. \tag{6}$$

Nulovou hypotézu budeme opět posuzovat na základě p-hodnoty, která se vyhodnocuje stejně jako v případě dílčích t-testů. Čili je-li p-hodnota menší než zvolená hladina významnosti testu $\alpha = 0,05$, pak zamítneme nulovou hypotézu H_0, v opačném případě hypotézu H_0 nelze zamítnout.

2 Druhy kódování kategoriálních proměnných

Kategoriální (neboli kvalitativní) **proměnnou** Z se rozumí proměnná, u které nemůžeme zjistit měřitelné hodnoty. Jde například o proměnné typu: pohlaví, dosažené vzdělání, roční období, apod. V této kapitole si ukážeme různé druhy kódování kategoriální proměnné. Ty se využívají především pro jinou interpretaci odhadů parametrů β nezávislé kategoriální proměnné Z. Na samotný výsledek F-testu či indexu determinace, nemá zvolený druh kódování žádný vliv. Druhy kódování budou představeny na jednoduchém regresním modelu s jedinou kategoriální vysvětlující proměnnou, která nabývá k úrovní. Vzhledem k použití absolutního členu a zachování lineární nezávislosti sloupců designové matice $\mathbf{X}$ je třeba vynechat z modelu jednu úroveň kategoriální proměnné. Takovouto úroveň nazýváme **referenční**. Po představení všech druhů kódování bude na konci kapitoly uveden příklad, který srovná odhady parametrů $\boldsymbol{\beta}$ a jejich interpretace při užití různého druhu kódování kategoriální proměnné. [2]

2.1 Indikátory dummy

Nejpoužívanějším typem kódování v praxi je kódování pomocí tzv. dummy indikátorů. S tímto typem kódování se čtenář jistě setkal v mnoha případech, jen pravděpodobně nevěděl, že se jedná právě o dummy kódování. Dummy indikátory jsou založeny na intuitivním principu přiřadit pozorovanému subjektu hodnotu 1, jestliže patří do sledované úrovně, a 0 v případě, kdy do dané úrovně nespadá. Samozřejmě lze hodnoty 0 a 1 prohodit. Tedy v případě kdy do dané úrovně sledovaná veličina nenáleží, označit 1 a v případě náležení 0. Je ale nutné mít na zřeteli, že se touto změnou změní i interpretace výsledků. [2] [3]

Uvažujme jednoduchý model, kde nám náhodná veličina Y závisí na jediné kategoriální proměnné Z, nabývající dvou úrovní. Např. na proměnná „pohlaví", nabývající úrovní muž a žena. Spadá-li i-té pozorování do 1.úrovně (např. je-li objektem pozorování muž) označíme proměnnou $z_i = 1$. V opačném případě označíme proměnnou $z_i = 0$. V případě, kdy uvažujeme model s absolutním členem zavádíme do modelu pouze kategoriální proměnnou pro jedno pohlaví

a to z důvodu zachování lineární nezávislosti sloupců designové matice $\mathbf{X}$. Výsledný model obsahující absolutní člen μ je poté tvaru:

$$Y_i = \mu + \beta_1 z_i + e_i, \qquad i = 1, \ldots, n.$$

Parametr β_1 značí parametr kategoriální proměnné pro zvolené pohlaví (v našem případě pro 1.úroveň muže). Parametr β_2 kategoriální proměnné pro 2. úroveň (úroveň žena) je z modelu vynechán, jak již bylo řečeno kvůli zachování lineární nezávislosti sloupců designové matice $\mathbf{X}$. Říkáme, že úroveň "žena"byla zvolena jako referenční úroveň. Dosazením hodnot za z_i bychom dostali pro každou úroveň jiný model.

$$\begin{array}{ll} Y_i = \mu + \beta_1 + e_i, & i = 1, \ldots, n_1, \text{ pro pozorování z 1.úrovně} \\ Y_j = \mu + e_j, & j = 1, \ldots, n_2, \text{ pro pozorování z 2.úrovně} \end{array}$$

Zapíšeme-li model maticově, designová matice, kdy bez újmy na obecnosti předpokládejme, že n_1 prvních pozorování odpovídá 1. úrovni, bude tvaru:

$$\mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1 \text{ pozorování 1. úrovně} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2 \text{ pozorování 2. úrovně} \end{matrix}$$

Odhad parametrů $\mu,\ \beta_1$ vypočteme metodou nejmenších čtverců.

$$(\hat{\mu}, \hat{\beta}_1)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{Y}.$$

Podíváme se nejprve na tvar matice $\mathbf{C}$, kterou nazýváme maticí kontrastů:

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{pmatrix} n & n_1 \\ n_1 & n_1 \end{pmatrix}^{-1} \mathbf{X}' = \begin{pmatrix} \frac{1}{n-n_1} & -\frac{1}{n-n_1} \\ -\frac{1}{n-n_1} & \frac{n}{n_1(n-n_1)} \end{pmatrix} \begin{pmatrix} 1 & 1 & \ldots & 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 & 0 & \ldots & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & \ldots & 0 & \frac{1}{n-n_1} & \cdots & \frac{1}{n-n_1} \\ \frac{1}{n_1} & \ldots & \frac{1}{n_1} & -\frac{1}{n-n_1} & \cdots & -\frac{1}{n-n_1} \end{pmatrix}$$

Jelikož platí vztah $n = n_1 + n_2$ lze matici kontrastů $\mathbf{C}$ vyjádřit takto:

$$\mathbf{C} = \begin{pmatrix} 0 & \dots & 0 & \frac{1}{n_2} & \dots & \frac{1}{n_2} \\ \frac{1}{n_1} & \dots & \frac{1}{n_1} & -\frac{1}{n_2} & \dots & -\frac{1}{n_2} \end{pmatrix}.$$

Po dosazení do vzorce odhadu parametrů $(\hat{\mu}, \hat{\beta}_1)'$ dostaneme

$$(\hat{\mu}, \hat{\beta}_1)' = \mathbf{CY} = \begin{pmatrix} 0 & \dots & 0 & \frac{1}{n_2} & \dots & \frac{1}{n_2} \\ \frac{1}{n_1} & \dots & \frac{1}{n_1} & -\frac{1}{n_2} & \dots & -\frac{1}{n_2} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} =$$

$$\begin{pmatrix} \frac{1}{n_2}(Y_{n_1+1} + \dots + Y_{n_1+n_2}) \\ \frac{1}{n_1}(Y_1 + \dots + Y_{n_1}) - \frac{1}{n_2}(Y_{n_1+1} + \dots + Y_{n_1+n_2}) \end{pmatrix} = \begin{pmatrix} \bar{Y}_{2\cdot} \\ \bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} \end{pmatrix},$$

kde $\bar{Y}_{2\cdot}$ značíme průměr přes všechna pozorování 2. úrovně a $\bar{Y}_{1\cdot}$ značíme průměr přes všechna pozorování 1.úrovně.

Odhad parametru β_1 při použití dummy kódování kategoriální proměnné nám udává rozdíl mezi průměrnou hodnotou pozorování dané úrovně (v našem případě muž) a průměrnou hodnotou pozorování referenční úrovně (v našem případě žena). Parametr μ udává průměrnou hodnotu pozorování referenční úrovně (v našem případě žena). Parametru β_2 referenční úrovně je roven nule.

Nyní si představíme dummy kódování na složitějším modelu, kde uvažujeme k úrovní kategoriální proměnné Z. Zaveďme označení $\mathbf{1}_{n_i}$ je sloupcový vektor n_i jedniček. Dále $\mathbf{0}_{n_i}$ nechť je sloupcový vektor n_i nul, kde n_i, $i = 1, \dots k$ značí počet pozorování i-té úrovně kategoriální vysvětlující proměnné. Jelikož opět uvažujeme model s absolutním členem zvolíme jednu referenční úroveň, kterou z modelu vynecháme z důvodu zachování lineární nezávislosti sloupců designové matice $\mathbf{X}$. Jako referenční úroveň volíme poslední k-tou úroveň. Bez újmy na obecnosti opět uvažujme uspořádaný výběr, tedy že n_i pozorování všech k úrovní jde postupně za sebou. Tedy n_1 prvních pozorování odpovídá 1. úrovni, n_2 následujících pozorování odpovídá 2. úrovni, apod. Designovou matici $\mathbf{X}$ pak

můžeme zapsat ve tvaru:

$$\mathbf{X}_{n\times k} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \dots & \mathbf{0}_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \dots & \mathbf{1}_{n_{k-1}} \\ \mathbf{1}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \dots & \mathbf{0}_{n_k} \end{pmatrix}.$$

Hodnoty odhadnutých parametrů $\mu, \boldsymbol{\beta}$ odhadneme opět pomocí metody nejmenších čtverců (2). Nejprve vypočtěme matici kontrastů $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$:

$$\mathbf{C}_{k\times n} = \begin{pmatrix} & \mathbf{0}'_{n_1} & & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \dots & & \mathbf{0}'_{n_{k-1}} & \frac{1}{n_k}\mathbf{1}'_{n_k} \\ \frac{1}{n_1}\mathbf{1}'_{n_1} & & & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \dots & & \mathbf{0}'_{n_{k-1}} & -\frac{1}{n_k}\mathbf{1}'_{n_k} \\ & \mathbf{0}'_{n_1} & \frac{1}{n_2}\mathbf{1}'_{n_2} & & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \dots & & \mathbf{0}'_{n_{k-1}} & -\frac{1}{n_k}\mathbf{1}'_{n_k} \\ & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ & \mathbf{0}'_{n_1} & & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \dots & \frac{1}{n_{k-1}}\mathbf{1}'_{n_{k-1}} & -\frac{1}{n_k}\mathbf{1}'_{n_k} & \mathbf{1}'_{n_k} \end{pmatrix}.$$

Dosadíme do vzorce (2) pro výpočet odhadů parametru $\boldsymbol{\beta}$ a dostaneme.

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_{k-1} \end{pmatrix} = \begin{pmatrix} \bar{Y}_{k\cdot} \\ \bar{Y}_{1\cdot} - \bar{Y}_{k\cdot} \\ \bar{Y}_{2\cdot} - \bar{Y}_{k\cdot} \\ \bar{Y}_{3\cdot} - \bar{Y}_{k\cdot} \\ \vdots \\ \bar{Y}_{k-1\cdot} - \bar{Y}_{k\cdot} \end{pmatrix}.$$

Značením $\bar{Y}_{i\cdot}$ rozumíme průměr přes všechna pozorování i-té úrovně kategoriální vysvětlující proměnné. Odhady parametrů kvalitativní proměnné u dummy kódování představují rozdíl průměrné hodnoty napozorovaných dat konkrétní úrovně s průměrnou hodnotou napozorovaných dat tzv. referenční úrovně. V našem případě byla jako referenční úroveň zvolena k-tá úroveň, to znamená, že s ní porovnáváme všechny ostatní úrovně kategoriální proměnné. Parametr referenční úrovně je roven nule.[2] [3]

2.2 Indikátory simple

Indikátory simple, stejně jako další indikátory, které budou postupně představeny, se již nekódují tak intuitivně jako dummy indikátory. To je také jeden

z hlavních důvodů, proč jsou ostatní typy kódování tak velmi málo využívané a známé. [2] [3]

Uvažujeme regresní model s kategoriální proměnnou Z o k úrovních

$$Y_{ij} = \mu+\beta_1 z_{1j}+\beta_2 z_{2j}+\cdots+\beta_{k-1} z_{(k-1)j}+e_{ij}, \text{ pro } \forall i = 1,\ldots,k,\ j = 1,\ldots,n_i. \quad (7)$$

Symbol Y_{ij} značí j-té pozorování i-té úrovně pro $j = 1,\ldots,n_i,\ i = 1,\ldots,k$. Uvažujeme opět model s absolutním členem, tedy bylo třeba zvolit jednu referenční úroveň, která byla z modelu vynechána z důvodu zachování lineární nezávislosti sloupců designové matice $\mathbf{X}$. Jako referenční úroveň byla zvolena poslední k-tá úroveň.

Kódování proměnné $z_{ij},\ i = 1,\ldots,k-1,\ j = 1,\ldots,n_i$ pomocí indilátorů simple provádíme následovně: Pokud j-té pozorování spadá do i-té úrovně pak označme:

$$z_{ij} = \begin{cases} (1-\frac{1}{k}) & \text{pokud j-té pozorování patří do i-té úrovně,} \\ -\frac{1}{k} & \text{pokud j-té pozorování nepatří do i-té úrovně.} \end{cases}$$

Po dosazení hodnot z_{ij} do uvažovaného regresního modelu o k úrovních, dostáváme k různých modelů.

$$\begin{array}{ll}
Y_{1j} = \mu + \beta_1(1-\frac{1}{k}) - \frac{1}{k}\beta_2 - \frac{1}{k}\beta_3 + \ldots - \frac{1}{k}\beta_{k-1} + e_{1j}, & j = 1,\ldots,n_1, \\
Y_{2j} = \mu - \frac{1}{k}\beta_1 + \beta_2(1-\frac{1}{k}) - \frac{1}{k}\beta_3 + \ldots - \frac{1}{k}\beta_{k-1} + e_{2j}, & j = 1,\ldots,n_2, \\
\vdots & \\
Y_{(k-1)j} = \mu - \frac{1}{k}\beta_1 - \frac{1}{k}\beta_2 + \ldots - \frac{1}{k}\beta_{k-2} + \beta_{k-1}(1-\frac{1}{k}) + e_{(k-1)j}, & j = 1,\ldots,n_{k-1}, \\
Y_{kj} = \mu - \frac{1}{k}\beta_1 - \frac{1}{k}\beta_2 + \ldots - \frac{1}{k}\beta_{k-2} - \frac{1}{k}\beta_{k-1} + e_{kj}, & j = 1,\ldots,n_k.
\end{array}$$

Po úpravě dostáváme zjednodušený tvar modelů

$$\begin{aligned}
Y_{ij} &= \mu + \beta_i - \tfrac{1}{k}\sum_{l=1}^{k-1}\beta_l + e_{ij}, \quad \text{pro} \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,k-1, \\
Y_{kj} &= \mu - \tfrac{1}{k}\sum_{l=1}^{k-1}\beta_l + e_{kj}, \qquad \text{pro} \quad j = 1,\ldots,n_k.
\end{aligned}$$

Designová matice pro idikátory simple:

$$\mathbf{X}_{n\times k} = \begin{pmatrix} \mathbf{1}_{n_1} & (1-\frac{1}{k})\,\mathbf{1}_{n_1} & -\frac{1}{k}\,\mathbf{1}_{n_1} & -\frac{1}{k}\,\mathbf{1}_{n_1} & \dots & -\frac{1}{k}\,\mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & -\frac{1}{k}\,\mathbf{1}_{n_2} & (1-\frac{1}{k})\,\mathbf{1}_{n_2} & -\frac{1}{k}\,\mathbf{1}_{n_2} & \dots & -\frac{1}{k}\,\mathbf{1}_{n_2} \\ \mathbf{1}_{n_3} & -\frac{1}{k}\,\mathbf{1}_{n_3} & -\frac{1}{k}\,\mathbf{1}_{n_3} & (1-\frac{1}{k})\,\mathbf{1}_{n_3} & \dots & -\frac{1}{k}\,\mathbf{1}_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_{k-1}} & -\frac{1}{k}\,\mathbf{1}_{n_{k-1}} & -\frac{1}{k}\,\mathbf{1}_{n_{k-1}} & -\frac{1}{k}\,\mathbf{1}_{n_{k-1}} & \dots & (1-\frac{1}{k})\,\mathbf{1}_{n_{k-1}} \\ \mathbf{1}_{n_k} & -\frac{1}{k}\,\mathbf{1}_{n_k} & -\frac{1}{k}\,\mathbf{1}_{n_k} & -\frac{1}{k}\,\mathbf{1}_{n_k} & \dots & -\frac{1}{k}\,\mathbf{1}_{n_k} \end{pmatrix}.$$

Tvar matice kontrastů:

$$\mathbf{C}_{k\times n} = \begin{pmatrix} \frac{1}{kn_1}\,\mathbf{1}'_{n_1} & \frac{1}{kn_2}\,\mathbf{1}'_{n_2} & \frac{1}{kn_3}\,\mathbf{1}'_{n_3} & \frac{1}{kn_4}\,\mathbf{1}'_{n_4} & \dots & \frac{1}{kn_{k-1}}\,\mathbf{1}'_{n_{k-1}} & \frac{1}{kn_k}\,\mathbf{1}'_{n_k} \\ \frac{1}{n_1}\,\mathbf{1}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \dots & \mathbf{0}'_{n_{k-1}} & -\frac{1}{n_k}\,\mathbf{1}'_{n_k} \\ \mathbf{0}'_{n_1} & \frac{1}{n_2}\,\mathbf{1}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \dots & \mathbf{0}'_{n_{k-1}} & -\frac{1}{n_k}\,\mathbf{1}'_{n_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \dots & \frac{1}{n_{k-1}}\,\mathbf{1}'_{n_{k-1}} & -\frac{1}{n_k}\,\mathbf{1}'_{n_k} \end{pmatrix}. \qquad (8)$$

Odhady parametrů $\mu,\ \boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$ odhadnuté metodou nejmenších čtverců:

$$\hat{\mu} = \bar{Y}_{\cdot\cdot}, \quad \hat{\beta}_1 = \bar{Y}_{1\cdot} - \bar{Y}_{k\cdot}, \quad \dots, \quad \hat{\beta}_{k-1} = \bar{Y}_{k-1\cdot} - \bar{Y}_{k\cdot}.$$

Odhady parametrů $\beta_1, \dots, \beta_{k-1}$ zůstávají stejné jako v případě dummy indikátorů a udávájí rozdíl průměrné hodnoty pozorování dané úrovně s průměrnou hodnotou pozorování referenční úrovně. Změní se ovšem odhad absolutního členu μ, který nyní udává průměrnou hodnotu všech pozorování. Parametr β_k referenční úrovně je stejně jako v případě dummy kódování roven nule. Při užití kódování simple známe sice celkovou průměrnou hodnotu pozorování a rozdíl průměrné hodnoty referenční úrovně s průměrnými hodnotami ostatních úrovní, ale samotnou průměrnou hodnotu referenční úrovně neznáme.

2.3 Indikátory effect

Indikátory effect se používají v situacích, kdy je pro interpretaci parametrů výhodné vliv jednotlivých úrovní na hodnotu vysvětlované proměnné v součtu kompenzovat, tzn. položíme součet parametrů $\beta_1, \dots, \beta_k$ roven nule. [2] [3]. Kódování indikátorů effect je velmi podobné kódování indikátorů dummy s výjimkou kódování referenční, v našem případě k-té úrovně.

Uvažujeme regresní model s kategoriální proměnnou Z o k úrovních opět ve tvaru (7). Kódování proměnné $z_{ij},\ i = 1, \ldots, k-1$ pomocí indikátorů effect provádíme následovně:

$$z_{ij} = \begin{cases} 1 \text{ pokud j-té pozorování patří do i-té úrovně,} \\ 0 \text{ pokud j-té pozorování nepatří do i-té úrovně,} \\ -1 \text{ pokud j-té pozorování patří do k-té úrovně.} \end{cases}$$

Po dosazení hodnot proměnné z_{ij} při kódování effect dostaneme k různých modelů. Opět se od ostatních výrazně liší model pro pozorování spadající do referenční k-té úrovně.

$$\begin{array}{ll} Y_{1j} = \mu + \beta_1 + e_{1j}, & j = 1, \ldots, n_1, \\ Y_{2j} = \mu + \beta_2 + e_{2j}, & j = 1, \ldots, n_2, \\ \vdots & \\ Y_{(k-1)j} = \mu + \beta_{k-1} + e_{(k-1)}, & j = 1, \ldots, n_{k-1}, \\ Y_{kj} = \mu - \beta_1 - \beta_2 - \ldots - \beta_{k-1} + e_{kj}, & j = 1, \ldots, n_k. \end{array}$$

Po úpravě dostáváme:

$$\begin{array}{ll} Y_{ij} = \mu + \beta_i + e_{ij}, & j = 1, \ldots, n_i \quad i = 1, \ldots, k-1, \\ Y_{kj} = \mu - \sum_{l=1}^{k-1} \beta_l + e_{kj}, & j = 1, \ldots, n_k. \end{array}$$

Designová matice kategoriální proměnné při kódování effect:

$$\mathbf{X}_{n \times k} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \ldots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \ldots & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \ldots & \mathbf{0}_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \ldots & \mathbf{1}_{n_{k-1}} \\ \mathbf{1}_{n_k} & -\mathbf{1}_{n_k} & -\mathbf{1}_{n_k} & -\mathbf{1}_{n_k} & \ldots & -\mathbf{1}_{n_k} \end{pmatrix}.$$

Z metody nejmenších čtverců, plyne tvar matice kontrastů

$$\mathbf{C}_{k \times n} = \begin{pmatrix} \frac{1}{kn_1}\mathbf{1}'_{n_1} & \frac{1}{kn_2}\mathbf{1}'_{n_2} & \frac{1}{kn_3}\mathbf{1}'_{n_3} & \ldots & \frac{1}{kn_{k-1}}\mathbf{1}'_{n_{k-1}} & \frac{1}{kn_k}\mathbf{1}'_{n_k} \\ (1-\frac{1}{k})\frac{1}{n_1}\mathbf{1}'_{n_1} & -\frac{1}{kn_2}\mathbf{1}'_{n_2} & -\frac{1}{kn_3}\mathbf{1}'_{n_3} & \ldots & -\frac{1}{kn_{k-1}}\mathbf{1}'_{n_{k-1}} & -\frac{1}{kn_k}\mathbf{1}'_{n_k} \\ -\frac{1}{kn_1}\mathbf{1}'_{n_1} & (1-\frac{1}{k})\frac{1}{n_2}\mathbf{1}'_{n_2} & -\frac{1}{kn_3}\mathbf{1}'_{n_3} & \ldots & -\frac{1}{kn_{k-1}}\mathbf{1}'_{n_{k-1}} & -\frac{1}{kn_k}\mathbf{1}'_{n_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{kn_1}\mathbf{1}'_{n_1} & -\frac{1}{kn_2}\mathbf{1}'_{n_2} & -\frac{1}{kn_3}\mathbf{1}'_{n_3} & \ldots & (1-\frac{1}{k})\frac{1}{n_{k-1}}\mathbf{1}'_{n_{k-1}} & -\frac{1}{kn_k}\mathbf{1}'_{n_k} \end{pmatrix}.$$

Odhady parametrů $\mu, \beta_1, \ldots, \beta_{k-1}$ odhadnuté metodou nejmenších čtverců:

$$\hat{\mu} = \bar{Y}_{\cdot\cdot}, \quad \hat{\beta}_1 = \bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot}, \quad \ldots, \quad \hat{\beta}_{k-1} = \bar{Y}_{k-1\cdot} - \bar{Y}_{\cdot\cdot}.$$

Parametr pro každou úroveň pak představuje rozdíl průměrné hodnoty napozorovaných dat konkrétní úrovně s průměrnou hodnotou napozorovaných dat pro všechny úrovně kategoriální proměnné, tzv. efekt i-té úrovně. Jelikož součet parametrů $\sum_{i=1}^{k} \hat{\beta}_i = 0$ je roven nule, parametr pro k-tou úroveň lze snadno dopočítat.

$$\hat{\beta}_k = -\sum_{i=1}^{k-1} \hat{\beta}_i = (k-1)\bar{Y}_{\cdot\cdot} - \sum_{i=1}^{k-1} \bar{Y}_{i\cdot} = (k-1)\bar{Y}_{\cdot\cdot} - \sum_{i=1}^{k-1} \bar{Y}_{i\cdot} - \bar{Y}_{k\cdot} + \bar{Y}_{k\cdot} + \bar{Y}_{\cdot\cdot} - \bar{Y}_{\cdot\cdot} =$$
$$= k\bar{Y}_{\cdot\cdot} - \sum_{i=1}^{k} \bar{Y}_{i\cdot} + \bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot} = k\tfrac{1}{k} \sum_{i=1}^{k} \bar{Y}_{i\cdot} - \sum_{i=1}^{k} \bar{Y}_{i\cdot} + \bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot} = \bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot}.$$

Z výpočtu vyplývá, že také odhad parametru β_k představuje srovnání průměrné hodnoty pozorování k-té úrovně vysvětlované proměnné s celkovou průměrnou hodnotou pozorování, tedy s průměrnou hodnotou pozorování všech k úrovní.

2.4 Helmertovy indikátory

Úkolem Helmertových indikátorů je srovnání průměrné hodnoty napozorovaných dat konkrétní úrovně s průměrnou hodnotou napozorovaných dat všech následujících úrovní. Proto má smysl počítat s těmito indikátory pouze v případě, kdy existuje nějaké logické uspořádání úrovní kategoriální proměnné. Například seřazení krajů dle počtu obyvatel, agregace kvantitativní proměnné věk (11-20, 21-30, 31-40,...) apod. [2] [3]

Uvažujeme regresní model s kategoriální proměnnou Z o k úrovních ve tvaru (7), kde jako referenční úroveň byla zvolena poslední k-tá úroveň.

