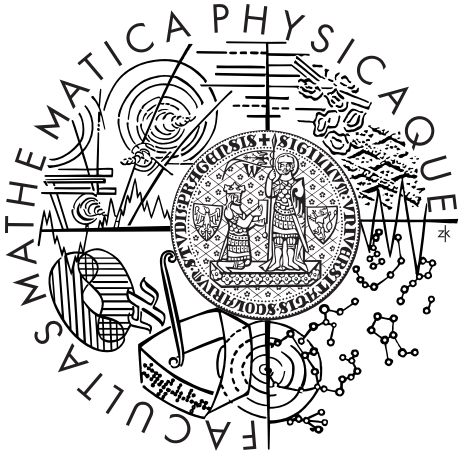
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Pavel Poul

Ekologická regrese

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Karel Zvára, CSc.
Studijní program: Matematika
Studijní obor: PMSE

Praha 2015

Poděkování:
Rád bych na tomto místě poděkoval doc. RNDr. Karlu Zvárovi, CSc. za vedení mé diplomové práce, za veškerou vynaloženou energii a čas, který mi věnoval.

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Ekologická regrese

Autor: Pavel Poul

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Karel Zvára, CSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá problémem při analýze dat, který vzniká agregací veličin do jednotlivých souborů, pro které známe pouze průměry původních veličin a počty jednotek, ze kterých tento soubor vznikl. Jedná se o problém nedostatečného množství informací, který způsobuje nepřesné stanovení vztahů mezi původními veličinami. Tato práce si klade za cíl detailně seznámit čtenáře s dopady agregace dat, představit jednotlivé možnosti přístupu k problému a představit takové modely a předpoklady, které povedou ke správnému stanovení vztahů mezi původními veličinami. Práce je zakončena praktickým použitím jednotlivých přístupů na reálných datech. Výpočty jsou prováděny v software R.

Klíčová slova: agregovaná data, ekologická korelace, skupinový jev, Goodmanova regrese, hierarchický model

Title: Ecological Regression

Author: Pavel Poul

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Karel Zvára, CSc.,
Department of Probability and Mathematical Statistics

Abstract: This paper considers the problem of analyzing data which arises by aggregating the original individual-level variables to separate groups knowing only averages of the original values and numbers of units, that the groups are based on. It is a problem of insufficient amount of information, which causes inaccurate determination of the relations between the original variables. This work aims to make reader familiar with a detailed impact of data aggregation, introduce the possibility of approaching the problem and introduce such models and assumptions that will lead to well established relationships among the original variables. Finally we show the practical use of different approaches to a real data. Calculations are performed in R software.

Keywords: Aggregated data, Ecological correlation, Cross-level effect, Goodman regression, Hierarchical model

Obsah

Úvod

Ekologická regrese je statistická metoda, kdy používáme regresi na agregovaná data (typicky se jedná o průměry v rámci různých geografických oblastí) a odhady takto získané interpretujeme jako vztahy na úrovni neagregovaných jednotek, jednotlivců. Po celou dobu, co se ekologická regrese používá, je znám fakt, že samotná regrese závisí na různých předpokladech, které však v praxi jsou často netestovatelné (viz Robinson [28], Goodman [13], [14] a Duncan [8]), pokud nemáme přistup k neagregovaným datům. Nicméně je zde snaha tyto předpoklady co možná nejlépe zkontrolovat z existujících agregovaných dat. Pro mnoho aplikací je ekologická regrese přirozenou cestou, pokud však nedokážeme předpoklady modelu řádně otestovat, interpretace získaných výsledků je velice složitá. Například Freedman [10] ukázal, že konvenční statistiky pro posuzování modelů, jako je například reziduální součet čtverců, nejsou vhodné pro posuzování kvality ekologických modelů.

Ekologická regresní analýza se často používá pro odhad chování hlasování různých skupin, například etnických skupin, pokud nejsou údaje z průzkumů k dispozici. Typickou aplikací ekologické regrese je analýza voleb mezi dvěma kandidáty, kde cílem je odhadnout podíly hlasů pro jednotlivé kandidáty za jednotlivé skupiny. Pokud neexistuje průzkum, velmi často máme k dispozici jen údaje za oblasti, územní celky, skupiny, pro které známe zastoupení jednotlivých etnik a celkové zisky kandidátů. Tyto skupiny budeme nadále označovat za soubory.

Další pole využití nachází ekologické regrese například v epidemiologii a biostatistice u studií zabývajících se rakovinou a dalšími chronickými onemocněními. Od padesátých let dvacátého století, kdy se tyto studie začaly objevovat, je zde snaha o využití odpovídající komplexní a přesné metodiky. Právě z tohoto důvodu se využívá ekologické regrese i pro tyto úlohy (viz Breslow a Day [5], Morgernstern [23]).

V této práci se budeme věnovat převážně analýze volebních dat pro Spojené státy, kde je standardní systém dvou stran a populace je složena z bělochů a afroameričanů (v určitých oblastech i z hispánců, kterým se však detailně nebudeme věnovat). Zevrubnou analýzu provedl například King [18].

V první kapitole se budeme věnovat ekologické regresi obecných náhodných vektorů a definici základních pojmů s tím spojených. Ukážeme si vztahy mezi parametry pro agregovaná data a parametry původních neagregovaných dat v základním regresním modelu. Probereme důsledky agregace a zanalyzujeme vznik vychýlení v důsledku agregace.

V druhé kapitole si ukážeme jednotlivé přístupy k analýze dat a to pro nula-jedničkové proměnné. Zavedeme si základní model a ukážeme si specifické vztahy, které pro tyto veličiny platí. Specifikujeme některé náležitosti modelu. Tuto část následně zakončíme teoretickým příkladem, který nám lépe ilustruje problémy, kterým musíme při analýze dat čelit.

Dále se seznámíme se základním Goodmanovým modelem, který posléze zobecníme. Zavedeme nutné předpoklady modelu, které zaručí nevychýlení odhadů parametrů modelu. Předvedeme si, jak je možné naložit s porušením předpokladů pomocí další vysvětlující proměnné.

V kapitole 2.7 si představíme neparametrický přístup k analýze dat, který je

znám jako metoda hranic. Následně v kapitole 2.8 si představíme neighbourhood model, který používá rozdílné předpoklady oproti Goodmanově modelu.

Dále si zavedeme Kingův model, který vychází z předpokladu, že parametry modelu se řídí useknutým dvourozměrným normálním rozdělením. Definujeme si jednotlivé předpoklady modelu, projdeme si způsob, jakým je model zkonstruován, a ukážeme si, jak lze v tomto modelu odhadnout jednotlivé parametry.

Nakonec si představíme hierarchické modely, které předpokládají jiná rozdělení regresních koeficientů.

V poslední kapitole budeme prezentovat dosažené výsledky na simulovaných a reálných datech.

1. Důsledky agregace dat

O pojmu ekologická korelace se začalo hovořit zřejmě teprve se článkem Robinson [28], který se zabýval korelačním koeficientem mezi procentem černochů v populaci a podílem negramotných. Zjistil, že korelační koeficient vychází značně rozdílně, pokud použil data za celé Spojené státy, než když použil data z příslušných podílů v devíti oblastech, do kterých pro statistické účely USA dělí.

Slovo ekologický v tomto smyslu je historické označení pro pojem agregovaný. Jak velký dopad může tento problém, v literatuře velmi často uváděný též jako ekologický blud, způsobit, ilustruje následující jednoduchý hypotetický příklad, který uvádí Piantadosi [26].

Mějme n pozorování rozdělených do m skupin o velikosti k, navíc předpokládejme pro jednoduchost, že n/k je celé číslo. Mějme proměnnou X, která nabývá hodnot $k(i-1)+1$ až ki pro i-tou skupinu, kde $i = 1, \ldots, m$. Dále mějme proměnnou Y mající opačnou, sestupnou sadu hodnot v rámci každé skupiny, tj. ki až $k(i-1)+1$ pro i-tou skupinu, kde $i = 1, \ldots, m$. Vše je pro přehled zobrazeno v tabulce 1.1. Je evidentní, že průměry veličin X a Y jsou shodné, jak pro jednotlivé skupiny, tak pro všechna pozorování (tyto průměry budeme nadále značit jako $\bar{x}_{i\bullet}$ a $\bar{Y}_{i\bullet}$ pro jednotlivé skupiny $i = 1, \ldots, m$, resp. $\bar{x}_{\bullet\bullet}$ a $\bar{Y}_{\bullet\bullet}$). To samé platí i pro rozptyly. Koeficienty vnitrotřídní korelace ρ_i jsou rovny -1, pro každou z m skupin. Dále je zřejmé, že koeficient meziskupinové (ekologické) korelace ρ_e je roven 1. Budeme-li však počítat celkovou korelaci ρ mezi X a Y, dostaneme následující

$$\rho = \frac{n^2 + 1 - 2k^2}{n^2 - 1}. \tag{1.1}$$

Tudíž platí, že $\rho = \rho_i$ pouze pro $k = n$. Naopak $\rho = \rho_e$ v případě, že $k = 1$ nebo $n \to \infty$ a zároveň $k \ll \infty$. V případě, že $m = 2, 5, 10$, koeficient ρ vždy přesáhne hodnotu 0,5, resp. 0,92 a 0,98, a to bez ohledu na n.

Tabulka 1.1: Hypotetický příklad ilustrující ekologický klam.

X	Y	Skupina
1	k	1
2	$k-1$	1
3	$k-2$	1
$\vdots$	$\vdots$	$\vdots$
k	1	1
$k+1$	$2k$	2
$k+2$	$2k-1$	2
$k+3$	$2k-2$	2
$\vdots$	$\vdots$	$\vdots$
n	$n-k+1$	m

V případech, jako je tento, v nichž je znatelný skupinový jev, t.j. podmíněná střední hodnota Y při X není stejná pro všechny skupiny, není naším zájmem ani ρ, ani ρ_e, ale průměrná (vážená) vnitro-skupinová korelace ρ_w, která je v tomto příkladě rovna -1, protože všechny jednotlivé ρ_i jsou totožné a to rovny -1.

Označme si klasický lineární model

$$\boldsymbol{Y} = \alpha \mathbf{1}_N + \beta \boldsymbol{X} + \boldsymbol{e}, \qquad \boldsymbol{e} \sim (\mathbf{0}_N, \sigma_1^2 \boldsymbol{I}_N) \tag{1.2}$$

a jeho ekologickou verzi

$$\bar{\boldsymbol{Y}} = \alpha_e \mathbf{1}_m + \beta_e \bar{\boldsymbol{X}} + \boldsymbol{e}_e, \qquad \boldsymbol{e}_e \sim (\mathbf{0}_m, \sigma_2^2 \boldsymbol{I}_m), \tag{1.3}$$

kde jsme si označili

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1k} \\ \vdots \\ Y_{m1} \\ \vdots \\ Y_{mk} \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1k} \\ \vdots \\ x_{m1} \\ \vdots \\ x_{mk} \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1k} \\ \vdots \\ e_{m1} \\ \vdots \\ e_{mk} \end{pmatrix}$$

a

$$\bar{\boldsymbol{Y}} = \begin{pmatrix} \bar{Y}_{1\bullet} \\ \bar{Y}_{2\bullet} \\ \vdots \\ \bar{Y}_{m\bullet} \end{pmatrix}, \qquad \bar{\boldsymbol{X}} = \begin{pmatrix} \bar{x}_{1\bullet} \\ \bar{x}_{2\bullet} \\ \vdots \\ \bar{x}_{m\bullet} \end{pmatrix}, \qquad \bar{\boldsymbol{e}}_e = \begin{pmatrix} e_{e1} \\ e_{e2} \\ \vdots \\ e_{em} \end{pmatrix}.$$

Odhady parametrů v modelech (1.2) a (1.3) metodou nejmenších čtverců (dále jen OLS odhady) jsou $b = \rho$ a $b_e = \rho_e$. Tento příklad krásně demonstruje maximální rozpor mezi ekologickou korelací a korelací celkovou.

1.1 Ekologická regrese

Předpokládejme nyní klasický normální lineární model, jehož data pochází z m disjunktních skupin. Uvažujme posloupnost nezávislých náhodných veličin Y_{it} s konečnými druhými momenty, kde $t = 1, \ldots, n_i$ a $i = 1, \ldots, m$, a posloupnost nenáhodných hodnot x_{ijt}, kde $t = 1, \ldots, n_i$ a $i = 1, \ldots, m$. Index $j = 1, \ldots, h$ značí j-tý regresor. První index vyjadřuje příslušnost k dané skupině, poslední index rozlišuje pozorování uvnitř skupiny. Celkově máme $n = \sum_{i=1}^{m} n_i$ pozorování. V případě ekologické regrese, jak bylo demonstrováno na příkladu, máme místo jednotlivých pozorování k dispozici pouze průměry za jednotlivé skupiny, které značíme $\bar{Y}_{i\bullet}$ a $\bar{x}_{ij\bullet}$. Pro účely zkoumání vlivu agregace dat na odhady parametrů modelu závislosti Y_{it} na x_{ijt} budeme uvažovat *rozšířený model*

$$Y_{it} = \alpha + \sum_{j=1}^{h} \beta_j x_{ijt} + \sum_{j=1}^{h} \gamma_j \bar{x}_{ij\bullet} + e_{it}, \qquad e_{it} \sim \mathsf{N}(0, \sigma^2), \tag{1.4}$$

přičemž náhodné veličiny e_{it} a tedy také Y_{it} jsou nezávislé.

Budeme-li chtít zapsat model (1.4) v maticovém tvaru, musíme upravit značení používané v předchozím příkladu

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{m1} \\ \vdots \\ Y_{mn_m} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} x_{111} & x_{121} & \cdots & x_{1h1} \\ x_{112} & x_{122} & \cdots & x_{1h2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{11n_1} & x_{12n_1} & \cdots & x_{1hn_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m11} & x_{m21} & \cdots & x_{mh1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1n_m} & x_{m2n_m} & \cdots & x_{mhn_m} \end{pmatrix}, \boldsymbol{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \\ \vdots \\ e_{m1} \\ \vdots \\ e_{mn_m} \end{pmatrix}.$$

Vektor $\boldsymbol{Y}$ má n složek, matice $\boldsymbol{X}$ má n řádků a h sloupců. Stejně jako v příkladu zavedeme ještě m-členný vektor průměrů závisle proměnné a také matici průměrů regresorů o m řádcích a h sloupcích jako

$$\bar{\boldsymbol{Y}} = \begin{pmatrix} \bar{Y}_{1\bullet} \\ \bar{Y}_{2\bullet} \\ \vdots \\ \bar{Y}_{m\bullet} \end{pmatrix}, \qquad \bar{\boldsymbol{X}} = \begin{pmatrix} \bar{x}_{11\bullet} & \bar{x}_{12\bullet} & \cdots & \bar{x}_{1h\bullet} \\ \bar{x}_{21\bullet} & \bar{x}_{22\bullet} & \cdots & \bar{x}_{2h\bullet} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_{m1\bullet} & \bar{x}_{m2\bullet} & \cdots & \bar{x}_{mh\bullet} \end{pmatrix}.$$

Zaveďme ještě matici $\boldsymbol{P}$ tvaru

$$\boldsymbol{P} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \vdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \vdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_m} & \mathbf{0}_{n_m} & \vdots & \mathbf{1}_{n_m} \end{pmatrix}.$$

Zřejmě platí

$$(\boldsymbol{P}'\boldsymbol{P})^{-1}\boldsymbol{P}'\boldsymbol{X} = \bar{\boldsymbol{X}}, \qquad (\boldsymbol{P}'\boldsymbol{P})^{-1}\boldsymbol{P}'\boldsymbol{Y} = \bar{\boldsymbol{Y}}.$$

Dále v práci budeme předpokládat, že matice $(\mathbf{1}, \boldsymbol{X})$ i matice $(\mathbf{1}, \bar{\boldsymbol{X}})$ mají lineárně nezávislé sloupce. Speciálně to znamená, že musí platit $m \geq h$, budeme však předpokládat dokonce ostrou nerovnost.

Modelový vztah (1.4) můžeme nyní zapsat jako

$$\boldsymbol{Y} = \mathbf{1}_n\alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}\bar{\boldsymbol{X}}\boldsymbol{\gamma} + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \mathsf{N}(\mathbf{0}, \sigma^2\boldsymbol{I}). \tag{1.5}$$

Zaveďme značení pro projekční matice:

$$\boldsymbol{H} = \boldsymbol{P}(\boldsymbol{P}'\boldsymbol{P})^{-1}\boldsymbol{P}', \qquad \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}.$$

Matice $\boldsymbol{H}$ promítá n-členné vektory do lineárního obalu sloupců matice $\boldsymbol{P}$, matice $\boldsymbol{M}$ promítá n-členné vektory do ortogonálního doplňku zmíněného obalu. Proto platí také následující vztahy

$$\boldsymbol{H}\mathbf{1}_n = \mathbf{1}_n, \qquad \boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}, \qquad \boldsymbol{H}\boldsymbol{M} = \mathbf{0}.$$

Vztah (1.5) lze zapsat jako

$$\boldsymbol{Y} = \mathbf{1}_n\alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{e}.$$

Nyní využijeme skutečnosti, že $\boldsymbol{I} = \boldsymbol{H} + \boldsymbol{M}$ a dostaneme tak následující tvar

$$\boldsymbol{Y} = \mathbf{1}_n\alpha + \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{X}(\boldsymbol{\beta}+\boldsymbol{\gamma}) + \boldsymbol{e}. \tag{1.6}$$

Zajímá nás vztah modelu (1.6) k modelu založeném pouze na průměrech. Vynásobíme-li (1.6) zleva maticí $(\boldsymbol{P}'\boldsymbol{P})^{-1}\boldsymbol{P}'$ a použijeme pravidlo pěti matic (viz Anděl [2], str. 324), dostaneme

$$\bar{\boldsymbol{Y}} = \mathbf{1}_m\alpha + \bar{\boldsymbol{X}}(\boldsymbol{\beta}+\boldsymbol{\gamma}) + \bar{\boldsymbol{e}}, \qquad \bar{\boldsymbol{e}} \sim (\mathbf{0}_m, \sigma^2(\boldsymbol{P}'\boldsymbol{P})^{-1}). \tag{1.7}$$

Klasický model, který je založený jen na průměrech, má tvar

$$\bar{\boldsymbol{Y}} = \mathbf{1}_m\alpha_e + \bar{\boldsymbol{X}}\boldsymbol{\beta}_e + \bar{\boldsymbol{e}}_e, \qquad \bar{\boldsymbol{e}}_e \sim \mathsf{N}(\mathbf{0}_m, \sigma_e^2\boldsymbol{I}). \tag{1.8}$$

Obecně $\sigma_e^2 \neq \sigma^2$, oba parametry však předpokládáme kladné. Odhad $\boldsymbol{b}_e$ vektoru $\boldsymbol{\beta}_e$ z (1.8) metodou nejmenších čtverců je nestranným odhadem vektoru $\boldsymbol{\beta}+\boldsymbol{\gamma}$ z (1.7). Pokud by navíc velikosti skupin $n_1, \ldots, n_m$ byly všechny stejné, byl by tento odhad také nejlepším nestranným lineárním odhadem. Vzhledem k tomu, že velikosti jednotlivých skupin nemusí být apriori stejné, nabízí se další model, který lépe modeluje reziduální složku

$$\bar{\boldsymbol{Y}} = \mathbf{1}_m\alpha_e + \bar{\boldsymbol{X}}\boldsymbol{\beta}_e + \bar{\boldsymbol{e}}_e, \qquad \bar{\boldsymbol{e}}_e \sim \mathsf{N}(\mathbf{0}_m, \sigma^2(\boldsymbol{P}'\boldsymbol{P})^{-1}). \tag{1.9}$$

Zpravidla nás zajímá především závislost na hodnotách samotných regresorů, tedy klasický model

$$\boldsymbol{Y} = \mathbf{1}_n\alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \mathsf{N}(\mathbf{0}, \sigma^2\boldsymbol{I}), \tag{1.10}$$

tedy podmodel modelu (1.5). Použijeme znovu vztah $\boldsymbol{I} = \boldsymbol{H} + \boldsymbol{M}$. Model (1.10) můžeme zapsat jako

$$\boldsymbol{Y} = \mathbf{1}_n\alpha + \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \mathsf{N}(\mathbf{0}, \sigma^2\boldsymbol{I}). \tag{1.11}$$

Okamžitě je vidět, že tento model je identický s modelem (1.6) jen v případě, že $\boldsymbol{\gamma} = \mathbf{0}$.