Kódování Helmertových indikátorů provádíme následujícím způsobem:

$$z_{ij} = \begin{cases} -\frac{1}{k-i+1} & \text{pokud j-té pozorování patří do l-té úrovně,} \\ & \text{kde } i < l,\ l = 1, \ldots, k-1, \\ 1 - \frac{1}{k-i+1} & \text{pokud j-té pozorování patří do i-té úrovně,} \\ 0 & \text{pokud j-té pozorování nepatří do s-té úrovně,} \\ & \text{kde } s < i,\ s = 1, \ldots, k-1. \end{cases}$$

Dosazením hodnot z_i do regresního modelu dostáváme k modelů:

$$\begin{array}{ll}
Y_{1j} = \mu + \beta_1(1 - \frac{1}{k}) + e_{1j}, & j = 1, \ldots, n_1, \\
Y_{2j} = \mu + \beta_1(-\frac{1}{k}) + \beta_2(1 - \frac{1}{k-1}) + e_{2j}, & j = 1, \ldots, n_2, \\
\vdots & \\
Y_{(k-1)j} = \mu - \frac{1}{k}\beta_1 - \frac{1}{k-1}\beta_2 + \cdots + \beta_{k-1}(1 - \frac{1}{2}) + e_{(k-1)j}, & j = 1, \ldots, n_{k-1}, \\
Y_{kj} = \mu - \frac{1}{k}\beta_1 - \frac{1}{k-1}\beta_2 + \ldots - \frac{1}{2}\beta_{k-1} + e_{kj}, & j = 1, \ldots, n_k.
\end{array}$$

Po úpravách lze modely zapsat jako:

$$\begin{array}{ll}
Y_{ij} = \mu + (1 - \frac{1}{k-i+1})\beta_i - \sum\limits_{l<i} \frac{1}{k+1-l}\beta_l + e_{ij}, & j = 1, \ldots, n_i,\ i = 1, \ldots, k-1, \\
Y_{kj} = \mu - \sum\limits_{l=1}^{k-1} \frac{1}{k+1-l}\beta_l + e_{kj}, & j = 1, \ldots, n_k.
\end{array}$$

Designová matice Helmertových indikátorů

$$\mathbf{X}_{n\times k} = \begin{pmatrix}
\mathbf{1}_{n_1} & (1-\frac{1}{k})\,\mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \ldots & \mathbf{0}_{n_1} \\
\mathbf{1}_{n_2} & -\frac{1}{k}\,\mathbf{1}_{n_2} & (1-\frac{1}{k-1})\,\mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \ldots & \mathbf{0}_{n_2} \\
\mathbf{1}_{n_3} & -\frac{1}{k}\,\mathbf{1}_{n_3} & -\frac{1}{k-1}\,\mathbf{1}_{n_3} & (1-\frac{1}{k-2})\,\mathbf{1}_{n_3} & \ldots & \mathbf{0}_{n_3} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{1}_{n_{k-1}} & -\frac{1}{k}\,\mathbf{1}_{n_{k-1}} & -\frac{1}{k-1}\,\mathbf{1}_{n_{k-1}} & -\frac{1}{k-2}\,\mathbf{1}_{n_{k-1}} & \ldots & (1-\frac{1}{2})\,\mathbf{1}_{n_{k-1}} \\
\mathbf{1}_{n_k} & -\frac{1}{k}\,\mathbf{1}_{n_k} & -\frac{1}{k-1}\,\mathbf{1}_{n_k} & -\frac{1}{k-2}\,\mathbf{1}_{n_k} & \ldots & -\frac{1}{2}\,\mathbf{1}_{n_k}
\end{pmatrix}.$$

Matice kontrastů $\mathbf{C}$

$$\mathbf{C}_{k\times n} = \begin{pmatrix}
\frac{1}{kn_1}\,\mathbf{1}'_{n_1} & \frac{1}{kn_2}\,\mathbf{1}'_{n_2} & \ldots & \frac{1}{kn_{k-2}}\,\mathbf{1}'_{n_{k-2}} & \frac{1}{kn_{k-1}}\,\mathbf{1}'_{n_{k-1}} & \frac{1}{kn_k}\,\mathbf{1}'_{n_k} \\
\frac{1}{n_1}\,\mathbf{1}'_{n_1} & -\frac{1}{(k-1)n_2}\,\mathbf{1}'_{n_2} & \ldots & -\frac{1}{(k-1)n_{k-2}}\,\mathbf{1}'_{n_{k-2}} & -\frac{1}{(k-1)n_{k-1}}\,\mathbf{1}'_{n_{k-1}} & -\frac{1}{(k-1)n_k}\,\mathbf{1}'_{n_k} \\
\mathbf{0}'_{n_1} & \frac{1}{n_2}\,\mathbf{1}'_{n_2} & \ldots & -\frac{1}{(k-2)n_{k-2}}\,\mathbf{1}'_{n_{k-2}} & -\frac{1}{(k-2)n_{k-1}}\,\mathbf{1}'_{n_{k-1}} & -\frac{1}{(k-2)n_k}\,\mathbf{1}'_{n_k} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \ldots & \frac{1}{n_{k-2}}\,\mathbf{1}'_{n_{k-2}} & -\frac{1}{2n_{k-1}}\,\mathbf{1}'_{n_{k-1}} & -\frac{1}{2n_k}\,\mathbf{1}'_{n_k} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \ldots & \mathbf{0}'_{n_{k-1}} & \frac{1}{n_{k-1}}\,\mathbf{1}'_{n_{k-1}} & -\frac{1}{n_k}\,\mathbf{1}'_{n_k}
\end{pmatrix}.$$

Odhady parametrů $\boldsymbol{\beta}$ vypočtené metodou nejmenších čtverců

$$\hat{\mu} = \bar{Y}_{\cdot\cdot},\ \hat{\beta}_1 = \bar{Y}_{1\cdot} - \frac{\bar{Y}_{2\cdot} + \cdots + \bar{Y}_{k\cdot}}{k-1},\ \hat{\beta}_2 = \bar{Y}_{2\cdot} - \frac{\bar{Y}_{3\cdot} + \cdots + \bar{Y}_{k\cdot}}{k-2}, \ldots,\ \hat{\beta}_{k-1} = \bar{Y}_{k-1\cdot} - \bar{Y}_{k\cdot}.$$

Odhad parametru β_k je roven nule, neboť odhady paramatrů kategoriální proměnné kódované Helmertovými indikátory srovnává průměrné hodnoty napozorovaných dat konkrétní úrovně s průměrnou hodnotou napozorovaných dat všech

následujících úrovní. Jinými slovy průměr napozorovaných hodnot k-té úrovně nemá žádné další úrovně, které by za ní následovali v nějakém logickém sledu, proto odhad parametru β_k položíme rovno nule. Odhad absolutního členu udává průměrnou hodnotu všech napozorovaných dat pro k úrovní.

2.5 Inverzní Helmertovy indikátory

Jak již název napovídá, odhady kategoriální proměnné kódované pomocí inverzních Helmertových indikátorů mají opačný význam oproti odhadům kategoriální proměnné kódované pomocí Helmertových indikátorů. Odhady v této kapitole tedy představují rozdíl mezi průměrnou hodnotou napozorovaných dat konkrétní úrovně s průměrnou hodnotou napozorovaných dat všech úrovní této úrovni předcházejících. Stejně jako u Helmertových indikátorů, má smysl počítat odhady parametrů pomocí inverzních Helmertových indikátorů pouze v případě, kdy existuje nějaké logické uspořádání úrovní. Z podstaty významu inverzních Helmertových indikátorů bude proto v této kapitole zvolena za referenční úroveň 1.úroveň a nikoli k-tá, jako tomu bylo u předchozích kapitol. [2] [3]
Uvažujeme regresní model s kategoriální proměnnou Z o k úrovních ve tvaru

$$Y_{ij} = \mu + \beta_2 z_{2j} + \beta_3 z_{3j} + \cdots + \beta_k z_{kj} + e_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i.$$

Kódování proměnné $z_{ij},\ i = 2, \ldots, k$ pomocí inverzních Helmertových indikátorů:

$$z_{ij} = \begin{cases} -\frac{1}{i} & \text{pokud j-té pozorování patří do 1. úrovně,} \\ 0 & \text{pokud j-té pozorování patří do l-té úrovně,} \\ & \text{kde } i < l,\ l = 2, \ldots, k, \\ 1 - \frac{1}{i} & \text{pokud j-té pozorování patří do i-té úrovně,} \\ -\frac{1}{i} & \text{pokud j-té pozorování patří do s-té úrovně,} \\ & \text{kde } s < i,\ s = 2, \ldots, k-1. \end{cases}$$

Dosazením hodnot z_{ij} do regresního modelu dostáváme k modelů:

$$\begin{aligned}
&Y_{1j} = \mu - \tfrac{1}{2}\beta_2 - \tfrac{1}{3}\beta_3 + \ldots - \tfrac{1}{k}\beta_k + e_{1j}, && j = 1, \ldots, n_1, \\
&Y_{2j} = \mu + (1 - \tfrac{1}{2})\beta_2 - \tfrac{1}{3}\beta_3 + \ldots - \tfrac{1}{k}\beta_k + e_{2j}, && j = 1, \ldots, n_2, \\
&Y_{3j} = \mu + (1 - \tfrac{1}{3})\beta_3 - \tfrac{1}{4}\beta_4 + \ldots - \tfrac{1}{k}\beta_k + e_{3j}, && j = 1, \ldots, n_3, \\
&\vdots \\
&Y_{(k-1)j} = \mu + (1 - \tfrac{1}{k-1})\beta_{k-1} - \tfrac{1}{k}\beta_k + e_{(k-1)j}, && j = 1, \ldots, n_{k-1}, \\
&Y_{kj} = \mu + (1 - \tfrac{1}{k})\beta_k + e_{kj}, && j = 1, \ldots, n_k.
\end{aligned}$$

Modely lze také zapsat ve tvaru:

$$Y_{1j} = \mu - \sum_{i=1}^{k-1} \tfrac{1}{1+i}\beta_i + e_{1j}, \qquad j = 1, \ldots, n_1,$$
$$Y_{ij} = \mu + \beta_{i-1} - \sum_{l=i}^{k} \tfrac{1}{l}\beta_{l-1} + e_{ij}, \quad i = 2, \ldots, k, \ j = 1, \ldots, n_i.$$

Designová matice modelu:

$$\mathbf{X}_{n\times k} = \begin{pmatrix} \mathbf{1}_{n_1} & -\frac{1}{2}\,\mathbf{1}_{n_1} & -\frac{1}{3}\,\mathbf{1}_{n_1} & \ldots & -\frac{1}{k-1}\,\mathbf{1}_{n_1} & -\frac{1}{k}\,\mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & (1-\frac{1}{2})\,\mathbf{1}_{n_2} & -\frac{1}{3}\,\mathbf{1}_{n_2} & \ldots & -\frac{1}{k-1}\,\mathbf{1}_{n_2} & -\frac{1}{k}\,\mathbf{1}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & (1-\frac{1}{3})\,\mathbf{1}_{n_3} & \ldots & -\frac{1}{k-1}\,\mathbf{1}_{n_3} & -\frac{1}{k}\,\mathbf{1}_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \ldots & (1-\frac{1}{k-1})\,\mathbf{1}_{n_{k-1}} & -\frac{1}{k}\,\mathbf{1}_{n_{k-1}} \\ \mathbf{1}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \ldots & \mathbf{0}_{n_k} & (1-\frac{1}{k})\,\mathbf{1}_{n_k} \end{pmatrix}.$$

Matice kontrastů $\mathbf{C}$ je pak tvaru:

$$\mathbf{C}_{k\times n} = \begin{pmatrix} \frac{1}{kn_1}\,\mathbf{1}_{n_1} & \frac{1}{kn_2}\,\mathbf{1}_{n_2} & \frac{1}{kn_3}\,\mathbf{1}_{n_3} & \ldots & \frac{1}{kn_{k-1}}\,\mathbf{1}_{n_{k-1}} & \frac{1}{kn_k}\,\mathbf{1}_{n_k} \\ -\frac{1}{n_1}\,\mathbf{1}_{n_1} & \frac{1}{n_2}\,\mathbf{1}_{n_2} & \mathbf{0}_{n_3} & \ldots & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_k} \\ -\frac{1}{2n_1}\,\mathbf{1}_{n_1} & -\frac{1}{2n_2}\,\mathbf{1}_{n_2} & \frac{1}{n_3}\,\mathbf{1}_{n_3} & \ldots & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{(k-2)n_1}\,\mathbf{1}_{n_1} & -\frac{1}{(k-2)n_2}\,\mathbf{1}_{n_2} & -\frac{1}{(k-2)n_3}\,\mathbf{1}_{n_3} & \ldots & \mathbf{1}_{n_{k-1}} & \mathbf{0}_{n_k} \\ -\frac{1}{(k-1)n_1}\,\mathbf{1}_{n_1} & -\frac{1}{(k-1)n_2}\,\mathbf{1}_{n_2} & -\frac{1}{(k-1)n_3}\,\mathbf{1}_{n_3} & \ldots & -\frac{1}{(k-1)n_{k-1}}\,\mathbf{1}_{n_{k-1}} & \mathbf{1}_{n_k} \end{pmatrix}.$$

Odhady parametrů jsou tvaru:

$$\hat{\mu} = \bar{Y}_{\cdot\cdot}, \quad \hat{\beta}_2 = \bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}, \quad \hat{\beta}_3 = \bar{Y}_{3\cdot} - \frac{\bar{Y}_{1\cdot} + \bar{Y}_{2\cdot}}{2}, \ldots, \quad \hat{\beta}_k = \bar{Y}_{k\cdot} - \frac{\bar{Y}_{1\cdot} + \cdots + \bar{Y}_{(k-1)\cdot}}{k-1}.$$

Odhad parametru β_1 referenční úrovně je opět roven nule. Vyplývá to z podstaty odhadů parametrů inverzních Helmertových indikátorů. Ty porovnávají průměrnou hodnotu napozorovaných dat dané úrovně s průměrnou hodnotou napozorovaných dat všech předcházejících úrovní.

2.6 Indikátory Profile

Indikátory profile vyjadřují srovnání průměrných hodnot napozorovaných dat v sousedních úrovních. Proto stejně jako u Helmertových a inverzních Helmertových indikátorů mají význam pouze v případě, kdy existuje nějaké libovolné uspořádání úrovní kategoriální proměnné. [2] [3]

Uvažujeme regresní model s kategoriální proměnnou Z o k úrovních ve tvaru (7). Jako referenční úroveň byla zvolena poslední k-tá úroveň.

Kódování proměnné $z_{ij},\ i = 1, \ldots, k-1$ pomocí indikátorů Profile:

$$z_{ij} = \begin{cases} 1 - \frac{l}{k} & \text{pokud j-té pozorování patří do l-té úrovně,} \\ & \text{kde } i < l,\ l = 2, \ldots, k-1, \\ 1 - \frac{i}{k} & \text{pokud j-té pozorování patří do i-té úrovně,} \\ -\frac{s}{k} & \text{pokud j-té pozorování patří do s-té úrovně,} \\ & \text{kde } s < i,\ s = 3, \ldots, k. \end{cases}$$

Dosazením hodnot z_i do regresního modelu dostáváme k modelů:

$$\begin{array}{ll} Y_{1j} = \mu + (1 - \frac{1}{k})\beta_1 + (1 - \frac{2}{k})\beta_2 + \cdots + (1 - \frac{k-1}{k})\beta_{k-1} + e_{1j}, & j = 1, \ldots, n_1, \\ Y_{2j} = \mu - \frac{1}{k}\beta_1 + (1 - \frac{2}{k})\beta_2 + \cdots + (1 - \frac{k-1}{k})\beta_{k-1} + e_{2j}, & j = 1, \ldots, n_2, \\ \vdots & \\ Y_{(k-1)j} = \mu - \frac{1}{k}\beta_1 - \frac{2}{k}\beta_2 + \ldots - \frac{k-2}{k}\beta_{k-2} + (1 - \frac{k-1}{k})\beta_{k-1} + e_{(k-1)j}, & j = 1, \ldots, n_{k-1}, \\ Y_{kj} = \mu - \frac{1}{k}\beta_1 - \frac{2}{k}\beta_2 + \ldots - \frac{k-2}{k}\beta_{k-2} - \frac{k-1}{k}\beta_{k-1} + e_{kj}, & j = 1, \ldots, n_k. \end{array}$$

Po úpravě můžeme modely zapsat ve tvaru:

$$\begin{aligned} Y_{ij} &= \mu + \sum_{l=i}^{k-1} \beta_l - \sum_{l=1}^{k-1} \frac{l}{k}\beta_l + e_{ij}, \quad i = 1, \ldots, k-1, \quad j = 1, \ldots, n_i, \\ Y_{kj} &= \mu - \sum_{i=1}^{k-1} \frac{i}{k}\beta_i + e_{kj}, \qquad\qquad j = 1, \ldots, n_k. \end{aligned}$$

Designová matice pro indikátory profile je tvaru:

$$\mathbf{X}_{n\times k} = \begin{pmatrix} \mathbf{1}_{n_1} & (1-\frac{1}{k})\,\mathbf{1}_{n_1} & (1-\frac{2}{k})\,\mathbf{1}_{n_1} & \ldots & (1-\frac{k-2}{k})\,\mathbf{1}_{n_1} & (1-\frac{k-1}{k})\,\mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & -\frac{1}{k}\,\mathbf{1}_{n_2} & (1-\frac{2}{k})\,\mathbf{1}_{n_2} & \ldots & (1-\frac{k-2}{k})\,\mathbf{1}_{n_2} & (1-\frac{k-1}{k})\,\mathbf{1}_{n_2} \\ \mathbf{1}_{n_3} & -\frac{1}{k}\,\mathbf{1}_{n_3} & -\frac{2}{k}\,\mathbf{1}_{n_3} & \ldots & (1-\frac{k-2}{k})\,\mathbf{1}_{n_3} & (1-\frac{k-1}{k})\,\mathbf{1}_{n_3} \\ \mathbf{1}_{n_4} & -\frac{1}{k}\,\mathbf{1}_{n_4} & -\frac{2}{k}\,\mathbf{1}_{n_4} & \ldots & (1-\frac{k-2}{k})\,\mathbf{1}_{n_4} & (1-\frac{k-1}{k})\,\mathbf{1}_{n_4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_{k-2}} & -\frac{1}{k}\,\mathbf{1}_{n_{k-2}} & -\frac{2}{k}\,\mathbf{1}_{n_{k-2}} & \ldots & (1-\frac{k-2}{k})\,\mathbf{1}_{n_{k-2}} & (1-\frac{k-1}{k})\,\mathbf{1}_{n_{k-2}} \\ \mathbf{1}_{n_{k-1}} & -\frac{1}{k}\,\mathbf{1}_{n_{k-1}} & -\frac{2}{k}\,\mathbf{1}_{n_{k-1}} & \ldots & -\frac{k-2}{k}\,\mathbf{1}_{n_{k-1}} & (1-\frac{k-1}{k})\,\mathbf{1}_{n_{k-1}} \\ \mathbf{1}_{n_k} & -\frac{1}{k}\,\mathbf{1}_{n_k} & -\frac{2}{k}\,\mathbf{1}_{n_k} & \ldots & -\frac{k-2}{k}\,\mathbf{1}_{n_k} & -\frac{k-1}{k}\,\mathbf{1}_{n_k} \end{pmatrix}.$$

Matice kontrastů $\mathbf{C}$ je tvaru:

$$\mathbf{C}_{k\times n} = \begin{pmatrix} \frac{1}{kn_1}\,\mathbf{1}'_{n_1} & \frac{1}{kn_2}\,\mathbf{1}'_{n_2} & \frac{1}{kn_3}\,\mathbf{1}'_{n_3} & \ldots & \frac{1}{kn_{k-1}}\,\mathbf{1}'_{n_{k-1}} & \frac{1}{kn_k}\,\mathbf{1}'_{n_k} \\ \frac{1}{n_1}\,\mathbf{1}'_{n_1} & -\frac{1}{n_2}\,\mathbf{1}'_{n_2} & \mathbf{0}'_{n_3} & \ldots & \mathbf{0}'_{n_{k-1}} & \mathbf{0}'_{n_k} \\ \mathbf{0}'_{n_1} & \frac{1}{n_2}\,\mathbf{1}'_{n_2} & -\frac{1}{n_3}\,\mathbf{1}'_{n_3} & \ldots & \mathbf{0}'_{n_{k-1}} & \mathbf{0}'_{n_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \ldots & \frac{1}{n_{k-1}}\,\mathbf{1}'_{n_{k-1}} & -\frac{1}{n_k}\,\mathbf{1}'_{n_k} \end{pmatrix}.$$

Odhady parametrů indikátoru profile jsou tvaru:

$$\hat{\mu} = \bar{Y}_{\cdot\cdot}, \quad \hat{\beta}_1 = \bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}, \quad \ldots, \quad \hat{\beta}_{k-1} = \bar{Y}_{k-1\cdot} - \bar{Y}_{k\cdot}.$$

Parametr β_k referenční úrovně je opět roven nule, neboť neexistuje žádná další úroveň, se kterou by se porovnala.

2.7 Inverzní indikátory Profile

Inverzní indikátory profile také vyjadřují srovnání průměrných hodnot napozorovaných dat v sousedních úrovních kvalitativní proměnné. Tentokrát ale udávají rozdíl průměrné hodnoty dané úrovně s průměrnou hodnotou úrovně předcházející. [3][2]

Uvažujeme regresní model s kategoriální proměnnou Z o k úrovních ve tvaru (7). Referenční úrovní volíme opět k-tou úroveň. Kódování proměnné z_{ij}, $i = 1, \ldots, k-1$ pomocí inverzních indikátorů Profile:

$$z_{ij} = \begin{cases} -(1-\frac{i}{k}) & \text{pokud j-té pozorování patří do l-té úrovně,} \\ & \text{kde } i < l,\ l = 2, \ldots, k-1, \\ -(1-\frac{i}{k}) & \text{pokud j-té pozorování patří do i-té úrovně,} \\ \frac{i}{k} & \text{pokud j-té pozorování patří do s-té úrovně,} \\ & \text{kde } s < i,\ s = 3, \ldots, k. \end{cases}$$

Kódování pomocí inverzních indikátorů profile je velmi podobné kódování pomocí indikátorů profile, liší se pouze opačnými znaménky. Dosazením hodnot z_{ij} do regresního modelu dostáváme k modelů

$$\begin{aligned}
&Y_{1j} = \mu - \beta_1(1-\tfrac{1}{k}) - \beta_2(1-\tfrac{2}{k}) - \ldots - \beta_{k-1}(1-\tfrac{k-1}{k}) + e_{1j}, && j = 1, \ldots, n_1, \\
&Y_{2j} = \mu + \beta_1\tfrac{1}{k} - \beta_2(1-\tfrac{2}{k}) - \ldots - \beta_{k-1}(1-\tfrac{k-1}{k}) + e_{2j}, && j = 1, \ldots, n_2, \\
&\vdots \\
&Y_{(k-1)j} = \mu + \beta_1\tfrac{1}{k} + \beta_2\tfrac{2}{k} + \cdots + \beta_{k-2}\tfrac{k-2}{k} - \beta_{k-1}(1-\tfrac{k-1}{k}) + e_{(k-1)j}, && j = 1, \ldots, n_{k-1}, \\
&Y_{kj} = \mu + \beta_1\tfrac{1}{k} + \beta_2\tfrac{2}{k} + \cdots + \beta_{k-2}\tfrac{k-2}{k} + \beta_{k-1}\tfrac{k-1}{k} + e_{(k)j}, && j = 1, \ldots, n_k.
\end{aligned}$$

Po aritmetických úpravách dostáváme k modelů tvaru:

$$\begin{aligned}
&Y_{ij} = \mu - \sum_{l=i}^{k-1} \beta_l + \sum_{l=1}^{k-1} \tfrac{l}{k}\beta_l + e_{ij}, \quad i = 1, \ldots, k-1, \quad j = 1, \ldots, n_i, \\
&Y_{kj} = \mu + \sum_{i=1}^{k-1} \tfrac{i}{k}\beta_i + e_{kj}, \qquad\qquad j = 1, \ldots, n_k.
\end{aligned}$$

Designová matice pro inverzní indikátory profile.

$$\mathbf{X}_{n\times k} = \begin{pmatrix} \mathbf{1}_{n_1} & -(1-\frac{1}{k})\,\mathbf{1}_{n_1} & -(1-\frac{2}{k})\,\mathbf{1}_{n_1} & \dots & -(1-\frac{k-2}{k})\,\mathbf{1}_{n_1} & -(1-\frac{k-1}{k})\,\mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & \frac{1}{k}\,\mathbf{1}_{n_2} & -(1-\frac{2}{k})\,\mathbf{1}_{n_2} & \dots & -(1-\frac{k-2}{k})\,\mathbf{1}_{n_2} & -(1-\frac{k-1}{k})\,\mathbf{1}_{n_2} \\ \mathbf{1}_{n_3} & \frac{1}{k}\,\mathbf{1}_{n_3} & \frac{2}{k}\,\mathbf{1}_{n_3} & \dots & -(1-\frac{k-2}{k})\,\mathbf{1}_{n_3} & -(1-\frac{k-1}{k})\,\mathbf{1}_{n_3} \\ \mathbf{1}_{n_4} & \frac{1}{k}\,\mathbf{1}_{n_4} & \frac{2}{k}\,\mathbf{1}_{n_4} & \dots & -(1-\frac{k-2}{k})\,\mathbf{1}_{n_4} & -(1-\frac{k-1}{k})\,\mathbf{1}_{n_4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_{k-2}} & \frac{1}{k}\,\mathbf{1}_{n_{k-2}} & \frac{2}{k}\,\mathbf{1}_{n_{k-2}} & \dots & -(1-\frac{k-2}{k})\,\mathbf{1}_{n_{k-2}} & -(1-\frac{k-1}{k})\,\mathbf{1}_{n_{k-2}} \\ \mathbf{1}_{n_{k-1}} & \frac{1}{k}\,\mathbf{1}_{n_{k-1}} & \frac{2}{k}\,\mathbf{1}_{n_{k-1}} & \dots & \frac{k-2}{k}\,\mathbf{1}_{n_{k-1}} & -(1-\frac{k-1}{k})\,\mathbf{1}_{n_{k-1}} \\ \mathbf{1}_{n_k} & \frac{1}{k}\,\mathbf{1}_{n_k} & \frac{2}{k}\,\mathbf{1}_{n_k} & \dots & \frac{k-2}{k}\,\mathbf{1}_{n_k} & \frac{k-1}{k}\,\mathbf{1}_{n_k} \end{pmatrix}.$$

Matici kontrastů $\mathbf{C}$ zapíšeme ve tvaru:

$$\mathbf{C}_{k\times n} = \begin{pmatrix} \frac{1}{kn_1}\,\mathbf{1}'_{n_1} & \frac{1}{kn_2}\,\mathbf{1}'_{n_2} & \frac{1}{kn_3}\,\mathbf{1}'_{n_3} & \dots & \frac{1}{kn_{k-1}}\,\mathbf{1}'_{n_{k-1}} & \frac{1}{kn_k}\,\mathbf{1}'_{n_k} \\ -\frac{1}{n_1}\,\mathbf{1}'_{n_1} & \frac{1}{n_2}\,\mathbf{1}'_{n_2} & \mathbf{0}'_{n_3} & \dots & \mathbf{0}'_{n_{k-1}} & \mathbf{0}'_{n_k} \\ \mathbf{0}'_{n_1} & -\frac{1}{n_2}\,\mathbf{1}'_{n_2} & \frac{1}{n_3}\,\mathbf{1}'_{n_3} & \dots & \mathbf{0}'_{n_{k-1}} & \mathbf{0}'_{n_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \dots & -\frac{1}{n_{k-1}}\,\mathbf{1}'_{n_{k-1}} & \frac{1}{n_k}\,\mathbf{1}'_{n_k} \end{pmatrix}.$$

Při bližším prozkoumání matice kontrastů a designové matice indikátorů profile s týmiž maticemi inverzních indikátorů profile, zjistíme, že se liší pouze znaménky. Tedy až na první řádek u kontrastní matice a první sloupec u designové matice. Ty zůstávají i nadále beze změny, jelikož zastupují absolutní člen.

Odhady parametrů vypočtené metodou nejmenších čtverců

$$\hat{\mu} = \bar{Y}_{\cdot\cdot}, \quad \hat{\beta}_1 = \bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}, \quad \dots, \quad \hat{\beta}_{k-1} = \bar{Y}_{k\cdot} - \bar{Y}_{k-1,\cdot}.$$

Odhad parametru β_k referenční úrovně je opět roven nule, neboť stejně jako v případě kódování pomocí indikátorů profile již nezbyla žádná sousední dvojice uspořádaných úrovní, jejíž průměry by nebyly porovnány.

2.8 Srovnání druhů kódování na konkrétních datech

V předchozích kapitolách bylo představeno sedm typů kódování kategoriální proměnné (Dummy, Simple, Effect, Helmert, inverzní Helmert, Profile a Inverzní profile) a vysvětleny různé interpretace odhadů parametrů kategoriální proměnné,

kdy jsme si ukázali, že každý typ kódování je trochu jiný a dává nám jiné interpretace odhadů parametrů kategoriální proměnné. Tato kapitola má za úkol představit a vzájemně srovnat již zmíněné druhy kódování na konkrétních datech. Jako datový soubor byla vybrána data z ČSÚ, která jsou obsahem přiloženého CD.

Budeme zkoumat počty cizinců v krajích ČR v rozmezí let $1996 - 2009$. V ČR máme 14 krajů, tedy kategoriální proměnná modelu bude nabývat 14 úrovní zastupující právě tyto kraje. Vzhledem k tomu, že pro kódování typu Helmert, Inverzní Helmert, Profile a Inverzní profile musí být hodnoty kategoriální proměnné uspořádány, jsou v tabulce kraje seřazeny sestupně dle počtu obyvatel v daném kraji. Výběr referenční úrovně (úrovně kategoriální proměnné, ke které chceme vztahovat naše výsledky) provedeme dle typů kódování následovně. Pro první tři typy kódování (Dummy, Simple, Effect) si můžeme zvolit libovolný z krajů, zvolme tedy Olomoucký kraj, který je nám nejbližší. U ostatních druhů kódování je však vhodné volit referenční úroveň tak, aby to bylo pro dané kódování nejvýhodnější. Pro kódování typu Helmert, Profile a Inverzní Profile zvolíme tedy poslední úroveň (v našem případě Karlovarský kraj) a pro kódování pomocí inverzních Helmertových indikátorů zvolíme 1. úroveň (v našem případě Moravskoslezský kraj).

Kraje	Počet obyvatel	$\bar{Y}_{i\cdot}$	Odhady parametrů pro kódování						
			DUM.	SIMP.	EFFECT	HELM.	I. HEL.	PROF.	I. PRO.
Absolutní člen			7 460	19 680	19 680	19 680	19 680	19 680	19 680
Moravskoslezský	1 249 356	21 079	13 619	13 619	1 399	1 507	0	−64 072	64 072
Hlavní město Praha	1 242 956	85 151	77 691	77 691	65 471	71 044	64 071	49 095	−49 095
Středočeský	1 239 673	36 055	28 596	28 596	16 376	23 944	−17 059	12 609	−12 609
Jihomoravský	1 150 009	23 447	15 987	15 987	3 767	12 469	−23 981	2 408	−2 408
Ústecký	836 128	21 039	13 580	13 580	1 360	11 179	−20 394	13 580	−13 580
Olomoucký	641 945	7 460	0	0	−12 220	−2 701	−29 895	−3 460	3 460
Jihočeský	637 015	10 919	3 460	3 460	−8 760	867	−21 453	3 684	−3 684
Zlínský	591 303	7 236	−224	−224	−12 444	−3 286	−22 072	−6 778	6 778
Plzeňský kraj	571 199	14 013	6 553	6 553	−5 667	4 190	−12 535	3 130	−3 130
Královéhradecký	554 511	10 883	3 423	3 423	−8 797	1 325	−14 273	3 493	−3 493
Pardubický	515 868	7 390	−70	−70	−12 290	−2 891	−16 338	1 372	−1 372
Vysočina	515 326	6 018	−1 442	−1 442	−13 662	−6 395	−16 225	−5 323	5 323
Liberecký	438 238	11 341	3 881	3 881	−8 339	−2 143	−9 550	−2 143	2 143
Karlovarský	307 962	13 484	6 025	6 025	−6 195	0	−6 672	0	0

Tabulka 1: Srovnání typů kódování kategoriální proměnné interpretované na datech udávající počet cizinců v krajích ČR v rozmezí let 1996 – 2009.

Kraje	Směrodatná odchylka (chyba) odhadu $s(\hat{\beta}_j)$						
	DUM.	SIMP.	EFFECT	HELM.	I. HEL.	PROF.	I. PRO.
Abs.čl.	2 693, 5	719, 9	719, 9	719, 9	719, 9	719, 9	719, 9
Mor.sl.	3 809, 2	3 809, 2	2 595, 6	2 795, 2	–	3 809, 2	3 809, 2
Praha	3 809, 2	3 809, 2	2 595, 6	2 803, 5	3 809, 2	3 809, 2	3 809, 2
Stř.čs.	3 809, 2	3 809, 2	2 595, 6	2 813, 3	3 298, 2	3 809, 2	3 809, 2
Jihom.	3 809, 2	3 809, 2	2 595, 6	2 825, 0	3 110, 2	3 809, 2	3 809, 2
Ústec.	3 809, 2	3 809, 2	2 595, 6	2 839, 2	3 011, 5	3 809, 2	3 809, 2
Olom.	–	–	2 595, 6	2 856, 9	2 950, 6	3 809, 2	3 809, 2
Jihočs	3 809, 2	3 809, 2	2 595, 6	2 879, 5	2 909, 4	3 809, 2	3 809, 2
Zlín	3 809, 2	3 809, 2	2 595, 6	2 909, 4	2 879, 5	3 809, 2	3 809, 2
Plzeň	3 809, 2	3 809, 2	2 595, 6	2 950, 6	2 856, 9	3 809, 2	3 809, 2
Kr.hr.	3 809, 2	3 809, 2	2 595, 6	3 011, 5	2 839, 2	3 809, 2	3 809, 2
Pard.	3 809, 2	3 809, 2	2 595, 6	3 110, 2	2 825, 0	3 809, 2	3 809, 2
Vys.	3 809, 2	3 809, 2	2 595, 6	3 298, 9	2 813, 3	3 809, 2	3 809, 2
Lib.	3 809, 2	3 809, 2	2 595, 6	3 809, 2	2 803, 5	3 809, 2	3 809, 2
Karl.	3 809, 2	3 809, 2	2 595, 6	–	2 795, 2	–	–

Tabulka 2: Srovnání směrodatných odchylek odhadů regresních parametrů při užití různých typů kódování kategoriální proměnné interpretované na datech udávající počet cizinců v krajích ČR v rozmezí let 1996 – 2009.

Výpočet odhadů parametrů μ, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{14})'$, jejich směrodatných odchylek (4) a p-hodnot byl proveden pomocí statistického programu `R`. Výsledné odhady jsou uvedeny v tabulce 1, směrodatné chyby odhadů jsou pak uvedeny v tabulce 2, p-hodnoty dílčích t-testů jsou pak v tabulce 3. Pro práci s `R` je důležité si uvědomit, že při výpočtu odhadů parametrů se automaticky dodává do zadaného modelu absolutní člen a jako referenční úroveň vybírá automaticky poslední k - tou úroveň. Pokud si tedy vybereme referenční úroveň, ke které chceme vztahovat naše výsledky, musíme upravit data, tak aby námi vybraná referenční úroveň byla umístěna jako poslední. Směrodatné odchylky (chyby) odhadů regresních parametrů $\mu, \boldsymbol{\beta}$ získáváme společně s výpočtem samotných odhadů parametrů v `R` příkazem `summary()`. Hodnoty směrodatné odchylky jsou uloženy ve druhém sloupci s názvem `Std. Error`. Ještě než si popíšeme interpretace jednotlivých odhadů, podívejme se blíže na tabulku 2 popisující právě směrodatné odchylky odhadů. Jak můžeme vidět, nejmenší odchylkou (chybou) $s(\hat{\beta}_j) = 2\ 595, 6$ disponují odhady regresních parametrů při užití kódování effect.

Kraje	P-hodnoty dílčích t-testů						
	DUM.	SIMP.	EFFECT	HELM.	I. HEL.	PROF.	I. PRO.
Abs.čl.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Mor.sl.	0,0477	0,0477	0,5904	0,5904	–	0,0000	0,0000
Praha	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Stř.čs.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0011	0,0011
Jihom.	0,0097	0,0097	0,1484	0,0000	0,0000	0,5281	0,5281
Ústec.	0,0488	0,0488	0,6010	0,0001	0,0000	0,0005	0,0005
Olom.	–	–	–	0,3457	0,0000	0,3650	0,3650
Jihočs	0,1155	0,1155	0,0009	0,7636	0,0000	0,3348	0,3348
Zlín	0,5016	0,5016	0,0000	0,2602	0,0000	0,0769	0,0769
Plzeň	0,1026	0,1026	0,0303	0,1573	0,0000	0,4123	0,4123
Kr.hr.	0,8898	0,8898	0,0009	0,6606	0,0000	0,3604	0,3604
Pard.	0,4956	0,4956	0,0000	0,3539	0,0000	0,7190	0,7190
Vys.	0,1114	0,1114	0,0000	0,0541	0,0000	0,1640	0,1640
Lib.	0,0515	0,0515	0,0016	0,5744	0,0008	0,5744	0,5744
Karl.	0,5744	0,5744	0,0180	–	0,0180	–	–

Tabulka 3: Srovnání p-hodnot dílčích t-testů odhadů regresních parametrů při užití různých typů kódování kategoriální proměnné interpretované na datech udávající počet cizinců v krajích ČR v rozmezí let 1996 – 2009.

Naopak o dost hůře $s(\hat{\beta}_j) = 3\ 809,2$ jsou na tom ty typy kódování, jejichž odhady parametrů β porovnávají průměry pouze dvou úrovní. Jedná se o kódování typu dummy, simple, profile a inverzní profile. Co se týče odhadu absolutního členu, jsou na tom s hodnotou $s(\hat{\mu}) = 719,9$ mnohem lépe ty druhy kódování, při jejichž užití je odhad absolutního členu roven průměrné hodnotě napozorovaných dat pro všechny úrovně (v našem případě kraje). Tedy všechny druhy kódování až na kódování typu dummy. Z výsledků směrodatných odchylek odhadů tedy vidíme, že nejpřesnějších odhadů dosáhneme užitím kódování effect, kdežto s dummy kódováním si o dost pohoršíme.

Z výsledků p-hodnot (viz tabulka 3) pro dílčí t-testy vidíme, které kraje jsou pro model na hladině významnosti $\alpha = 0,05$ významné (p-hodnota$< \alpha$). Podívejme se nyní blíže na jednotlivé odhady (viz tabulka 1), jejich interpretace a p-hodnoty dílčích t-testů (viz tabulka 3).

DUMMY: Vypočtené odhady parametrů μ, $\boldsymbol{\beta}$ kategoriální proměnné Z zastupující kraje v ČR, udávají rozdíl průměrného počtu cizinců v jednotlivých krajích

oproti Olomouckému kraji, který jsme si vybrali jako referenční úroveň. Odhad absolutního členu μ udává průměrný počet cizinců referenční úrovně. V průběhu 13-ti let se v Olomouckém kraji průměrně zdržuje 7 460 cizinců. Z vypočtených dat můžeme dále například vyčíst, že ve dvou sousedních krajích Olomouckého kraje se v průběhu let 1996 - 2009 zdržuje mnohem více cizinců, než v samotném Olomouckém kraji. A to o 13 619 cizinců více v Moravskoslezském kraji a o 15 987 cizinců více v Jihomoravském kraji.

Pro model s dummy kódováním je významný pouze Moravskoslezský, Středočeský a Jihomoravský kraj společně s krajem Hlavní město Praha. To znamená, že průměrný počet cizinců v těchto krajích se výrazně liší od průměrného počtu cizinců v Olomouckém kraji, který byl zvolen jako referenční. Pro Ústecký a Liberecký kraj jsme dostali p-hodnoty velmi blízké hladině významnosti α, je tedy na zvážení jak se k významnosti těchto krajů pro model zachovat. Při bližším prozkoumání odhadů předpokládejme, že Ústecký kraj je pro model významný, kdežto Liberecký kraj pro model významný není, jelikož průměrný počet cizinců se od Olomouckého průměru příliš neliší.

Jelikož jsme vybírali jako referenční úroveň Olomoucký kraj, bylo třeba data upravit a přesunout pozorování této úrovně na konec datového souboru.

```
Kraje_dat<-read.table("kraje.dat",header=TRUE,sep=";")
Ol_ref<-data.frame(t(Kraje_dat[1:5,]),t(Kraje_dat[7:14,]),t(Kraje[6,]))
Cizinci_ol=(c(as.matrix(Ol_ref)))
f_ol=c("Moravskoslezský","Praha","Středočeský","Jihomoravský","Ústecký",
"Jihočeský","Zlínský","Plzeňský","Královéhradecký","Pardubický","Vysočina",
"Liberecký","Karlovarský","Olomoucký")
Kraje_ol=gl(k,ni,k*ni,factor(f_ol))

#Ukázka nadefinování sloupců designové matice
Mor_d=as.integer(Kraje_ol==(f[1]))
Pha_d=as.integer(Kraje_ol==(f[2]))
Strc_d=as.integer(Kraje_ol==(f[3]))
...
Xdum=matrix(c(Mor_d,Pha_d,Strc_d,Jihm_d,Úst_d,Jizc_d,Zli_d,Plz_d,Krá_d,Par_d,
Vys_d,Lib_d,Kar_d,Olo_d),ncol=14)
```

SIMPLE: Vypočtené odhady parametrů $\boldsymbol{\beta}$ mají stejný význam jako u dummy proměnných. Zastupují rozdíl průměrného počtu cizinců v jednotlivých krajích oproti Olomouckému kraji, který jsme si vybrali jako referenční úroveň. Rozdíl je pouze v interpretaci odhadu absolutního členu μ. Ten nyní zastupuje průměrný počet cizinců v krajích v období let $1996 - 2009$. Průměrně se tedy v jednotlivých krajích ČR vyskytuje kolem 19680 cizinců.

Výsledky p-hodnot pro dílčí t-testy jsou pro model s kódováním simple totožné s p-hodnotami pro dílčí t-testy s kódováním dummy. Čili od Olomouckého kraje se průměrný počet cizinců výrazně liší pouze v Moravskoslezském, Středočeském, Jihomoravském a Ústeckém kraji společně s krajem Hlavní město Praha.

K nadefinování designové matice pro kódování pomocí indikátorů Simple, můžeme využít designovou matici pro dummy kódování:

```
Xsim=Xdum-1/k
```

EFFECT: Vypočtené odhady parametrů nyní zastupují rozdíl průměrného počtu cizinců v jednotlivých krajích ČR oproti průměrnému počtu cizinců v kraji. Interpretace odhadu absolutního členu zůstává stejná jako u indikátorů simple, tzn. zastupuje průměrný počet cizinců v krajích v období $1996 - 2009$. Z vypočtených hodnot dále plyne, že pouze v pěti krajích je průměrně více cizinců, než je průměr počtu cizinců v kraji ČR. A to v Praze, Středočeském kraji, Ústeckém kraji, Jihomoravském kraji a Moravskoslezském kraji, kde ovšem Ústecký, Jihomoravský ani Moravskoslezský kraj nejsou pro model významné. V dalších krajích je již jejich výskyt menší. Konkrétně v Olomouckém kraji je přibližně o 12 220 cizinců méně, než kolik je průměr počtu cizinců v jednotlivých krajích.

Z tabulky p-hodnot vidíme, že pro model s kódováním effect jsou významné všechny kraje až na Moravskoslezský, Jihomoravský a Ústecký kraj. To znamená, že průměrný počet cizinců v těchto třech krajích se výrazně neliší od celkového průměru cizinců všech krajů.

K nadefinování designové matice pro kódování pomocí indikátorů Effect, opět můžeme využít designovou matici pro dummy kódování:

```
ef=c(rep(0,(k-1)*ni),rep(1,k)) #pomocný vektor
Xeff=Xdum-ef
```

HELMERT: Odhad absolutního členu opět jako u předešlých indikátorů (simple, effect) představuje průměrný počet cizinců v kraji ČR během let 1996 – 2009. Vypočtené odhady parametrů $\beta_1, \ldots, \beta_{k-1}$ nyní zastupují rozdíl průměrného počtu cizinců v konkrétním kraji s průměrným počtem cizinců v krajích s menším počtem obyvatel. Konkrétně v Praze je v průměru o 71 044 cizinců více než je průměr počtu cizinců v krajích s menším počtem obyvatel. Ve Středočeském kraji je oproti tomu v průměru jen o 23 944 cizinců více než je průměr počtu cizinců v krajích s menším počtem obyvatel.

Při pohledu na tabulku p-hodnot dílčích t-testů je ihned vidět, že pro model s kódováním Helmert jsou významné pouze Středočeský, Jihomoravský a Ústecký kraj společně s krajem Hlavní město Praha. Tedy průměrný počet cizinců v těchto čtyřech krajích se výrazně liší od průměrného počtu cizinců všech krajů s nižším počtem obyvatel.

V případě Helmertových indikátorů vybíráme jako referenční úroveň poslední úroveň, tedy Karlovarský kraj, tudíž pracujeme s původními neupravenými daty. Ukázka nadefinování sloupců designové matice:

```
Mor_h=as.integer(Kraje==(f[1]))-1/k
Pha_h=as.integer(Kraje==(f[2]))-c(rep(0,ni),rep(1/(k-1),(k-1)*ni))
Strc_h=as.integer(Kraje==(f[3]))-c(rep(0,2*ni),rep(1/(k-2),(k-2)*ni))
...
Lib_h=as.integer(Kraje==(f[13]))-c(rep(0,12*ni),rep(1/(k-12),(k-12)*ni))
Kar_h=as.integer(Kraje==(f[14]))-c(rep(0,13*ni),rep(1/(k-13),(k-13)*ni))
```

INVERZNÍ HELMERT: Odhad absolutního členu μ opět jako u předešlých indikátorů (simple, effect, Helmert) zasupuje průměrný počet cizinců v kraji ČR za léta 1996 – 2009. Další vypočtené odhady parametrů $\beta_2, \ldots, \beta_k$ nyní zastupují rozdíl průměrného počtu cizinců v kraji oproti průměrnému počtu cizinců v krajích s větším počtem obyvatel. Konkrétně v Praze je o 64 071 cizinců více než je průměrně v krajích s větším počtem obyvatel (tedy než je průměrně v Moravskoslezském kraji). Naproti tomu v Olomouckém kraji je až o 29 895 cizinců

méně než je průměrně v krajích s větším počtem obyvatel (tedy v Moravskoslezském kraji). Překvapivě v Karlovarském kraji, který je krajem s nejmenším počtem obyvatel je v průměru jen o 6 672 cizinců méně, než je průměr cizinců v krajích s větším počtem obyvatel, tedy ve všech ostatních krajích ČR.

Co se týče významnosti proměnných, jsou pro model s kódováním pomocí Inverzních Helmertových indikátorů významné všechny kraje. Tedy průměrný počet cizinců v kraji se výrazně liší od průměrného počtu cizinců v krajích s větším počtem obyvatel.

V případě Inverzních Helmertových indikátorů vybíráme jako referenční úroveň první úroveň, tedy Moravskoslezský kraj. Opět pracujeme s původními daty, jen do designové matice posuneme sloupec pro první úroveň nakonec matice. Ukázka nadefinování sloupců designové matice:

```
Pha_ih=as.integer(Kraje==(f[2]))-c(rep(1/2,2*ni),rep(0,(k-2)*ni))
Strc_ih=as.integer(Kraje==(f[3]))-c(rep(1/3,3*ni),rep(0,(k-3)*ni))
...
Kar_ih=as.integer(Kraje==(f[14]))-c(rep(1/14,14*ni))
Mor_ih=as.integer(Kraje==(f[1]))-c(rep(1,ni),rep(0,(k-1)*ni))
```

PROFILE: Vypočtené odhady parametrů $\beta_1, \ldots, \beta_{k-1}$ nyní zastupují rozdíl průměrného počtu cizinců ve dvou sousedních krajích uspořádaných sestupně, dle počtu obyvatel. V našem případě porovnáváme rozdíl průměrné hodnoty počtu cizinců v kraji s větším počtem obyvatel s průměrnou hodnotou počtu cizinců v kraji s následujícím menším počtem obyvatel. Konkrétně v Moravskoslezském kraji, což je kraj s největším počtem obyvatel, je o 64 072 cizinců méně než v Praze, což je kraj s druhým největším počtem obyvatel. V porovnání se Středočeským krajem, který je s počtem jako třetí v pořadí, je v Praze v průměru už jen o 49 095 cizinců více. Odhad absolutního členu stále jako v předchozích případech (simple, effect, Helmert) zastupuje průměrný počet cizinců v kraji během let $1996 - 2009$.

Z výsledků p-hodnot vidíme, že při užití profile kódování jsou pro model významné pouze Moravskoslezský, Středočeský a Ústecký kraj společně s krajem Hlavní město Praha. Tedy průměrný počet cizinců v těchto krajích se výrazně liší od průměrného počtu cizinců v následujícím kraji s menším počtem obyvatel.

V případě indikátorů Profile vybíráme jako referenční úroveň poslední úroveň, tedy Moravskoslezský kraj. Můžeme tedy pracovat s původními daty. Ukázka nadefinování sloupců designové matice:

```
Mor_p=c(rep(1-1/k,ni),rep(-1/k,(k-1)*ni))
Pha_p=c(rep(1-2/k,2*ni),rep(-2/k,(k-2)*ni))
...
Lib_p=c(rep(1-(13/k),13*k),rep(-(13/k),(k-13)*ni))
Kar_p=c(rep(1-(14/k),14*k),rep(-(14/k),(k-14)*ni))
```

INVERZNÍ PROFILE: Z výpočtu odhadu parametrů je zřejmé, že dostáváme stejné hodnoty jako u výpočtu odhadů parametrů profile, pouze se liší znaménkem. Vypočtené odhady parametrů $\beta_1, \ldots, \beta_{k-1}$ nyní zastupují rozdíl průměrného počtu cizinců sousedního následujícího kraje s daným krajem. Kraje jou uspořádány sestupně dle počtu obyvatel. V našem případě tedy porovnáváme rozdíl průměrné hodnoty počtu cizinců v kraji s menším počtem obyvatel s průměrnou hodnotou počtu cizinců v kraji s následujícím větším počtem obyvatel. Interpretace odhadů je tedy totožná s interpretací odhadů při užití kódování Profile. Konkrétně například v Praze, což je kraj s druhým největším počtem obyvatel je o 64 072 cizinců více, než v Moravskoslezkém kraji, což je kraj s největším počtem obyvatel. Odhad absolutního členu stále jako v předchozích případech (simple, effect, Helmert) zastupuje průměrný počet cizinců v kraji během let 1996 – 2009.