Pokud máme k dispozici veškerá data, můžeme běžným způsobem provést test modelu (1.11) jako podmodelu (1.4).

Výskyt *skupinového jevu*, tj. vychýlení odhadů v důsledku agregace dat, pozorujeme pomocí parametru $\boldsymbol{\gamma}$. Existence tohoto jevu se však velmi těžko dokazuje, nejsou-li k dispozici veškerá potřebná data. Sám Firebaugh [9] zkoumá možnosti odstranění tohoto jevu a to v závislosti na způsobu vzniku tohoto jevu. V prvém případě se předpokládá závislost j-té vysvětlující proměnné na nějaké jiné vysvětlující proměnné $\boldsymbol{Z} = (z_{11}, \ldots, z_{mn_m})'$, kterou je poté možno přidat do modelu. Autor tuto situaci nazývá *nepravým skupinovým jevem*. Dalším případem je lineární závislost $\bar{x}_{i\bullet}$ se skupinovými průměry druhé veličiny, $\bar{z}_{i\bullet}$. Oproti prvnímu případu, kdy vztah mezi vysvětlovanou a vysvětlující proměnnou je ovlivněn jinou proměnou, která je na úrovni jednotlivých pozorování, v druhém případě je tato dodatečná proměnná na úrovni jednotlivých souborů a nelze ji rozdělit na jednotlivá pozorování uvnitř souborů. Analogicky můžeme říci, že v prvním případě se jedná o mikro proměnné a ve druhém o makro proměnné na úrovni jednotlivých souborů. Třetí uvažovaná možnost je nejzajímavější. Jedná se o nejhůře rozklíčovatelný předpoklad, že $\bar{x}_{i\bullet}$ popisuje celou sadu jevů, tzv. *atmosféru*.

Jedná se o případ, kdy vyšetřujeme například politickou orientaci na základě rasové příslušnosti, kdy voličstvo je rozděleno do správních celků (viz King [18],[20]). Každá správní jednotka se chová specificky a chování jednotlivých ras je ovlivněno i jejich poměrným zastoupením v rámci každého správního celku. Tudíž v tomto případě se skupinový jev nedá odstranit. Velice dobře je rozpracován skupinový jev v [4].

1.2 Odhad parametrů

Upravíme ještě jednou rovnici rozšířeného modelu (1.6). Protože vektor $\mathbf{1}_n$ patří do lineárního obalu sloupců matice $\boldsymbol{P}$, tedy platí $\boldsymbol{H}\mathbf{1} = \mathbf{1}$, připojíme ještě násobek vektoru jedniček k absolutnímu členu. Zavedeme novou matici

$$\boldsymbol{H}_0 = \boldsymbol{H} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}',$$

která promítá n-členný vektor do centrovaného lineárního obalu sloupců matice $\boldsymbol{P}$.

Model (1.6) můžeme tak upravit na

$$\boldsymbol{Y} = \mathbf{1}_n(\alpha + \bar{\boldsymbol{x}}'\boldsymbol{\gamma} + \bar{\boldsymbol{x}}'\boldsymbol{\beta}) + \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}_0\boldsymbol{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) + \boldsymbol{e}. \tag{1.12}$$

Vektor $\bar{\boldsymbol{x}}$ obsahuje průměry jednotlivých regresorů spočítané přes veškerá pozorování, tj. $\bar{\boldsymbol{x}}' = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{X}$.

Střední hodnota náhodného vektoru $\boldsymbol{Y}$ je nyní vyjádřena jako součet tří vektorů, které jsou vůči sobě ortogonální. Proto odhady regresních koeficientů v těchto sčítancích metodou nejmenších čtverců jsou navzájem nezávislé. Matice normální soustavy rovnic je kvazidiagonální s tím, že diagonální bloky odpovídají po řadě koeficientům $\alpha + \bar{\boldsymbol{x}}'\boldsymbol{\gamma} + \bar{\boldsymbol{x}}'\boldsymbol{\beta}$, $\boldsymbol{\beta}$ a $\boldsymbol{\beta} + \boldsymbol{\gamma}$. Prvnímu bloku odpovídá $\mathbf{1}'\mathbf{1} = n$ a odpovídající člen na pravé straně je $\mathbf{1}'\boldsymbol{Y} = n\bar{Y}_{\bullet\bullet}$. Zajímavější je diagonální blok odpovídající koeficientu $\boldsymbol{\beta}$:

$$\boldsymbol{W}^{xx} = \boldsymbol{X}'\boldsymbol{M}\boldsymbol{X}.$$

Na pozici uv této matice je součet součinů

$$w_{uv}^{xx} = \sum_{i=1}^{m}\sum_{t=1}^{n_i}(x_{iut} - \bar{x}_{iu\bullet})(x_{ivt} - \bar{x}_{iv\bullet}),$$

který vyjadřuje variabilitu regresorů uvnitř skupin. Podobně vyjádříme vztah regresorů a vysvětlované veličiny uvnitř skupin pomocí vektoru $\boldsymbol{w}^{xy}$, jehož u-tá složka je

$$w_u^{xy} = \sum_{i=1}^{m}\sum_{t=1}^{n_i}(x_{iut} - \bar{x}_{iu\bullet})(Y_{it} - \bar{Y}_{i\bullet}).$$

Odhad vektoru $\boldsymbol{\beta}$ je dán vztahem

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{b}}_w = (\boldsymbol{W}^{xx})^{-1}\boldsymbol{w}^{xy}. \tag{1.13}$$

Index w u značení odhadu používáme jakožto explicitní označení způsobu výpočtu, podle anglického slova *within-group*.

Podobně zavedeme matici $\boldsymbol{B}^{xx}$ a vektor $\boldsymbol{b}^{xy}$, které vyjadřují variabilitu mezi průměrnými hodnotami:

$$\boldsymbol{B}^{xx} = \boldsymbol{X}'\boldsymbol{H}_0\boldsymbol{X}, \qquad \boldsymbol{b}^{xy} = \boldsymbol{X}'\boldsymbol{H}_0\boldsymbol{Y}.$$

Matice $\boldsymbol{B}^{xx}$ má na pozici uv

$$b^{xx}_{uv} = \sum_{i=1}^{m} n_i(\bar{x}_{iu\bullet} - \bar{x}_{\bullet u\bullet})(\bar{x}_{iv\bullet} - \bar{x}_{\bullet v\bullet}),$$

podobně vektor $\boldsymbol{b}^{xy}$ má u-tou složku danou výrazem

$$b^{xy}_{u} = \sum_{i=1}^{m} n_i(\bar{x}_{iu\bullet} - \bar{x}_{\bullet u\bullet})(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}).$$

Odhadem vektoru $\boldsymbol{\beta} + \boldsymbol{\gamma}$ je podobně jako výše vektor

$$\widehat{\boldsymbol{\beta} + \boldsymbol{\gamma}} = \hat{\boldsymbol{b}}_b = (\boldsymbol{B}^{xx})^{-1}\boldsymbol{b}^{xy}.$$

Odtud je odhadem vektoru $\boldsymbol{\gamma}$ metodou nejmenších čtverců vektor

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= \widehat{\boldsymbol{\beta} + \boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{b}}_b - \hat{\boldsymbol{b}}_w \\ &= (\boldsymbol{B}^{xx})^{-1}\boldsymbol{b}^{xy} - (\boldsymbol{W}^{xx})^{-1}\boldsymbol{w}^{xy}. \end{aligned} \tag{1.14}$$

Z kvazidiagonálního charakteru matice soustavy normálních rovnic plyne, že odhady uvedené v (1.13) a (1.14) jsou nezávislé, takže varianční matice odhadu $\hat{\boldsymbol{\gamma}}$ je rovna

$$\operatorname{var}(\hat{\boldsymbol{\gamma}}) = \sigma^2\left((\boldsymbol{W}^{xx})^{-1} + (\boldsymbol{B}^{xx})^{-1}\right),$$

což lze použít při testování podmodelu (1.11).

Vraťme se nyní zpět ke klasickému modelu (1.10) s maticí $\boldsymbol{X}$. Když centrujeme sloupce této matice, dostaneme

$$\boldsymbol{X}_0 = \boldsymbol{X} - \boldsymbol{1}\bar{\boldsymbol{x}}',$$

můžeme dále klasický model (1.10) přepsat na

$$\boldsymbol{Y} = \boldsymbol{1}(\alpha + \bar{\boldsymbol{x}}'\boldsymbol{\beta}) + \boldsymbol{M}\boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \mathsf{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Řešení soustavy normálních rovnic tohoto modelu můžeme vyjádřit pomocí matice $\boldsymbol{T}^{xx}$ a vektor $\boldsymbol{t}^{xy}$, jejichž uv-tý a u-tý prvek jsou dány vztahy

$$t^{xx}_{uv} = \sum_{i=1}^{m}\sum_{t=1}^{n_i}(x_{iut} - \bar{x}_{\bullet u\bullet})(x_{ivt} - \bar{x}_{\bullet v\bullet}),$$

$$t^{xy}_{u} = \sum_{i=1}^{m}\sum_{t=1}^{n_i}(x_{iut} - \bar{x}_{\bullet u\bullet})(Y_{it} - \bar{Y}_{\bullet\bullet}).$$

Vzhledem k tomu, že platí $\boldsymbol{T}^{xx} = \boldsymbol{B}^{xx} + \boldsymbol{W}^{xx}$ a $\boldsymbol{t}^{xy} = \boldsymbol{b}^{xy} + \boldsymbol{w}^{xy}$, můžeme výsledný odhad $\hat{\boldsymbol{\beta}}$ postupně upravit na

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{T}^{xx})^{-1}\boldsymbol{t}^{xy} \\ &= (\boldsymbol{T}^{xx})^{-1}\boldsymbol{B}^{xx}(\boldsymbol{B}^{xx})^{-1}\boldsymbol{b}^{xy} + (\boldsymbol{T}^{xx})^{-1}\boldsymbol{W}^{xx}(\boldsymbol{W}^{xx})^{-1}\boldsymbol{w}^{xy} \\ &= (\boldsymbol{B}^{xx} + \boldsymbol{W}^{xx})^{-1}\boldsymbol{B}^{xx}\boldsymbol{b}_b + (\boldsymbol{B}^{xx} + \boldsymbol{W}^{xx})^{-1}\boldsymbol{W}^{xx}\boldsymbol{b}_w. \qquad (1.15)\end{aligned}$$

Odhad je tedy váženým průměrem odhadů (po složkách) založených na variabilitě mezi jednotlivými soubory a variabilitě uvnitř souborů. V literatuře [23], či [24] se můžeme běžně setkat s rozdělením vychýlení v důsledku existence ekologického jevu na dvě části, na tzv. *vychýlení v důsledku agregace*, tedy rozdíl mezi $\hat{\boldsymbol{\beta}}$ a $\boldsymbol{b}_b$, a na *vychýlení v důsledku specifikace*, tedy rozdíl mezi $\hat{\boldsymbol{\beta}}$ a $\boldsymbol{b}_w$. Tímto nám však nevznikají dva separátní problémy k řešení. Z rovnice (1.15) je po přepsání patrné, že existuje-li jeden typ vychýlení, existuje i druhý

$$\boldsymbol{B}^{xx}(\boldsymbol{b}_b - \hat{\boldsymbol{\beta}}) = \boldsymbol{W}^{xx}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_w).$$

Pokud budeme pracovat s modelem vážených průměrů (1.9), bude odhadem vektoru $\boldsymbol{\beta}$ již definovaný vektor $\boldsymbol{b}_b$

1.3 Hierarchický model

Další možný pohled na ekologickou regresi je možný pomocí analýzy rozptylu (viz Anděl [2]), respektive problém budeme modelovat pomocí dvoustupňového hierarchického modelu, což je podrobně popsáno v [27]. Pro jednoduchost budeme pracovat s jediným regresorem.

Nechť je i-tý soubor charakterizován dvojicí náhodných veličin

$$\begin{pmatrix}\xi_i \\ \tau_i\end{pmatrix} \sim \left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\omega_1^2 & \vartheta\omega_1\omega_2 \\ \vartheta\omega_1\omega_2 & \omega_2^2\end{pmatrix}\right), \qquad i = 1, 2, \ldots, m.$$

Náhodné veličiny X_{it}, Y_{it} vyhovují vztahům

$$X_{it} = \xi_i + \epsilon_{it}, \qquad X_{it} = \xi_i + \varphi_{it},$$

kde

$$\begin{pmatrix}\epsilon_{it} \\ \varphi_{it}\end{pmatrix} \sim \left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right), \qquad t = 1, 2, \ldots, n_i, i = 1, 2, \ldots, m.$$

Předpokládáme, že náhodné vektory $(\epsilon_{it}, \varphi_{it})'$ a $(\xi_i, \tau_i)'$ jsou pro $t = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, m$ nezávislé. Z předpokladů dostaneme

$$\begin{pmatrix}X_{it} \\ Y_{it}\end{pmatrix} \sim \left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\omega_1^2 + \sigma_1^2 & \vartheta\omega_1\omega_2 + \rho\sigma_1\sigma_2 \\ \vartheta\omega_1\omega_2 + \rho\sigma_1\sigma_2 & \omega_2^2 + \sigma_2^2\end{pmatrix}\right).$$

Regresní funkce je podmíněná střední hodnota, v našem případě $\mathsf{E}(Y_{it}|X_{it} = x_{it})$. Nás zajímá především citlivost střední hodnoty na změnu nezávisle proměnné, tedy směrnice regresní přímky. Podle Anděl [2], Věta 4.12, je tato směrnice dána vztahem

$$\begin{aligned}\beta_{y|x} &= \frac{\vartheta\omega_1\omega_2 + \rho\sigma_1\sigma_2}{\omega_1^2 + \sigma_1^2} \\ &= \vartheta\frac{\omega_2}{\omega_1}\frac{\omega_1^2}{\omega_1^2 + \sigma_1^2} + \rho\frac{\sigma_2}{\sigma_1}\frac{\sigma_1^2}{\omega_1^2 + \sigma_1^2} \\ &= \beta_b\frac{\omega_1^2}{\omega_1^2 + \sigma_1^2} + \beta_w\frac{\sigma_1^2}{\omega_1^2 + \sigma_1^2}. \qquad (1.16)\end{aligned}$$

Při poslední úpravě jsme zavedli označení β_b pro koeficient popisující závislost na úrovni skupin a β_w pro koeficient popisující závislost uvnitř skupin. Porovnáním zlomků v tomto vyjádření (1.16) docházíme k závěru, že koeficient $\beta_{y|x}$ je váženým průměrem koeficientů β_b a β_w. Jedná se o bezprostřední obdobu vztahu (1.15).

2. Nula-jedničkové proměnné

Tato kapitola vychází z Achen a Shively [1] a Gelman a kol. [12]. Zatím jsme se zaobírali ekologickou regresí pro veličiny X a Y s normálním rozdělením. Nyní však přejdeme k binárním datům, tedy k případům, kdy veličiny X a Y nabývají hodnot $\{0, 1\}$. Takováto data jsou primárním zdrojem při různých výzkumných činnostech, které pracují s ekologickou složkou. Začneme upřesněním *individual-level* modelu, tj. modelu, kde nám vstupují neagregovaná data za jednotlivce. Nechť Y_{it} je vysvětlovaná umělá proměnná, kde první index vyjadřuje příslušnost k danému souboru a druhý určuje danou jednotku uvnitř souboru. Nechť X_{it} je vysvětlující umělá proměnná se stejnou indexací. Realizaci této proměnné budeme značit x_{it}.

Jako příklad si lze uvést zkoumání závislosti výskytu jevu A na výskytu jevu B, kdy $Y = 1$ značí stav, kdy jev A nastal a podobně $X = 1$ v případě, že nastal jev B, nula jinak. Za jev A můžeme považovat například volbu demokratů v USA při prezidentských volbách. Za jev B v tomto případě uvažujeme černou barvu pleti (viz [18]) nebo například volbu demokratů v předchozích volbách (viz [1]).

Dále pro daný soubor i zavádíme β_i^1 jakožto pravděpodobnost $\mathsf{P}(Y_{it} = 1|X_{it} = 1)$, podobně pak $\beta_i^0 = \mathsf{P}(Y_{it} = 1|X_{it} = 0)$. V literatuře se s nimi setkáme jako s *transition rates* (viz [1]) nebo jako *quantities of interest* (viz [18]). V této práci o nich budeme mluvit jako o pravděpodobnostech, parametrech modelu či parametrech zájmu. Dle potřeby budeme používat značení $\boldsymbol{\beta}_i = (\beta_i^1, \beta_i^0)'$. Tyto pravděpodobnosti se mohou lišit podle souboru, uvnitř souborů je považujeme za konstantní. Někdy do modelu vstupuje navíc prostorová (*spatial*) autokorelace, která analýzu dat komplikuje, viz například Papalia [25].

Regresní funkci můžeme vyjádřit pomocí podmíněných pravděpodobností náhodné veličiny Y:

$$\mathsf{E}(Y_{it}|X_{it} = x_{it}) = \beta_i^1 x_{it} + \beta_i^0(1 - x_{it}), \tag{2.1}$$

kde $t = 1, \ldots, n_i$, $i = 1, \ldots, m$.[1]

Vzhledem k charakteru jednotlivých veličin můžeme po úpravě vyjádřit podmíněný rozptyl Y_{it} jako

$$\mathrm{var}(Y_{it}|X_{it} = x_{it}) = \beta_i^1(1-\beta_i^1)x_{it}+\beta_i^0(1-\beta_i^0)(1-x_{it}), t = 1, \ldots, n_i, i = 1, \ldots, m. \tag{2.2}$$

Kdybychom pro zvolené i znali všechny hodnoty Y_{it}, X_{it} a platilo by $n_i > 1$ pro každé i, mohli bychom parametry $\boldsymbol{\beta}_i$ odhadnout metodou nejmenších čtverců. Vztah

$$y = \beta_i^0 + (\beta_i^1 - \beta_i^0)x, \qquad i = 1, \ldots, m, \tag{2.3}$$

určuje rovnici regresní přímky, která prochází těžištěm, tedy bodem $(\bar{Y}_{i\bullet}, \bar{x}_{i\bullet})$, takže odhad $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_i^1, \hat{\beta}_i^0)'$ vektoru $\boldsymbol{\beta}_i$ vyhovuje vztahu

$$\bar{Y}_{i\bullet} = \hat{\beta}_i^1 \bar{x}_{i\bullet} + \hat{\beta}_i^0(1 - \bar{x}_{i\bullet}), \qquad i = 1, \ldots, m. \tag{2.4}$$

[1] Nadále budeme považovat x_{it} za pevné.

V ekologické regresi však neznáme všech n_i pozorování uvnitř souborů. K dispozici máme pouze ony průměry $(\bar{Y}_{i\bullet}, \bar{x}_{i\bullet})$. Víme jen, že odhad vektoru $\boldsymbol{\beta}_i$ splňuje (2.4), to ale k jednoznačnému určení nestačí. Více o tomto vztahu je v kapitole 2.7.

Vzhledem k charakteru dat, se velice často setkáváme s tabulkovým zápisem 2.1. Jednotlivé pravděpodobnosti β_i^1 a β_i^0 zapisujeme dovnitř tabulky. Do spodního řádku zapisujeme vysvětlovanou proměnnou a do posledního sloupce známou vysvětlující proměnnou.

Tabulka 2.1: Značení souborových průměrů a parametrů pro i-tý soubor.

	$Y=1$	$Y=0$	
$X=1$	β_i^1	$1-\beta_i^1$	$\bar{x}_{i\bullet}$
$X=0$	β_i^0	$1-\beta_i^0$	$1-\bar{x}_{i\bullet}$
	$\bar{Y}_{i\bullet}$	$1-\bar{Y}_{i\bullet}$	1

Někdy se můžeme setkat se zápisem pomocí marginálních četností, namísto relativních četností pro jednotlivé soubory. Tyto veličiny budeme značit $Y_{i\bullet}$ a $x_{i\bullet}$. Tento zápis je uveden v tabulce 2.2.

Tabulka 2.2: Značení souborových (marginálních) četností a parametrů pro i-tý soubor.