Výsledky p-hodnot dílčích t-testů při užití kódování Inverzní profile, je naprosto totožné s p-hodnotami dílčích t-testů při užití kódování Profile. Liší se pouze interpretací. Tedy průměrný počet cizinců v Moravskoslezském, Středočeském, Ústeckém kraji a v Hlavním městě Praha se výrazně liší od průměrného počtu cizinců v předcházejícím kraji s větším počtem obyvatel.

Pro nadefinování designové matice inverzních indilátorů profile můžeme použít již nadefinovanou designovou matici pro indikátory Profile.

```
XProfINV=-1*XProf
```

2.9 Vlastní typ kódování kategoriální proměnné

V případě potřeby specifické interpretace odhadů parametrů $\mu, \beta_1, \ldots, \beta_k$ si lze vytvořit vlastní styl kódování kategoriální proměnné. Navíc se znalostí kon-

trastních matic předchozích typů kódování je vytvoření vlastního kódovacího schéma velice jednoduché. Použijeme data z příkladu pro porovnání počtu cizinců v jednotlivých krajích. Vytvořme příklad kódování, kde interpretace odhadu parametrů bude následující. Odhad absolutního členu μ bude udávat průměrný počet cizinců ve všech krajích. [3]

Odhady parametrů $\beta_1, \beta_2, \beta_3$ budou udávat rozdíl počtu cizinců v Moravskoslezkém kraji, Hlavním městě Praha a ve Středočeském kraji oproti Olomouckému kraji, který si zvolíme jako referenční úroveň. Další parametry $\beta_4, \beta_5, \beta_6$ budou naopak udávat rozdíl v průměrném počtu cizinců Jihomoravského, Ústeckého a Olomouckého kraje s celorepublikovým průměrem počtu cizinců v jednom kraji. Parametry β_7, β_8 prozradí rozdíl mezi průměry počtů cizinců v Jihočeském a Zlínském kraji, a rozdíl mezi průměry počtů cizinců ve Zlínském a Plzeňském kraji. Parametry $\beta_9, \beta_{10}, \beta_{11}$ budou interpretovat rozdíly Plzeňského, Královéhradeckého a Pardubického kraje oproti průměru všech následujících krajů, tedy krajů s menším průměrným počtem cizinců. A poslední odhady parametrů β_{12}, β_{13} určí postupně rozdíly mezi průměrným počtem cizinců v Libereckém a Karlovarském kraji oproti průměrnému počtu cizinců ve všech krajích s větším počtem obyvatel. Odhad parametru β_{14} bude vzhledem k zachování lineární nezávislosti sloupců nulový.

Při vytváření vlastního kódování kategoriální proměnné nejprve vytvoříme kontrastní matici a z té následně pomocí matematického softwaru vypočteme designovou matici. Zvolené interpretace odhadu parametrů jsou v podstatě kombinací předchozích typů indikátorů postupně (Dummy(simple), Effect, Profile, Helmertovy indikátory a Inverzní Helmertovy indikátory). Matice kontrastů bude tedy kombinací odpovídajích řádků matic kontrastů předchozích typů kódování.

$$\mathbf{C}_{k\times n} = \begin{pmatrix}
\frac{1}{kn_1}\mathbf{1}'_{n_1} & \frac{1}{kn_2}\mathbf{1}'_{n_2} & \frac{1}{kn_3}\mathbf{1}'_{n_3} & \frac{1}{kn_4}\mathbf{1}'_{n_4} & \frac{1}{kn_5}\mathbf{1}'_{n_5} & \frac{1}{kn_6}\mathbf{1}'_{n_6} & \frac{1}{kn_7}\mathbf{1}'_{n_7} & \frac{1}{kn_8}\mathbf{1}'_{n_8} & \frac{1}{kn_9}\mathbf{1}'_{n_9} & \frac{1}{kn_{10}}\mathbf{1}'_{n_{10}} & \frac{1}{kn_{11}}\mathbf{1}'_{n_{11}} & \frac{1}{kn_{12}}\mathbf{1}'_{n_{12}} & \frac{1}{kn_{13}}\mathbf{1}'_{n_{13}} & \frac{1}{kn_{14}}\mathbf{1}'_{n_{14}} \\
\frac{1}{n_1}\mathbf{1}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \frac{-1}{n_6}\mathbf{1}'_{n_6} & \mathbf{0}'_{n_7} & \mathbf{0}'_{n_8} & \mathbf{0}'_{n_9} & \mathbf{0}'_{n_{10}} & \mathbf{0}'_{n_{11}} & \mathbf{0}'_{n_{12}} & \mathbf{0}'_{n_{13}} & \mathbf{0}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \frac{1}{n_2}\mathbf{1}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \frac{-1}{n_6}\mathbf{1}'_{n_6} & \mathbf{0}'_{n_7} & \mathbf{0}'_{n_8} & \mathbf{0}'_{n_9} & \mathbf{0}'_{n_{10}} & \mathbf{0}'_{n_{11}} & \mathbf{0}'_{n_{12}} & \mathbf{0}'_{n_{13}} & \mathbf{0}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \frac{1}{n_3}\mathbf{1}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \frac{-1}{n_6}\mathbf{1}'_{n_6} & \mathbf{0}'_{n_7} & \mathbf{0}'_{n_8} & \mathbf{0}'_{n_9} & \mathbf{0}'_{n_{10}} & \mathbf{0}'_{n_{11}} & \mathbf{0}'_{n_{12}} & \mathbf{0}'_{n_{13}} & \mathbf{0}'_{n_{14}} \\
\frac{-1}{kn_1}\mathbf{1}'_{n_1} & \frac{-1}{kn_2}\mathbf{1}'_{n_2} & \frac{-1}{kn_3}\mathbf{1}'_{n_3} & \frac{k-1}{kn_4}\mathbf{1}'_{n_4} & \frac{-1}{kn_5}\mathbf{1}'_{n_5} & \frac{-1}{kn_6}\mathbf{1}'_{n_6} & \frac{-1}{kn_7}\mathbf{1}'_{n_7} & \frac{-1}{kn_8}\mathbf{1}'_{n_8} & \frac{-1}{kn_9}\mathbf{1}'_{n_9} & \frac{-1}{kn_{10}}\mathbf{1}'_{n_{10}} & \frac{-1}{kn_{11}}\mathbf{1}'_{n_{11}} & \frac{-1}{kn_{12}}\mathbf{1}'_{n_{12}} & \frac{-1}{kn_{13}}\mathbf{1}'_{n_{13}} & \frac{-1}{kn_{14}}\mathbf{1}'_{n_{14}} \\
\frac{-1}{kn_1}\mathbf{1}'_{n_1} & \frac{-1}{kn_2}\mathbf{1}'_{n_2} & \frac{-1}{kn_3}\mathbf{1}'_{n_3} & \frac{-1}{kn_4}\mathbf{1}'_{n_4} & \frac{k-1}{kn_5}\mathbf{1}'_{n_5} & \frac{-1}{kn_6}\mathbf{1}'_{n_6} & \frac{-1}{kn_7}\mathbf{1}'_{n_7} & \frac{-1}{kn_8}\mathbf{1}'_{n_8} & \frac{-1}{kn_9}\mathbf{1}'_{n_9} & \frac{-1}{kn_{10}}\mathbf{1}'_{n_{10}} & \frac{-1}{kn_{11}}\mathbf{1}'_{n_{11}} & \frac{-1}{kn_{12}}\mathbf{1}'_{n_{12}} & \frac{-1}{kn_{13}}\mathbf{1}'_{n_{13}} & \frac{-1}{kn_{14}}\mathbf{1}'_{n_{14}} \\
\frac{-1}{kn_1}\mathbf{1}'_{n_1} & \frac{-1}{kn_2}\mathbf{1}'_{n_2} & \frac{-1}{kn_3}\mathbf{1}'_{n_3} & \frac{-1}{kn_4}\mathbf{1}'_{n_4} & \frac{-1}{kn_5}\mathbf{1}'_{n_5} & \frac{k-1}{kn_6}\mathbf{1}'_{n_6} & \frac{-1}{kn_7}\mathbf{1}'_{n_7} & \frac{-1}{kn_8}\mathbf{1}'_{n_8} & \frac{-1}{kn_9}\mathbf{1}'_{n_9} & \frac{-1}{kn_{10}}\mathbf{1}'_{n_{10}} & \frac{-1}{kn_{11}}\mathbf{1}'_{n_{11}} & \frac{-1}{kn_{12}}\mathbf{1}'_{n_{12}} & \frac{-1}{kn_{13}}\mathbf{1}'_{n_{13}} & \frac{-1}{kn_{14}}\mathbf{1}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \mathbf{0}'_{n_6} & \frac{1}{n_7}\mathbf{1}'_{n_7} & \frac{-1}{n_8}\mathbf{1}'_{n_8} & \mathbf{0}'_{n_9} & \mathbf{0}'_{n_{10}} & \mathbf{0}'_{n_{11}} & \mathbf{0}'_{n_{12}} & \mathbf{0}'_{n_{13}} & \mathbf{0}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \mathbf{0}'_{n_6} & \mathbf{0}'_{n_7} & \frac{1}{n_8}\mathbf{1}'_{n_8} & \frac{-1}{n_9}\mathbf{1}'_{n_9} & \mathbf{0}'_{n_{10}} & \mathbf{0}'_{n_{11}} & \mathbf{0}'_{n_{12}} & \mathbf{0}'_{n_{13}} & \mathbf{0}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \mathbf{0}'_{n_6} & \mathbf{0}'_{n_7} & \mathbf{0}'_{n_8} & \frac{1}{n_9}\mathbf{1}'_{n_9} & \frac{-1}{n_{10}}\mathbf{1}'_{n_{10}} & \mathbf{0}'_{n_{11}} & \mathbf{0}'_{n_{12}} & \mathbf{0}'_{n_{13}} & \mathbf{0}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \mathbf{0}'_{n_6} & \mathbf{0}'_{n_7} & \mathbf{0}'_{n_8} & \mathbf{0}'_{n_9} & \frac{1}{n_{10}}\mathbf{1}'_{n_{10}} & \frac{-1}{4n_{11}}\mathbf{1}'_{n_{11}} & \frac{-1}{4n_{12}}\mathbf{1}'_{n_{12}} & \frac{-1}{4n_{13}}\mathbf{1}'_{n_{13}} & \frac{-1}{4n_{14}}\mathbf{1}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \mathbf{0}'_{n_6} & \mathbf{0}'_{n_7} & \mathbf{0}'_{n_8} & \mathbf{0}'_{n_9} & \mathbf{0}'_{n_{10}} & \frac{1}{n_{11}}\mathbf{1}'_{n_{11}} & \frac{-1}{3n_{12}}\mathbf{1}'_{n_{12}} & \frac{-1}{3n_{13}}\mathbf{1}'_{n_{13}} & \frac{-1}{3n_{14}}\mathbf{1}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \mathbf{0}'_{n_6} & \mathbf{0}'_{n_7} & \mathbf{0}'_{n_8} & \mathbf{0}'_{n_9} & \mathbf{0}'_{n_{10}} & \mathbf{0}'_{n_{11}} & \frac{1}{n_{12}}\mathbf{1}'_{n_{12}} & \frac{-1}{2n_{13}}\mathbf{1}'_{n_{13}} & \frac{-1}{2n_{14}}\mathbf{1}'_{n_{14}} \\
\mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \mathbf{0}'_{n_3} & \mathbf{0}'_{n_4} & \mathbf{0}'_{n_5} & \mathbf{0}'_{n_6} & \mathbf{0}'_{n_7} & \mathbf{0}'_{n_8} & \mathbf{0}'_{n_9} & \mathbf{0}'_{n_{10}} & \mathbf{0}'_{n_{11}} & \mathbf{0}'_{n_{12}} & \frac{1}{n_{13}}\mathbf{1}'_{n_{13}} & \frac{-1}{n_{14}}\mathbf{1}'_{n_{14}}
\end{pmatrix}.$$

Výsledné odhady parametrů jsou uvedeny v následující tabulce:

Kraje	Počet obyvatel	Parametry	Odhady parametrů	Vypočtené hodnoty
		$\hat{\mu}$	$\bar{Y}_{\cdot\cdot}$	19 680
Moravskoslezský kraj	1 249 356	$\hat{\beta}_1$	$\bar{Y}_{1\cdot} - \bar{Y}_{6\cdot}$	13 619
Hlavní město Praha	1 242 956	$\hat{\beta}_2$	$\bar{Y}_{2\cdot} - \bar{Y}_{6\cdot}$	77 691
Středočeský kraj	1 239 673	$\hat{\beta}_3$	$\bar{Y}_{3\cdot} - \bar{Y}_{6\cdot}$	28 595
Jihomoravský kraj	1 150 009	$\hat{\beta}_4$	$\bar{Y}_{4\cdot} - \bar{Y}_{\cdot\cdot}$	3 767
Ústecký kraj	836 128	$\hat{\beta}_5$	$\bar{Y}_{5\cdot} - \bar{Y}_{\cdot\cdot}$	1 359
Olomoucký kraj	641 945	$\hat{\beta}_6$	$\bar{Y}_{6\cdot} - \bar{Y}_{\cdot\cdot}$	−12 220
Jihočeský kraj	637 015	$\hat{\beta}_7$	$\bar{Y}_{7\cdot} - \bar{Y}_{8\cdot}$	3 683
Zlínský kraj	591 303	$\hat{\beta}_8$	$\bar{Y}_{8\cdot} - \bar{Y}_{9\cdot}$	−6 778
Plzeňský kraj	571 199	$\hat{\beta}_9$	$\bar{Y}_{9\cdot} - \frac{\bar{Y}_{10\cdot}+\cdots+\bar{Y}_{14\cdot}}{5}$	3 130
Královéhradecký kraj	554 511	$\hat{\beta}_{10}$	$\bar{Y}_{10\cdot} - \frac{\bar{Y}_{11\cdot}+\cdots+\bar{Y}_{14\cdot}}{4}$	1 325
Pardubický kraj	515 868	$\hat{\beta}_{11}$	$\bar{Y}_{11\cdot} - \frac{\bar{Y}_{12\cdot}+\cdots+\bar{Y}_{14\cdot}}{3}$	−2 891
Kraj Vysočina	515 326	$\hat{\beta}_{12}$	$\bar{Y}_{13\cdot} - \frac{\bar{Y}_{1\cdot}+\cdots+\bar{Y}_{12\cdot}}{12}$	−6 395
Liberecký kraj	438 238	$\hat{\beta}_{13}$	$\bar{Y}_{14\cdot} - \frac{\bar{Y}_{1\cdot}+\cdots+\bar{Y}_{13\cdot}}{13}$	−2 143
Karlovarský kraj	307 962	$\hat{\beta}_{14}$		0

Tabulka 4: Odhady parametrů při užití vlastího kódování kategoriální proměnné, interpretované na datech udávající počet cizinců v jednotlivých krajích ČR v rozmezí let 1996 – 2009

3 Metoda umělých proměnných v regresní analýze

Metoda umělých proměnných umožňuje zahrnout do regresního modelu i vliv kategoriálních (kvalitativních) znaků. Kvalitativní znaky přirozeným způsobem rozdělují data podle zvolené úrovně do dvou a více souborů podle počtu úrovní. Nabízí se tyto soubory řešit samostatně, čímž ale dochází k velké ztrátě informací. Metoda umělých proměnných nám umožňuje spojit tyto soubory do jednoho modelu. Umělé proměnné představují vhodně zvolené náhradní vysvětlující proměnné diskrétního typu, ve většině případů se jedná o binární proměnné (dummy kódování). S metodou umělých proměnných se můžeme setkat například v ekonometrii, časových řadách, při porovnávání shodnosti modelů, v analýze rozptylu, a spousty dalších. Připomeňme, že v následující kapitole budeme využívat výhradně dummy kódování (viz kapitola 2.1).

3.1 Dichotomické (binární) umělé proměnné v modelu regresních přímek

Nejprve si metodu umělých proměnných představme na nejjednoduším případě. Do regresního modelu budeme zavádět jedinou kvalitativní proměnnou, která nabývá pouze dvou úrovní. Kvalitativní proměnnou pak do modelu zavádíme pomocí umělé proměnné, která nabývá hodnot 0 a 1. Takovouto umělou proměnnou nazýváme proměnnou dichotomickou neboli binární.

Uvažujme nejprve model s jednou kvantitativní vysvětlující proměnnou, do kterého budeme navíc zavádět jednou kategoriální (kvalitativní) vysvětlující proměnnou. Pro lepší porozumění si další postup předveďme na ilustrativním příkladě. Uvažujme hypotetický příklad zkoumání vztahu mezi průměrnou mzdou (veličina Y) a délkou praxe v oboru uvedenou v letech (veličina x). K dispozici máme i informace o pohlaví, o kterém lze předpokládat, že může mít také vliv na průměrnou mzdu. Z tohoto důvodu zavedeme do modelu umělou proměnnou z_i nabývající hodnot:

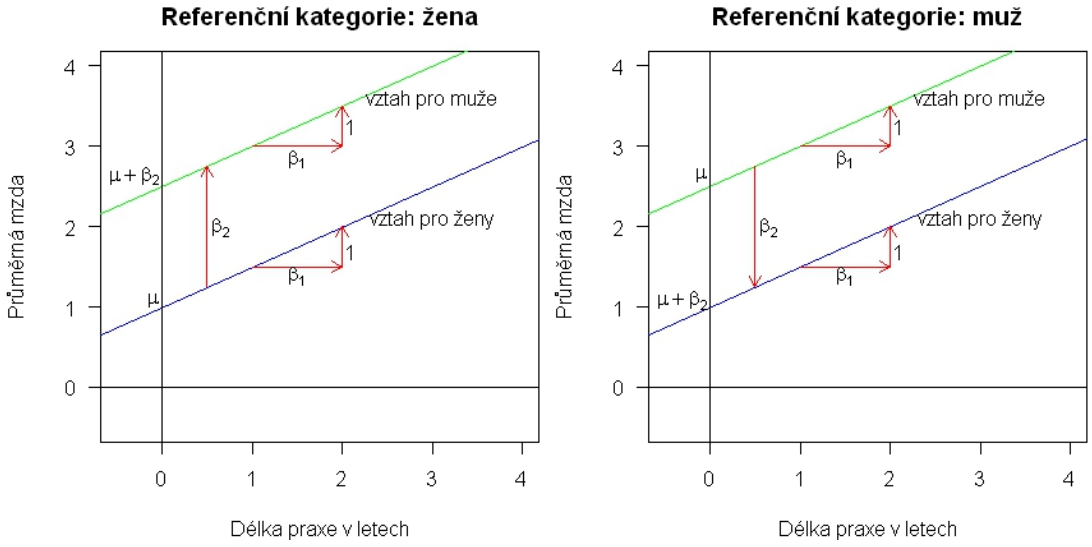
$$z_i = \begin{cases} 0 & \text{je-li předmětem } i\text{-tého pozorování žena,} \\ 1 & \text{je-li předmětem } i\text{-tého pozorování muž.} \end{cases} \tag{9}$$

Jelikož uvažujeme regresní model s absolutním členem μ, pak (jak víme z předchozích kapitol) je třeba zvolit referenční úroveň (viz kapitola 2). V našem případě je referenční úrovní zvolena úroveň „žena“. Zkoumáme tedy model:

$$Y_i = \mu + \beta_1 x_i + \beta_2 z_i + e_i, \quad \text{pro } i = 1, \ldots, n. \tag{10}$$

Dosazením hodnot za umělou proměnnou můžeme zkoumat dva separované modely

$$\begin{array}{ll} Y_i = \mu + \beta_1 x_i + e_i, & \text{model pro vyhodnocení údajů žen,} \\ Y_i = (\mu + \beta_2) + \beta_1 x_i + e_i, & \text{model pro vyhodnocení údajů mužů.} \end{array} \tag{11}$$



Obrázek 1: Modely regresních přímek s binární umělou proměnnou označující pohlaví při změně referenční úrovně.

Znázorněme si nastalou situaci pomocí regresních přímek. Jak je vidět na obrázku 1 vlevo, modely popisují dvě regresní přímky, které jsou rovnoběžné. Je to dáno tím, že se oba modely (11) liší pouze v hodnotě tzv. úrovňové konstanty (resp. absolutního členu). Parametr μ udává průměrnou mzdu žen v případě, kdy

je délka praxe nebo parametr β_1 roven 0. Směrnice přímek (tedy parametr β_1) je pro oba modely stejná. Parametr β_1 uvádí o kolik se liší průměrná mzda při změně délky praxe o jednotku. Také rozptyl náhodné chyby je pro obě přímky stejný. Parametr β_2 je měřítkem určujícím rozdíl v průměrné mzdě mezi muži a ženami při stejné délce praxe. Vertikální vzdálenost mezi paralelními regresními přímkami vyjadřuje vliv pohlaví na průměrnou mzdu. V modelu pro muže je absolutní člen ovlivněn parametrem indikujícím mužské pohlaví pozorované vysvětlující proměnné.

Pro zkoumání vlivu vysvětlujících proměnných na proměnnou vysvětlovanou pomocí separovaných regresní modelů rozdělených dle úrovní kvalitativního znaku však není příliš vhodný. Ač se tento přístup zdá být rozumný, má svá omezení: Konstrukce regrese pro odhady a testování vlivu pohlaví na průměrnou mzdu je velmi obtížná. Proto je preferován právě model s umělými proměnnými (10), jelikož nabízí jediný komplexní výsledek. Pokud tedy předpokládáme paralelní regresi pro ženy a muže, je efektivnější provádět odhady směrnice (parametru β_1), který je pro oba modely totožný, z modelu, který spojuje všechna pozorovaná data. Odhady parametrů μ, β_1, β_2 se provádějí stejně jako u regresních modelů s kvantitativními vysvětlujícími proměnnými pomocí metody nejmenších čtverců (MNČ) (2).

Podobné výsledky bychom samozřejmě získali, pokud bychom jako referenční úroveň zvolili muže. Jinými slovy, umělá proměnná bude nabývat hodnot

$$z_i = \begin{cases} 0 & \text{je-li předmětem } i\text{-tého pozorování muž,} \\ 1 & \text{je-li předmětem } i\text{-tého pozorování žena.} \end{cases} \tag{12}$$

Znázorněme si tuto situaci pomocí regresních přímek.

Situace je znázorněna na obrázku 1 napravo. Hodnota parametru β_2 je záporná, protože nyní představuje rozdíl v úrovňové konstantě mezi ženami a muži, ale její velikost zůstává stejná. Parametr μ nyní zobrazuje průměrnou mzdu mužů v případě, kdy je parametr β_1 nebo délka praxe rovna 0. Jak je vidět není nijak důležité, která skupina bude kódována 1 a která bude kódována 0. Analogicky není důležité, která úroveň bude vybrána jako referenční v případě užití abso-

lutního členu v regresním modelu. Je pouze nutné, abychom na základě zvoleného kódování byli schopní interpretovat výsledky.

Představili jsme si vytvoření umělé proměnné v modelu s jediným kvantitativním regresorem. Uvedená metoda umělých proměnných může být zcela analogicky aplikována na regresní modely s libovolným počtem kvantitativních vysvětlujících proměnných. Pro představu, pokud bychom sestrojili model s k kvantitativními vysvětlujícími proměnnými

$$Y_i = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \beta_{k+1} z_i + e_i,$$

potom po dosazení hodnot $0, 1$ za dichotomickou umělou proměnnou z_i bychom dostali dva separované modely:

$$\begin{aligned} Y_i &= \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i, && \text{pro } z_i = 0, \\ Y_i &= (\mu + \beta_{k+1}) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i, && \text{pro } z_i = 1. \end{aligned}$$

3.1.1 Dichotomická umělá proměnná v regresním modelu s interakcemi

V předchozí kapitole jsme zkoumali dva separované regresní modely s rozdílnou hodnotou úrovňové konstanty. Nyní se zaměříme na separované regresní modely, které se liší i v hodnotě směrnice. Tuto situaci lze namodelovat právě pomocí užití interakce $x_i z_i$. Co si ale pod pojmem interakce představit? Důležité je nezaměňovat interakci s korelací. Korelace zjišťuje závislost mezi vysvětlujícími proměnnými, interakce naopak zjišťuje, zda a jakým způsobem ovlivňuje kombinace vysvětlujících proměnných hodnotu vysvětlované proměnné. Model s interakcemi zapíšeme následovně:

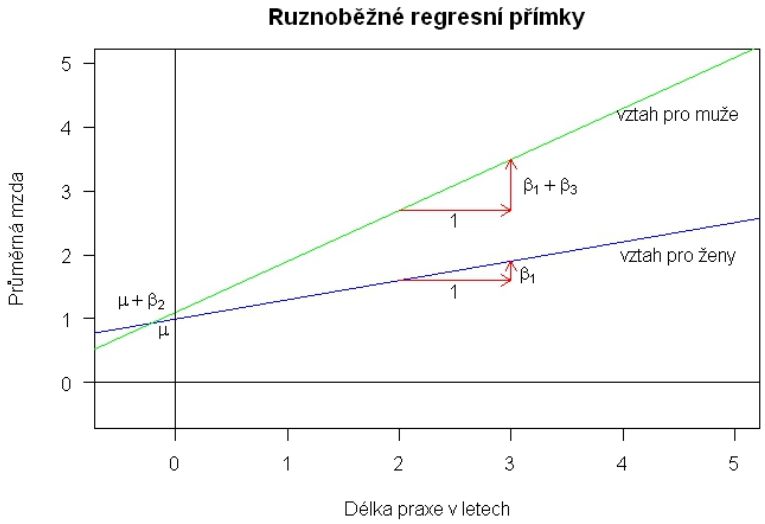
$$Y_i = \mu + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + e_i. \tag{13}$$

Zde stejně jako v předešlé kapitole je z_i umělá proměnná. Jelikož opět uvažujeme model s absolutním členem volíme referenční úroveň, například úroveň „žena“. Pak umělou proměnnou kódujeme dle (9) a dosazením hodnot za umělou proměnnou dostáváme opět dva separované regresní modely, tentokrát i s různými

hodnotami pro směrnice:

$$\begin{aligned} Y_i &= \mu + \beta_1 x_i + e_i, && \text{model pro ženy,} \\ Y_i &= (\mu + \beta_2) + (\beta_1 + \beta_3) x_i + e_i, && \text{model pro muže.} \end{aligned} \tag{14}$$

Z modelu je zřejmé, že parametr β_2 opět ovlivňuje hodnotu abslutního členu a nový parametr β_3 ovlivňuje směrnici modelu. Znázorněme si nastalou situaci pomocí regresních přímek.



Obrázek 2: Model regresních přímek s binární umělou proměnnou a interakcemi. Referenční úrovní je úroveň „žena“.

Jak je vidět z obrázku 2 regresní přímky již nejsou paralelní, tak jako v předchozím případě. Je to samozřejmě dáno odlišnou směrnicí v regresních modelech. Je zřejmé, že směrnice regresní přímky pro muže je větší než směrnice regresní přímky pro ženy. Vzhledem k tomu můžeme říct, že vliv délky praxe se liší podle pohlaví, tzn. délka praxe a pohlaví se ovlivňují. Tím pádem se vliv pohlaví na hodnotu vysvětlované proměnné liší v závislosti na hodnotě délky praxe a stejně tak se liší vliv délky praxe na hodnotu vysvětlované proměnné dle pohlaví. V předchozí kapitole představoval parametr β_2 konstantní dílčí vliv pohlaví na průměrnou mzdu při stejné délce praxe. Směrnice β_1 udávala jednotný vliv délky praxe na průměrnou mzdu pro muže i ženy. V modelu s interakcemi již tyto interpretace parametrů neplatí. V modelu pro muže je totiž hodnota

směrnice ovlivňována parametrem β_3 indikujícím možnou závislost délky praxe a mužského pohlaví. Dále je třeba mít na paměti, že při testování hypotéz v modelu s interakcemi je potřeba dodržet princip marginality. Čili nejdříve testujeme hypotézy $H_0 : \beta_3 = 0$ pro parametry u proměnných vyššího řádu, v našem případě pro parametry u interakcí. Pokud hypotézu H_0 nelze zamítnout, pak je parametr β_3 pro model nevýznamný a po úpravě modelu můžeme pokračovat v testování a testovat novou hypotézu $H_0 : \mu = \beta_1 = \beta_2 = 0$. Případně dílčí hypotézy $H_0 : \mu = 0$, $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$. Ovšem v případě, kdy původní hypotézu $H_0 : \beta_3 = 0$ zamítáme, tedy na základě testovacího kritéria zjišťujeme, že parametr β_3 je pro model významný, již nelze testovat hypotézu $H_0 : \beta_1 = 0$.

3.2 Vícekategoriální kvalitativní znaky

Zatím jsme se zabývali situací, kdy jsme do modelu zaváděli kvalitativní vysvětlující proměnnou, která nabývala pouze dvou úrovní. Tuto proměnnou jsme do modelu vkládali pomocí jediné dichotomické umělé proměnné z_i. Nyní si tento přístup kódování poněkud zobecníme. Tentokrát uvažujme kategoriální proměnnou, která nabývá více než dvou úrovní, např. vzdělání s úrovněmi ZŠ, SŠ, VŠ. V takovémto případě, nám již jedna umělá proměnná nestačí. Intuitivně zaveďme kódování následujícím způsobem:

	z_{i1}	z_{i2}	z_{i3}
ZŠ	1	0	0
SŠ	0	1	0
VŠ	0	0	1

Model bez absolutního členu by tedy byl tvaru:

$$Y_i = \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 z_{i3} + e_i. \tag{15}$$

Jak již víme z předchozích kapitol, pro model s absolutním členem obecně platí, že pro m-kategoriální proměnnou je potřeba $m-1$ umělých proměnných. Pro tří kategoriální kvalitativní proměnnou tedy zvolíme dummy kódování 2.1. Úroveň VŠ byla zvolena jako referenční.

	z_{i1}	z_{i2}
ZŠ	1	0
SŠ	0	1
VŠ	0	0

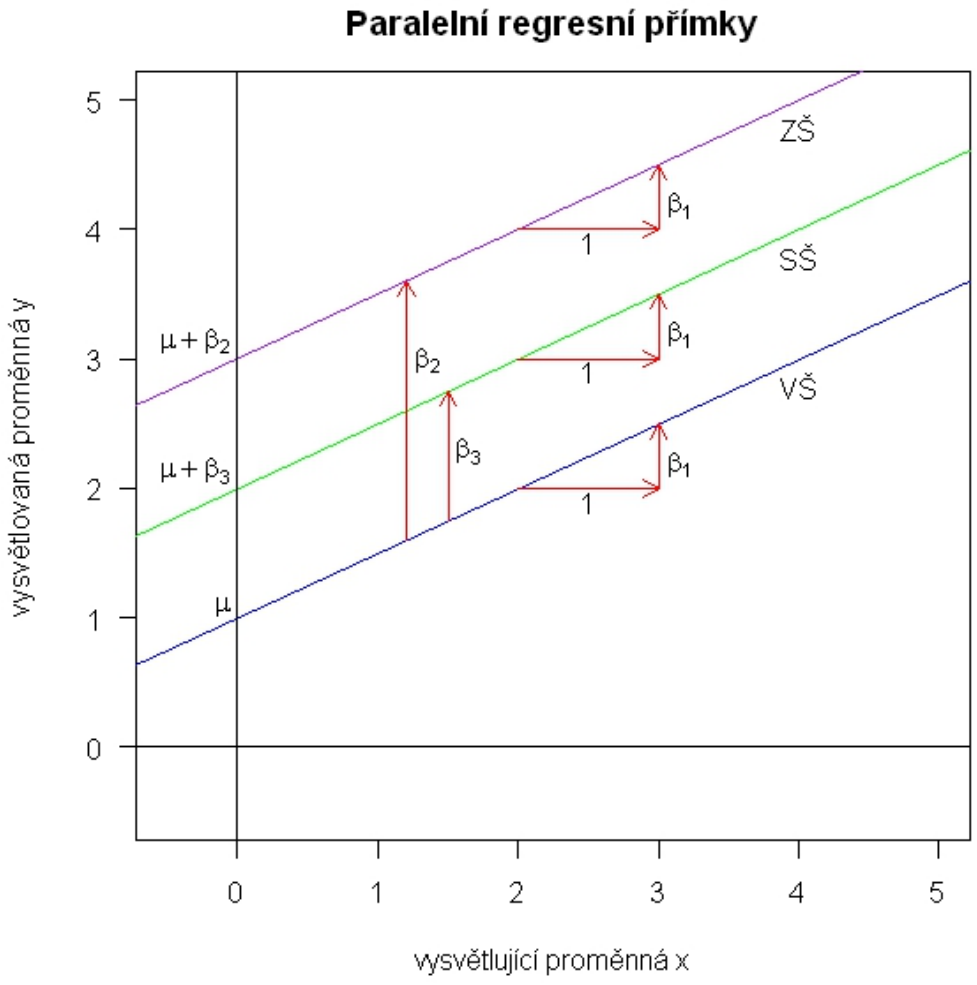
Uvažujme opět nejjednodušší regresní model s jednou kvantitativní a jednou tří krategoriální kvalitativní proměnnou

$$Y_i = \mu + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + e_i. \tag{16}$$

Dosazením za umělé proměnné dostáváme tři separované regresní rovnice

$$\begin{aligned} Y_i &= (\mu + \beta_2) + \beta_1 x_i + e_i, \text{ pro ZŠ}, \\ Y_i &= (\mu + \beta_3) + \beta_1 x_i + e_i, \text{ pro SŠ}, \\ Y_i &= \mu + \beta_1 x_i + e_i, \qquad \text{pro VŠ}. \end{aligned} \tag{17}$$

Jelikož se jedná o model bez interakcí, tak se jedná o paralelní přímky. Parametr β_2 vyjadřuje konstantní vzdálenost mezi regresní přímkou ZŠ a VŠ, parametr β_3 naopak vyjadřuje konstantní vzdálenost mezi regresními přímkami SŠ a VŠ. Jinak řečeno parametr β_2 udává rozdíl vysvětlované proměnné mezi ZŠ a VŠ. Parametr β_3 udává rozdíl vysvětlované proměnné mezi SŠ a VŠ. Absolutní člen modelu pro základní školy je ovlivněn hodnotou parametru β_2 indikujícím dosažené základní vzdělání. Absolutní člen modelu pro střední školy je zase ovlivněn hodnotou parametru β_3 indikujícím dosažené středoškolské vzdělání.



Obrázek 3: Model regresních přímek pro 3-kategoriální proměnnou vzdělání. Referenční úroveň je úroveň „VŠ“.

3.2.1 Interakce s vícekategoriálními kvalitativními proměnnými

Pro zlepšení kvality modelu předpokládejme model s interakcemi. Stejně jako u dichotmockých umělých proměnných chceme zjistit, zda se nějakým způsobem nemění vliv vzdělání na vysvětlovanou proměnnou v závislosti na kvantitativní vysvětlující proměnné. Model bude ve tvaru

$$Y_i = \mu + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 x_i z_{i1} + \beta_5 x_i z_{i2} + e_i, \tag{18}$$

s kódováním

	z_{i1}	z_{i2}
ZŠ	1	0
SŠ	0	1
VŠ	0	0

Po dosazení za umělé proměnné dostáváme opět tři separované regresní modely, které se liší v úrovňové konstantě a navíc i ve směrnici.

$$\begin{aligned} Y_i &= (\mu + \beta_2) + (\beta_1 + \beta_4)x_i + e_i, \text{ pro ZŠ}, \\ Y_i &= (\mu + \beta_3) + (\beta_1 + \beta_5)x_i + e_i, \text{ pro SŠ}, \\ Y_i &= \mu + \beta_1 x_i + e_i, \qquad\qquad \text{pro VŠ}. \end{aligned} \tag{19}$$

Směrnice modelu pro základní školy je ovlivněna hodnotou parametru β_4 idnikujícím možný vliv mezi dosaženým základním vzděláním a kvantitativní proměnnou x_i. Analogicky směrnice modelu pro střední školy je ovlivněna hodnotou parametru β_5 idnikujícím možný vliv mezi dosaženým středoškolským vzděláním a kvantitativní proměnnou x_i. Zcela analogicky by se pracovalo i s kvalitativními proměnnými, které mají více než 3 úrovně. Jejich kódování je naznačeno v kapitole 2.

3.3 Modely s více kategoriálními proměnnými

Dosud jsme si popsali pouze případy, kdy jsme do modelu zaváděli jedinou kategoriální (kvalitativní) proměnnou. Nyní se zaměříme na případ, kdy budeme do modelu zavádět více kategoriálních proměnných. Pro ukázku uvažujme případ, kdy zjišťujeme vliv jedné kvantitativní a dvou kategoriálních (přitom jedna bude

dichotomická a druhá víceúrovňová) proměnných na hodnotu vysvětlované proměnné. Jelikož opět uvažujeme model s absolutním členem, je třeba vybrat pro každou kategoriální proměnnou jednu refrenční úroveň. Uvažujeme například čistě hypotetický model, kde zkoumáme vliv délky praxe (x), pohlaví (z_1) a dosaženého vzdělání (z_2, z_3) na průměrnou mzdu (y). Pro kategoriální proměnnou pohlaví volme referenční úroveň „muž“, pro proměnnou dosažené vzdělání volme referenční úroveň „ZŠ“. Model zapíšeme následovně:

$$Y_i = \mu + \beta_1 x_{i1} + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 z_{i3} + e_i, \tag{20}$$

kde pro regresor z_{i1} platí:

$$z_{i1} = \begin{cases} 0 & \text{je-li předmětem } i\text{-tého pozorování muž,} \\ 1 & \text{je-li předmětem } i\text{-tého pozorování žena.} \end{cases} \tag{21}$$

Regresory z_{i2}, z_{i3} jsou regresory tří-úrovňové kvalitativní proměnné dosažené vzdělání, pro něž platí

	z_{i2}	z_{i3}
VŠ	1	0
SŠ	0	1
ZŠ	0	0

Po dosazení hodnot za umělé proměnné dostáváme 6 separovaných regresních modelů

Pohlaví	Vzdělání	
M	VŠ	$Y_i = (\mu + \beta_3) + \beta_1 x_{i1} + e_i$
M	SŠ	$Y_i = (\mu + \beta_4) + \beta_1 x_{i1} + e_i$
M	ZŠ	$Y_i = \mu + \beta_1 x_{i1} + e_i$
Ž	VŠ	$Y_i = (\mu + \beta_2 + \beta_3) + \beta_1 x_{i1} + e_i$
Ž	SŠ	$Y_i = (\mu + \beta_2 + \beta_4) + \beta_1 x_{i1} + e_i$
Ž	ZŠ	$Y_i = (\mu + \beta_2) + \beta_1 x_{i1} + e_i$

Modely se liší pouze v hodnotě absolutního členu. U prvního modelu je absolutní člen ovlivněn parametrem β_3 indikujícím dokončené vysokoškolské vzdělání. Absolutní člen druhého modelu je naopak ovlivněn parametrem β_4 indikující dokončené středoškolské vzdělání. U třetího modelu není absolutní člen ovlivněn žádným jiným parametrem, neboť se jedná o model pro obě referenční úrovně

kategoriálních proměnných pohlaví a dosažené vzdělání. U třetího až šestého modelu jde o analogii pro první tři modely. Absolutní členy jsou ale navíc ovlivněny i parametrem β_2 indikující ženské pohlaví pozorované náhodné veličiny. Abychom docílili lepšího popisu modelu, pak musíme do modelu zavést také interakce zjišťující, zda a jakým způsobem se mění vliv pohlaví a dosažené vzdělání v závislosti na délce praxe na hodnotu průměrné mzdy. Uvažujme tedy model s interakcemi.

$$Y_i = \mu + \beta_1 x_{i1} + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 z_{i3} + \beta_5 x_{i1} z_{i1} + \beta_6 x_{i1} z_{i2} + \beta_7 x_{i1} z_{i3} + e_i. \quad (22)$$

To však stále nepokrývá všechny situace. Ještě je do modelu třeba zavést interakce mezi kvalitativními proměnnými, které zjistí, zda má pohlaví vliv na dosažené vzdělání. Uvažujeme tedy model

$$\begin{aligned} Y_i = \mu &+ \beta_1 x_{i1} + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 z_{i3} + \beta_5 x_{i1} z_{i1} + \beta_6 x_{i1} z_{i2} + \beta_7 x_{i1} z_{i3} + \\ &+ \beta_8 z_{i1} z_{i2} + \beta_9 z_{i1} z_{i3} + e_i. \end{aligned} \quad (23)$$

Dosazením za umělé proměnné dostáváme separované proměnné tvaru

Pohlaví	Vzdělání	
M	VŠ	$Y_i = (\mu + \beta_3) + (\beta_1 + \beta_6) x_{i1} + e_i$
M	SŠ	$Y_i = (\mu + \beta_4) + (\beta_1 + \beta_7) x_{i1} + e_i$
M	ZŠ	$Y_i = \mu + \beta_1 x_{i1} + e_i$
Ž	VŠ	$Y_i = (\mu + \beta_2 + \beta_3 + \beta_8) + (\beta_1 + \beta_5 + \beta_6) x_{i1} + e_i$
Ž	SŠ	$Y_i = (\mu + \beta_2 + \beta_4 + \beta_9) + (\beta_1 + \beta_5 + \beta_7) x_{i1} + e_i$
Ž	ZŠ	$Y_i = (\mu + \beta_2) + (\beta_1 + \beta_5) x_{i1} + e_i$

Jak vidíme, situace už se poněkud horší. Pro posouzení modelu s jednou kvantitativní a pouhými dvěmi kvalitativními (jednou dvou kategoriální a jednou tří kategoriální) proměnnými je třeba odhadnout 10 parametrů. Při užití interakcí se v modelech liší nejen hodnoty absolutního členu, ale i směrnice regresní přímky. Podívejme se na absolutní členy a směrnice regresní přímky blíže. U prvního modelu je hodnota absolutního členu ovlivněna navíc parametrem β_3 indikující dokončené vysokoškolské vzdělání a hodnota směrnice regresní přímky je ovlivněna parametrem β_6 indikující závislost délky praxe na dokončeném vysokoškolském vzdělání. U druhého modelu je hodnota absolutního členu ovlivněna

parametrem β_4 indikující dokončené středoškolské vzdělání a hodnota směrnice regresní přímky je ovlivněna parametrem β_7 indikující závislost délky praxe na dokončeném středoškolském vzdělání. Modely pro ženy jsou o něco složitější analogií modelů pro muže, jelikož nám do modelu vstupují i parametry indikující ženské pohlaví pozorované veličiny. Absolutní člen čtvrtého modelu je stejně jako u prvního modelu ovlivňován parametrem β_3 indikujícím dokončené vysokoškolské vzdělání navíc je ale ovlivňován i parametrem β_2 indikujícím ženu a parametrem β_8 indikujícím možnou závislot mezi ženským pohlavím a dokončeným vysokoškolským vzděláním. Směrnice téhož modelu je pak ovlivněna kromě parametru β_6 indikujícím možnou závislot mezi dokočeným vysokoškolským vzděláním a délkou praxe i parametrem β_5, který indikuje možnou závislot mezi ženským pohlavím a délkou praxe.

4 Aplikace metody umělých proměnných

Dosud získané teoretické znalosti o metodě umělých proměnných si nyní představme a více rozveďme na konkrétních případech užití této metody.

4.1 Porovnávání separovaných regresních modelů

Jednou z oblastí, kde nám může umělá proměnná významně pomoci je porovnávání separovaných modelů jedné datové množiny. V předchozích kapitolách jsme zjistili, že některá data lze pomocí umělých proměnných rozlišit podle úrovní nějaké kvalitativní proměnné a lze tak pro každou z úrovní získat samostatný separovaný model. V této kapitole se zaměříme na poněkud odlišnou situaci. Na základě dat obdržíme regresní modely rozlišené dle některé úrovně. Našim úkolem je tyto modely vzájemně porovnat a zjistit, zda nejsou totožné. K tomuto využijeme informace získané z předchozích kapitol. Pro názornost budeme opět uvažovat nejjednodušší regresní modely.

4.1.1 Modely s odlišným absolutním členem a rozdílnou směrnicí

Uvažujme separované regresní modely, které jsme dostali na základě hypotetického příkladu zkoumajícího závislost průměrné mzdy na délce praxe a pohlaví:

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 x_i + e_i, \quad \text{model pro muže,} \\ Y_i &= \delta_0 + \delta_1 x_i + e_i, \quad \text{model pro ženy.} \end{aligned} \tag{24}$$

Našim úkolem je tyto modely porovnat. Nejprve vyslovme hypotézu, která otestuje, zda jsou separované modely (24) identické

$$H_0: \quad \gamma_0 = \delta_0 \ \wedge \gamma_1 = \delta_1.$$

Test takovéto hypotézy by se prováděl příliš složitě, a proto nám mohou pomoci umělé proměnné. Z předchozích kapitol víme, že takovéto separované modely jsme dostali při užití dummy regresního modelu s interakcemi

$$Y_i = \mu + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + e_i, \tag{25}$$

Původní separované modely můžeme tedy přepsat na

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 x_i + e_i = \mu + \beta_1 x_i + e_i, && z_i = 0, \\ Y_i &= \delta_0 + \delta_1 x_i + e_i = (\mu + \beta_2) + (\beta_1 + \beta_3) x_i + e_i, && z_i = 1. \end{aligned} \tag{26}$$

Jelikož je model (25) ekvivalentní s oběma separovanými modely, lze hypotézu vyslovit i takto

$$H_0 : \beta_2 = \beta_3 = 0.$$

Tato hypotéza se testuje pomocí F-testu pro testování stability podmodelu. Testovací statistika je

$$F = \frac{n-k}{k-l} \frac{R^2}{1-R^2} \sim F_{k-l,n-k}, \tag{27}$$

kde R^2 je index determinace modelu (25), n je počet pozorování, k je počet parametrů modelu (25) a l je počet parametrů modelu (25) za platnosti nulové hypotézy $H_0 : \beta_2 = \beta_3 = 0$. V našem případě $k = 4$ a $l = 2$.

4.1.2 Modely se stejnou směrnicí a rozdílnou úrovňovou konstantou

Nyní uvažujme situaci, kdy máme důvod předpokládat, že regresní modely mají stejnou směrnici. Uvažujeme tedy modely

$$\begin{aligned} Y_i &= \gamma_0 + \beta_1 x_i + e_i, \quad \text{model pro muže}, \\ Y_i &= \delta_0 + \beta_1 x_i + e_i, \quad \text{model pro ženy}. \end{aligned} \tag{28}$$

Opět chceme dané modely porovnat a zjistit zda nejsou identické. V tomto případě chceme otestovat hypotézu

$$H_0 : \gamma_0 = \delta_0$$

Vytvoříme proto opět regresní model s umělými proměnnými, tentokrát již ale bez interakce:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i, \tag{29}$$

Pro separované modely (28) poté platí

$$\begin{aligned} Y_i &= \gamma_0 + \beta_1 x_i + e_i = \beta_0 + \beta_1 x_i + e_i, && z_i = 0, \\ Y_i &= \delta_0 + \beta_1 x_i + e_i = (\beta_0 + \beta_2) + \beta_1 x_i + e_i, && z_i = 1. \end{aligned} \tag{30}$$

Můžeme proto testovat ekvivalentí hypotézu

$$H_0 : \beta_2 = 0.$$

Opět pro otestování hypotézy můžeme použít F-test pro testování stability podmodelu. Ekvivalentně lze užít i t-test. Testovací statistika pro t-test je

$$T = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 v_{33}}} \sim t_{n-3},$$

kde v_{33} je diagonální prvek na pozici 3, 3 matice $V = (X'X)^{-1}$.

4.1.3 Modely s totožnou úrovňovou konstantou a rozdílnou směrnicí

Zaměřme se nyní na porovnání modelů, které mají totožnou úrovňovou konstantu, ale odlišnou směrnici. Tedy řešíme modely

$$\begin{aligned} Y_i &= \mu + \gamma_1 x_i + e_i, \quad \text{model pro muže,} \\ Y_i &= \mu + \delta_1 x_i + e_i, \quad \text{model pro ženy.} \end{aligned} \tag{31}$$

Opět testujeme hypotézu o shodnosti modelů. $H_0 : \gamma_1 = \delta_1$. Stejně jako v případě 4.1.1 zkonstruujeme dummy regresní model s interakcemi:

$$Y_i = \mu + \beta_1 x_i + \beta_2 x_i z_i + e_i. \tag{32}$$

Původní separované modely přepíšeme

$$\begin{aligned} Y_i &= \mu + \gamma_1 x_i + e_i = \mu + \beta_1 x_i + e_i, && z_i = 0, \\ Y_i &= \mu + \delta_1 x_i + e_i = \mu + (\beta_1 + \beta_2) x_i + e_i, && z_i = 1. \end{aligned} \tag{33}$$

Nyní můžeme vyslovit ekvivalentní hypotézu $H_0 : \beta_2 = 0$. Test této hypotézy lze jako v předešlém případě provést buď F-testem (27) nebo t-testem (5).

$$T = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 v_{33}}} \sim t_{n-3},$$

kde v_{33} je třetí diagonální prvek matice $V = (X'X)^{-1}$. Zcela analogicky by se testy prováděli i pro vícektagoriální kvalitativní proměnnou.