	$Y=1$	$Y=0$	
$X=1$	n_i^1	.	$x_{i\bullet}$
$X=0$	n_i^0	.	$n_i - x_{i\bullet}$
	$Y_{i\bullet}$	$n_i - Y_{i\bullet}$	n_i

Literatura se zpravidla zabývá odhadem parametrů, které by měly charakterizovat celou populaci a to bez ohledu na jednotlivé skupiny. K tomu se zavádějí parametry β^1 a β^0 jakožto vážené průměry odpovídajících skupinových parametrů, přičemž váhami jsou počty hodnot 1 a 0 nezávislé proměnné x,

$$\beta^1 = \frac{\sum_{i=1}^m n_i \bar{x}_{i\bullet} \beta_i^1}{\sum_{i=1}^m n_i \bar{x}_{i\bullet}}, \tag{2.5}$$

$$\beta^0 = \frac{\sum_{i=1}^m n_i (1-\bar{x}_{i\bullet}) \beta_i^0}{\sum_{i=1}^m n_i (1-\bar{x}_{i\bullet})} \tag{2.6}$$

za předpokladu, že bereme β_i^1 a β_i^0 jako neznámé konstantní parametry.

V případě, že β_i^1 a β_i^0 modelujeme jako náhodné veličiny, je nutné (2.5) a (2.6) pozměnit a aplikovat střední hodnotu na parametry β_i^1 a β_i^0, získáváme tak následující předpis

$$\beta^1 = \frac{\sum_{i=1}^m n_i \bar{x}_{i\bullet} \mathsf{E}(\beta_i^1)}{\sum_{i=1}^m n_i \bar{x}_{i\bullet}}, \tag{2.7}$$

$$\beta^0 = \frac{\sum_{i=1}^m n_i (1-\bar{x}_{i\bullet}) \mathsf{E}(\beta_i^0)}{\sum_{i=1}^m n_i (1-\bar{x}_{i\bullet})}. \tag{2.8}$$

Střední hodnota $\mathsf{E}(\beta_i^j)$, pro $j = 0, 1$, je konečná, neboť nosič distribuční funkce, ze kterého jsou β_i^0 a β_i^0 brány, je omezený na množině $[0, 1]^2$. Tyto parametry budeme označovat jako *populační pravděpodobnosti*. Slovo populační zde používáme, abychom odlišili fakt, že se jedná o parametry vztahující se ke všem souborům, celé populaci, a ne pouze k jednomu souboru.

Zkusme interpretovat právě uvedené definice. Kdybychom na moment předpokládali, že náhodné jsou hodnoty regresoru X i příslušnost k jednotlivým skupinám (s pravděpodobnostmi π_i), mohli bychom psát

$$\begin{aligned}
\mathsf{P}(Y = 1|X = 1) &= \frac{\mathsf{P}(Y = 1, X = 1)}{\mathsf{P}(X = 1)} \\
&= \frac{\sum_i \mathsf{P}(Y = 1, X = 1|i)\pi_i}{\sum_i \mathsf{P}(X = 1|i)\pi_i} \\
&= \frac{\sum_i \mathsf{P}(Y = 1|X = 1, i)\,\mathsf{P}(X = 1|i)\pi_i}{\sum_i \mathsf{P}(X = 1|i)\pi_i}.
\end{aligned}$$

Když pravděpodobnosti $\mathsf{P}(X = 1|i)$ odhadujeme pomocí $\bar{x}_{i\bullet}$ a pravděpodobnost π_i nahradíme relativní velikostí souboru n_i/n, dostaneme β^1 z (2.5). Podobně bychom dostali také β^0.

Velmi často se budeme omezovat na případ, že $\bar{x}_{\bullet\bullet} < \frac{1}{2}$, tedy, že $X = 1$ bude minoritní jev. V této souvislosti hovoříme o parametru β^1 jako o *minority cohesion*, kdežto β^0 chápeme jako *majority defection*.

Definice (2.5) a (2.6) nejsou jediné, se kterými se můžeme v literatuře setkat. Jako váhy můžeme vzít n_i. Pak dostáváme alternativní celosouborové pravděpodobnosti[2] jako

$$\mathcal{B}^j = \sum_{i=1}^{m} \frac{n_i \beta_i^j}{n} \qquad j = 0, 1.$$

Případně můžeme vzít nevážený průměr

$$\mathfrak{B}^j = \sum_{i=1}^{m} \frac{\beta_i^j}{m} \qquad j = 0, 1. \tag{2.9}$$

Rozdíl mezi P(Y = 1|X = 1) a P(Y = 1|X = 0) značíme $p_i = \beta_i^1 - \beta_i^0$ a nazýváme ho *polarizací* pro každý soubor i. Obdobně můžeme i zde definovat celkovou polarizaci jako $p = \beta^1 - \beta^0$.

Dále nazveme $h_i = \bar{x}_{i\bullet}(1 - \bar{x}_{i\bullet})$ *heterogenitou* v rámci každého souboru.

Jako poslední si připomeneme definici vychýlení odhadu $\hat{\beta}^1$ parametru β^1 jakožto

$$\mathrm{Bias}(\hat{\beta}^1) = \mathsf{E}(\hat{\beta}^1) - \beta^1,$$

obdobně definujeme $\mathrm{Bias}(\hat{\beta}^0)$ a $\mathrm{Bias}(\hat{p})$.

Nyní se můžeme pustit do předmětu ekologické regrese, která se snaží parametry v modelu (2.4) stanovit přidáním různých podmínek a využitím zákonitostí. Tou je kupříkladu fakt, že oba parametry, jak β_i^0 tak β_i^1, musí ležet v intervalu

[2] Pravděpodobnosti $\mathcal{B}^j$ a $\mathfrak{B}^j$, pro $j = 0, 1$ nevyjadřují skutečné pravděpodobnosti jevu $(Y = 1|X = 1)$ v celé populaci, ale slouží jako alternativní charakteristiky vztahu Y a X.

$[0,1]$ pro každé $i = 1, \ldots, m$. Rychlý přehled několika metod je možné nalézt například v [30].

Příklad 2.1. Na závěr této kapitoly si ještě předvedeme důsledek agregace dat, kterým je nejednoznačnost v odhadech parametrů β_i^0 a β_i^1 v modelu (2.4). Uvažujme, že x_i jsou realizace náhodné veličiny, která má beta rozdělení $\mathrm{B}(2,5)$. Nechť β_i^0 a β_i^1 jsou realizace z dvourozměrného normálního rozdělení $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde

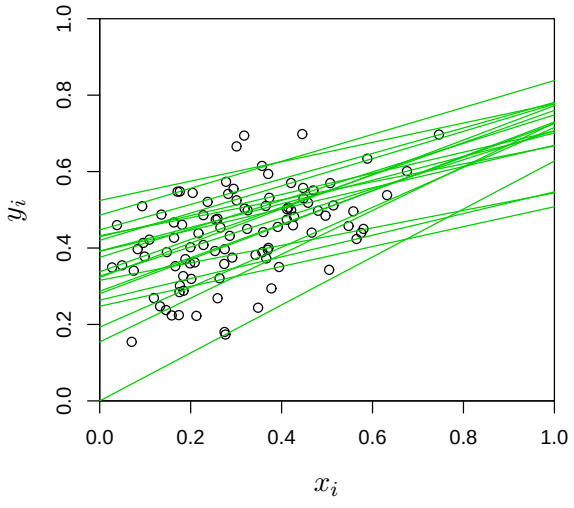
$$\boldsymbol{\mu} = \begin{pmatrix} 0{,}7 \\ 0{,}3 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 0{,}010 & 0{,}005 \\ 0{,}005 & 0{,}020 \end{pmatrix},$$

přičemž hodnoty mimo interval $[0,1]$ upravíme na nejbližší hodnotu tohoto intervalu. K hodnotám x_i, β_i^1 a β_i^0 dopočteme y_i jako

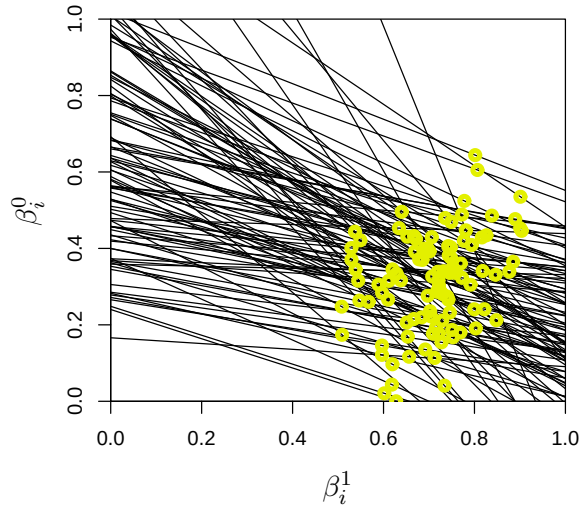
$$y_i = \beta_i^1 x_i + \beta_i^0 (1 - x_i). \tag{2.10}$$

Celkem provedeme 100 simulací. Výsledky simulací y_i a x_i pro $i = 1, \ldots, 100$ jsou znázorněny na obr. 2.1, k těmto hodnotám je zobrazeno prvních 20 přímek s interceptem β_i^0 a sklonem $\beta_i^1 - \beta_i^0$, tyto přímky můžeme chápat jako regresní přímky neagregovaného modelu (2.1).

Výsledky simulací β_i^1 a β_i^0 pro $i = 1, \ldots, 100$ jsou znázorněny na obr. 2.2 jako světlé body. Běžně však tyto hodnoty neznáme a jsme schopni pozorovat pouze úsečky, které určují všechny možné kombinace parametrů β_i^1 a β_i^0 na základě (2.10). Tento typ grafu, kdy na osách máme odhadované parametry modelu, označujeme jako *tomogram*.



Obrázek 2.1: Bodový diagram pozorovaných hodnot y_i a x_i a 20 vybraných regresních přímek určených podle skutečných hodnot parametrů β_i^1 a β_i^1.



Obrázek 2.2: Porovnání skutečných hodnot parametrů β_i^1 a β_i^1 a možných kombinací parametrů β_i^1 a β_i^1 na základě pozorovaných hodnot y_i a x_i.

2.2 Goodmanův přístup

2.2.1 Jednoduchý Goodmanův model

V roce 1953 v [13] publikoval Leo A. Goodman svůj model, který má i dnes své místo v ekologické regresi a ze kterého se dodnes vychází při tvorbě nových modelů. Velmi často je to jediný model, který se pro vyhodnocování dat používá v různých základních analýzách. Značným problémem však zůstává neustálá ignorace předpokladů tohoto modelu, bez kterých není model obecně platný a výsledky, které dává, mohou být velmi zavádějící.

Model vychází ze vztahu (2.12), nicméně nyní považujeme jednotlivé pravděpodobnosti β_i^1 a β_i^0 za konstantní parametry napříč soubory, tj.

$$\beta_i^1 \equiv \beta^1 \qquad \beta_i^0 \equiv \beta^0, \qquad i = 1, \ldots, m. \tag{2.11}$$

Model (2.1) si spolu s podmínkou (2.11) přepíšeme

$$Y_{it} = \beta^1 x_{it} + \beta^0(1 - x_{it}) + e_{it}, \qquad t = 1, \ldots, n_i, i = 1, \ldots, m. \tag{2.12}$$

Nyní zprůměrujeme obě strany rovnice přes jednotlivé soubory, jak závislou proměnou, tak nezávislou proměnou spolu s reziduální složkou. Takto získáme *jednoduchý Goodmanův model*:

$$\bar{Y}_{i\bullet} = \beta^1 \bar{x}_{i\bullet} + \beta^0(1 - \bar{x}_{i\bullet}) + \bar{e}_i, \qquad i = 1, \ldots, m, \tag{2.13}$$

kde $\bar{e}_i$ je náhodná veličina splňující požadavek $\mathsf{E}\,\bar{e}_i = 0, i = 1, \ldots, m$. V literatuře se běžně předpokládá konstantní rozptyl (nezávislý na i). Tento model jsme si už zavedli v obecné podobě v (1.8). V rámci Goodmanova přístupu nazýváme tento model *konstantním modelem*[3], jak je uvedeno například v [12].

Přesnější by bylo počítat rozptyl $\bar{Y}_{i\bullet}$ na základě (2.2) jako

$$\operatorname{var} \bar{Y}_{i\bullet} = \left(\beta^1(1 - \beta^1)\bar{x}_{i\bullet} + \beta^0(1 - \beta^0)(1 - \bar{x}_{i\bullet})\right)/n_i, \qquad i = 1, \ldots, m. \tag{2.14}$$

Jako kompromis či výchozí odhady pro iterační zpřesňování je možné předpokládat $\operatorname{var} \bar{e}_i = \sigma^2/n_i$.

Model (2.13) lze pomocí polarizace přepsat jako

$$\bar{Y}_{i\bullet} = \beta^0 + p\bar{x}_{i\bullet} + \bar{e}_i, \qquad i = 1, \ldots, m,$$

Standardní metoda vážených nejmenších čtverců pak vede k odhadům

$$\hat{p} = \frac{\sum_{i=1}^m n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})}{\sum_{i=1}^m n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2}, \qquad \hat{\beta}^0 = \bar{Y}_{\bullet\bullet} - \hat{p}\bar{x}, \qquad \hat{\beta}^1 = \hat{p} + \hat{\beta}^0, \tag{2.15}$$

kde pro připomenutí $p = \beta^1 - \beta^0$ je již jednou zmíněná polarizace a

$$\bar{x}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^m n_i \bar{x}_{i\bullet}, \qquad \bar{Y}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^m n_i \bar{Y}_{i\bullet}.$$

[3] Model je pojmenován podle konstantních parametrů β^1 a β^0.

Za platnosti (2.11) jsou odhady tohoto modelu $\hat{\beta}^1$ a $\hat{\beta}^0$ nestrannými odhady koeficientů β_i^0 a β_i^1, nazýváme je *Goodmanovy odhady*. Problém tohoto modelu je fakt, že předpokládáme neměnnost chování odhadovaných koeficientů β_i^1 a β_i^0 pro různé soubory[4]. Tento předpoklad není obecně platný. Odhady populačních parametrů β^1 a β^0 definovaných v (2.5) a (2.6) pak nejsou nestranné a vzniká vychýlení. Navíc tento model nikterak nerespektuje fakt, že koeficienty musí ležet v intervalu $[0, 1]$ a tak se běžně může stát, že dostaneme záporný odhad některého koeficientu.

2.2.2 Sofistikovaný Goodmanův model

Nyní předpokládejme, stejně jako v [1], že pro každý soubor jsou pravděpodobnosti β_i^1 a β_i^0 vybrány z dvourozměrného rozdělení s distribuční funkcí $F_i(\cdot, \cdot)$, která se může lišit podle souboru. Rádi bychom nyní odhadli β^1 a β^0 určené dle (2.7) a (2.8). Předpokládejme nyní, že platí

$$\bar{Y}_{i\bullet} = \beta_i^1 \bar{x}_{i\bullet} + \beta_i^0 (1 - \bar{x}_{i\bullet}) + \bar{e}_i, \qquad i = 1, \ldots, m, \tag{2.16}$$

kde $\mathsf{E}\,\bar{e}_i = 0$, $\operatorname{var}\bar{e}_i = \sigma^2/n_i$. Abychom se přiblížili modelu (2.13), přesuneme rozdíly $\beta_i^1 - \beta^1$ a $\beta_i^0 - \beta^0$ do reziduální složky. Uvažujme také, že regresor $\bar{\boldsymbol{X}}$ není násobkem jedničkového vektoru. Dostáváme tak *sofistikovaný Goodmanův model* (v originále Sophisticated Goodman model, viz [1] str. 51):

$$\bar{Y}_{i\bullet} = \beta^1 \bar{x}_{i\bullet} + \beta^0 (1 - \bar{x}_{i\bullet}) + u_i, \qquad i = 1, \ldots, m, \tag{2.17}$$

kde $u_i = (\beta_i^1 - \beta^1)\bar{x}_{i\bullet} + (\beta_i^0 - \beta^0)(1 - \bar{x}_{i\bullet}) + \bar{e}_i$. Distribuční funkce $F_i(\cdot)$ se může lišit podle souboru a obecně tak neplatí, že $\mathsf{E}\,\beta_i^1 = \beta^1$, resp. $\mathsf{E}\,\beta_i^0 = \beta^0$, tudíž ani pro reziduální složku obecně nemusí platit, že $\mathsf{E}\,u_i = 0$. Na druhou stranu opět budeme předpokládat, že $\operatorname{var} u_i = \sigma^2/n_i$.

Nyní je zapotřebí stanovit, za jakých okolností můžeme pomocí metody nejmenších čtverců získat nestranné odhady parametrů β^1 a β^0. K tomu poslouží následující tvrzení 2.3.

Tvrzení 2.3. *Nechť platí*

$$Y_i = \alpha + \beta(x_i - \bar{x}) + u_i, \qquad var\, u_i = \sigma^2/n_i, \tag{2.18}$$

kde $i = 1, \ldots, m$. *Dále nechť* $x_1, \ldots, x_m$ *jsou známe konstanty splňující* $T_{xx} = \sum_{i=1}^m n_i (x_i - \bar{x})^2 > 0$, $n_1, \ldots, n_m$ *jsou známé kladné konstanty,* $n = \sum_{i=1}^m n_i$ *a* $\bar{x} = \sum_{i=1}^m n_i x_i / n$. *Odhady parametrů* α *a* β *váženou metodou nejmenších čtverců jsou nestranné právě tehdy, když současně platí*

$$\sum_{i=1}^n n_i \, \mathsf{E}(u_i) = 0 \tag{2.19}$$

$$\sum_{i=1}^n n_i x_i \, \mathsf{E}(u_i) = 0. \tag{2.20}$$

[4] V literatuře [1] a dalších se můžeme dočíst o neměnnosti chování odhadovaných koeficientů β_i^1 a β_i^0 při různých hodnotách $\bar{x}_{i\bullet}$. Tento předpoklad vychází z myšlenky, že v souborech, kde je stejná hodnota veličiny $\bar{x}_{i\bullet}$, je chování koeficientů β_i^1 a β_i^0 totožné a tak stačí požadovat pouze stejné chování pro soubory s různými hodnotami regresorů.

Důkaz. Minimalizace výrazu

$$\sum_{i=1}^{m} n_i(Y_i - (\alpha + \beta(x_i - \bar{x})))^2$$

vede na soustavu lineárních rovnic (použijeme $\sum_{i=1}^{m} n_i(x_i - \bar{x}) = 0$)

$$\begin{aligned} \sum_{i=1}^{m} n_i Y_i &= n\hat{\alpha}, \\ \sum_{i=1}^{m} n_i(x_i - \bar{x})Y_i &= \hat{\beta} \sum_{i=1}^{m} n_i(x_i - \bar{x})^2. \end{aligned}$$

Odhad parametrů α a β je

$$\begin{aligned} \hat{\alpha} &= \bar{Y}, \\ \hat{\beta} &= \frac{\sum_{i=1}^{m} n_i(x_i - \bar{x})Y_i}{\sum_{i=1}^{m} n_i(x_i - \bar{x})^2} = \frac{T_{xy}}{T_{xx}}, \end{aligned}$$

kde $\bar{Y} = \sum_{i=1}^{m} \frac{n_i}{n} Y_i$ a $T_{xy} = \sum_{i=1}^{m} n_i(x_i - \bar{x})(Y_i - \bar{Y})$. Dosaďme nyní do obou rovnic za Y_i podle (2.18). Dostaneme (opět využijeme $\sum_{i=1}^{m} n_i(x_i - \bar{x}) = 0$)

$$\begin{aligned} \hat{\alpha} &= \alpha + \frac{1}{n} \sum_{i=1}^{m} n_i u_i, \\ \hat{\beta} &= \frac{1}{T_{xx}} \beta \sum_{i=1}^{m} n_i(x_i - \bar{x})^2 + \frac{1}{T_{xx}} \sum_{i=1}^{m} n_i(x_i - \bar{x})u_i \\ &= \beta + \frac{1}{T_{xx}} \sum_{i=1}^{m} n_i(x_i - \bar{x})u_i. \end{aligned}$$

Nestrannost odhadu $\hat{\alpha}$ je ekvivalentní s požadavkem (2.19). Nestrannost odhadu $\hat{\beta}$ je ekvivalentní s rovnicí

$$\sum_{i=1}^{m} n_i(x_i - \bar{x}) \, \mathsf{E}\, u_i = 0. \tag{2.21}$$

Je zřejmé, že dvojice vztahů (2.19) a (2.20) je ekvivalentní s dvojicí (2.19) a (2.21). Tím je tvrzení dokázáno. □

Poznámka 2.4. Tvrzení 2.3 neříká na rozdíl od Gaussovy-Markovovy věty (viz Zvára [37], str. 13) nic o eficienci odhadu parametrů modelu, pouze udává, za jakých předpokladů je odhad nestranný.