4.2 Agregace kvantitativní proměnné

Umělé proměnné využíváme i v případě agregace (neboli seskupení) kvantitativní proměnné, která obsahuje příliš mnoho hodnot. Nejčastějším příkladem agregace proměnné je věk nebo peněžní příjem, kde nám stačí rozlišit pozorované osoby do věkových nebo příjmových skupin. Nejprve všechny hodnoty, kterých může kvantitativní proměnná nabývat rozdělíme do disjunktních skupin. Uvažujme například kvantitativní proměnnou věk. Rozdělíme ji do skupin $0-10, 11-20, 21-30, 31$ a víc. Použijeme-li dummy kódování, pak umělé proměnné budou nabývat hodnot

$$z_1 = \begin{cases} 1 & 0-10 \\ 0 & \text{jinak} \end{cases} \qquad z_2 = \begin{cases} 1 & 11-20 \\ 0 & \text{jinak} \end{cases}$$

$$z_3 = \begin{cases} 1 & 21-30 \\ 0 & \text{jinak} \end{cases} \qquad z_4 = \begin{cases} 1 & \geq 31 \\ 0 & \text{jinak} \end{cases}$$

Existuje ještě další způsob kódování, který jsme si zatím nepředváděli:

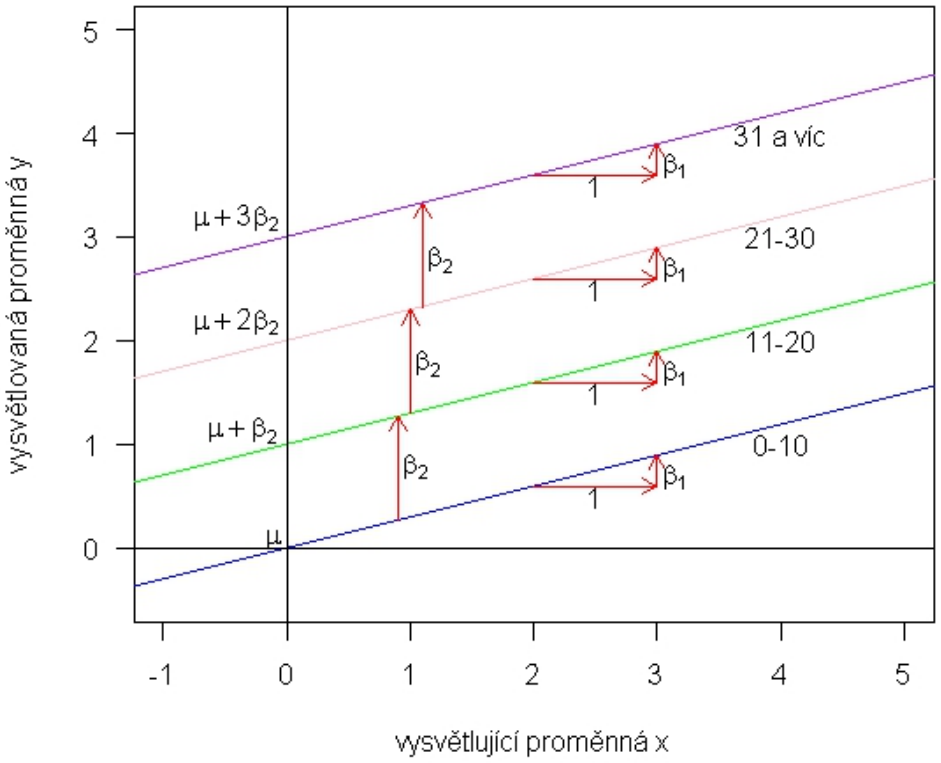
$$z = \begin{cases} 0 & 0-10 \\ 1 & 11-20 \\ 2 & 21-30 \\ 3 & \geq 31 \end{cases}$$

Tento způsob kódování však má své značné nedostatky. Lze totiž použít pouze v případě, kdy lze úrovně kvalitativní (popř. agregované kvantitativní) proměnné uspořádat. Pokud zkoumáme regresní model s jednou kvantitativní a jednou agregovanou kvantitativní proměnnou bez interakcí

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i,$$

pak po dosazení hodnot za umělou proměnnou z dostáváme separované modely ve tvaru:

$$\begin{array}{ll} Y_i = \beta_0 + \beta_1 x_i + e_i, & z_i = 0, \\ Y_i = (\beta_0 + \beta_2) + \beta_1 x_i + e_i, & z_i = 1, \\ Y_i = (\beta_0 + 2\beta_2) + \beta_1 x_i + e_i, & z_i = 2, \\ Y_i = (\beta_0 + 3\beta_2) + \beta_1 x_i + e_i, & z_i = 3. \end{array} \tag{34}$$



Obrázek 4: Model regresních přímek s konstantní změnou úrovňové konstanty a stejnou směrnicí. Referenční úroveň „0-10“.

Znázorněme si situaci pomocí regresních přímek. Jak můžeme vidět na obrázku 4, parametr β_2 je konstantní rozdíl mezi sousedními uspořádanými úrovněmi. Takové chování je v reálném světě nepravděpodobné.

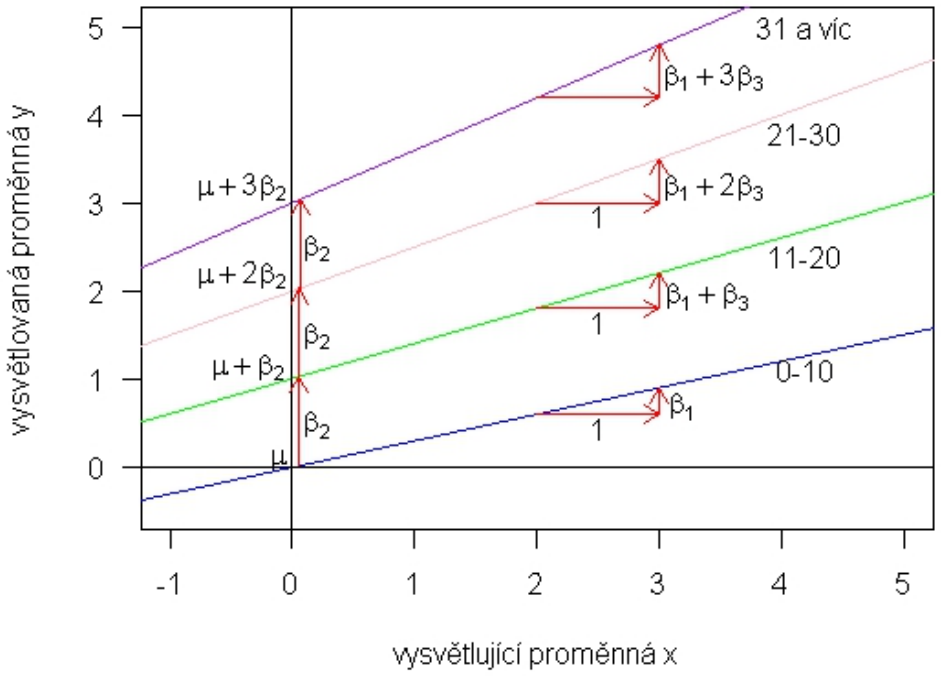
Uvažujme nyní model s interakcemi:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 z_i x_i + e_i,$$

Po dosazení hodnot za umělou proměnnou z dostáváme separované modely ve tvaru:

$$\begin{array}{ll} Y_i = \beta_0 + \beta_1 x_i + + e_i, & z_i = 0, \\ Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i + e_i, & z_i = 1, \\ Y_i = (\beta_0 + 2\beta_2) + (\beta_1 + 2\beta_3) x_i + e_i, & z_i = 2, \\ Y_i = (\beta_0 + 3\beta_2) + (\beta_1 + 3\beta_3) x_i + e_i, & z_i = 3. \end{array} \tag{35}$$

Znázorněme si situaci pomocí regresních přímek. Jak můžeme vidět na obrázku 5, parametr β_2 je stále konstantní rozdíl mezi sousedními uspořádanými úrovněmi



Obrázek 5: Model regresních přímek s konstantní změnou úrovňové konstanty a směrnice.Referenční úroveň „0-10“.

a směrnice regresních přímek se pro každou úroveň zvětšuje o parametr β_3. I takovéto chování je v reálném světě nepravděpodobné.

5 Umělé proměnné v ANOVĚ

Další oblastí, kde se můžeme s metodou umělých proměnných setkat je analýza rozptylu neboli ANOVA. Uvažujme, že máme k ($k \geq 2$) vzájemně nezávislých náhodných výběrů z normálního rozdělení se stejným rozptylem:

$$\begin{aligned} &Y_{11}, \ldots, Y_{1n_1} \text{ výběr z } N(\mu_1, \sigma^2) \\ &\vdots \\ &Y_{k1}, \ldots, Y_{kn_k} \text{ výběr z } N(\mu_k, \sigma^2) \end{aligned}$$

Úkolem analýzy rozptylu je otestovat hypotézu o shodnosti středních hodnot $H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_k$ oproti alternativní hypotéze, že některá z rovností neplatí. Jak již vyplynulo z předešlé kapitoly při porovnávání modelů je výhodnější a přesnější pokusit se všechny modely shrnout do jednoho. K tomu nám pomohou umělé proměnné. Pokud sepíšeme všechna pozorování pod sebe, pak můžeme pomocí dummy kódování (kapitola 2.1) jednoduše sestrojit regresní model

$$Y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_{k-1} x_{(k-1)j} + e_{ij}, \quad i = 1, \ldots, k, \ j = 1, \ldots, n_i, \tag{36}$$

kde e_{ij} je náhodná veličina vyjadřující chybu modelu, která má normální rozdělní s nulovou střední hodnotou a rozptylem σ^2. Referenční úrovní byl zvolen $k-$tý výběr $Y_{kj} = \beta_0 + e_{kj}, \ j = 1, \ldots, n_k$. V tomto modelu máme $k-1$ umělých proměnných $x_{1j}, \ldots, x_{(k-1)j}$, pro která platí:

$$x_{ij} = \begin{cases} 1 & \text{pozorování } y_{ij} \text{ patří do } i\text{-tého výběru,} \\ 0 & \text{jinak.} \end{cases}$$

Nyní si předveďme, jak bude vypadat model při užití kódování effect (kapitola 2.3), kde je k-tý výběr opět vybrán jako referenční úroveň.

$$\begin{aligned} Y_{ij} &= \mu_j + e_{ij}, & i = 1, \ldots, k-1, \quad j = 1, \ldots, n_i, \\ Y_{kj} &= -\sum_{l=1}^{k-1} \mu_l + e_{ij}, & j = 1, \ldots, n_k. \end{aligned}$$

Pro shrnutí do jednoho modelu opět použijeme umělé proměnné, tentokrát pro pro-

měnné x_{ij} platí:

$$x_{ij} = \begin{cases} 1 & \text{pozorování } y_{ij} \text{ patří do } i\text{-tého výběru}, i = 1, \ldots, k-1, \\ 0 & \text{jinak, pro } i = 1, \ldots, k-1, \\ -1 & \text{pozorování } y_{ij} \text{ patří do } k\text{-tého výběru.} \end{cases}$$

Celkový model bude tvaru:

$$Y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_{k-1} x_{(k-1)j} + e_{ij}, \quad i = 1, \ldots, k, \; j = 1, \ldots, n_i. \tag{37}$$

U kódování efect se musí myslet také na výpočet odhadu parametru referenční úrovně dle vztahu $\sum_1^k \hat{\beta}_i = 0$.

Před provedením samotné analýzy rozptylu je třeba ještě ověřit předpoklady, tedy normalitu a homoskedasticitu pro všech k výběrů. Normalitu ověříme například pomocí Shapiro-Wilkova testu normality. Pro ověření homoskadasticity použijeme Bartlettův nebo Cochranův test shody rozptylů. V případě, že jsou předpoklady splněny můžeme přejít k samotné analýze rozptylu. Nyní můžeme vyslovit ekvivalentní hypotézu pro srovnání shodnosti středních hodnot jednotlivých výběrů $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$. Hypotézu H_0 můžeme opět testovat pomocí F-testu (27) stability podmodelu

$$F = \frac{(n-k)}{(k-1)} \frac{R^2}{1-R^2} \sim F_{k-1,n-k},$$

kde $n = n_1 + \cdots + n_k$ je celkový počet pozorování. Pro porovnání provedeme výpočet analýzy rozptylu na konkrétním příkladě nejprve pomocí dummy kódování a následně pomocí kódování effect (viz 9.1). V literatuře se dá setkat i dalšími typy kódování uvedenými v kapitole 2.

6 Analýza kovariance ANCOVA

Analýza kovariance je metoda, která kombinuje analýzu rozptylu (ANOVU) a regresní analýzu (RA). Cílem analýzy kovariance je očištění studované závislosti vysvětlované (měřené) proměnné na zvolených kategoriálních proměnných od „zavádějícího“ působení doprovodných vlivů (označovaných jako **kovariáty**). Působení doprovodných proměnných na vysvětlované proměnné je sice podstatné, ale není v dané úloze přímým předmětem zájmu. V analýze kovariance se uplatňuje složitější sestava proměnných [14],[15]:

1. Jedna nebo několik vysvětlujících kategoriálních proměnných $z_j, j = 1, \ldots, p$, (stejně jako u analýzy rozptylu), které mohou nabývat až k úrovní. Pro kategoriální proměnnou z_j zapíšeme úrovně $z_{j1}, z_{j2}, \ldots, z_{jk}$.

2. Jedna nebo více vysvětlovaných proměnných $Y_1, \ldots, Y_n$, jejichž závislost na kategoriálních proměnných chceme prokázat.

3. Jedna nebo více doprovodných proměnných, tzv. **kovariát** $x_1, x_2, \ldots, x_q$, které zahrnujeme do modelu z důvodu očištění závislosti vysvětlovaných proměnných Y_i od vlivu kategoriálních proměnných z_j.

Jelikož je analýza kovariance složením metod ANOVY a regresní analýzy, musí model pro analýzu kovariance splňovat několik podmínek. Stejně jako u ANOVY uvažujeme k vzájemně nezávislých náhodných výběrů z normálního rozdělení se stejným rozptylem. Navíc musí existovat lineární závislost mezi náhodnou veličinou $\mathbf{y}$ a kovariátou (regresní proměnnou) $\mathbf{x}$. Regresní přímky pro všechny úrovně kategoriální proměnné $\mathbf{z}$ musí být rovnoběžné. Jinými slovy regresní koeficienty (udávají směrnici regresní přímky) musí být totožné $\beta_1 = \cdots = \beta_k$.

Připomeňme si tvary jednoduchých modelů pro regresní analýzu a analýzu rozptylu:

Regresní model	ANOVA
$Y_i = \beta_0 + \beta_1 x_i + e_i$	$Y_{ij} = \mu + \alpha_i + e_{ij}$

Přidáním regresní proměnné x_{ij} do modelu analýzy rozptylu narazíme na problém s absolutním členem. U ANOVY je absolutní člen roven střední hodnotě celkového výběru, kdežto u regresního modelu nikoli. Proto regresní proměnnou budeme centrovat a použijeme proměnnou $x^*_{ij} = x_{ij} - \bar{x}_{..}$. Výrazem $\bar{x}_{..}$ označujeme průměr přes všechny regresní proměnné.

V analýze kovariance užíváme pro kódování kategoriální proměnné $\mathbf{z}$ kódování effect uvedené v kapitole 2.3. Uvažujme nejjednodušší případ modelu pro analýzu kovariance, tedy model s jednou kategoriální proměnnou $\mathbf{z}$ nabývající k úrovní a jednou doprovodnou regresní proměnnou $\mathbf{x}$:

$$Y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + e_{ij}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, n_i. \tag{38}$$

Y_{ij} značí j-tou hodnotu vysvětlované proměnné pro i-tou úroveň kategoriální proměnné $\mathbf{z}$, μ značí střední hodnotu výběru, α_i je tzv. effekt i-té úrovně kategoriální proměnné $\mathbf{z}$. Platí pro ně vztah $\sum_{i=1}^{k} n_i\alpha_i = 0$. Jde o analogii vztahu $\sum_1^k \hat{\gamma}_i = 0$ pro odhady parametrů kategoriální proměnné kódované pomocí kódování effect. Platí $\alpha_i = \frac{1}{n_i}\hat{\gamma}_i = \frac{1}{n_i}(\bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot})$. Parametr β je regresní koeficient, x_{ij} je i-tá hodnota kovariáty pro j-tou úroveň kategoriální proměnné $\mathbf{z}$, $\bar{x}$ označuje průměr kovariáty a e_{ij} označuje j-tou hodnotu náhodné chyby pro i-tou úroveň kategoriální proměnné $\mathbf{z}$.

Pokud odhadneme parametr β, můžeme „očistit“ závislost vysvětlované proměnné $\mathbf{y}$ na kategoriální proměnné $\mathbf{z}$ od závislosti kovariáty $\mathbf{x}$ tím, že vypočítáme opravené hodnoty Y_{ij} vztahem

$$Y^*_{ij} = Y_{ij} - \hat{\beta}(x_{ij} - \bar{\mathbf{x}}_{..}). \tag{39}$$

Opravená hodnota vysvětlované proměnné $\mathbf{y}$ je vlastně hodnota Y_{ij} vysvětlované proměnné $\mathbf{y}$ přepočítaná na průměrnou hodnotu $\bar{\mathbf{x}}_{..}$ kovariáty $\mathbf{x}$. Tedy pomocí analýzy kovariance standardizujeme původní výběrové průměry pro jednotlivé úrovně kategoriální proměnné $\mathbf{z}$ pomocí regresího vztahu mezi vysvětlovanou proměnnou $\mathbf{y}$ a kovariátou $\mathbf{x}$ na úroveň průměrné hodnoty kovariáty $\mathbf{x}$.

Dosazením do (38) získáme model analýzy kovariance

$$Y^*_{ij} = \mu + \alpha_i + e_{ij}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, n_i. \tag{40}$$

Z tvaru rovnice (40) je ihned vidět, že jde o model analýzy rozptylu. Jinými slovy analýza kovariance je analýzou rozptylu pro opravené hodnoty $\mathbf{y}^*$ náhodného výběru $\mathbf{y}$.

Postup analýzy kovariance:

1. Ověření předpokladu normality a homoskedasticity jednotlivých výběrů. V práci je použit standardní Shapiro-Wilkův test normality a Bartlettův test shody rozptylů.

2. Ověření lineární závislosti vysvětlované proměnné $\mathbf{y}$ na předpokládané kovariátě $\mathbf{x}$. Tedy ověření, zda je lineární regresní model významný

3. Ověření, zda jsou lineární modely pro všechny úrovně paralelní. Ověřujeme hypotézu $H_0 : \beta_1 = \cdots = \beta_k$. Pokud tuto hypotézu nebude možné zamítnout, pak můžeme použít analýzu kovariance. Pokud je hypotéza zamítnuta, je zde možnost použít jiné postupy, které již ale nejsou součástí této práce (viz [16])

4. Pokud jsou ověřeny všechny podmínky pro analýzu kovariance, pak lze provést vlastní analýzu kovariance, tj. srovnáváme opravené průměry (na průměrnou hodnoty kovariáty).

Analýza kovariance na konkrétních datech je uvedena v kapitole 9.2.

7 Umělé proměnné v ekonometrických modelech

Ekonometrie je vědní disciplína představující systém poznatků z matematiky, statistiky a ekonomických teorií. Zabývá se matematickým modelováním složitých ekonomických jevů a systémů, analýzou a verifikací těchto modelů, predikcí a optimálním rozhodováním. Využívá zejména statistické a optimalizační metody. Pro metodu umělých proměnných se v ekonometrii najde hned několik uplatnění. Jedná se především o analogie již zmíněného užití umělých proměnných.

7.1 Posun lineárního modelu v prostoru

Uvažujme model, který popisuje poptávku ve městě a na venkově.

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 t + e_t, \quad \text{model poptávky ve městě,} \\ Y_t &= \alpha_0 + \beta_1 t + e_t, \quad \text{model poptávky na vesnici.} \end{aligned}$$

Oba modely popisující poptávku ve městě a na vesnici sloučíme do jednoho modelu pomocí umělé proměnné z,

$$Y_t = \beta_0 + \beta_1 t + \beta_2 z + e_t, \tag{41}$$

kde parametr β_2 můžeme uvažovat jako $\beta_2 = \alpha_0 - \beta_0$ a proměnná z nabývá hodnot:

$$z = \begin{cases} 0 & \text{pro poptávku ve městě,} \\ 1 & \text{pro poptávku na vesnici.} \end{cases}$$

Jak je na první pohled zřejmé, jedná se o modely se stejnou směrnicí a odlišnou hodnotou úrovňové konstanty. Touto situací jsme se již zabývali v kapitole 3.1.1.

7.2 Posun lineárního modelu v čase

Tentokrát uvažujme regresní modely, které se od sebe liší v závislosti na čase.

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 x_t + e_t, \quad t \in T_1, \\ Y_t &= \alpha_0 + \beta_1 x_t + e_t, \quad t \in T_2, \end{aligned}$$

kde T_1 zastupuje období s neuspokojenou poptávkou, T_2 zastupuje období s uspokojenou poptávkou, x_t značí peněžní příjem, Y_t předtavuje spotřebu, β_1 je mezní sklon ke spotřebě a $\alpha_0 > \beta_0$.

Opět se jedná o regresní modely s různou hodnotou úrovňové konstanty a totožnou směrnicí. Po vzoru kapitoly 3.1.1 vytvoříme sloučený model s absolutním členem

$$Y_t = \beta_0 + \beta_1 x_t + \beta_2 z + e_t, \tag{42}$$

kde $\beta_2 = \alpha_0 - \beta_0$ a pro umělou proměnnou z platí

$$z = \begin{cases} 0 & t \in T_1, \\ 1 & t \in T_2. \end{cases}$$

7.3 Změna regresních parametrů v prostoru i čase

Uvažujme separované regresní modely, které se liší jak v hodnotě absolutního členu tak směrnici:

$$\begin{aligned} Y_{1t} &= \alpha_1 + \alpha_2 x_{1t} + e_{1t}, \quad t \in T_1, \\ Y_{2t} &= \beta_1 + \beta_2 x_{2t} + e_{2t}, \quad t \in T_2. \end{aligned}$$

Sloučením obou modelů dostáváme

$$Y_t = \alpha_1 + (\beta_1 - \alpha_1) z_{1t} + \alpha_2 x_t + (\beta_2 - \alpha_2) x_t z_{1t} + e_t,$$

kde pro proměnné z_{1t}, z_{2t} platí

$$z_{1t} = \begin{cases} 0 & t \in T_1, \\ 1 & t \in T_2. \end{cases}$$

7.4 Umělé proměnné a sezónnost

S metodou umělých proměnných se setkáváme často i při analýze časových řad, konkrétně v případě namodelování sezónní složky modelu. Sezónní složka modeluje existenci sezónních vlivů, se kterými se setkáváme téměř vždy při analýze časových řad s periodicitou kratší než jeden rok. Tyto vlivy mohou způsobovat pravidelné výkyvy oproti normálnímu vývoji. Pokud se obdobné vlivy opakují

v intervalu delším než jeden rok, hovoříme o cyklické složce. Sezónní a cyklická složka spolu tvoří periodickou složku.

Uvažujme model

$$Y_{ij} = T_{ij} + S_j + e_{ij}, \;\; i = 1, \ldots, p, \;\; j = 1, \ldots, r, \tag{43}$$

kde p je celkový počet let, r udává počet sezón ($r = 12$ pro měsíční sezónnost, $r = 4$ pro čtvrtletní sezónnost, apod.). T_{ij} značí trend časové řady a S_j je sezónní složka. Po výběru referenční sezóny (vyberme jako referenční sezónu poslední $r-$tou sezónu), lze sezónní složka vyjádřit pomocí $(r-1)$ umělých proměnných z_{kj} vztahem:

$$S_j = \alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_{(r-1)} z_{r-1j}, \; j = 1, \ldots, r, \tag{44}$$

kde pro z_{kj} platí:

$$z_{kj} = \begin{cases} 0 & k \neq j, \\ 1 & k = j. \end{cases}$$

Dále předpokládejme lineární trend, model (43) pak lze přepsat na tvar:

$$Y_{ij} = \beta_0 + \beta_1 t + \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_{(r-1)} z_{r-1} + e_{ij}, \; i = 1, \ldots, p, \;\; j = 1, \ldots, r, \tag{45}$$

pro $t = 1, \ldots, n$, kde n je celkový počet pozorování a platí vztahy $t = (i-1)r + j$ a $n = pr$. Pro model sestavíme designovou matici, kde první sloupec bude tvořen jedničkami pro výpočet odhadu absolutního členu, druhý sloupec bude tvořen hodnotami $1, \ldots, n$ pro proměnnou T a dalších $k = r - 1$ sloupců bude tvořeno dummy umělými proměnnými zastupujícími sezónnost, kde

$z_1 = (\overbrace{1, 0, 0, \ldots, 0}^{1.\text{rok}}, \overbrace{1, 0, 0, \ldots, 0}^{2.\text{rok}}, \ldots, \overbrace{1, 0, 0, \ldots, 0}^{p.\text{rok}})'$, kde 1 je pro 1. sezónu v roce.

$z_2 = (\overbrace{0, 1, 0, \ldots, 0}^{1.\text{rok}}, \overbrace{0, 1, 0, \ldots, 0}^{2.\text{rok}}, \ldots, \overbrace{0, 1, 0, \ldots, 0}^{p.\text{rok}})'$, kde 1 je pro 2. sezónu v roce.

Pokračujeme analogicky až poslední sloupec bude tvaru:

$z_{r-1} = (\overbrace{0, 0, \ldots, 1, 0}^{1.\text{rok}}, \overbrace{0, 0, \ldots, 1, 0}^{2.\text{rok}}, \ldots, \overbrace{0, 0, \ldots, 1, 0}^{p.\text{rok}})'$, kde 1 je pro $(r-1)$. sezónu v roce.

8 Po částech spojitý regresní model

Po částech spojitý regresní model je další typ modelu, kde se můžeme setkat s metodou umělých proměnných.

Při analýze vztahu mezi vysvětlovanou proměnnou Y a vysvětlující proměnnou x můžeme narazit na situaci, kdy pro různé rozsahy x, mohou nastat různé lineární vztahy. V těchto případech samozřejmě nemůže jeden lineární model poskytnout dostatečný popis modelu a nelineární model nemusí být také příliš vhodný. Výsledek nám poskytne právě po částech lineární regrese, která umožňuje spojit více lineárních modelů, které jsou vhodné k výsledným údajům pro různé rozsahy x. Hraniční hodnoty těchto lineárních modelů jsou hodnoty x, kde se mění směrnice lineární funkce.