Nyní ověříme splnění předpokladů pro náš model (2.17), který si pro tyto účely upravíme na

$$\begin{aligned} \bar{Y}_{i\bullet} &= (\beta^0 + p\bar{x}_{\bullet\bullet}) + p(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) + u_i \\ &= \alpha + \beta(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) + u_i, \qquad\qquad i = 1, \ldots, m. \end{aligned}$$

Podmínka (2.19) platí, neboť

$$\begin{aligned}\sum_{i=1}^{m} n_i \,\mathsf{E}\, u_i &= \sum_{i=1}^{m} \mathsf{E}[(\beta_i^1 - \beta^1)\bar{x}_{i\bullet} n_i + (\beta_i^0 - \beta^0)(1 - \bar{x}_{i\bullet}) n_i] \\ &= \sum_{i=1}^{m} n_i \bar{x}_{i\bullet} \,\mathsf{E}(\beta_i^1 - \beta^1) + \sum_{i=1}^{m} n_i (1 - \bar{x}_{i\bullet}) \,\mathsf{E}(\beta_i^0 - \beta^0) \\ &= 0,\end{aligned}$$

což plyne z definice parametrů β^1 a β^0 podle (2.7) a (2.8). Podmínka (2.20) však nutně nemusí být splněna. Rozepíšeme-li si tuto podmínku pro model (2.17), dostaneme

$$\sum_{i=1}^{m} n_i \bar{x}_{i\bullet}^2 \,\mathsf{E}(\beta_i^1 - \beta^1) + \sum_{i=1}^{m} n_i \bar{x}_{i\bullet} (1 - \bar{x}_{i\bullet}) \,\mathsf{E}(\beta_i^0 - \beta^0) = 0, \tag{2.22}$$

což obecně neplatí. Aby tato rovnost platila, musíme předpokládat, že střední hodnoty veličin β_i^1 a β_i^0 jsou při speciálním případu vážení rovny v součtu nule[5]. Označme si

$$\mathsf{E}_x(\cdot)) = \frac{\sum_{it} x_{it} \,\mathsf{E}(\cdot)}{\sum_{it} x_{it}} = \frac{\sum_i n_i \bar{x}_{i\bullet} \,\mathsf{E}(\cdot)}{\sum_i n_i \bar{x}_{i\bullet}}$$

a

$$\mathsf{E}_{1-x}(\cdot)) = \frac{\sum_{it} (1 - x_{it}) \,\mathsf{E}(\cdot)}{\sum_{it} (1 - x_{it})} = \frac{\sum_i n_i (1 - \bar{x}_{i\bullet}) \,\mathsf{E}(\cdot)}{\sum_i n_i (1 - \bar{x}_{i\bullet})}.$$

Z definice (2.7) plyne, že $\mathsf{E}_x(\beta_i^1) = \beta^1$, stejně tak $\mathsf{E}_{1-x}(\beta_i^0) = \beta^0$. Předpokládejme nyní, že platí

$$\mathsf{E}_x \left((\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\beta_i^1 - \beta^1) \right) = 0, \tag{2.23}$$

$$\mathsf{E}_{1-x} \left((\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\beta_i^0 - \beta^0) \right) = 0. \tag{2.24}$$

Tyto předpoklady můžeme ještě upravit, když si uvědomíme, že

$$\mathsf{E}_x \left((\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\beta_i^1 - \beta^1) \right) = \mathsf{E}_x \left(\bar{x}_{i\bullet} (\beta_i^1 - \beta^1) \right),$$

obdobně můžeme upravit i (2.24).

Nyní si rozepíšeme první část výrazu nalevo z (2.22) jako

$$\begin{aligned}\sum_{i=1}^{m} n_i \bar{x}_{i\bullet}^2 \,\mathsf{E}(\beta_i^1 - \beta^1) &= \sum_{i=1}^{m} \sum_{t=1}^{n_i} x_{it} \bar{x}_{i\bullet} \,\mathsf{E}(\beta_i^1 - \beta^1)/n \\ &= \frac{\sum_{i=1}^{m} \sum_{t=1}^{n_i} x_{it}}{n} \frac{\sum_{i=1}^{m} \sum_{t=1}^{n_i} x_{it} \bar{x}_{i\bullet} \,\mathsf{E}(\beta_i^1 - \beta^1)}{\sum_{i=1}^{m} \sum_{t=1}^{n_i} x_{it}} \\ &= k_x \mathsf{E}_x \left(\bar{x}_{i\bullet} (\beta_i^1 - \beta^1) \right) \\ &= 0,\end{aligned}$$

kde $k_x = \sum_{it} x_{it}/n$. Obdobně postupujeme v případě β_i^0, tedy

$$\begin{aligned}\sum_{i=1}^{m} n_i \bar{x}_{i\bullet} (1 - \bar{x}_{i\bullet}) \,\mathsf{E}(\beta_i^0 - \beta^0) &= \sum_{i=1}^{m} \sum_{t=1}^{n_i} (1 - x_{it}) \bar{x}_{i\bullet} \,\mathsf{E}(\beta_i^0 - \beta^0)/n \\ &= k_{1-x} \mathsf{E}_{1-x} \left(\bar{x}_{i\bullet} (\beta_i^0 - \beta^0) \right) \\ &= 0,\end{aligned}$$

[5] V [1] se v této souvislosti mluví o podmíněné kovarianci parametrů β_i^1 a β_i^0 a původní vysvětlující proměnné X_{it}.

kde $k_{1-x} = \sum_{it}(1 - x_{it})/n$.

Tvrzení 2.5. *Mějme model* (2.17) *a, přičemž platí podmínky* (2.23) *a* (2.24). *Potom jsou odhady parametrů* β^1 *a* β^0 *metodou nejmenších čtverců nestranné.*

Důkaz. Důsledek diskuze výše a tvrzení 2.3. □

Nejedná se o jediný možný předpoklad, který se ke Goodmanově modelu pojí. Pod názvem *zero-correlation model* (viz [16] a [12]) se běžně používá model (2.17) spolu s předpoklady

$$\mathsf{E}(\beta_i^1 | \bar{X}_{i\bullet} = \bar{x}_{i\bullet}) = \beta^1, \qquad \mathsf{E}(\beta_i^0 | \bar{X}_{i\bullet} = \bar{x}_{i\bullet}) = \beta^0. \tag{2.25}$$

Tento předpoklad je silnější než (2.23) a (2.24), diskuze na téma nutných předpokladů je například v [17]. Nejsilnějším zde uváděným předpokladem je předpoklad jednoduchého Goodmanova modelu.

Následující odstavec je převzat z [3]. V případě porušení těchto předpokladů se směr vychýlení řídí následujícím tvrzením 2.6, které ve zkratce říká, že vychýlení odhadů obou parametrů jsou v opačném směru.

Tvrzení 2.6. *Nechť platí* $0 < \bar{x}_{\bullet\bullet} < 1$ *a* $\hat{\beta}^1$ *a* $\hat{\beta}^0$ *jsou odhady modelu* (2.17) *metodou vážených nejmenších čtverců, potom*

$$\mathrm{Bias}(\hat{\beta}^0) = -\frac{\bar{x}_{\bullet\bullet}}{1 - \bar{x}_{\bullet\bullet}} \mathrm{Bias}(\hat{\beta}^1).$$

Důkaz. Vyjdeme z vlastnosti odhadů $\hat{\beta}^1$ a $\hat{\beta}^0$, o těch je známo, že procházejí těžištěm a je tedy splněna rovnice

$$\bar{Y}_{\bullet\bullet} = \hat{\beta}^1 \bar{x}_{\bullet\bullet} + \hat{\beta}^0 (1 - \bar{x}_{\bullet\bullet}).$$

Dále víme, že platí

$$\bar{Y}_{\bullet\bullet} = \beta^1 \bar{x}_{\bullet\bullet} + \beta^0 (1 - \bar{x}_{\bullet\bullet}) + u,$$

kde $\mathsf{E}\, u = \mathsf{E} \sum_{i=1}^{m} u_i n_i / n = 0$. Nyní stačí od sebe obě rovnice odečíst, aplikovat střední hodnotu

$$0 = (\mathsf{E}\,\hat{\beta}^1 - \beta^1)\bar{x}_{\bullet\bullet} + (\mathsf{E}\,\hat{\beta}^0 - \beta^0)(1 - \bar{x}_{\bullet\bullet}).$$

Úpravou získáme tvrzení. □

Budeme-li předpokládat, že $\bar{x}_{\bullet\bullet} < \frac{1}{2}$, pak $0 < \frac{\bar{x}_{\bullet\bullet}}{1-\bar{x}_{\bullet\bullet}} < 1$ a tudíž nám předchozí věta říká, že pokud existuje vychýlení jako takové, pak vychýlení odhadu koeficientu β^1 je větší, než vychýlení β^0. Tudíž budeme-li mít vychýlení v celém modelu, pak Goodmanův přistup není vhodným pro odhad minoritního koeficientu. Dalším důsledkem předchozí věty je vztah vychýlení polarizace a odhadovaných koeficientů

$$\mathrm{Bias}(\hat{\beta}^1) = (1 - \bar{x}_{\bullet\bullet})\mathrm{Bias}(\hat{p}), \qquad \mathrm{Bias}(\hat{\beta}^0) = -\bar{x}_{\bullet\bullet}\mathrm{Bias}(\hat{p}).$$

Více o vychýleních v rámci ekologické regrese je možné nalézt v [3].

Alternativou vzhledem ke Goodmanovu modelu spolu s předpokladem (2.25) je modelování veličin β_i^1 a β_i^0 jakožto lineárně závislých na vysvětlující proměnné $\bar{x}_{i\bullet}$, předpokládejme tedy, že platí

$$\mathsf{E}[\beta_i^1 | \bar{X}_{i\bullet} = \bar{x}_{i\bullet}] = \lambda_0 + \lambda_1 \bar{x}_{i\bullet}, \qquad i = 1, \ldots, m,$$

a

$$\mathsf{E}[\beta_i^0 | \bar{X}_{i\bullet} = \bar{x}_{i\bullet}] = \eta_0 + \eta_1 \bar{x}_{i\bullet}, \qquad i = 1, \ldots, m, .$$

Přidáme-li předpoklady o rozptylu β_i^1 a β_i^0, dostaneme

$$\beta_i^1 = \lambda_0 + \lambda_1 \bar{x}_{i\bullet} + v_i$$

a

$$\beta_i^0 = \eta_0 + \eta_1 \bar{x}_{i\bullet} + w_i$$

pro každý soubor $i = 1, \ldots, m$, kde veličiny v_i a w_i představují bílý šum (tj. veličiny s nulovou střední hodnotou, navzájem nezávislé a s konstantním rozptylem) a jsou nezávislé na $\bar{x}_{i\bullet}$. Předpoklad lineární závislosti β_i^1 a β_i^0 vůči $\bar{x}_{i\bullet}$ a h_i pozměňuje předpoklad nezávislosti β_i^1 a β_i^0 na $\bar{x}_{i\bullet}$ do podoby nezávislosti pouze jejich chyb u_i a v_i vůči $\bar{x}_{i\bullet}$ a h_i. Tyto předpoklady vedou ke kvadratickému modelu odvozenému z (2.16) za ponechání chybové složky e_i, který vypadá následovně

$$\begin{aligned} \bar{Y}_{i\bullet} &= (\lambda_0 + \lambda_1 \bar{x}_{i\bullet} + v_i)\bar{x}_{i\bullet} + (\eta_0 + \eta_1 \bar{x}_{i\bullet} + w_i)(1 - \bar{x}_{i\bullet}) + e_i \\ &= (\lambda_0 + \eta_1)\bar{x}_{i\bullet} + \eta_0(1 - \bar{x}_{i\bullet}) + (\lambda_1 - \eta_1)\bar{x}_{i\bullet}^2 + [w_i + (v_i - w_i)\bar{x}_{i\bullet} + e_i] \\ &= \eta_0 + (\lambda_0 + \lambda_1 - \eta_1)\bar{x}_{i\bullet} + (\lambda_1 - \eta_1)h_i + u_i^* \qquad i = 1, \ldots, m, \end{aligned} \tag{2.26}$$

kde u_i^* je chybová složka.

Tento model obsahuje čtyři neznámé parametry (*contextual effects parameters*), ale pouze tři regresní koeficienty. Je zde tudíž přítomna nejednoznačnost a musí být přidán nějaký další omezující předpoklad. Předpoklad rovnosti $\lambda_1 = \eta_1$ však nepomůže, neboť potom dostáváme tři neznámé parametry při dvou koeficientech. V případ volby $\lambda_1 = \eta_1 = 0$, dostáváme model (2.17) s předpokladem (2.25). Existují jisté možnosti, jak zvolit nastavení parametrů, ale vycházejí hlavně z externí informace o pozorovaných objektech. Diskuze na téma maximálního vychýlení odhadů populačních parametrů tj. $\mathrm{P}(Y = 1|X = 1)$, resp. $\mathrm{P}(Y = 1|X = 0)$, v případě kvadratického modelu je možné nalézt v [1], kap. 5.

Kvadratický model (2.26) je nicméně užitečným předstupněm zobecnění modelu (2.13). Uvažujme nyní, že se parametry β_i^0 a β_i^1 mohou měnit nelineárně v závislosti na x_i, t.j.

$$\beta_i^1 = \beta^1(\bar{x}_{i\bullet}) + \epsilon_i^1, \qquad i = 1, \ldots, m,$$

a

$$\beta_i^0 = \beta^0(\bar{x}_{i\bullet}) + \epsilon_i^0, \qquad i = 1, \ldots, m,$$

kde $\beta^j(\cdot)$, $j = 0, 1$ je nespecifikovaná nelineární funkce a ϵ_i^j, $j = 0, 1$, mají nulovou podmíněnou střední hodnotu vzhledem k $\bar{x}_{i\bullet}$, dále předpokládáme konstantní rozptyl a nulové kovariance pro jednotlivé $j = 0, 1$. Vložením do vztahu (2.16) získáváme následující model

$$\bar{Y}_{i\bullet} = \delta_i^c + \delta_i^s \bar{x}_{i\bullet} + \epsilon_i, \qquad i = 1, \ldots, m, \tag{2.27}$$

kde

$$\begin{aligned} \delta_i^c &= \beta^0(\bar{x}_{i\bullet}), \\ \delta_i^s &= \beta^1(\bar{x}_{i\bullet}) - \beta^0(\bar{x}_{i\bullet}), \\ \epsilon_i &= \epsilon_i^1 \bar{x}_{i\bullet} + \epsilon_i^0(1 - \bar{x}_{i\bullet}) + \bar{e}_i, \qquad i = 1, \ldots, m. \end{aligned}$$

Ignorujeme-li zjevnou heteroskedasticitu v reziduální složce modelu (2.27), můžeme tento model odhadnout pomocí neparametrické regrese. Chambers a Steel [16] navrhují lokální lineární vyhlazování pomocí jádrové funkce $K(\cdot)$, kdy δ_i^c a δ_i^s určíme pomocí LWLS (*locally weighted least squares*) z rovnice

$$\sum_{j=1}^{m} K\left(\frac{\bar{x}_{j\bullet} - \bar{x}_{i\bullet}}{\kappa_i}\right)(\bar{Y}_{j\bullet} - \delta_i^c - \delta_i^s \bar{x}_{j\bullet})\begin{pmatrix} 1 \\ \bar{j}_{i\bullet} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad i = 1, \ldots, m, \tag{2.28}$$

kde κ_i je vhodně zvolená šířka vyhlazovacího okna. Tímto způsobem získáme odhady veškerých koeficientů β_i^1 a β_i^0 pro každé $i = 1, \ldots, m$

$$\hat{\beta}_i^0 = \hat{\delta}_i^c \tag{2.29}$$

a

$$\hat{\beta}_i^1 = \hat{\delta}_i^c + \hat{\delta}_i^s \tag{2.30}$$

Z tohoto pak dostáváme požadované celkové odhady β^0 a β^1 jako

$$\hat{\beta}^0 = \frac{\sum_{i=1}^m n_i \hat{\beta}_i^0 (1 - \bar{x}_{i\bullet})}{\sum_{i=1}^m n_i (1 - \bar{x}_{i\bullet})}$$

a

$$\hat{\beta}^1 = \frac{\sum_{i=1}^m n_i \hat{\beta}_i^1 \bar{x}_{i\bullet}}{\sum_{i=1}^m n_i \bar{x}_{i\bullet}}.$$

Pokud je κ_i dostatečně malé, pak heteroskedasticita v reziduální složce nezpůsobuje tak velké problémy. Je možné s heteroskedasticitou v rovnici (2.28) pracovat, bude to však na úkor robustnosti a celkově zvýší náročnost výpočtu.

2.6.1 Goodmanův model s pomocnou proměnnou

Tato kapitola vychází z [32]. Goodmanův předpoklad (2.25) je velice silný a pravdou je, že je často porušován, vesměs na základě struktury dat a s tím spojené prostorové autokorelace. Vzniklo tak mnoho snah, jak toto porušení předpokladů napravit a zároveň pracovat stále s modelem 2.17. Jednou takovou snahou je i posun k proměnlivé regresi (*switching regression*).

Předpokládejme nyní, že je možné soubory i, kde $i \in I = \{1, \ldots, m\}$, rozdělit do více kategorií, pro jednoduchost pracujme jen se dvěma disjunktními množinami I_1 a I_2. Proměnlivá regrese pak pracuje se dvěma modely

$$\bar{Y}_{i\bullet} = \beta_1^1 \bar{x}_{i\bullet} + \beta_1^0(1 - \bar{x}_{i\bullet}) + e_{i1}, \qquad i \in I_1, \tag{2.31}$$

$$\bar{Y}_{i\bullet} = \beta_2^1 \bar{x}_{i\bullet} + \beta_2^0(1 - \bar{x}_{i\bullet}) + e_{i2}, \qquad i \in I_2, \tag{2.32}$$

kde e_{i1} a e_{i2} jsou nezávislé veličiny z normálního rozdělení s nulovou střední hodnotou a konstantním rozptylem. Nechť D_i, $i = 1, \ldots, m$, je zatím neznámá proměnná, která určuje, který model pro daný i-tý soubor platí, přičemž $D_i = 0$ znamená, že pro i-tý soubor platí model (2.31), a pro $D_i = 1$ platí model (2.32). Oba modely nyní můžeme zapsat do jedné rovnice jako

$$\begin{aligned} \bar{Y}_{i\bullet} &= [(1 - D_i)(\beta_1^1 - \beta_1^0) + D_i(\beta_2^1 - \beta_2^0)]\bar{x}_{i\bullet} + (1 - D_i)\beta_1^0 + D_i\beta_2^0 + \\ &\quad +(1 - D_i)e_{1i} + D_i e_{2i}, \qquad i \in I. \end{aligned} \tag{2.33}$$

Pro odhad parametrů β_1^1, β_1^0, β_2^1 a β_2^0 je nejprve nutné stanovit hodnoty neznámých parametrů D_i, $i = 1, \ldots, m$, respektive určit rozhodovací algoritmus. K tomu nám poslouží vektory vysvětlujících proměnných $\mathbf{z}_i = (z_i^1, \ldots, z_i^s)'$, $i = 1, \ldots, m$. K tomu, abychom získali požadované hodnoty D_i použijeme *logit* model (viz [7], str. 168). Uvažujme vysvětlovanou proměnou D_i^* v lineárním modelu

$$D_i^* = \mathbf{z}_i'\boldsymbol{\gamma} + \epsilon_i, \qquad i \in I,$$

kde ϵ_i jsou nezávislé stejně rozdělené náhodné veličiny s nulovou střední hodnotou. $\boldsymbol{\gamma}$ je vektor neznámých parametrů. Vysvětlovaná proměnná D_i je pak pro každé $i \in I$ určena vztahem

$$D_i = \begin{cases} 1 & \text{pro } D_i^* > 0, \\ 0 & \text{pro } D_i^* \leq 0. \end{cases}$$

Pak platí

$$\mathrm{P}(D_i = 1|\mathbf{z}_i, \boldsymbol{\gamma}) = 1 - F(\mathbf{z}_i'\boldsymbol{\gamma}) = 1 - \frac{e^{-\mathbf{z}_i'\boldsymbol{\gamma}}}{1 + e^{-\mathbf{z}_i'\boldsymbol{\gamma}}}, \qquad i \in I,$$

kde $F(\cdot)$ je distribuční funkce logistického rozdělení. Tento práh poskytuje rámec pro jednoznačné rozlišení skupin. Odhad modelu (2.33) s binární vysvětlovanou proměnnou se provede metodou maximální věrohodnosti. Ne vždy je nutné uvažovat D_i za diskrétní. K tomu slouží hybridní přístup, kdy D_i je určena vztahem.

$$D_i = 1 - \frac{e^{-\mathbf{z}_i'\boldsymbol{\gamma}}}{1 + e^{-\mathbf{z}_i'\boldsymbol{\gamma}}}, \qquad i \in I.$$

Tento vztah pak můžeme dosadit do modelu (2.33).

Nyní máme $6 + s$ (počítáme-li rozptyly chybových složek σ_1 a σ_2, β_1^1, β_1^0, β_2^1 a β_2^0 z modelů (2.31) a (2.32) a kde s je velikost vektoru $\mathbf{z}_i$) neznámých parametrů v modelu (2.33), za předpokladu, že veličiny y_i mají normální rozdělení a počet souborů je dostatečně velký, jsou opět veškeré parametry odhadnutelné metodou maximální věrohodnosti.

Pro určení, zdali je pomocná proměnná $\mathbf{z}_i$ vhodná pro zachycení změny koeficientů v modelu, použijeme testy stability, konkrétně *CUSUM-statistiky*, které se používají například při analýzách časových řad (převzato z [7], str. 130). Vzhledem k tomu, že se zde nejedná o časovou řadu, použijeme pro seřazení souborů (resp. jednotlivých pozorování) právě pomocné proměnné $\mathbf{z}_i$, které v případě $s > 1$ řadíme po složkách. Dále však budeme uvažovat případ $s = 1$.

Pro přehlednost si označme $Y_i = \bar{Y}_{i\bullet}$, $\boldsymbol{x}_i = (1, \bar{x}_{i\bullet})'$ a $\boldsymbol{X}_t = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)'$ pro $t = 2, \ldots, m$. Předpokládejme, že $\boldsymbol{X}_2$ má plnou hodnost.

Pozorování si uspořádáme vzestupně podle proměnné z_i a provedeme regresi $Y_i = \boldsymbol{\beta}'\boldsymbol{x}_i + e_i$ pro $i = 1, \ldots, t-1$. Pomocí metody nejmenších čtverců získáme prvky posloupnosti *rekurentních odhadů* označených jako $\boldsymbol{b}_{t-1}$, předpovědí $\hat{y}_t = \boldsymbol{b}'_{t-1}\boldsymbol{x}_t$ a chyb predikce

$$f_t = Y_t - \hat{Y}_t, \qquad t = 3, \ldots, m.$$

Za předpokladu normality je pak v klasickém modelu lineární regrese

$$e_t = \frac{f_t}{\sqrt{k_t}} \sim N(0, \sigma^2), \qquad t = 3, \ldots, m,$$

kde

$$k_t = 1 + \boldsymbol{x}'_t(\boldsymbol{X}'_{t-1}\boldsymbol{X}_{t-1})^{-1}\boldsymbol{x}_t.$$

O reziduích e_t mluvíme jako o *rekurentních reziduích. CUSUM-statistiky* (kumulativní součty) v jednotlivých časech t mají přibližně normální rozdělení

$$CUSUM_t = \sum_{u=3}^{t} \frac{e_u}{s} \sim N(0, t-2), \qquad t = 3, \ldots, m,$$

kde s je odhad směrodatné odchylky reziduální složky σ. Testujeme tak, zdali na hladině významnosti α přesáhne v absolutní hodnotě pro dané pozorování statistika $CUSUM_t$ práh $\Phi^{-1}(1-\alpha/2)\sqrt{t-2}$, kde $\Phi^{-1}(\cdot)$ je kvantil normovaného normálního rozdělení. V takovém případě zamítáme hypotézu o neměnnosti parametru $\boldsymbol{\beta}$.

Pokud chceme testovat neměnnost parametru σ, můžeme použít *CUSUMQ-statistiku* založenou na čtvercích rekurentních reziduí

$$CUSUMQ_t = \frac{\sum_{u=3}^{t} e_u^2}{\sum_{u=3}^{m} e_u^2} \sim \frac{\chi^2_{t-2}}{m-2}, \qquad t = 3, \ldots, m.$$

Při překročení kritické hodnoty je výsledkem *CUSUMQ-testu* zamítnutí hypotézy o neměnnosti rozptylu.

2.7 Metoda hranic

Doposud uvedené metody naprosto ignorovaly fakt, že odhadované parametry mají omezený obor hodnot a že pro ně platí další omezující pravidla vzhledem k tomu, že se jedná o pravděpodobnosti $\mathsf{P}(Y_{it} = 1|X_{it} = 1)$, resp. $\mathsf{P}(Y_{it} = 1|X_{it} = 0)$. Tento způsob analýzy dat nazýváme metodou hranic (*method of bounds*, Duncan [8]). Tato metoda tak dokáže zúžit definiční obor a pro určité typy dat může tato metoda dostatečně přesně odhadnout parametry β_i^1 a β_i^0. Z rovnice (2.4) dostáváme následující

$$\hat{\beta}_i^1 = \frac{\bar{Y}_{i\bullet}}{\bar{x}_{i\bullet}} - \frac{1 - \bar{x}_{i\bullet}}{\bar{x}_{i\bullet}}\hat{\beta}_i^0$$

respektive

$$\hat{\beta}_i^0 = \frac{\bar{Y}_{i\bullet}}{1 - \bar{x}_{i\bullet}} - \frac{\bar{x}_{i\bullet}}{1 - \bar{x}_{i\bullet}}\hat{\beta}_i^1$$

pro $i = 1, \ldots, m$. Obě tyto rovnosti definují podprostor v prostoru $[0, 1]^2$, tzv. *tomografickou úsečku*, viz obrázek 2.2, pro každé $i = 1, \ldots, m$. Takto získáváme následující omezení

$$L_i^1 = \max\left\{0, \frac{\bar{Y}_{i\bullet} - (1 - \bar{x}_{i\bullet})}{\bar{x}_{i\bullet}}\right\} \leq \hat{\beta}_i^1 \leq \min\left(1, \frac{\bar{Y}_{i\bullet}}{\bar{x}_{i\bullet}}\right) = U_i^1 \tag{2.34}$$

a

$$L_i^0 = \max\left\{0, \frac{\bar{Y}_{i\bullet} - \bar{x}_{i\bullet}}{1 - \bar{x}_{i\bullet}}\right\} \leq \hat{\beta}_i^0 \leq \min\left(1, \frac{\bar{Y}_{i\bullet}}{1 - \bar{x}_{i\bullet}}\right) = U_i^0, \tag{2.35}$$

kde pro $j = 0, 1$ nazýváme L_i^j dolní hranicí odhadu parametru β_i^j, dále pak U_i^j nazýváme horní hranicí odhadu parametru β_i^j. Mezi horní a dolní hranicí odhadů parametrů β_i^1 a β_i^0 platí následující vztah

$$U_i^1 = \frac{\bar{Y}_{i\bullet}}{\bar{x}_{i\bullet}} - \frac{1 - \bar{x}_{i\bullet}}{\bar{x}_{i\bullet}}L_i^0,$$

a

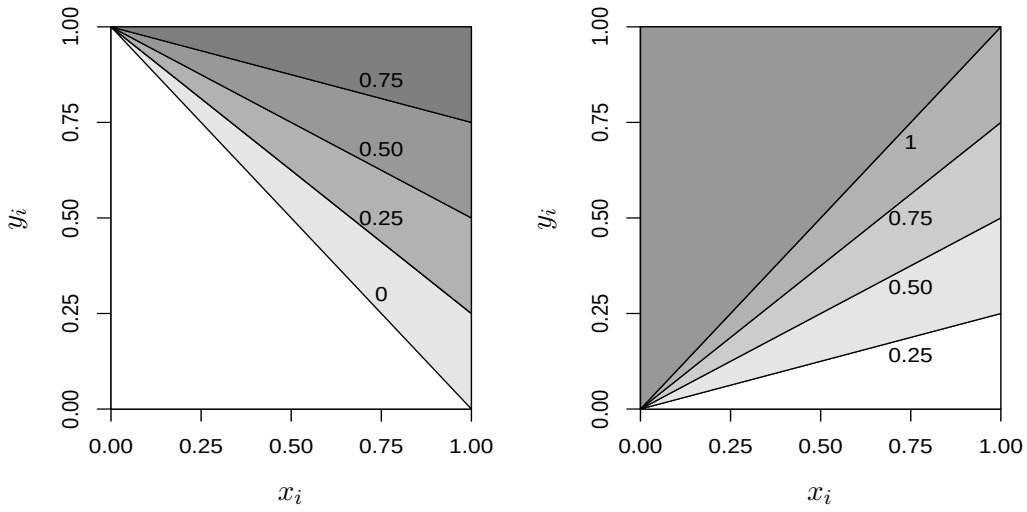
$$L_i^1 = \frac{\bar{Y}_{i\bullet}}{\bar{x}_{i\bullet}} - \frac{1 - \bar{x}_{i\bullet}}{\bar{x}_{i\bullet}}U_i^0.$$

Pro ilustraci chování mezních hodnot $\hat{\beta}_i^1$ z (2.34) a $\hat{\beta}_i^1$ z (2.34) převezmeme zobrazení popsané v [18]. Jak vypadají jednotlivé hranice pro odhady parametru β_i^1, je zde obr. 2.3. Levý graf dává v závislosti na $\bar{x}_{i\bullet}$ a $\bar{Y}_{i\bullet}$ hodnotu L_i^1. Jak můžeme vidět, s rostoucími hodnotami $\bar{x}_{i\bullet}$ a $\bar{Y}_{i\bullet}$ roste i L_i^1. Pro lepší orientaci jsou zde zvýrazněny vrstevnice L_i^1 s hodnotami 0, 0,25, 0,5 a 0,75. Pravý graf zobrazuje tytéž hodnoty pro horní hranici U_i^1. Obrázek 2.4 popisuje tu samou situaci pro druhý odhadovaný parametr β_i^0. Jedná se o symetrické zobrazení.

Pro lepší pochopení, jak se vyvíjí délka intervalu (L_i^1, U_i^1), resp. (L_i^0, U_i^0), je zde obr. 2.5. Zpřesňování jednoho parametru vede k nárůstu nejistoty u druhého parametru. Budeme-li znát realizace $\bar{Y}_{i\bullet}$ a hodnoty $\bar{x}_{i\bullet}$ pro $i = 1, \ldots, m$, stačí tyto body znázornit do obrázku 2.5 a ihned vidíme, pro jaké soubory jsou omezení daná metodou hranic aktivní a pro jaké soubory je tato metoda nedostatečná.

Chambers a Steel [16] dále zapracovávají tato omezení do (2.29) a (2.30) pomocí následující rovnosti

$$\hat{\beta}_i^1 = a_i - b_i\hat{\beta}_i^0, \tag{2.36}$$



Obrázek 2.3: Dolní (obrázek vlevo) a horní (obrázek vpravo) hranice pro odhady parametru β_i^1. Znázorněné úsečky spojují body, které dávají stejné hodnoty hranic. Tyto hodnoty jsou zobrazeny vedle jednotlivých úseček. Plocha vlevo dole (levý obrázek) představuje oblast, pro kterou metoda hranic neposkytuje dodatečnou informaci, dolní hranice v případě obrázku vlevo je rovna 0 (v případě obrázku vpravo se jedná o oblast vlevo nahoře, horní hranice je zde rovna jedné).

kde

$$a_i = \frac{U_i^1 U_i^0 - L_i^1 L_i^0}{U_i^0 - L_i^0}$$

a

$$b_i = \frac{U_i^1 - L_i^1}{U_i^0 - L_i^0},$$

která musí platit pro $i = 1, \ldots, m$. Z (2.29) a (2.30) dosadíme do (2.36), dostáváme tak následující identitu

$$\hat{\delta}_i^s = a_i - (b_i + 1)\hat{\delta}_i^c, qquadi = 1, \ldots, m, \tag{2.37}$$

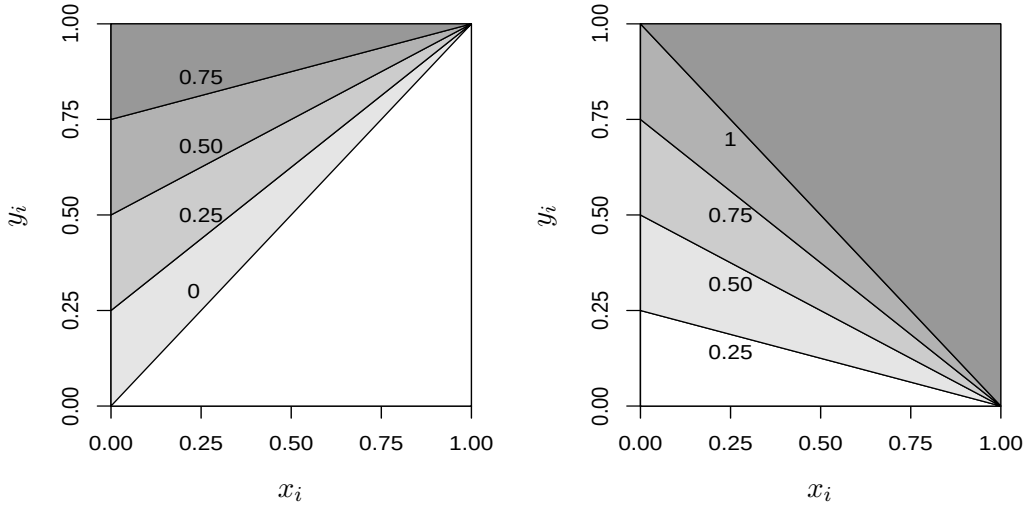
což nám umožňuje nahradit (2.28) jednodušší rovnicí

$$\sum_{j=1}^{m} K\left(\frac{\bar{x}_{j\bullet} - \bar{x}_{i\bullet}}{b_i}\right) V_{ij}(U_{ij} - \delta_i^c V_{ij}) = 0, \qquad i = 1, \ldots, m, \tag{2.38}$$

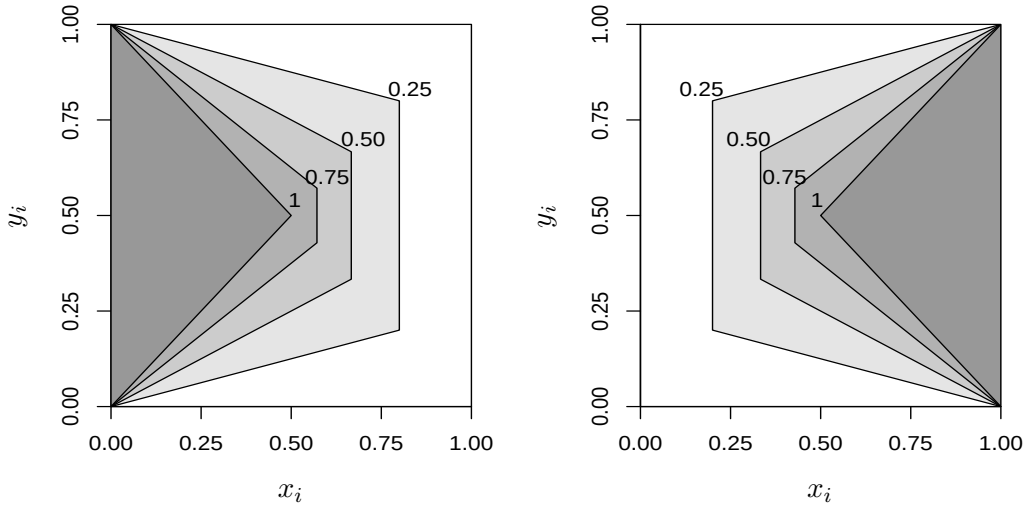
kde $U_{ij} = \bar{Y}_j - a_i \bar{x}_{j\bullet}$ a $V_{ij} = 1 - (b_i + 1)\bar{x}_{j\bullet}$. Výsledkem (2.38) však není požadovaný odhad parametru δ_i^c, nýbrž $\tilde{\delta}_i^c$. Po odhadu vyžadujeme, aby padl do intervalu $[L_i^0, U_i^0]$. Finální odhad $\hat{\delta}_i^c$ získáme následující úpravou

$$\hat{\delta}_i^c = \min\{\max(\tilde{\delta}_i^c, L_i^0), U_i^0\}, \qquad i = 1, \ldots, m. \tag{2.39}$$

Konečné odhady parametrů β_i^1 a β_i^0 jsou pak definovány kombinací rovnice (2.39) s rovnicí (2.37) a rovnic (2.29) a (2.30).



Obrázek 2.4: Dolní (obrázek vlevo) a horní (obrázek vpravo) hranice pro odhady parametru β_i^0. Znázorněné úsečky spojují body, které dávají stejné hodnoty hranic. Tyto hodnoty jsou zobrazeny vedle jednotlivých úseček. Plocha vpravo dole (levý obrázek) představuje oblast, pro kterou metoda hranic neposkytuje dodatečnou informaci, dolní hranice v případě obrázku vlevo je rovna 0 (v případě obrázku vpravo se jedná o oblast vpravo nahoře, horní hranice je zde rovna jedné).



Obrázek 2.5: Délka intervalu (L_i^1, U_i^1) (levý graf) a (L_i^0, U_i^0) (pravý graf). Jednotlivé křivky spojují body, které dávají stejně široký interval. Tato šířka intervalu je napsána vedle jednotlivých křivek. Nejmenší trojúhelník představuje oblast, pro kterou omezení neposkytuje dodatečnou informaci a délka intervalu je rovna jedné.

2.8 Neighbourhood model

Předpokládejme na moment, že jsou hodnoty regresoru X náhodné. Dále předpokládejme, že vysvětlovaná proměnná Y je nezávislá na X uvnitř každého souboru i. Předpokládáme tudíž ze platí následující rovnost

$$\mathrm{P}(Y = 1|X = j, i) = \mathrm{P}(Y = 1|i), \tag{2.40}$$

pro $j = 0, 1$ a $i = 1, \ldots, m$. Přirozeným odhadem koeficientu β_i^j je v tomto případě $\bar{Y}_{i\bullet}$, pro $j = 0, 1$. Všimněme si, že to neznamená, že jednotlivé hodnoty X a Y jsou nezávislé na úrovni celé populace, neboť obecně neplatí rovnost $\mathrm{P}(Y = 1|X = j) = \mathrm{P}(Y = 1)$.

Vyjdeme-li z (2.5) a (2.6) dostáváme odhady populačních pravděpodobností jako

$$\hat{\beta}^0 = \frac{\sum_{i=1}^m n_i \bar{Y}_{i\bullet}(1 - \bar{x}_{i\bullet})}{\sum_{i=1}^m n_i(1 - \bar{x}_{i\bullet})} \tag{2.41}$$

a

$$\hat{\beta}^1 = \frac{\sum_{i=1}^m n_i \bar{Y}_{i\bullet}\bar{x}_{i\bullet}}{\sum_{i=1}^m n_i \bar{x}_{i\bullet}}. \tag{2.42}$$

Tyto odhady $\hat{\beta}^1$ a $\hat{\beta}^0$ nejsou obecně stejné jako celkový průměr $\bar{Y}_{\bullet\bullet}$, nemusí být ani navzájem shodné.
Tento přístup vytvořený Freedmanem a kol. [10] nazýváme *sousedským modelem*, ale raději zůstaneme u anglického názvu neighbourhood model. Myšlenka tohoto modelu spočívá v představě, že agregovaná vysvětlující proměnná $\bar{x}_{i\bullet}$ nedává žádnou informaci o chování agregované vysvětlované proměnné $\bar{Y}_{i\bullet}$. Toto chování se ovšem předpokládá pouze pro agregovaná data a nemusí platit pro jednotlivá pozorování Y_{it} a X_{it}. Jedná se o logický doplněk chování proměnných vůči jednoduchému Goodmanově modelu. Neighbourhood model je speciálním případem Goodmanova kvadratického modelu (2.26) s parametry, pro která navíc platí, že $\lambda_1 = \eta_1$ a $\lambda_0 = \eta_0$.