Hodnoty zlomu mohou, ale nemusí být známy před prováděnou analýzou, proto je třeba tyto zlomy odhadnout. Regresní funkce na zlomu může být přerušovaná, ale model může být sestaven tak, že funkce je spojitá ve všech bodech včetně bodů krajních. Předpokládejme spojitý model s jediným hraničním bodem $x = c$. Model musí být sestaven tak, aby byl v hraničním bodě spojitý. Zapíšeme jej pomocí separovaných regresních modelů:

$$\begin{aligned} Y_i &= \alpha_0 + \alpha_1 x_i + e_i, \text{ pro } x_i \leq c, \\ Y_i &= \beta_0 + \beta_1 x_i + e_i, \text{ pro } x_i > c. \end{aligned} \tag{46}$$

Hodnota regresních rovnic musí být v hraničním bodě stejná. Dosaďme $x_i = c$, pak

$$\alpha_0 + \alpha_1 c = \beta_0 + \beta_1 c$$

Vyjádřeme si nyní parametr β_0

$$\beta_0 = \alpha_0 + c(\alpha_1 - \beta_1)$$

Dosazením do 2. regresní rovnice (46) dostáváme:

$$\begin{aligned} Y_i &= \alpha_0 + c\alpha_1 - c\beta_1 + \beta_1 x_i + e_i = \alpha_0 + c\alpha_1 + \beta_1(x_i - c) + \alpha_1 x_i - \alpha_1 x_i + e_i = \\ &= \alpha_0 + \alpha_1 x_i + (\beta_1 - \alpha_1)(x_i - c) + e_i. \end{aligned}$$

Čili po dosazení a úpravě druhé rovnice, dostáváme separované rovnice tvaru:

$$\begin{aligned} Y_i &= \alpha_0 + \alpha_1 x_i + e_i, && \text{pro } x_i \leq c, \\ Y_i &= \alpha_0 + \alpha_1 x_i + \alpha_2(x_i - c) + e_i, && \text{pro } x_i > c, \end{aligned} \tag{47}$$

kde $\alpha_2 = (\beta_1 - \alpha_1)$. Nyní sestrojme celkový model pomocí umělé proměnné z_i:

$$Y_i = \alpha_0 + \alpha_1 x_i + \alpha_2(x_i - c)z_i + e_i, \tag{48}$$

kde pro umělou proměnnou z_i platí

$$z_i = \begin{cases} 0 & x_i \leq c, \\ 1 & x_i > c. \end{cases}$$

Uvažujme nyní separované modely, které jsou v hraničním bodě posunuty o skok γ. Opět předpokládejme model s jediným hraničním bodem $x_i = c$. Model zapíšeme pomocí separovaných regresních modelů:

$$\begin{aligned} Y_i &= \alpha_0 + \alpha_1 x_i + e_i, \text{ pro } x_i \leq c, \\ Y_i &= \beta_0 + \beta_1 x_i + e_i, \text{ pro } x_i > c. \end{aligned} \tag{49}$$

Víme, že hodnota regresních rovnic je v hraničním bodě posunutá o skok γ. Dosaďme $x_i = c$, pak

$$\alpha_0 + \alpha_1 c + \gamma = \beta_0 + \beta_1 c$$

Vyjádřeme si nyní parametr β_0:

$$\beta_0 = \alpha_0 + \alpha_1 c + \gamma - \beta_1 c$$

Dosazením do 2. regresní rovnice (49) dostáváme:

$$\begin{aligned} Y_i &= \alpha_0 + \alpha_1 c + \gamma - \beta_1 c + \beta_1 x_i + e_i = \\ &= \alpha_0 + \alpha_1 c + \gamma + \beta_1(x_i - c) + \alpha_1 x_i - \alpha_1 x_i + e_i = \\ &= \alpha_0 + \alpha_1 x_i + (\beta_1 - \alpha_1)(x_i - c) + \gamma + e_i \end{aligned}$$

Čili po dosazení a úpravě druhé rovnice, dostáváme separované rovnice tvaru:

$$\begin{aligned} Y_i &= \alpha_0 + \alpha_1 x_i + e_i, && \text{pro } x_i \leq c, \\ Y_i &= \alpha_0 + \alpha_1 x_i + \alpha_2(x_i - c) + \gamma + e_i, && \text{pro } x_i > c, \end{aligned} \tag{50}$$

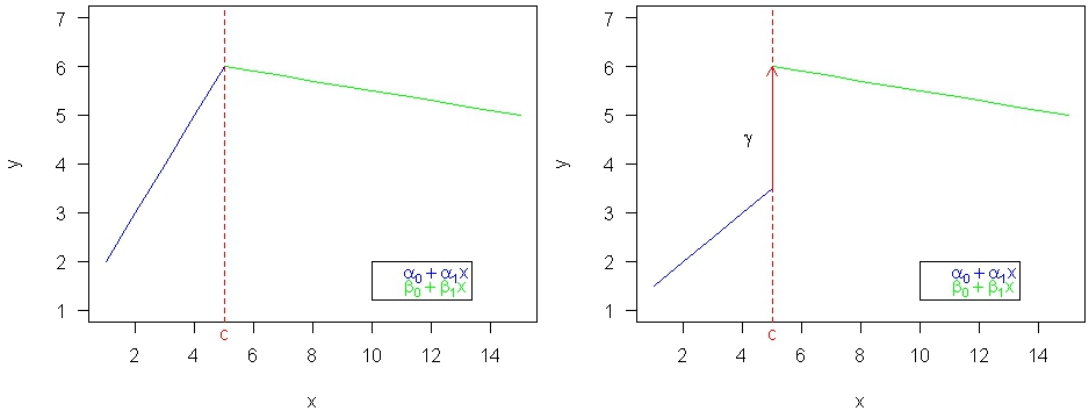
kde $\alpha_2 = (\beta_1 - \alpha_1)$. Nyní sestrojme celkový model pomocí umělé proměnné z.

$$Y_i = \alpha_0 + \alpha_1 x_i + (\alpha_2(x_i - c) + \gamma) z_i + e_i, \tag{51}$$

kde pro umělou proměnnou z_i platí

$$z_i = \begin{cases} 0 & x_i \leq c, \\ 1 & x_i > c. \end{cases}$$

Oba typy po částech spojitého regresního modelu jsou vykresleny na obrázku 6.



Obrázek 6: Po částech spojité regresní modely. Vlevo regresní model spojitý v bodě zlomu c, vpravo regresní model se skokem o velikosti γ v bodě zlomu.

9 Příklady

Tato kapitola má za cíl ukázat vybrané teoretické postupy na konkrétních datech. Všechna následující data jsou spracovávána pomocí statistického programu R. K jednotlivým spracováním jsou přiloženy vybrané části kódu společně s vysvětlivkami některých funkcí.

9.1 ANOVA s kódováním dummy a effect

Data byla stažena z internetových stránek http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/owan/frames/frame.html a reprezentují náklady v tisících dolarech pro uvedení do provozu podnikatelských odvětví. Cílem je ověřit hypotézu o rovnosti středních hodnot pro jednotlivé výběry. Pozorování bylo prováděno pro 5 odvětví: Pizza, Baker and Donuts, Shoe store, Gift shops a Pet stores. Nejprve si upravme data pro jejich další spracování. Načteme data z datového souboru. Jelikož jsou data uspořádána do pěti sloupců, kde každý ze sloupců obsahuje data pro jednu úroveň, vytvoříme vektor, kde budou data z jednolivých odvětví uspořádána za sebou. (Prvních n_1 dat bude náležet první úroveň, dalších n_2 dat bude náležet druhé úrovni, atd.). V neposlední řadě je důležité vytvořit si faktorový vektor s názvy úrovní odpovídající vektoru pozorování, tedy kde prvních n_1 hodnot bude nést název 1. úrovně, dalších n_2 hodnot bude nést název druhé úrovně, atd.

```
Business<-read.table("ANOVA2.dat",header=TRUE,sep=";",
+ col.names=c("Pizza","Baker","Shoe","Gift","Pet"))
k=5                                    #Počet úrovní
náklady=c(as.matrix(Business))         #Vysvětlovaná proměnná Y
ni=length(Business[,1])                #Počet pozorování jedné úrovně
n=length(náklady)                      #Celkový počet pozorování
f=c("Pizza","Baker","Shoe","Gift","Pet")#Pomocný vektor s názvy úrovní
odvětví=gl(k,ni,k*ni,factor(f))        #Vektor úrovní odpovídající vektoru
                                       dat pozorování
```

Funkce `gl(k,ni,n,factor())` slouží k vytvoření faktorového vektoru s názvy úrovní dle zadaných parametrů:

`k`	počet úrovní, které chceme postupně nakopírovat,
`ni`	počet kopií jedné úrovni než přejde na další,
`n`	délka výsledného vektoru,
`factor(f)`	faktorový vektor s názvy úrovní, které má nakopírovat.

Nyní provedeme test normality pomocí Shapiro-Wilkova testu `shapiro.test()`, pro otestování všech úrovní najednou použijeme příkaz `tapply()`. Tento příkaz přiřadí data z vektoru `náklady` napozorovaných počátečních nákladů ke správným výběrům a provede zadanou funkci pro každý výběr (úroveň) zvlášť.

```
> tapply(náklady,odvětví,shapiro.test)
$Pizza
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.95893, p-value = 0.7371

$Baker
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.94296, p-value = 0.5559

$Shoe
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.92403, p-value = 0.3918

$Gift
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.94227, p-value = 0.5786

$Pet
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.92084, p-value = 0.174
$
```

Jak je vidět z výstupu z `R` všechny p-hodnoty (p-value) jsou větší než hladina významnosti $\alpha = 0,05$ z čehož plyne, že nulovou hypotézu o normálním rozdělení jednotlivých výběrů nelze zamítnout. Dále provedeme pomocí Bartlettova testu `bartlett.test(y,x)`, test shody rozptylů jednotlivých výběrů. Opět je p-value$= 0,7768$ větší než hladina významnosti $\alpha = 0,05$ z čehož plyne, že nulovou hypotézu $H_0 : \sigma_1^2 = \cdots = \sigma_5^2$ nelze zamítnout. Můžeme tedy přistoupit

k samotné analýze rozptylu. Nejprve si nadefinujeme designovou matici, kde jsme kategoriální proměnnou kódovali pomocí dummy kódování. Pomocí porovnání příkazem `==` vytvoříme vektory logických hodnot `TRUE, FALSE`, kde hodnota `TRUE` značí, na kterých pozicích vektoru `náklady` se vyskytují data z konkrétních úrovní. Příkaz `as.integer()` převede hodnoty `TRUE, FALSE` na hodnoty 1, 0.

```
#Nadefinování vektorů jedniček a nul dle principu dummy kódování
p_bus=as.integer(odvětví==("Pizza"))
b_bus=as.integer(odvětví==("Baker"))
s_bus=as.integer(odvětví==("Shoe"))
g_bus=as.integer(odvětví==("Gift"))
pet_bus=as.integer(odvětví==("Pet"))
#Designová matice
dummyX=matrix(c(p_bus,b_bus,s_bus,g_bus,pet_bus),ncol=5,nrow=n)
```

Pro vytvoření designové matice za použití effect kódování postupujeme obdobně, posledních n_k míst každého vektoru ale nahradíme hodnotou -1.

```
#Nadefinování vektorů jedniček a nul dle principu kódování effect
p_bus_eff=c(as.integer(odvětví_ef==("Pizza")),rep(-1,ni))
b_bus_eff=c(as.integer(odvětví_ef==("Baker")),rep(-1,ni))
s_bus_eff=c(as.integer(odvětví_ef==("Shoe")),rep(-1,ni))
g_bus_eff=c(as.integer(odvětví_ef==("Gift")),rep(-1,ni))
pet_bus_eff=-1*as.integer(odvětví==("Pet"))
#Designová matice
effectX=matrix(c(p_bus_eff,b_bus_eff,s_bus_eff,g_bus_eff,pet_bus_eff),ncol=5,
nrow=n)
```

Designové matice nejsou tvořeny s absolutním členem. To proto, že příkaz pro výpočet odhadů parametrů a samotné analýzy roztylu si absolutní člen přidává sám a jako referenční úroveň volí vždy poslední k-tou úrovní. Příkaz pro samotnou analýzu roztylu `anova()` nám nenabízí náhled na odhady parametrů β_i jednotlivých úrovní. Ty získáme pomocí příkazu `summary(lm())`.

```
summary(lm(náklady~dummyX))
summary(lm(náklady~effectX))
```

Výsledné odhady parametrů jsou uvedeny v tabulce 5. Z hodnot směrodatných odchylek jednotlivých odhadů vidíme, že přesnější odhady nabízí užití kódování effect. Podívejme se nyní na konkrétní odhady. Odhad absolutního členu v případě

Parametry	Dummy			Effect		
	Odhad	sm. odch.	p-hodnota	Odhad	sm. odch.	p-hodnota
Abs. člen	51,6	8,3	0	77,2	4,4	0
Pizza	31,4	12,4	0,014	5,8	8,4	0,491
Baker	40,5	13,0	0,003	14,9	8,9	0,100
Shoe	20,7	13,4	0,128	−4,9	9,2	0,597
Gift	35,4	13,4	0,011	9,8	9,2	0,293
Pet	0	–	–	−25,6	–	–

Tabulka 5: Srovnání odhadů parametrů pro jednotlivé typy obchodů s užitím kódování dummy a effect, včetně p-hodnot, dílčích t-testů a směrodatných odchylek odhadů.

dummy kódování je roven průměrným nákladům na zahájení provozu obchodu s potřebami pro domácí mazlíčky, kdežto v případě kódování effect je roven průměrným nákladům na zahájení provozu všech uvažovaných typů obchodů. Z odhadu parametru pro pizzerii můžeme vyčíst, že průměrné náklady na uvedení pizzerie do provozu jsou o 34 130 dolarů větší než průměrné náklady na zahájení provozu obchodu pro domácí mazlíčky v případě dummy kódování. V případě kódování effect odhad parametru pro pizzerii udává, že průměrné náklady na uvedení pizzerie do provozu jsou o 5 800 větší než jsou průměrné náklady na uvedení do provozu všech uvažovaných podniků. Pokud se zaměříme na odhady pomocí dummy kódování, pak vidíme, že nejvíce se od průměrné hodnoty nákladů na uvedení do provozu obchodu pro domácí mazlíčky liší průměrné náklady pekařství a to o 40 500 dolarů. Z odhadů při užití kódování effect, lze vyčíst, že nejvíce se od průměrných nákladů na zahájení provozu všech uvažovaných podniků liší právě průměrné náklady na uvedení do provozu obchodu s potřebami pro domácí mazlíčky a to o 25 600 dolarů. Při užití dummy kódování je nevýznamná pouze úroveň pro obchod s botami, kdežto při užití kódování effect jsou pro model nevýznamné všechny úrovně kategoriální proměnné. Hypotézu o rovnosti středních hodnot výběrů otestujeme pomocí příkazu
`anova()`.

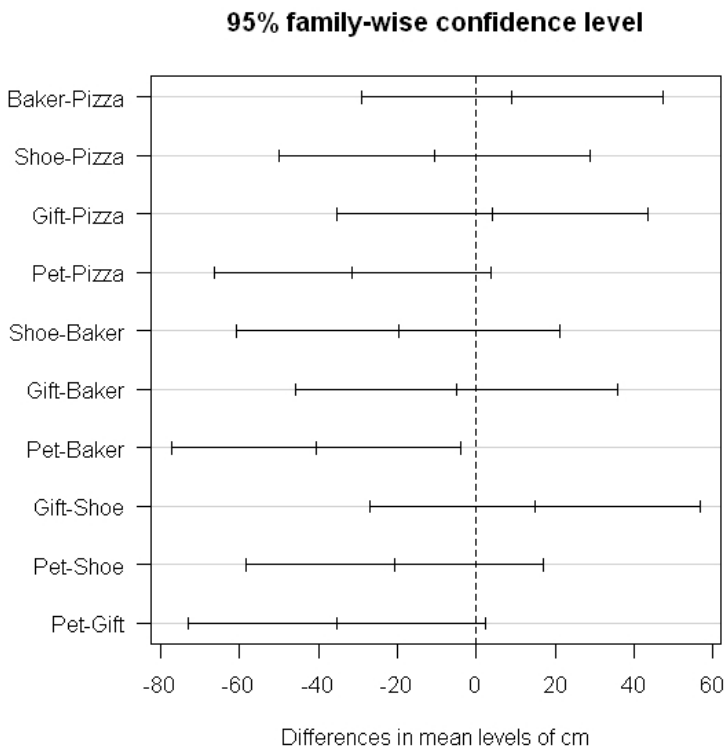
```
anova(lm(náklady~dummyX))
anova(lm(náklady~effectX))
```

U obou příkladů je F-statistika rovna hodnotě 3,2463 s hodnotou p-value=0,02. Jelikož je hodnota p-value menší než zvolená hladina významnosti testu $\alpha = 0,05$, hypotézu H_0 zamítáme. Tedy existuje alespoň jedna střední hodnota, která je

odlišná od střední hodnoty jiného výběru. O které dvojice středních hodnot jde, nám prozradí Tukeyův test `TukeyHSD()`.

```
> TukeyHSD(bus,"odvětví", ordered = FALSE)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = náklady ~ odvětví)
$odvětví
                  diff       lwr       upr     p adj
Baker-Pizza   9.090909 -29.24903 47.430852 0.9622775
Shoe-Pizza  -10.700000 -50.06462 28.664623 0.9390656
Gift-Pizza    4.000000 -35.36462 43.364623 0.9984797
Pet-Pizza   -31.375000 -66.31969  3.569690 0.0982134
Shoe-Baker  -19.790909 -60.68186 21.100037 0.6522120
Gift-Baker   -5.090909 -45.98186 35.800037 0.9966432
Pet-Baker   -40.465909 -77.12143 -3.810387 0.0235175
Gift-Shoe    14.700000 -27.15322 56.553218 0.8584474
Pet-Shoe    -20.675000 -58.40098 17.050981 0.5379762
Pet-Gift    -35.375000 -73.10098  2.350981 0.0761112
```

Z výsledků p-hodnot je vidět že se významně liší pouze střední hodnoty úrovní pro obchod se zvířaty a pekárnu. Přičemž ze záporné hodnoty $-40,46$ můžeme říci, že obchod se zvířaty má mnohem menší náklady na zahájení provozu než pekárna. Z odhadu parametrů při dummy kódování uvedených v tabulce 5 vidíme, že se liší o 40 500 dolarů. Výsledek Tukeyova testu je vykreslen na obrázku 7.



Obrázek 7: Výsledek Tukeyova testu

9.2 ANCOVA

Následující data byla stažena z h internetových stránek http://www.statlab.uni-heidelberg.de/data/ancova/ancovadata.html. Data byla sesbírána na základě pokusu, který se prováděl na šesti komerčních kozích farmách k určení, zda je standardní forma krmného programu dostačující. Ke každému pokusu bylo použito čtyřicet koz. Dvacet zcela náhodně vybraných koz bylo krmeno podle standardního programu, zatímco zbývajících dvacet bylo krmeno častěji. Kozy byly jednotlivě označeny, a zváženy na začátku a na konci celoroční studie. Na každé farmě ve studii byly výsledné zisky tělesné hmotnosti uvedeny spolu s počátečními tělesnými hmotnostmi. V každém pokusu bylo hlavním cílem porovnání váhového přírůstku mezi oběma ošetřeními. Toto srovnání by mohlo být provedeno s použitím ANOVA. Nicméně běžně vypozorovaný biologický jev, kdy lehčí zvířata mívají větší váhový přírůstek než těžší zvířata, nám umožňuje zvýšení přesnosti analýzy. Vzhledem k tomu, že je možné předpokládat výskyt tohoto jevu v obou zkoumaných sku-

pinách, je vhodné použití ANCOVY. Odpovídající srovnání ANCOVY u daných průměrných váhových přírůstků bude tedy obecně citlivější než srovnání ANOVA.

Datový soubor obsahuje proměnné:

- Weightgain - závislá proměnná udávající váhový přírůstek koz zjištěný na konci pokusu.
- Treatment - kategoriální proměnná s úrovněmi standard a intensive, určující typ krmného programu.
- Initialwt - nezávislá proměnná udávající počáteční váhu kozy na začátku pokusu.

Nejprve je třeba ověřit předpoklady normality `shapiro.test()` a homoskedasticity `bartlett.test()`. Shapirův test nám dává p-hodnoty: $0,3558$ pro výsledky pozorování intenzivního krmného programu a $0,7083$ pro výsledky pozorování standardního krmného programu. Čili ani v jednom z provedených testů normality nelze zamítnout nulovou hypotézu. Pro Bartlettův test shody rozptylů dostáváme p-hodnotu$= 0,8594$, tedy opět nelze zamítnout nulovou hypotézu.

Podívejme se jak by vypadalo řešení ANOVY pro model bez zavedeného vlivu kovariáty, v našem případě počáteční váhy koz. Uvažujeme tedy model, kdy zkoumáme vztah váhového přírůstku na základě zvoleného krmného programu.

```
> summary(goat_anova<-aov(goat$Weightgain~goat$Treatment))
               Df Sum Sq Mean Sq F value Pr(>F)
goat$Treatment  1   16.9  16.900    4.13 0.0492 *
Residuals      38  155.5   4.092
```

Jak můžeme vidět z výsledků F-testu analýzy rozptylu, p-hodnota je rovna $0,0492$ což je velmi blízko hladině $\alpha = 0,05$. Z tohoto důvodu je na uvážení, zda nulovou hypotézu o shodnosti středních hodnot váhových přírůstků pro jednotlivé typy krmného programu zamítnout, či nikoli.

Nyní prozkoumejme, zda existuje závislost váhového přírůstku kozy na její počáteční váze bez ohledu na zvolený krmný program.

```
>summary(lm(goat$Weightgain~goat$Initialwt))
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.39581    1.85047   7.780 2.22e-09 ***
goat$Initialwt -0.35403    0.07906  -4.478 6.68e-05 ***

Residual standard error: 1.723 on 38 degrees of freedom
Multiple R-squared:  0.3454,    Adjusted R-squared:  0.3282
F-statistic: 20.05 on 1 and 38 DF,  p-value: 6.681e-05
```

Z výsledků velmi nízkých p-hodnot je zřejmé, že počáteční váha kozy má vliv na její váhový přírůstek. Podívejme se nyní na to, zda závěry z provedené ANOVY a regresní analýzy dokáže opravit užití ANCOVY.

Před provedením ANCOVY samotné je třeba ověřit poslední předpoklad, a to předpoklad o paralelních regresních přímkách pro jednotlivé druhy krmného programu. Uvažujme tedy model s interakcí krmný program a počáteční váha.

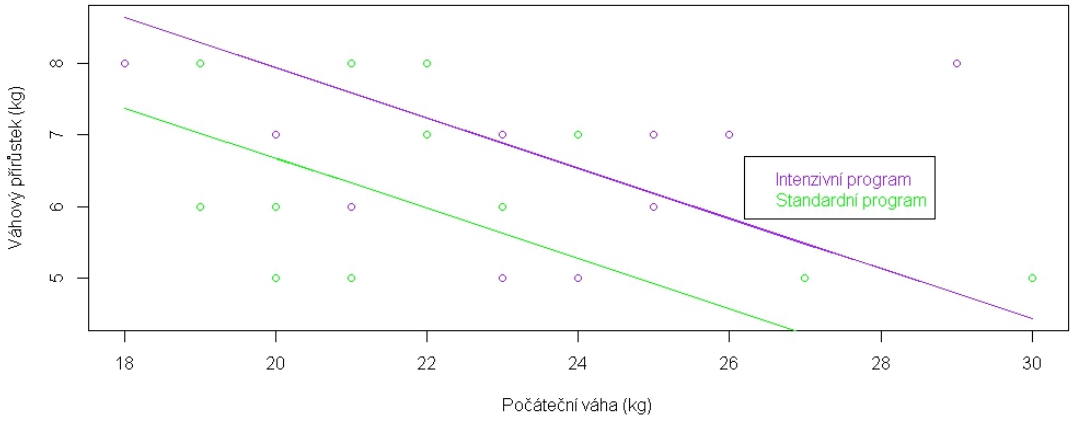
```
> summary(A<-lm(Weightgain~Treatment+Initialwt+Treatment*Initialwt,data=goat))
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  14.37288    2.43047   5.914 9.06e-07 ***
Treatmentstandard            -0.02077    3.52029  -0.006  0.99533
Initialwt                    -0.32567    0.10401  -3.131  0.00345 **
Treatmentstandard:Initialwt -0.05374    0.15040  -0.357  0.72296

Residual standard error: 1.637 on 36 degrees of freedom
Multiple R-squared:  0.4402,    Adjusted R-squared:  0.3935
F-statistic: 9.435 on 3 and 36 DF,  p-value: 9.765e-05
```

Existuje statisticky průkazný vztah mezi váhovým přírůstkem a počáteční váhou kozy pro jednotlivé druhy krmného programu (p-hodnota$= 0,003$) a zároveň není významná interakce mezi jednotlivými druhy krmného programu a počáteční váhou (p-hodnota$= 0,723$). To znamená, že model váhového přírůstku kozy závisícího na počáteční váze kozy je pro oba druhy krmného programu stejný. Model ale není příliš vhodný jelikož nám vyšla závislost přírůstku hmotnosti pouze na počáteční váze. Vyzkoušejme proto otestovat model bez interakcí.

```
> summary(C<-lm(Weightgain~Treatment+Initialwt, data=goat))

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)       14.96661    1.75261   8.540 2.82e-10 ***
Treatmentstandard -1.26486    0.51169  -2.472   0.0182 *
Initialwt         -0.35137    0.07424  -4.733 3.21e-05 ***

Residual standard error: 1.618 on 37 degrees of freedom
Multiple R-squared:  0.4382,    Adjusted R-squared:  0.4078
F-statistic: 14.43 on 2 and 37 DF,  p-value: 2.331e-05
```

Jak je vidět z výsledků model bez interakcí je významný, regresní přímky lze považovat za paralelní. Pokud by byla interakce významná, pak by pro každý krmný program existoval model s jiným sklonem regresní přímky, tj. s různou intenzitou vztahu mezi váhovým přírůstkem a počáteční váhou kozy.