Tento model byl do značné míry kritizován Kingem [20] pro jeho předpoklad (2.40), který je, jak sám autor tvrdí, bezdůvodný a netestovatelný. Avšak na reálných datech v [16] dává tento model lepší výsledky (při znalosti skutečných hodnot) než velice sofistikovaný Kingův model, o kterém bude řeč v další kapitole. Obtížné ověření předpokladů platí pro oba modely.

Budeme-li zkoumat odhady polarizace p, dojdeme k závěru, že polarizace je v případě sousedského modelu menší než u standardní (Goodmanovy) ekologické regrese (viz [3]).

Tvrzení 2.9. *Nechť $\tilde{p}$ je odhad polarizace pomocí sousedského modelu a nechť $\hat{p}$ je odhad polarizace pomocí jednoduchého Goodmanova modelu* (2.15). *Potom platí, že*

$$\tilde{p} = \tau\hat{p},$$

kde $0 < \tau \leq 1$.

Důkaz. Označme $T_{xy} = \sum_{i=1}^m n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})$ a $T_{xx} = \sum_{i=1}^m n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2$.

Dosadíme za $\tilde{p}$ rozdíl (2.41) a (2.42)

$$\begin{aligned}
\tilde{p} &= \frac{\sum_{i=1}^{m} n_i \bar{Y}_{i\bullet} \bar{x}_{i\bullet}}{\sum_{i=1}^{m} n_i \bar{x}_{i\bullet}} - \frac{\sum_{i=1}^{m} n_i \bar{Y}_{i\bullet}(1-\bar{x}_{i\bullet})}{\sum_{i=1}^{m} n_i (1-\bar{x}_{i\bullet})} \\
&= \frac{(1-\bar{x}_{\bullet\bullet})T_{xy}/n + \bar{x}_{\bullet\bullet}\bar{Y}_{\bullet\bullet}(1-\bar{x}_{\bullet\bullet})}{\bar{x}_{\bullet\bullet}(1-\bar{x}_{\bullet\bullet})} - \frac{\bar{x}_{\bullet\bullet}\bar{Y}_{\bullet\bullet} - \bar{x}_{\bullet\bullet}T_{xy}/n - \bar{x}_{\bullet\bullet}^2\bar{Y}_{\bullet\bullet}}{\bar{x}_{\bullet\bullet}(1-\bar{x}_{\bullet\bullet})} \\
&= \frac{T_{xy}}{n\bar{x}_{\bullet\bullet}(1-\bar{x}_{\bullet\bullet})} \\
&= \frac{T_{xx}}{n\bar{x}_{\bullet\bullet}(1-\bar{x}_{\bullet\bullet})}\hat{p}.
\end{aligned}$$

Výraz před $\hat{p}$ pak odpovídá τ. Nerovnost $\tau > 0$ platí neboť předpokládáme $T_{xx} > 0$, tj. existenci heterogenních souborů. Druhou stranu nerovnosti dostaneme po úpravě výrazu

$$\frac{T_{xx}}{n\bar{x}_{\bullet\bullet}(1-\bar{x}_{\bullet\bullet})} = \frac{\sum_{i=1}^{m} n_i \bar{x}_{i\bullet}^2 - n\bar{x}_{\bullet\bullet}^2}{\sum_{i=1}^{m} n_i \bar{x}_{i\bullet} - n\bar{x}_{\bullet\bullet}^2}.$$

Vzhledem k tomu, že $0 \leq \bar{x}_{i\bullet} \leq 1$, platí $\sum_{i=1}^{m} n_i \bar{x}_{i\bullet}^2 \leq \sum_{i=1}^{m} n_i \bar{x}_{i\bullet}$. Tím je dokázána druhá část nerovnosti. □

Důležitou vlastností tohoto přístupu je automatické splnění omezení (2.34) a (2.35). Samozřejmě, předpoklad nezávislosti uvnitř souborů je obtížné zdůvodnit obecně. Nicméně, může být rozumné předpokládat, že neznámé hodnoty β_i^1 a $\bar{Y}_{i\bullet}$ jsou přibližně úměrné. Předpokládejme stejně jako v [16], že tento vztah proporcionality platí ve střední hodnotě, to jest

$$\mathrm{E}\hat{\beta}_i^1 = \beta_i^1 = \theta \mathrm{E}\bar{Y}_{i\bullet}, \tag{2.43}$$

kde parametr $\theta > 0$. Zmíněný sousedský model předpokládá $\theta = 1$, avšak obecně tomu tak být nemusí. Dále budeme uvažovat omezení na odhad parametru θ, které odpovídá logice použité v (2.34) (nebereme v potaz degenerovaný případ):

$$\theta_i^L \leq \hat{\theta} \leq \theta_i^U, \qquad i = 1, \ldots, m,$$

kde jsme si označili θ_i^L a θ_i^L jakožto

$$\theta_i^L = \max\left(0, \frac{1}{\bar{Y}_{i\bullet}} - \frac{1}{\bar{x}_{i\bullet}} - \frac{1}{\bar{x}_{i\bullet}\bar{Y}_{i\bullet}}\right)$$

a

$$\theta_i^U = \min\left(\frac{1}{\bar{x}_{i\bullet}}, \frac{1}{\bar{Y}_{i\bullet}}\right), \qquad i = 1, \ldots, m.$$

Vzhledem k tomu, že toto omezení má platit pro $i = 1, \ldots, m$, musí také platit

$$\max_i(\theta_i^L) = \theta^L \leq \hat{\theta} \leq \theta^U = \min_i(\theta_i^U) \tag{2.44}$$

Není těžké dokázat, že hodnota 1 leží v intervalu $[\theta^L, \theta^U]$, který můžeme chápat jako intervalový odhad parametru θ. K získání bodového odhadu parametru θ, je nutné přidat další podmínku. Například můžeme parametr θ chápat jako náhodnou veličinu, potom ovšem platí (2.43) s určitou změnou

$$\beta_i^1 = \mathrm{E}[\theta \bar{Y}_{i\bullet}].$$

Nemáme-li žádnou apriorní informaci o rozdělení parametru θ, zdá se rozumné předpokládat, že všechny hodnoty ve výše uvedeném intervalu $[\theta^L, \theta^U]$ jsou stejně pravděpodobné. V tomto případě máme odhad $\hat{\theta} = (\theta^U + \theta^L)/2$. Z toho plynoucí odhady parametrů β_i^1 a β_i^0 jsou

$$\hat{\beta}_i^1 = \hat{\theta}\bar{Y}_{i\bullet}$$

a

$$\hat{\beta}_i^0 = \frac{\bar{Y}_{i\bullet}(1 - \hat{\theta}\bar{x}_{i\bullet})}{1 - \bar{x}_{i\bullet}}.$$

Celkové odhady dostaneme pak z předpisu

$$\hat{\beta}^1 = \hat{\theta}\frac{\sum_{i=1}^m n_i \bar{Y}_{i\bullet}\bar{x}_{i\bullet}}{\sum_{i=1}^m n_i \bar{x}_{i\bullet}} \tag{2.45}$$

a

$$\hat{\beta}^0 = \frac{\sum_{i=1}^m n_i \bar{Y}_{i\bullet}(1 - \hat{\theta}\bar{x}_{i\bullet})}{\sum_{i=1}^m n_i (1 - \bar{x}_{i\bullet})}. \tag{2.46}$$

Vzhledem k definici omezení θ^L, θ^U a rozdělení θ, nabízí se zde velký prostor pro další modifikace. Jednou z možných změn je právě změna rozdělení. Neboť toto rozdělení je jen expertním odhadem, který není podložen žádnými relevantními údaji. Tak se může velice snadno stát, že naše odhady budou vychýlené. Nicméně je zde možnost jistého vylepšení a to pomocí výběrových kvantilů, které zužitkují informace o pořadí θ_i^L a θ_i^U a pomohou nám vyloučit vzdálená pozorování, které vychylují náš odhad. Můžeme tak zobecnit nerovnosti (2.44) do podoby

$$\text{kvantil}_{1-\alpha}(\theta_i^L) = \theta^L(\alpha) \leq \theta \leq \theta^U(\alpha) = \text{kvantil}_\alpha(\theta_i^U), \quad \alpha < \frac{1}{2}.$$

Platí také, že $[\theta^L(\alpha), \theta^U(\alpha)] \subseteq [\theta^L, \theta^U]$ pro $\alpha < \frac{1}{2}$. Jako bodový odhad $\hat{\theta}(\alpha)$ použijeme střed intervalu $[\theta^L(\alpha), \theta^U(\alpha)]$. Odhady celosouborových parametrů β^1 a β^0 pomocí (2.45) a (2.46), kde použijeme odhad $\hat{\theta}(\alpha)$ místo $\hat{\theta}$, budeme označovat jako odhady pomocí *theta modelu* s parametrem alfa.

Vzhledem ke konečné populaci (tj. $N < \infty$) bude platit, že pro $\alpha \to 0$ platí $\hat{\theta}(\alpha) \to \hat{\theta}$. Tímto zobecněním vzniká problém určení správné hodnoty α. Vzhledem k nevelkému množství dat je vhodné volit α blízké nule, pro vyloučení pouze vzdálených pozorování. Bude-li však α daleko od nuly, dá se očekávat nárůst vychýlení. Za zmínku také stojí fakt, že přechodem na výběrové kvantily už není splněna podmínka polohy odhadu koeficientů β_i^1 a β_i^0 uvnitř hranic dle (2.34), resp. (2.35), pro všechna $i = 1, \ldots, m$. Možností, jak toto napravit, je počítat odhad $\hat{b}_i^1$ a $\hat{b}_i^0$ pomocí $\tilde{\theta}_i = \frac{1}{2}(\theta_i^L + \theta_i^U)$, pokud $\hat{\theta} < \theta_i^L$ nebo $\hat{\theta} > \theta_i^U$.

2.10 Kingův přístup

Samostatnou kapitolou je přístup uvažovaný v [18] a [20]. Autor se snaží vytvořit takový model, který má ověřitelné předpoklady, které jsou zároveň co nejměkčí a přitom dostatečné. Navíc se snaží vytvořit model, který je dostatečně robustní a uplatnitelný.

V této části budeme pracovat pouze s pozorováními na úrovni jednotlivých souborů a proto použijeme zjednodušené značení. Nechť Y_i a X_i, $i = 1, \ldots, m$, jsou náhodné veličiny odpovídající $\bar{Y}_{i\bullet}$, $i = 1, \ldots, m$, resp. $\bar{X}_{i\bullet}$, $i = 1, \ldots, m$, z předchozí části práce. Nechť x_i, $i = 1, \ldots, m$, jsou nenáhodné konstanty odpovídající $\bar{x}_{i\bullet}$, $i = 1, \ldots, m$. Stále předpokládáme, že Y_i a x_i nabývají hodnot z intervalu $[0, 1]$.

Nechť parametry β_i^1 a β_i^0 pocházejí zatím z nespecifikovaného dvourozměrného rozdělení s nosičem na intervalu $[0, 1]^2$. Nechť dále platí deterministický vztah

$$Y_i = \beta_i^1 x_i + \beta_i^0 (1 - x_i), \qquad i = 1, \ldots, m. \tag{2.47}$$

Jedná se o deterministické vyjádření vysvětlované proměnné. Chybová složka je zde nahrazena náhodnými parametry β_i^1 a β_i^0. Nyní k tomuto modelu přidáme další tři předpoklady.

Předpoklad (1) Budeme předpokládat, že β_i^1 a β_i^0 pochází z useknutého dvourozměrného normální rozdělení. Tedy než abychom předpokládali konstantní chování parametrů napříč soubory, viz Goodmanův model, budeme pouze předpokládat, že mají stejné rozdělení. Dále je zde umožněna i korelace obou parametrů v rámci souborů. Formálně předpokládáme, že $\boldsymbol{\beta_i} = (\beta_i^1, \beta_i^0)'$ má rozdělení

$$\boldsymbol{\beta_i} \sim \mathrm{TN}^*(\mathfrak{B}, \boldsymbol{\Sigma}), \tag{2.48}$$

kde zkratkou TN* značíme useknuté dvourozměrné normální rozdělení (*truncated normal distribution*) s useknutím na $[0, 1]^2$, tj. platí, že hustota je mimo interval $[0, 1]^2$ nulová[6]. Parametry TN* jsou vektor středních hodnot

$$\mathfrak{B} = \begin{pmatrix} \mathfrak{B}^1 \\ \mathfrak{B}^0 \end{pmatrix}$$

a rozptylová matice

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{10} \\ \sigma_{10} & \sigma_{00} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{10} \\ \sigma_{10} & \sigma_0^2 \end{pmatrix}.$$

Tato parametrizace není obvyklá, avšak pomáhá objasnit některé další kroky. Standardní parametrizace je zavedena v (2.54).

Účelem tohoto modelu je určení všech parametrů β_i^1 a β_i^0 pro $i = 1, \ldots, m$. Parametry $\mathfrak{B}$ a $\boldsymbol{\Sigma}$ nejsou přímým bodem zájmu. Připomeňme si, že β_i^j jsou pravděpodobnosti $\mathsf{P}(Y_{it} = 1 | X_{it} = j)$, kde $j = 0, 1$, potom $\mathfrak{B}^j$ můžeme chápat jakožto střední hodnotu neváženého průměru definovaného v (2.9). Přidání střední hodnoty je nutné, neboť v (2.9) pracujeme s nenáhodnými parametry a zde mluvíme

[6] Useknutí tvoří další parametry rozdělení. Jelikož se v tomto případě bude vždy jednat o omezení se na jednotkový čtverec, nebudeme tento parametr přidávat do značení.

o náhodných parametrech.

Předpoklad (2) Předpokládejme, že vektor $\boldsymbol{\beta_i} = (\beta_i^1, \beta_i^0)'$ je ve střední hodnotě nezávislý na x_i, tj.

$$\mathsf{E}(\boldsymbol{\beta_i}|X_i = x_i) = \mathsf{E}\,\boldsymbol{\beta_i}. \tag{2.49}$$

Tento předpoklad, který je ekvivalentní požadavku na neexistenci vychýlení v důsledku agregace, je velmi důležitý pro získání konzistentních odhadů $\mathfrak{B}$ a $\boldsymbol{\Sigma}$. Ačkoli, jak King [18] uvádí, není tento předpoklad vždy nutný a případné vychýlení může být identifikováno a vyrušeno.

Předpoklad (3) Třetím předpokladem je nezávislost Y_i v různých souborech, kdy Y_i jsou navzájem nezávislé po podmínění x_i. Předpokládáme tak, že parametry β_i^1 a β_i^0 nejsou prostorově závislé a že zde není přítomna žádná autokorelace.

Pro určení potřebných pěti parametrů z (2.48) si sepišme střední hodnoty a rozptyl sledovaných veličin β_i^1 a β_i^0

$$\mathsf{E}\begin{pmatrix}\beta_i^1\\ \beta_i^0\end{pmatrix} = \begin{pmatrix}\mathfrak{B}^1\\ \mathfrak{B}^0\end{pmatrix} = \mathfrak{B}, \qquad \mathrm{var}\begin{pmatrix}\beta_i^1\\ \beta_i^0\end{pmatrix} = \begin{pmatrix}\sigma_1^2 & \sigma_{10}\\ \sigma_{10} & \sigma_0^2\end{pmatrix} = \boldsymbol{\Sigma}$$

a jejich vztah k souborovým parametrům

$$\beta_i^j = \mathfrak{B}^j + \epsilon_i^j, \qquad j = 0, 1,$$

kde náhodná složka ϵ_i^j má nulovou střední hodnotu a konstantní rozptyl pro jednotlivé $j = 0, 1$. Dále předpokládáme nezávislost ϵ_i^j pro $i = 1, \ldots, m$ při pevné hodnotě j . Nyní dosadíme tyto rovnosti do modelu (2.47), díky čemuž dostaneme

$$Y_i = \mathfrak{B}^1 x_i + \mathfrak{B}^0(1 - x_i) + \epsilon_i, \tag{2.50}$$

kde

$$\epsilon_i = \epsilon_i^1 x_i + \epsilon_i^0(1 - x_i). \tag{2.51}$$

Platí-li předpoklady (1)-(3), je zřejmé, že $\mathsf{E}(\epsilon_i|x_i) = 0$, proto nyní můžeme naši rovnici upravit do tvaru

$$\mathsf{E}(Y_i|X_i = x_i) = \mathfrak{B}^1 x_i + \mathfrak{B}^0(1 - x_i). \tag{2.52}$$

Z toho vyplývá, že parametr $\mathfrak{B}$ z (2.52) můžeme odhadnout pomocí obdoby Goodmanovy regrese (2.13). Zbývající tři parametry $\mathrm{var}(\beta_i^1) = \sigma_1^2$, $\mathrm{var}(\beta_i^0) = \sigma_0^2$ a $\mathrm{cov}(\beta_i^1, \beta_i^0) = \sigma_{10}$ ovlivňují námi sledovaná agregovaná data skrze podmíněný rozptyl, který je heteroskedastický, tj. mění se s vysvětlující proměnnou x_i. Úpravou rovnice (2.51) dostáváme

$$\begin{aligned}\mathrm{var}(Y_i|X_i = x_i) &= \sigma_1^2 x_i^2 + \sigma_0^2(1 - x_i)^2 + \sigma_{10} 2x_i(1 - x_i)\\ &= \sigma_0^2 + (2\sigma_{10} - 2\sigma_0^2)x_i + (\sigma_1^2 + \sigma_0^2 - 2\sigma_{10})x_i^2.\end{aligned} \tag{2.53}$$

Místo parametru σ_{10} budeme při zápisu parametrů používat korelační koeficient $\rho = \frac{\sigma_{10}}{\sigma_1\sigma_0}$. Máme zde pětici parametrů modelu, které si společně označíme jakožto

$\Psi = (\mathfrak{B}^1, \mathfrak{B}^0, \sigma_1, \sigma_0, \rho)'$. Ačkoli toto nejsou parametry našeho hlavního zájmu, mají přímou interpretaci jakožto charakteristiky rozdělení β_i^1 a β_i^0. Bohužel se parametry Ψ těžko přímo upotřebí v početní části. Problém tkví v tom, že odhady parametrů Ψ mohou být určeny pouze implicitně, kdežto my potřebujeme explicitní vzorec. Uvažujme proto alternativní vektor parametrů $\breve{\Psi} = (\breve{\mathfrak{B}}^1, \breve{\mathfrak{B}}^0, \breve{\sigma}_1, \breve{\sigma}_0, \breve{\rho})'$. Tento vektor odpovídá parametrům dvourozměrného normálního rozdělení $\mathrm{N}(\breve{\Psi})$. Není zde žádné useknutí a tudíž $\breve{\boldsymbol{\mathfrak{B}}} = (\breve{\mathfrak{B}}^1, \breve{\mathfrak{B}}^0)'$ není omezen intervalem $[0,1]^2$, tak jako tomu je u (2.48).

Abychom vytvořili hustotu požadovaného useknutého dvourozměrného normálního rozdělení pro β_i^1 a β_i^0, musíme provést dvě operace na naší zatím neuseknuté hustotě normálního rozdělení. První z nich je vynulování hustoty mimo čtverec $[0,1]^2$ na nulu a druhým krokem je přeškálování hustoty tak, aby byla opět hustotou. Dostáváme tak následující přepis hustoty rozdělení TN jako

$$f_{\mathrm{TN}}(\beta_i^1, \beta_i^0 | \breve{\boldsymbol{\mathfrak{B}}}, \breve{\boldsymbol{\Sigma}}) = f_{\mathrm{N}}(\beta_i^1, \beta_i^0 | \breve{\boldsymbol{\mathfrak{B}}}, \breve{\boldsymbol{\Sigma}}) \frac{\mathbf{1}(\beta_i^1, \beta_i^0)}{R(\breve{\boldsymbol{\mathfrak{B}}}, \breve{\boldsymbol{\Sigma}})}, \tag{2.54}$$

kde

$$R(\breve{\boldsymbol{\mathfrak{B}}}, \breve{\boldsymbol{\Sigma}}) = \int_0^1 \int_0^1 f_{\mathrm{N}}(\beta_i^1, \beta_i^0 | \breve{\boldsymbol{\mathfrak{B}}}, \breve{\boldsymbol{\Sigma}}) \mathrm{d}\beta^1 \mathrm{d}\beta^0,$$

$$\mathbf{1}(\beta_i^1, \beta_i^0) = \begin{cases} 1 & \text{pokud } (\beta_i^1, \beta_i^0)' \in [0,1]^2 \\ 0 & \text{jinak} \end{cases}$$

a $f_{\mathrm{N}}(\cdot)$ je hustota dvourozměrného normálního rozdělení. Pro parametrizaci $\breve{\Psi}$ budeme používat značení

$$\boldsymbol{\beta_i} \sim \mathrm{TN}(\breve{\boldsymbol{\mathfrak{B}}}, \breve{\boldsymbol{\Sigma}}). \tag{2.55}$$

Vektory Ψ a $\breve{\Psi}$ nejsou obecně shodné, ale v případě, kdy $R(\breve{\boldsymbol{\mathfrak{B}}}, \breve{\boldsymbol{\Sigma}})$ je blízko jedné, jsou jednotlivé parametry numericky blízko. K tomu dochází například tehdy, když $\breve{\mathfrak{B}}^1 = \breve{\mathfrak{B}}^0 = 0{,}5$ a jednotlivé rozptyly jsou dostatečně malé. Přesný vztah mezi Ψ a $\breve{\Psi}$ nelze jednoduše analyticky vyjádřit, ale lze jej určit pomocí simulací. Následující tvrzení, které čerpá z [22], dává alespoň základní vztah mezi $\mathfrak{B}$ a $\breve{\Psi}$.

Tvrzení 2.11. *Nechť $\boldsymbol{\beta_i}$ je náhodná veličina s rozdělením* (2.55)*, resp. při alternativním značení s rozdělením* (2.48)*. Vztah mezi parametry $\boldsymbol{\mathfrak{B}}$ a $\breve{\boldsymbol{\mathfrak{B}}}$ je dán následujícím předpisem*

$$\mathfrak{B}^j = \breve{\mathfrak{B}}^j + \breve{\sigma}_{j1}(f_1(0) - f_1(1)) + \breve{\sigma}_{j0}(f_0(0) - f_0(1)), \qquad j = 0, 1,$$

kde $f_j(x)$ je marginální hustota veličiny β_i^j a $\breve{\sigma}_{j1}$ je příslušná buňka matice $\breve{\boldsymbol{\Sigma}}$.

Důkaz. Důkaz je veden přes momentovou vytvořující funkci, které je v případě TN tvaru

$$\begin{aligned} m(\boldsymbol{t}) &= \mathsf{E}\left(e^{\boldsymbol{t}'\boldsymbol{\beta_i}}\right) \\ &= \int_{\mathbf{0}}^{\mathbf{1}} e^{\boldsymbol{t}'\boldsymbol{x}} f_{\mathrm{TN}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2\pi |\breve{\Sigma}|^{1/2} R(\breve{\boldsymbol{\mathfrak{B}}}, \breve{\boldsymbol{\Sigma}})} \int_{\mathbf{0}}^{\mathbf{1}} \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{x} - \breve{\boldsymbol{\mathfrak{B}}})' \breve{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \breve{\boldsymbol{\mathfrak{B}}}) - 2\boldsymbol{t}'\boldsymbol{x} \right] \right\} \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{2.56}$$

kde $\boldsymbol{t} = (t_1, t_0)'$, $\boldsymbol{x} = (x_1, x_0)'$ a $\int_{\mathbf{0}}^{\mathbf{1}}$ je dvojný integrál.

Předpokládejme nejprve, že $\breve{\boldsymbol{\mathfrak{B}}} = \mathbf{0}$. Exponent uvnitř integrálu upravíme do tvaru

$$\frac{1}{2}\boldsymbol{t}'\breve{\boldsymbol{\Sigma}}\boldsymbol{t} - \frac{1}{2}\left[(\boldsymbol{x} - \breve{\boldsymbol{\Sigma}}\boldsymbol{t})'\breve{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \breve{\boldsymbol{\Sigma}}\boldsymbol{t})\right],$$

Označme si $T = \frac{1}{2}\boldsymbol{t}'\breve{\boldsymbol{\Sigma}}\boldsymbol{t}$ a

$$\boldsymbol{\Phi}_{\breve{\Psi}} = \frac{1}{2\pi|\breve{\Sigma}|^{1/2}R(\breve{\mathfrak{B}}, \breve{\boldsymbol{\Sigma}})}\int_{-\breve{\boldsymbol{\Sigma}}\boldsymbol{t}}^{\mathbf{1}-\breve{\boldsymbol{\Sigma}}\boldsymbol{t}} \exp\left\{-\frac{1}{2}\boldsymbol{x}'\breve{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x}\right\}\mathrm{d}\boldsymbol{x}.$$

Dostáváme tak nový zápis pro momentovou vytvořující funkci jako

$$m(\boldsymbol{t}) = e^T\boldsymbol{\Phi}_{\breve{\Psi}}. \tag{2.57}$$

Nyní budeme derivovat výraz (2.57) podle t_j, kde $j = 1, 0$.

$$\frac{\partial m(\boldsymbol{t})}{\partial t_j} = e^T\frac{\partial \boldsymbol{\Phi}_{\breve{\Psi}}}{\partial t_j} + \boldsymbol{\Phi}_{\breve{\Psi}}\frac{\partial e^T}{\partial t_j}. \tag{2.58}$$

Jednotlivé výrazy se dále dají přepsat do tvaru

$$\frac{\partial e^T}{\partial t_j} = e^T(\breve{\sigma}_{j1}t_1 + \breve{\sigma}_{j0}t_0) \tag{2.59}$$

a

$$\begin{aligned}\frac{\partial \boldsymbol{\Phi}_{\breve{\Psi}}}{\partial t_j} &= \frac{\partial}{\partial t_j}\int_{a_0^*}^{b_0^*}\int_{a_1^*}^{b_1^*} f_{\mathrm{TN}}\mathrm{d}x_1\mathrm{d}x_0 \\ &= \breve{\sigma}_{j1}(f_1(a_1^*) - f_1(b_1^*)) + \breve{\sigma}_{j0}(f_1(a_0^*) - f_1(b_0^*)),\end{aligned} \tag{2.60}$$

kde $a_j^* = -(\breve{\sigma}_{j1}t_1 + \breve{\sigma}_{j0}t_0)$ a $b_j^* = 1 - (\breve{\sigma}_{j1}t_1 + \breve{\sigma}_{j0}t_0)$ pro $j = 0, 1$ a

$$f_1(x) = \int_{a_0^*}^{b_0^*} f_{TN}(x, y)\mathrm{d}y,$$

analogicky pro $f_0(x)$.

Pro $\boldsymbol{t} = \mathbf{0}$ máme $\boldsymbol{a^*} = \mathbf{0}$ a $\boldsymbol{b^*} = \mathbf{1}$. Tudíž $f_j(x)$ je marginální hustota jedné z veličin vektoru $\boldsymbol{\beta}$. Z (2.58), (2.59) a (2.60) dostáváme první moment rozdělení TN jako

$$\mathsf{E}(\beta_i^j) = \frac{\partial m(\boldsymbol{t})}{\partial t_j}|_{\boldsymbol{t}=\mathbf{0}} = \breve{\sigma}_{j1}(f_1(0) - f_1(1)) + \breve{\sigma}_{j0}(f_0(0) - f_0(1)).$$

Je-li $\breve{\boldsymbol{\mathfrak{B}}} \neq \mathbf{0}$, stačí si uvědomit, že $\mathsf{E}(\boldsymbol{\beta_i} + \breve{\boldsymbol{\mathfrak{B}}}) = \mathsf{E}\,\boldsymbol{\beta_i} + \breve{\boldsymbol{\mathfrak{B}}}$.

Je dobrá si uvědomit, že TN* a TN jsou dva různé zápisy téhož a jelikož platí, že $\boldsymbol{\mathfrak{B}} = \mathsf{E}\,\boldsymbol{\beta_i}$, je tvrzení dokázáno. □

Nyní si definujeme speciální popisné vlastnosti pro veličiny pocházející z useknutého rozdělení. *Upravenou střední hodnotou* náhodné veličiny, která má useknuté rozdělení, budeme chápat střední hodnotu náhodné veličiny, která má původní

neuseknuté rozdělení, tyto rozdělení se tak liší pouze v useknutí. Tuto nepravou střední hodnotu budeme značit $\breve{\mathsf{E}}(\cdot)$. Dále *upraveným rozptylem* náhodné veličiny budeme rozumět rozptyl náhodné veličiny, která pochází z původního neuseknutého rozdělení, značit ho budeme jako $\breve{\text{var}}(\cdot)$[7]. Tudíž platí, že $\breve{\mathsf{E}}\beta_i^1 = \breve{\mathfrak{B}}^1$ a $\breve{\text{var}}(\beta_i^1) = \breve{\boldsymbol{\sigma}}_1^2$. V situaci, kdy je toto useknutí malé, platí, že $\breve{\mathsf{E}}(\cdot) \doteq \mathsf{E}(\cdot)$ a $\breve{\text{var}}(\cdot) \doteq \text{var}(\cdot)$.

Paralelně k (2.52) a (2.53) dostáváme, že

$$\breve{\mathsf{E}}(Y_i|x_i) = \mu_i = \breve{\mathfrak{B}}^1 x_i + \breve{\mathfrak{B}}^0(1 - x_i) \tag{2.61}$$

a

$$\begin{aligned} \breve{\text{var}}(Y_i|x_i) &= \sigma_i^2 \\ &= \breve{\sigma}_1^2 x_i^2 + \breve{\sigma}_0^2(1 - x_i)^2 + \breve{\sigma}_{10} 2x_i(1 - x_i) \\ &= \breve{\sigma}_0^2 + (2\breve{\sigma}_{10} - 2\breve{\sigma}_0^2)x_i^2 + (\breve{\sigma}_1^2 + \breve{\sigma}_0^2 - 2\breve{\sigma}_{10})x_i^2. \end{aligned} \tag{2.62}$$

Parametry μ_i a σ_i^2 mají v případě nízkého stupně useknutí (tj. $R(\breve{\mathfrak{B}}, \breve{\boldsymbol{\Sigma}})$ je blízko jedné) význam podmíněné střední hodnoty a podmíněného rozptylu. Je-li však useknutí silnější, pak tato interpretace je nesprávná.

Nyní ponechme stranou fakt, že neznáme hodnoty parametrů $\breve{\Psi}$ a snažme se určit rozdělení parametrů zájmu β_i^1 a β_i^0. De facto stačí určit pouze jeden z dvojice, neboť druhý vyplyne z tautologie (2.47). Zajímá nás tudíž pouze hustota $f(\beta_i^1|T_i, \breve{\Psi})$. Tvar této hustoty nám dává následující tvrzení.

Tvrzení 2.12. *Za platnosti předpokladů (1)-(3) má neznámý parametr β_i^1 pro $x_i > 0$ následující podmíněnou hustotu*

$$\begin{aligned} f(\beta_i^1|Y_i, \breve{\Psi}) &= f_{TN}\left(\beta_i^1|\breve{\mathfrak{B}}^1 + \frac{\omega_i}{\sigma_i^2}\epsilon_i, \sigma_1^2 - \frac{\omega_i^2}{\sigma_i^2}\right) \\ &= f_N\left(\beta_i^1|\breve{\mathfrak{B}}^1 + \frac{\omega_i}{\sigma_i^2}\epsilon_i, \sigma_1^2 - \frac{\omega_i^2}{\sigma_i^2}\right) \frac{\boldsymbol{1}(\beta_i^1|Y_i)}{S(\breve{\mathfrak{B}}, \breve{\boldsymbol{\Sigma}})}, \end{aligned} \tag{2.63}$$

kde

$$\omega_i = \sigma_1^2 x_i + \sigma_{10}^2(1 - x_i), \tag{2.64}$$

$$\epsilon_i = Y_i - \breve{\mathfrak{B}}^1 x_i - \breve{\mathfrak{B}}^0(1 - x_i) \tag{2.65}$$

a σ_i^2 je definována rovnicí (2.62)*. Dále pak*

$$\boldsymbol{1}(\beta_i^1|Y_i) = \begin{cases} 1 & \textit{pokud } \beta_i^1 \in \left[\max\left(0, \frac{Y_i - (1-x_i)}{x_i}\right), \min\left(1, \frac{Y_i}{x_i}\right)\right] \\ 0 & \textit{jinak} \end{cases}$$

a

$$S(\breve{\mathfrak{B}}, \breve{\boldsymbol{\Sigma}}) = \int_{\max\left(0, \frac{Y_i-(1-x_i)}{x_i}\right)}^{\min\left(1, \frac{Y_i}{x_i}\right)} f_N\left(\beta_i^1|\breve{\mathfrak{B}}^1 + \frac{\omega_i}{\sigma_i^2}\epsilon_i, \sigma_1^2 - \frac{\omega_i^2}{\sigma_i^2}\right) \mathrm{d}\beta^1. \tag{2.66}$$

[7] Pro lepší pochopení uvažujme přepis (2.54). Náhodná veličina s hustotou $f_{\text{TN}}(\beta_i^1, \beta_i^0|\breve{\mathfrak{B}}, \breve{\boldsymbol{\Sigma}})$ získaná pomocí (2.54) z hustoty $f_{\text{N}}(\beta_i^1, \beta_i^0|\breve{\mathfrak{B}}, \breve{\boldsymbol{\Sigma}})$) bude mít upravenou střední hodnotu rovnu střední hodnotě náhodné veličiny s hustotou právě $f_{\text{TN}}(\beta_i^1, \beta_i^0|\breve{\mathfrak{B}}, \breve{\boldsymbol{\Sigma}})$. To samé platí pro upravený rozptyl.

Důkaz. Pro důkaz využijeme vlastnosti normálního rozdělení. Vycházíme z rozdělení hledaných parametrů β_i^1, β_i^0 dle (2.48). Pro jednoduchost nebudeme uvažovat useknuté dvourozměrné normální rozdělení, nýbrž neuseknuté normální rozdělení. Z (2.47) plyne, že Y_i je lineární kombinací β_i^1, β_i^1 a tudíž i Y_i, β_i^0 mají dvourozměrné normální rozdělení. Střední hodnota a rozptyl β_i^1 je $\mathfrak{B}^1$, resp. σ_1^2. Střední hodnotu a rozptyl Y_i určíme z (2.47), tedy

$$\mathsf{E}\, Y_i = \mu_i = \breve{\mathfrak{B}}^1 x_i + \breve{\mathfrak{B}}^0 (1 - x_i), \qquad \mathrm{var}(Y_i) = \sigma_i^2, \tag{2.67}$$

definovaného v rovnici (2.62). Kovariance Y_i a β_i^1 je rovna ω_i.

Nyní stačí spočítat podmíněné rozdělení β_i^1 při Y_i, které je také normálním rozdělením se střední hodnotou

$$\begin{aligned} \mathsf{E}(\beta_i^1 | Y_i) &= \mathsf{E}\, \beta_i^1 + \mathrm{corr}(\beta_i^1, Y_i) \sqrt{\frac{\mathrm{var}(\beta_i^1)}{\mathrm{var}(Y_i)}} (Y_i - \mathsf{E}\, Y_i) \\ &= \breve{\mathfrak{B}}^1 + \frac{\omega_i}{\sigma_i^2} \epsilon_i, \end{aligned}$$

a rozptylem

$$\begin{aligned} \mathrm{var}(\beta_i^1 | Y_i) &= \mathrm{var}(\beta_i^1)(1 - \mathrm{corr}(\beta_i^1, Y_i)^2) \\ &= \sigma_1^2 - \frac{\omega_i^2}{\sigma_i^2}. \end{aligned} \tag{2.68}$$

Useknutím hustoty parametru β_i^1 pomocí $\mathbf{1}(\beta_i^1 | Y_i)$ a úpravou zpět na hustotu pomocí $S(\breve{\mathfrak{B}}, \breve{\mathbf{\Sigma}})$ dostáváme (2.63).

□

Jakmile je rozdělení veličiny β_i^1 známé, spočítané, resp. odhadnuté a máme odhad tohoto parametru, respektive jeho podmíněnou střední hodnotu, pak můžeme dopočíst odhad β_i^0 přímo z identity

$$\hat{\beta}_i^0 = \frac{Y_i - \hat{\beta}_i^1 x_i}{1 - x_i}. \tag{2.69}$$

Tento postup je možné i otočit a nejprve odhadnout β_i^0 na základě střední hodnoty aposteriorního rozdělení. Ačkoli tento model obsahuje pouze pět parametrů $\breve{\Psi}$, odhadneme díky němu všech $2m$ parametrů β_i^1 a β_i^0, kde $i = 1, \ldots, m$.

Pro odhad neznámých pěti parametrů $\breve{\Psi}$ použijeme odhad metodou maximální věrohodnosti. Celá myšlenka tohoto odhadu je založena na informacích z agregovaných dat o parametrech zájmu β_i^1 a β_i^0 spolu s předpoklady modelu (1)-(3).

Pojmem *heterogenní* soubor chápeme soubor, pro který je hodnota x_i v intervalu $(0, 1)$. Jinak mluvíme o homogenních souborech.

Tvrzení 2.13. *Za předpokladu (1)-(3) pro heterogenní soubory je věrohodnostní funkce* $\mathrm{L}(\breve{\Psi}|\boldsymbol{y})$ *tvaru*

$$\begin{aligned}\mathrm{L}(\breve{\Psi}|(Y_1,\ldots,Y_m)') &\propto \prod_{x_i\in(0,1)} f(Y_i|\breve{\Psi}) \\ &= \prod_{x_i\in(0,1)} f_{\mathrm{N}}(Y_i|\mu_i,\sigma_i^2)\frac{S(\breve{\mathfrak{B}},\breve{\Sigma})}{R(\breve{\mathfrak{B}},\breve{\Sigma})}, \end{aligned} \tag{2.70}$$

kde μ_i *a* σ_i^2 *jsou definované dle* (2.61) *a* (2.62)*, dále*

$$R(\breve{\mathfrak{B}},\breve{\Sigma}) = \int_0^1\int_0^1 f_N(\beta_i^1,\beta_i^0|\breve{\mathfrak{B}},\breve{\Sigma})\mathrm{d}\beta^1\mathrm{d}\beta^0, \tag{2.71}$$

a

$$S(\breve{\mathfrak{B}},\breve{\Sigma}) = \int_{\max\left(0,\frac{Y_i-(1-x_i)}{x_i}\right)}^{\min\left(1,\frac{Y_i}{x_i}\right)} f_N\left(\beta_i^1|\breve{\mathfrak{B}}^1+\frac{\omega_i}{\sigma_i^2}\epsilon_i,\sigma_1^2-\frac{\omega_i^2}{\sigma_i^2}\right)\mathrm{d}\beta^1. \tag{2.72}$$

Důkaz. Vycházíme ze vztahu (2.47), který použijeme do (2.54). Platí, že podmíněné sdružené rozdělení dvojice $(\beta_i^1,Y_i)'$ je opět useknuté dvourozměrné normální rozdělení s hustotou tvaru

$$f(\beta_i^1,Y_i|\breve{\Psi}) = \frac{\mathbf{1}(\beta_i^1|Y_i)\mathbf{1}(Y_i)}{R(\breve{\mathfrak{B}},\breve{\Sigma})}f_{\mathrm{N}}\left(\beta_i^1,Y_i|\breve{\mathfrak{B}}^{\mathbf{1}},\mu_i,\breve{\sigma}_1^2,\sigma_i^2,\frac{\omega_i}{(\breve{\sigma}_1\sigma_i)}\right), \tag{2.73}$$

kde $\mathbf{1}(Y_i)$ je rovno jedné v případě, že $Y_i\in[0,1]$ a nula jinak. Dále

$$\mathbf{1}(\beta_i^1|Y_i) = \begin{cases} 1 & \text{pokud } \beta_i^1\in\left[\max\left(0,\frac{Y_i-(1-x_i)}{x_i}\right),\min\left(1,\frac{Y_i}{x_i}\right)\right], \\ 0 & \text{jinak.}\end{cases}$$

Nosičem hustoty je nyní kosočtverec. Parametry μ_i, σ_i a ω_i jsou definované v (2.61), (2.62) a (2.64). Z věty o podmíněné hustotě (viz Anděl [2] strana 56) dostáváme vyjádření

$$f(\beta_i^1,Y_i|\breve{\Psi}) \propto \frac{\mathbf{1}(Y_i)}{R(\breve{\mathfrak{B}},\breve{\Sigma})}f_{\mathrm{N}}\left(Y_i|\mu_i,\sigma_i^2\right)\times f_{\mathrm{N}}\left(\beta_i^1|\breve{\mathfrak{B}}+\frac{\omega_i}{\sigma_i^2}\epsilon_i,\sigma_1^2-\frac{\omega_i^2}{\sigma_i^2}\right)\mathbf{1}(\beta_i^1|Y_i).$$

Dalším krokem je integrace přes β_i^1, díky čemuž pak dostáváme požadovaný tvar

$$\begin{aligned} f(Y_i|\breve{\Psi},0<x_i<1) &\propto \int_{\max\left(0,\frac{Y_i-(1-x_i)}{x_i}\right)}^{\min\left(1,\frac{Y_i}{x_i}\right)} f((\beta^1,Y_i|\breve{\Psi})\mathrm{d}\beta_1 \\ &\propto \frac{\mathbf{1}(Y_i)}{R(\breve{\mathfrak{B}},\breve{\Sigma})}f_{\mathrm{N}}(Y_i|\mu_i,\sigma_i^2)S(\breve{\mathfrak{B}},\breve{\Sigma}), \end{aligned}$$

kde $S(\breve{\mathfrak{B}},\breve{\Sigma})$ je definované v (2.66). □

Podíl $S(\breve{\mathfrak{B}},\breve{\Sigma})/R(\breve{\mathfrak{B}},\breve{\Sigma})$ je tím, čím se liší tento maximálně věrohodný odhad od maximálně věrohodného odhadu normálního rozdělení. Pouze v případě, kdy není rozdělení β_i^1 a β_i^0 silně omezeno oseknutím na jednotkový čtverec, je tento podíl blízko jedné a může býti ignorován.