Obrázek 8: Regresní přímka mezi váhovým přírůstkem kozy, její počáteční váhou a druh léčby

Vykresleme si pro představu graf regresních přímek pro model bez inetrakcí (obrázek 8).

Zjistili jsme, že ani poslední předpoklad pro provedení ANCOVY nelze zamítnout. Vypočítejme nejprve opravené hodnoty závislé veličiny „váhový přírůstek“ přepočítané na průměrnou hodnotu kovariáty „počáteční váha“ dle vztahu (39). Hodnotu odhadu parametru $\hat{\beta} = -0,351$ jsme dostali z testu předpokladu paralelních regresních přímek pro model bez interakcí, jako odhad parametru proměnné `Initialwt`.

```
prum_Initialwt=sum(goat$Initialwt)/length(goat$Initialwt)
beta=C$coefficients[3]
Weightgain_anc=goat$Weightgain-beta*(goat$Initialwt-prum_Initialwt)
```

Následně provedeme analýzu rozptylu pro opravené hodnoty váhového přírůstku.

```
> summary(goat_ancova<-aov(Weightgain_anc~goat$Treatment))
               Df Sum Sq Mean Sq F value Pr(>F)
goat$Treatment  1  16.00  15.999   6.277 0.0166 *
Residuals      38  96.86   2.549
```

Jak je vidět z výledků p-hodnoty$= 0,02$, nyní již můžeme hypotézu o shodnosti středních hodnot váhového přírůstku bezpečně zamítnout. Užitím příkazu `TukeyHSD()`, zjišťujeme, že větší váhový přírůstek vykazovali kozy, na kterých byl testován intenzivní krmný program a to o $1,26$ kg.

```
> TukeyHSD(goat_ancova,"goat$Treatment", ordered = FALSE)
  Tukey multiple comparisons of means
    95% family-wise confidence level
                         diff       lwr        upr     p adj
standard-intensive -1.264863 -2.286903 -0.2428231 0.0166365
```

9.3 Regresní model s jednou 4-úrovňovou kategoriální proměnnou a jednou kvantitativní proměnnou

Data byla stažena z internetových stránek http://people.sc.fsu.edu/j̃burkardt-/datasets/regression/x06.txt a reprezentují data získaná na základě pokusu, který měl za cíl zjistit zda závisí délka rybičky na teplotě vody v akváriu. Rybičky byli chovány ve čtyřech nádobách, kde se v každé nádobě udržovala jiná teplota a to 25°C, 27°C, 29°C a 31°C. Po narození byli testované rybičky umístěny do nádrží a přeměřovány zhruba každých 14 dní. K dispozici máme i údaje o věku rybiček. Cílem je zjistit, zda má teplota vody v akváriu vliv na délku rybiček.

Datový soubor obsahuje proměnné:

- Length - závislá proměnná udávající délku rybiček. Bohužel údaj o jednotce měření chybí, předpokládejme, proto, že se jednalo o palce (inch).
- Temperature - kategoriální proměnná s úrovněmi 25°C, 27°C, 29°C a 31°C, určující udržovanou konstantní teplotu v akváriu. Jako referenční úroveň

volíme poslední úroveň 31°C. Další úrovně kódujeme dle následujícího schématu:

- 25°C - kódována 1 pokud je měřená rybička z akvária o teplotě vody 25°C, jinak 0

- 27°C - kódována 1 pokud je měřená rybička z akvária o teplotě vody 27°C, jinak 0

- 29°C - kódována 1 pokud je měřená rybička z akvária o teplotě vody 29°C, jinak 0

- Age - nezávislá proměnná udávající věk rybiček.

Pokusíme se najít nejlepší model popisující danou situaci. Nadefinujme si nejprve designovou matici:

```
x25=as.integer(Fish$Temperature==("25"))
x27=as.integer(Fish$Temperature==("27"))
x29=as.integer(Fish$Temperature==("29"))
Xdum=matrix(c(x25,x27,x29),ncol=3)
```

Nejprve vyzkoušejme model bez interakcí

```
> summary(lm(Length~Age+Xdummy,data=Fish))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  299.161    223.482   1.339 0.188440
Age           26.241      1.845  14.221  < 2e-16 ***
Xdummy1      736.273    229.716   3.205 0.002693 **
Xdummy2      842.636    229.716   3.668 0.000728 ***
Xdummy3      923.182    229.716   4.019 0.000259 ***

Residual standard error: 538.7 on 39 degrees of freedom
Multiple R-squared:  0.851,     Adjusted R-squared:  0.8357
F-statistic: 55.67 on 4 and 39 DF,  p-value: 1.333e-15
```

Z výsledků velmi nízkých p-hodnot pro dílčí t-testy zamítáme hypotézu o nevýznamnosti věku a teplotě vody akavária. Z indexu determinace vidíme, že se jedná o dobře postavený model, kdy věk rybiček a teplota vody v akváriu z 85% vysvětluje jejich délku. Nyní zkusme do modelu přidat interakce mezi věkem a teplotou vody v akváriu.

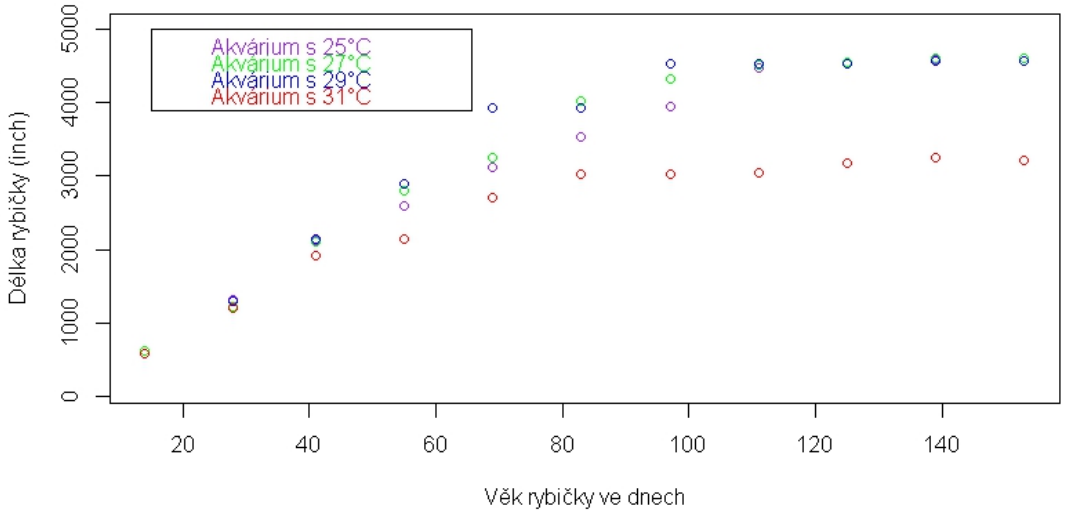
```
> summary(lm(Length~Age+Xdummy+Age*Xdummy,data=Fish))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1005.482    327.022   3.075  0.00401 **
Age           17.749      3.475   5.108 1.08e-05 ***
Xdummy1     -212.803    462.479  -0.460  0.64819
Xdummy2     -128.463    462.479  -0.278  0.78278
Xdummy3       18.075    462.479   0.039  0.96904
Age:Xdummy1   11.410      4.914   2.322  0.02602 *
Age:Xdummy2   11.674      4.914   2.376  0.02297 *
Age:Xdummy3   10.881      4.914   2.214  0.03324 *

Residual standard error: 507.3 on 36 degrees of freedom
Multiple R-squared:  0.878,     Adjusted R-squared:  0.8543
F-statistic: 37.02 on 7 and 36 DF,  p-value: 1.369e-14
```

Poněkud překvapivě existuje mezi teplotou vody v akváriu a věkem rybiček jakýsi vztah, který je pro model velmi významný. Samotná teplota vody v akváriu již pro model tolik významná není. Z indexu determinace vidíme, že se jedná zatím o nejlépe postavený model, kdy věk rybiček, teplota vody v akváriu a jejich vzájemná interakce vysvětluje jejich délku z 87,8%. Z hodnot odhadů kategoriální proměnné teplota vody v akváriu můžeme říci, že délka rybiček umístěných do nejchladnějšího akvária s 25°C je v průměru o 213 palců menší než délka rybiček v nejteplejším akváriu s 31°C, stejně tak rybičky umístěné do akvária s teplotou vody 27°C jsou v průměru o 128 palců menší než rybičky z akvária s 31°C. V průměru nejdelší rybičky jsou ale umístěny v akváriu s teplotou vody 29°C, jejich průměrná délka je větší o 18 palců oproti akváriu s 31°C teplotou vody. Zkusme nyní model, kde vypustíme z modelu s inetrakcemi teplotu vody v akváriu:

```
> summary(lm(Length~Age+Age:Xdummy,data=Fish))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  924.684    157.829   5.859 8.14e-07 ***
Age           18.508      2.158   8.575 1.64e-10 ***
Age:Xdummy1    9.411      2.219   4.242 0.000132 ***
Age:Xdummy2   10.468      2.219   4.718 3.03e-05 ***
Age:Xdummy3   11.051      2.219   4.981 1.33e-05 ***
```

```
Residual standard error: 489.6 on 39 degrees of freedom
Multiple R-squared:  0.8769,    Adjusted R-squared:  0.8643
F-statistic: 69.44 on 4 and 39 DF,  p-value: < 2.2e-16´
```

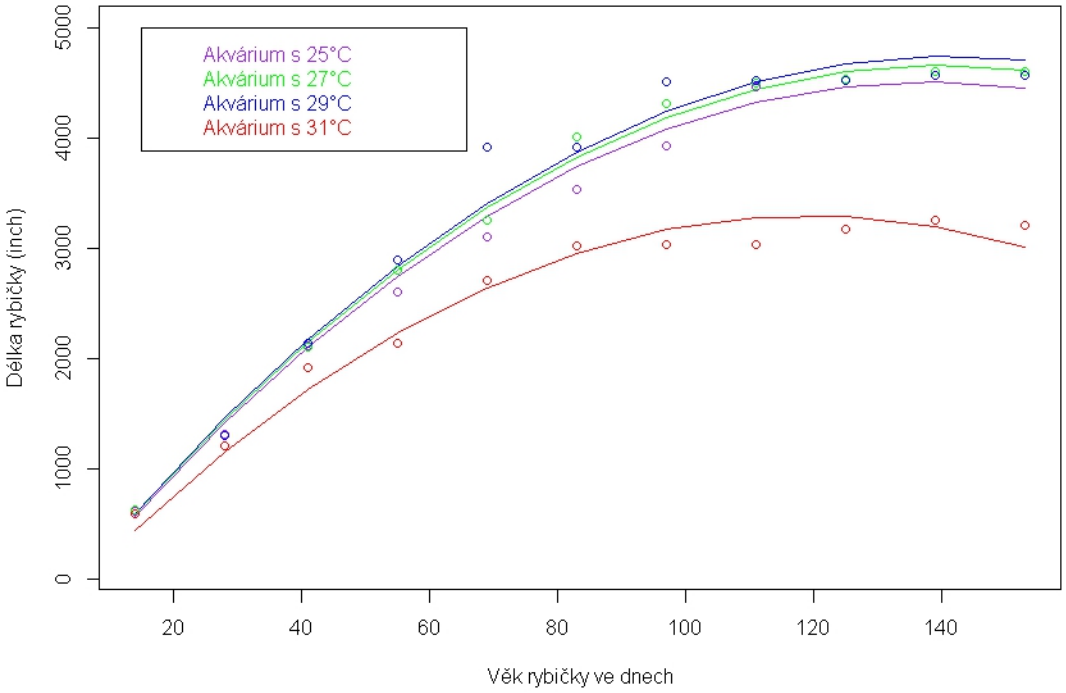
Pro tento model jsou již všechny proměnné velmi významné a index determinace se takřka nezměnil. Můžeme tedy říci, že délka rybiček závisí na teplotě vody s inerakcí na věk. Při pohledu na data (viz obrázek 9) se vkrádá myšlenka na zvo-



Obrázek 9: Vykreslení dat barevně odlišené pro jednotlivá teploty v akváriu

lení kvadratické závislosti na věku a interakcemi mezi věkem rybiček a teplotou vody v akváriu.

```
> summary(reg<-lm(Length~Age+diag(n)%*%Age^2+Age:Xdummy,data=Fish))
...
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -364.91117   87.50430   -4.17  0.00017 ***
Age                 61.31285    2.45514   24.97  < 2e-16 ***
diag(n) %*% Age^2   -0.25642    0.01409  -18.19  < 2e-16 ***
Age:Xdummy1          9.41096    0.72124   13.05  1.3e-15 ***
Age:Xdummy2         10.46787    0.72124   14.51  < 2e-16 ***
Age:Xdummy3         11.05083    0.72124   15.32  < 2e-16 ***

Residual standard error: 159.2 on 38 degrees of freedom
Multiple R-squared:  0.9873,    Adjusted R-squared:  0.9857
F-statistic: 591.8 on 5 and 38 DF,  p-value: < 2.2e-16
```



Obrázek 10: Model regresních přímek závisloti délky rybiček na věku a inetrakcí mezi věkem a teplotou vody v akváriu

Na základě vykreslených regresních přímek (obrázek 10), můžeme říci, že rybičkám nesvědčí pro jejich růst příliš teplá voda 31°C, ale ve vodě jen o dva stupně nižší už se jim co do vzrůstu daří nejlépe ze všech.

Závěr

Tato práce měla za cíl blíže seznámit s pojmem kategoriální proměnná a ukázat, jak s takovouto proměnnou pracovat.

V teoretické části byli ve druhé kapitole podrobně popsány ne příliš známé druhy kódování kategoriální proměnné včetně jejich interpretací, o kterých jsem měla na začátku psaní práce velmi málo zdrojů, které se dané problematiky dotýkaly jen velmi okrajově. Například pouze na základě informací o výsledném tvaru odhadu jednotlivých parametrů bylo potřeba odvodit nejen tvary designových maticí pro různé druhy kódování, ale také princip samotného kódování pro jednotlivé druhy indikátorů. Nakonec se mi ale s pomocí mé vedoucí práce podařilo sepsat vcelku zajímavou a ucelenou kapitolu včetně ukázky srovnání typů kódování na konkrétních datech, které se odborná literatura až na malé výjimky příliš nevěnuje.

Ve třetí kapitole se zabývám samotnou metodou umělých proměnných a další kapitoly jsou již věnovány aplikacím této metody v různých statistických odvětvích. Zde již bylo sice spoustu literatury, ze které se dalo čerpat, ale na základě omezeného rozsahu práce a faktu, že se některé poznatky v různých aplikacích metody opakovali, byli některé kapitoly zkráceny a omezeny pouze na fakta týkající se metody umělých proměnných. Pro podrobnější informace o uvedených statistických analýzách byli v každé kapitole uvedeny odkazy na odbornou literaturu.

V poslední deváté kapitole jsem se již věnovala ukázkám metody umělých proměnných na konkrétních datech pro vybrané analýzy. Zde jsem se nejvíce potýkala s nedostatkem vhodných dat. Existuje sice spousta data setů, ale v drtivé většině jsou velmi malého rozsahu a navíc nesplňují potřebné předpoklady (například předpoklad normality pro analýzu rozptylu a kovariance). Proto bylo potřeba otestovat několik datasetů, než jsem našla nějaký, který by předpoklady splňoval a který jsem dále mohla použít pro další zpracování.

Psaní práce mi poskytlo především nový pohled na způsob kódování kategoriálních proměnných, kdy člověk není vázán pouze velmi známým binárním dummy kódováním, ale má na výběr ze spousty jiných možností, díky nimž

získá nové výsledky a údaje z odhadů parametrů. Stejně tak zajímavý byl pohled „do zákulisí“ tvorby regresních přímek a jejich ovlivnění přidáním kategoriálních proměnných do modelu. Velmi také oceňuji, že díky psaní této práce, jsem objevila nové zajímavé funkce a způsoby programování ve statistickém programu `R`, o kterých jsem neměla do této doby ani tušení.

Seznam tabulek

1 Srovnání typů kódování kategoriální proměnné interpretované na datech udávající počet cizinců v krajích ČR v rozmezí let 1996−2009. 29
2 Srovnání směrodatných odchylek odhadů regresních parametrů při užití různých typů kódování kategoriální proměnné interpretované na datech udávající počet cizinců v krajích ČR v rozmezí let 1996−2009. 30
3 Srovnání p-hodnot dílčích t-testů odhadů regresních parametrů při užití různých typů kódování kategoriální proměnné interpretované na datech udávající počet cizinců v krajích ČR v rozmezí let 1996 − 2009. 31
4 Odhady parametrů při užití vlastího kódování kategoriální proměnné, interpretované na datech udávající počet cizinců v jednotlivých krajích ČR v rozmezí let 1996 − 2009 39
5 Srovnání odhadů parametrů pro jednotlivé typy obchodů s užitím kódování dummy a effect, včetně p-hodnot, dílčích t-testů a směrodatných odchylek odhadů. 72

Seznam obrázků

1 Modely regresních přímek s binární umělou proměnnou označující pohlaví při změně referenční úrovně. 41
2 Model regresních přímek s binární umělou proměnnou a interakcemi. Referenční úrovní je úroveň „žena“. 44
3 Model regresních přímek pro 3-kategoriální proměnnou vzdělání. Referenční úroveň je úroveň „VŠ“. 47
4 Model regresních přímek s konstantní změnou úrovňové konstanty a stejnou směrnicí. Referenční úroveň „0-10“. 56
5 Model regresních přímek s konstantní změnou úrovňové konstanty a směrnice.Referenční úroveň „0-10“. 57
6 Po částech spojité regresní modely. Vlevo regresní model spojitý v bodě zlomu c, vpravo regresní model se skokem o velikosti γ v bodě zlomu. 68
7 Výsledek Tukeyova testu 74
8 Regresní přímka mezi váhovým přírůstkem kozy, její počáteční váhou a druh léčby 77
9 Vykreslení dat barevně odlišené pro jednotlivá teploty v akváriu . 81
10 Model regresních přímek závisloti délky rybiček na věku a inetrakcí mezi věkem a teplotou vody v akváriu 82

Literatura

[1] MONTGOMERY, D.C., E.A. PECK a G.G. VINING. Introduction to linear regression analysis. Hoboken, New Jersey: John Wiley & sons, Inc, 2006. 4. ISBN 978-0-470-54281-1.

[2] PECÁKOVÁ, Iva. Kategoriální vysvětlující proměnné v lineárním modelu. In: Kategoriální vysvětlující proměnné v lineárním modelu [online]. Vysoká škola ekonomická v Praze, 2009. Dostupné z: http://panda.hyperlink.cz/cestapdf/pdf09c1/pecakova.pdf

[3] R Library: Contrast Coding Systems for categorical variables. UCLA: STATISTICAL CONSULTING GROUP. R Library: Contrast Coding Systems for categorical variables [online]. 2011. Dostupné z: http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm

[4] GILLESPIE, Maureen. Categorical Variables in Regression Analyses [online]. Northeastern University, 2010. Dostupné z: https://docs.google.com/viewer?a=v&q=cache:X8cd_q84dF0J:hlplab.files.wordpress.com/2011/02/codingtutorial.pdf+categorical+coding&hl=cs&gl=cz&pid=bl&srcid=ADGEESgOoJSLX3dd3rK2wLoCRd4Z3MBSPABTl33mqLQaBiQeUfqrhnQDgaTKD4AlT1L8ZbxnmMSYVRzygpwGUdm6pQI6giBM5-51eTBgZNqAUOSjkae26uNpTPxHAThWHfDBVmikKIE3&sig=AHIEtbSr1yiTz7MqzKkRoTSbsQELT0o2aQ. Prezentace. Northeastern University.

[5] KUNDEROVÁ, Pavla. Úvod do teorie pravděpodobnosti a matematické statistiky. 2. vyd. Olomouc: Univerzita Palackého, 2004. Skripta (Univerzita Palackého). ISBN 80-244-0843-0.

[6] GUPTA, Resmi. Coding Categorical Variables in Regression Models: Dummy and Effect Coding. Coding Categorical Variables in Regression Models: Dummy and Effect Coding. 2008, č. 72, s. 1-2. Dostupné z: https://docs.google.com/viewer?a=v&q=cache:7lDVqu42tHsJ:www.cscu.cornell.edu/news/statnews/stnews72.pdf+categorical+coding&hl=cs&gl=cz&pid=bl&srcid=ADGEESjrUTFVnIdJ38Yz7EY-hEtFJADrkVKmhP4wDvTXfxNgNv_ENBrYp-uQ65ukSM-g1PrWKa4d-p-B7wmJgSJOp2H03ph4wy6wbmQSGPm1sROv4tyw3p-frxOJZpxv9qyKHC1uyo5a&sig=AHIEtbST_9NSjyCHFuTURnfdjHXpJSEEtQ

[7] PARK, Hun Myoun. Using Dummy Variables in Regressio [online]. Indiana University at Bloomingto, 2005. Dostupné z: http://www.iuj.ac.jp/faculty/kucc625/documents/dummy.pdf. Článek. Indiana University at Bloomingto.

[8] http://dionysus.psych.wisc.edu/lit/Topics/Statistics/SAGESeries/SAGE_RegressionDummyVariables.pdf

[9] COHEN, Ayala. Dummy Variables in Stepwise Regression. Dummy Variables in Stepwise Regression [online]. 1991, Vol. 45, No. 3, s. 226-228. Dostupné z: http://www.jstor.org.proxy.k.utb.cz/stable/2684296?origin=crossref

[10] STARKWEATHER, Dr. Jon. Categorical Variables in Regression: Implementation and Interpretation. In: STARKWEATHER. Categorical Variables in Regression: Implementation and Interpretation [online]. University of North Texas, 2010. Dostupné z: http://www.unt.edu/rss/class/Jon/Benchmarks/CategoricalRegression_JDS_June2010.pdf

[11] BRANNICK, Michael T. Class Materials & Research Website. ANOVA 1: Categorical IVs: Dummy, Effect, & Orthogonal Coding [online]. 2007. Dostupné z: http://luna.cas.usf.edu/m̃brannic/files/regression/anova1.html

[12] Coding of Categorical Predictors and ANCOVA. In: NEWSOM, Jason. Coding of Categorical Predictors and ANCOVA [online]. Portland state university, 2010. Dostupné z: http://www.upa.pdx.edu/IOA/newsom/da2/ho_coding1.pdf

[13] FARAWAY, Julian J. Practical Regression and Anova using R. In: Practical Regression and Anova using R [online]. 2002. Dostupné z: http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

[14] HEBÁK, Petr, Jiří HUSTOPECKÝ a Iva MALÁ. Vícerozměrné statistické metody. 1. vyd. Praha: Informatorium, 2005, 239 s. ISBN 80-733-3036-9.

[15] DRÁPELA CSC., Ing. Karel. Analýza kovariance (ANCOVA) - její předpoklady a využití. In: Analýza kovariance (ANCOVA) - její předpoklady a využití [online]. Brno: Mendelova zemědělská a lesnická univerzita, fakulta lesnická a dřevařská, ústav hospodářské úpravy lesů, 2009. Dostupné z: https://docs.google.com/viewer?a=v&q=cache:HQCxvHp6Gs4J:user.mendelu.cz/drapela/Statisticke_metody/Tutorialy/Analyza_kovariance.doc+anal%C3%BDza+kovariance+Karel+dr%C3%A1pela&hl=en&gl=cz&pid=bl&srcid=ADGEESjicVBa7wuHDQlu3SpFDoDwmdzCbGl_53yli3pUMXrbcWbqn7VH2U0hJ1hIW2g7J6HVsuqyys1RWCkAEhlgMgHXcXENKG2mwHSRApYSUELwty1y6YuhQEzN2KorgCWt2G4u0oor&sig=AHIEtbSN7YzqRfU13l7BaJuGeJfeYZBTKA

[16] D'ALONZO, K. T. The Johnson-Neyman Procedure as an Alternative to ANCOVA. Western Journal of Nursing Research [online]. 2004, 26(7), 804-

812 [cit. 2016-04-26]. DOI: 10.1177/0193945904266733. ISSN 0193-9459. Dostupné z: http://wjn.sagepub.com/cgi/doi/10.1177/0193945904266733

[17] KABACOFF, Robert. R in action: data analysis and graphics with R. 2nd ed. Shelter Island, NY: Manning, 2015. ISBN 16-172-9138-2.

[18] ŠMILAUER, Petr. Moderní regresní metody. Http://regent.jcu.cz/ [online]. 2007, , 1-168 [cit. 2016-04-20]. Dostupné z: http://regent.jcu.cz/MRM.pdf

[19] CLEVELAND, Robert B., William S. CLEVELAND, Jean E MCRAE a Irma TERPENNING. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. Jurnal of Official Statistics. Statistics Sweden, 1990, 1990(1), 3-73.