Poznámka 2.14. Pro homogenní soubory i také potřebujeme marginální hustotu Y_i z (2.73). Pokud $x_i = 1$, potom $Y_i = \beta_i^1$, resp. pokud $x_i = 0$, potom $Y_i = \beta_i^0$. Pro $x_i = 1$ tak dostáváme

$$\begin{aligned} f(Y_i|\breve{\Psi}, x_i = 1) &= \int_0^1 f_{\mathrm{TN}}((Y_i, \beta^0|\breve{\Psi})\mathrm{d}\beta_0 \\ &= \frac{\mathbf{1}(Y_i)}{R(\breve{\mathfrak{B}}, \breve{\mathbf{\Sigma}})} f_{\mathrm{N}}(Y_i|\breve{\mathfrak{B}}^1, \sigma_1^2) \\ &\quad \times \int_0^1 f_{\mathrm{N}}\left(\beta^0|\breve{\mathfrak{B}}^0 + \breve{\rho}\frac{\breve{\sigma}_0}{\breve{\sigma}_1}(Y_i - \breve{\mathfrak{B}}^1), \breve{\sigma}_0^2(1 - \breve{\rho}^2)\right)\mathrm{d}\beta^0. \end{aligned}$$

Analogicky pro případ $x_i = 0$ platí

$$\begin{aligned} f(Y_i|\breve{\Psi}, x_i = 0) &= \int_0^1 f_{\mathrm{TN}}((\beta^1, Y_i|\breve{\Psi})\mathrm{d}\beta_1 \\ &= \frac{\mathbf{1}(Y_i)}{R(\breve{\mathfrak{B}}, \breve{\mathbf{\Sigma}})} f_{\mathrm{N}}(Y_i|\breve{\mathfrak{B}}^0, \sigma_0^2) \\ &\quad \times \int_0^1 f_{\mathrm{N}}\left(\beta^1|\breve{\mathfrak{B}}^1 + \breve{\rho}\frac{\breve{\sigma}_1}{\breve{\sigma}_0}(Y_i - \breve{\mathfrak{B}}^0), \breve{\sigma}_1^2(1 - \breve{\rho}^2)\right)\mathrm{d}\beta^1. \end{aligned}$$

Vzhledem k tomu, že homogenní soubory určí jeden z odhadovaných parametrů β_i^1 a β_i^0 jednoznačně, neboť v případě, že $x_i = 0$, je přirozeným odhadem β_i^0 hodnota Y_i. Pak pro daný parametr není zapotřebí odhadovat jeho hustotu. Homogenní soubory jsou nicméně velmi užitečné při určení useknutého dvourozměrného normálního rozdělení, které může být použito ke zpřesnění odhadů hustot heterogenních souborů.

Pro usnadnění maximalizace věrohodnostní funkce je nejlepším řešením pracovat s parametry, které mají rozdělení podobné normálnímu rozdělení. Mluvíme zde o parametrech modelu a jejich apriorních hustotách normálního rozdělení. Tato úprava dále pomáhá předejít numerickým komplikacím při výpočtu (2.71). Věrohodnostní funkce je invariantní vůči reparametrizaci, tudíž tato změna nezpůsobí změnu polohy maxima. Provedeme následující transformace parametrů:

$$\begin{aligned} \phi_1 &= \frac{\breve{\mathfrak{B}}^1 - 0{,}5}{\breve{\sigma}_1^2 + 0{,}25} \\ \phi_2 &= \frac{\breve{\mathfrak{B}}^0 - 0{,}5}{\breve{\sigma}_0^2 + 0{,}25} \\ \phi_3 &= \ln(\breve{\sigma}_1) \\ \phi_4 &= \ln(\breve{\sigma}_0) \\ \phi_5 &= 0{,}5\ln\left(\frac{1 + \breve{\rho}}{1 - \breve{\rho}}\right), \end{aligned} \tag{2.74}$$

kde ϕ_5 je Fisherova Z transformace. Označme si nyní tuto již třetí parametrizaci jako $\Phi = (\phi_1, \phi_2, \phi_3, \phi_4, \phi_5)'$. Nutno podotknout, že žádná ze zmíněných parametrizací není předmětem našeho zájmu. V praxi jsou parametry Φ počítány (odhadovány) jako první, z nich jsou inverzní operací k (2.75) dopočteny parametry $\breve{\Psi}$.

Poznámka 2.15. Pro usnadnění výpočtů zvolíme vhodné apriorní rozdělení ϕ_3, ϕ_4 a ϕ_5. Jelikož máme množství informací přímo v datech, apriorní rozdělení pro ϕ_1 a ϕ_2 nejsou volena krom případu, kdy máme nějakou dodatečnou externí informaci. Jelikož β_i^1 a β_i^0 jsou omezené intervalem $(0, 1)$, rozptyly $\breve{\sigma}_1$ a $\breve{\sigma}_0$ jsou pouze výjimečně větší než 0,5. Proto použijeme dle [18] apriorní *half-normal* rozdělení[8] pro $\breve{\sigma}_1$ a $\breve{\sigma}_0$ s nulovou střední hodnotou a rozptylem rovným 0,5. Nejméně informací máme o parametru ϕ_5. Nicméně jako vhodné apriorní rozdělení se jeví normální s nulovou střední hodnotou a rozptylem 0,25.

Věrohodnostní funkce (2.70) (z Bayesova pohledu aposteriorní rozdělení) je použita ke shrnutí veškerých informací, které máme o parametrech Φ. Abychom tohoto dosáhli, maximalizujeme věrohodnostní funkci a získáme odhad $\hat{\Phi}$ s příslušnou rozptylovou maticí $\hat{\mathrm{V}}(\hat{\Phi})$. Jednou možností, jak shrnout znalosti o Φ, je pomocí normálního rozdělení. Díky reparametrizaci parametrů Ψ na Φ, kterou jsme použili za účelem získání odhadů, můžeme uvažovat o jejich normalitě. Proto použijeme aproximaci normálním rozdělením

$$\Phi \sim \mathrm{N}\left(\hat{\Phi}, \hat{\mathrm{V}}(\hat{\Phi})\right). \tag{2.75}$$

S rostoucím počtem opakování bude $\hat{\Phi}$ konvergovat ke konstantě a odhadnutý rozptyl $\hat{\mathrm{V}}(\hat{\Phi})$ bude konvergovat k nule. Jelikož je aposteriorní rozdělení z centrální limitní věty asymptoticky normální, je tato aproximace vhodná. Případné problémy s nenormalitou vyřešíme pomocí *importance sampling* algoritmu. Detailní popis importance sampling algoritmu je možné nalézt v [11].

Další postup nyní shrneme do několika bodů. Výsledkem bude získání odhadu $\tilde{\beta}_i^1$ na základě aposteriorního rozdělení $(\beta_i^1|Y_i)$ za daný soubor i.

1. Vygenerujeme si náhodný vektor $\tilde{\Phi}$ z rozdělení, které je dáno rovnicí (2.75), kde $\hat{\Phi}$ je bodový odhad parametru Φ, který maximalizuje věrohodnostní funkci (2.70) a $\hat{\mathrm{V}}(\hat{\Phi})$ je odhad jeho rozptylové matice.

2. Reparametrizujeme $\tilde{\Phi}$ pomocí (2.75). Získáme tak odhad parametrů $\breve{\Psi}$, které označíme jako $\tilde{\breve{\Psi}}$.

3. Provedeme *importance sampling*, jehož důsledek bude zlepšení normality.

   (a) Spočteme významový poměr (*importance ratio*), což je podíl hodnoty věrohodnostní funkce $\mathrm{L}(\breve{\Psi}|y)$ v bodě $\tilde{\breve{\Psi}}$ a hodnoty hustoty normální aproximace $f_{\mathrm{N}}(\tilde{\Phi}, \hat{\mathrm{V}}(\hat{\Phi}))$ opět v bodě $\tilde{\breve{\Psi}}$.

   (b) Akceptujeme $\tilde{\breve{\Psi}}$ s pravděpodobností, která odpovídá významovému poměru. To můžeme provést pomocí vygenerování čísla r z rovnoměrného rozdělení na intervalu [0,1], které porovnáme se sledovaným významovým poměrem. Pokud je vygenerované číslo r vyšší než náš poměr, vracíme se do kroku 1.

4. Dopočítáme odhady parametrů σ_i^2, ω_i a ϵ_i pomocí (2.62), (2.64), (2.65) a odhadu $\tilde{\breve{\Psi}}$.

[8] Toto rozdělení vzniká aplikací absolutní hodnoty na veličinu mající normální rozdělení.

5. Vložíme $\tilde{\tilde{\Psi}}$ do podmíněné aposteriorní hustoty veličiny β_i^1, tj. $f(\beta_i^1|Y_i, \tilde{\tilde{\Psi}})$ dle rovnice (2.63). Z tohoto aposteriorního rozdělení náhodně vygenerujeme odhad $\tilde{\beta}_i^1$.

6. Kroky 1 až 5 provádíme, dokud nebudeme mít požadovaných K hodnot $\tilde{\beta}_i^1$ pro každý soubor i. King [18] udává hodnotu $K = 100$, nicméně záleží na uživateli, jak jemný odhad aposteriorní hustoty parametru β_i^1 chce získat.

7. Bodový odhad veličiny β_i^1 získáme jako

$$\hat{\beta}_i^1 = \frac{1}{K} \sum_{k=1}^{K} \tilde{\beta}_i^{1(k)},$$

   kde $\tilde{\beta}_i^{1(k)}$ je k-tá simulace veličiny β_i^1 z kroku 4. Odhad směrodatné odchylky získáme jako

$$SE(\hat{\beta}_i^1) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (\tilde{\beta}_i^{1(k)} - \hat{\beta}_i^1)^2}.$$

   Upřednostňujeme intervalový odhad, který konstruujeme tak, že používáme výběrové kvantily nasimulovaných hodnot $\tilde{\beta}_i^{1(k)}$. Nemusí nás tak trápit případná šikmost odhadnutého aposteriorního rozdělení, která by mohla způsobit, že bodový odhad není uprostřed intervalového odhadu

8. Odhady veličiny β_i^0 získáme pomocí $\tilde{\beta}_i^{1(k)}$ a identity (2.69).

9. Odhady celosouborových parametrů β^1 a β^0 získáme jako

$$\hat{\beta}^1 = \frac{1}{K} \frac{\sum_{k=1}^{K} \sum_{i=1}^{m} n_i x_i \tilde{\beta}_i^{1(k)}}{\sum_{i=1}^{m} n_i x_i}, \tag{2.76}$$

$$\hat{\beta}^0 = \frac{1}{K} \frac{\sum_{k=1}^{K} \sum_{i=1}^{m} n_i (1 - x_i) \tilde{\beta}_i^{0(k)}}{\sum_{i=1}^{m} n_i (1 - x_i)}. \tag{2.77}$$

Jednotlivé MCMC metody používané v této kapitole jsou přehledně zpracovány například v [6], [33], nebo [11].

Celý algoritmus je naprogramován v programu R [29] a to v knihovně EI, která je ke stažení na adrese `http://gking.harvard.edu/eiR`.

2.16 Hierarchické modely

Tato kapitola vychází z [19], kap. 1, a diskuzí nad tímto modelem v [34] a [36].

Mějme za úkol určit hustotu neznámého rozdělení, jímž se řídí náhodný vektor Y. Tuto hustotu si označme jako $p(y|\theta)$, kde θ je neznámý parametr, resp. vektor parametrů. Tato hustota však může být velice obtížně parametrizována. Hierarchické modely přistupují ke konstrukci hustoty $p(y|\theta)$ v několika separátních krocích. Jako příklad si můžeme představit, že máme podmíněnou hustotu $p_1(y|\beta)$, jíž se řídí $Y|\beta$. Dále uvažujme, že β není konstantní parametr, ale má rozdělení s hustotou $p_2(\beta|\theta)$. Požadovanou hustotu $p(y|\theta)$ pak dostáváme jako kombinaci hustot $p_1(y|\beta)$ a $p_2(\beta|\theta)$, respektive

$$p(y|\theta) = \int_{-\infty}^{\infty} p_1(y|\beta)p_2(\beta|\theta)\mathrm{d}\beta. \tag{2.78}$$

Tato myšlenka je už dlouhou dobu známá, avšak donedávna bylo spočtení integrálu v (2.78) obtížné. V dnešní době to není problém a to díky *simulaci Monte Carlo*. K určení hustoty $p(y|\theta)$ stačí vygenerovat hodnotu $\tilde{\beta}$ z rozdělení s hustotou $p_2(\beta|\theta)$, následně náhodně vygenerujeme hodnotu $\tilde{y}$ z rozdělení s hustotou $p_1(y|\tilde{\beta})$. Histogram takto vygenerovaných hodnot $\tilde{y}$ pak velice dobře odhaduje hustotu $p(y|\theta)$.

2.16.1 Beta-binomický model

V první fázi předpokládáme, že $Y_{i\bullet}$ se řídí binomickým rozdělením $\mathrm{Bi}(n_i, p_i)$, kde $p_i = \beta_i^1\bar{x}_{i\bullet} + (1-\bar{x}_{i\bullet})\beta_i^0$, kde pro tento moment považujeme parametry β_i^1 a β_i^0 za konstantní. Věrohodnostní funkce tohoto rozdělení pro i-tý soubor při pevném $Y_{i\bullet}$ je tvaru

$$p(\beta_i^1, \beta_i^0|Y_{i\bullet}, n_i) \propto (\beta_i^1\bar{x}_{i\bullet} + (1-\bar{x}_{i\bullet})\beta_i^0)^{Y_{i\bullet}}(1-\beta_i^1\bar{x}_{i\bullet} - (1-\bar{x}_{i\bullet})\beta_i^0)^{n_i-Y_{i\bullet}}. \tag{2.79}$$

Maximálně věrohodným odhadem dvojice parametrů β_i^1, β_i^0 pro i-tý soubor je bod ležící na úsečce v jednotkovém čtverci, která je definovaná vztahem

$$\beta_i^0 = \left(\frac{y}{1-x}\right) - \left(\frac{x}{1-x}\right)\beta_i^1.$$

V druhé fázi modelu předpokládáme, že se parametr β_i^1 řídí beta rozdělením $\mathrm{B}(c_1, d_1)$, kde $c_1, d_1 > 0$. Parametr β_i^0 modelujeme pomocí beta rozdělení $\mathrm{B}(c_0, d_0)$, kde $c_0, d_0 > 0$. Díky tomuto předpokladu bude vždy platit, že $\beta_i^1 \in [0,1]$ s pravděpodobností jedna, obdobně pro β_i^0. Toto je změna oproti Kingovu předpokladu (1) z předcházející kapitoly o modelaci náhodného vektoru $\boldsymbol{\beta}_i$ pomocí useknutého dvourozměrného normálního rozdělení (2.48), tj. unimodálního rozdělení. Navíc předpokládáme apriorní nezávislost parametrů zájmu β_i^1 a β_i^0. Za užití *Gibbsova algoritmu* se však později ukáže aposteriorní závislost parametrů β_i^1 a β_i^0.

Ve třetí a poslední fázi předpokládáme, že se neznámé parametry c_1, d_1, c_0 a d_0 řídí exponenciálním rozdělením $\mathrm{Ex}(\lambda)$ se střední hodnotou $1/\lambda$. King v [19] volí tuto střední hodnotu rovnu 2.

Dle Bayesovy věty, viz [15], je aposteriorní rozdělení, až na normovací konstantu, rovno součinu věrohodnostní funkce a apriorní hustoty. Tudíž dostáváme

$$\begin{aligned}
&\pi(\beta_i^1, \beta_i^0, c_1, d_1, c_0, d_0|\text{data}) \\
&\quad\propto\ f(\text{data}|(\beta_i^1, \beta_i^0), i = 1, \ldots, m) \\
&\qquad \times f((\beta_i^1, \beta_i^0), i = 1, \ldots, m|c_1, d_1, c_0, d_0) \times \pi(c_1, d_1, c_0, d_0) \\
&\quad= \prod_{i=1}^{m} (\beta_i^1 \bar{x}_{i\bullet} + (1 - \bar{x}_{i\bullet})\beta_i^0)^{Y_{i\bullet}} (1 - \beta_i^1 \bar{x}_{i\bullet} - (1 - \bar{x}_{i\bullet})\beta_i^0)^{n_i - Y_{i\bullet}} \\
&\qquad \times \prod_{i=1}^{m} \frac{\Gamma(c_1 + d_1)}{\Gamma(c_1)\Gamma(d_1)} (\beta_i^1)^{c_1 - 1} (1 - \beta_i^1)^{d_1 - 1} \prod_{i=1}^{m} \frac{\Gamma(c_0 + d_0)}{\Gamma(c_0)\Gamma(d_0)} (\beta_i^0)^{c_0 - 1} (1 - \beta_i^0)^{d_0 - 1} \\
&\qquad \times \exp(-\lambda c_1) \times \exp(-\lambda d_1) \times \exp(-\lambda c_0) \times \exp(-\lambda d_0),
\end{aligned}$$

kde $f(\cdot|\cdot)$ jsou jednotlivé podmíněné hustoty z první až třetí fáze, $\pi(\cdot)$ je apriorní rozdělení a daty označujeme pozorování $\{(\bar{Y}_{i\bullet}, \bar{x}_{i\bullet}, n_i)', i = 1, \ldots, m\}$. Marginální hustoty aposteriorního rozdělení získáme pomocí Gibbsova algoritmu. Abychom mohli použít Gibbsův algoritmus, je zapotřebí si vyjádřit podmíněné hustoty každého neznámého parametru vůči ostatním parametrům.

$$\begin{aligned}
f(\beta_i^1|\beta_i^0, c_1, d_1) \quad\propto\quad & (\beta_i^1 \bar{x}_{i\bullet} + (1 - \bar{x}_{i\bullet})\beta_i^0)^{Y_{i\bullet}} (1 - \beta_i^1 \bar{x}_{i\bullet} - (1 - \bar{x}_{i\bullet})\beta_i^0)^{n_i - Y_{i\bullet}} \\
& \times (\beta_i^1)^{c_1 - 1} (1 - \beta_i^1)^{d_1 - 1},
\end{aligned}$$

$$\begin{aligned}
f(\beta_i^0|\beta_i^1, c_0, d_0) \quad\propto\quad & (\beta_i^1 \bar{x}_{i\bullet} + (1 - \bar{x}_{i\bullet})\beta_i^0)^{Y_{i\bullet}} (1 - \beta_i^1 \bar{x}_{i\bullet} - (1 - \bar{x}_{i\bullet})\beta_i^0)^{n_i - Y_{i\bullet}} \\
& \times (\beta_i^0)^{c_0 - 1} (1 - \beta_i^0)^{d_0 - 1},
\end{aligned}$$

$$f(c_1|\beta_i^1, i = 1, \ldots, m, d_1) \quad\propto\quad \left(\frac{\Gamma(c_1 + d_1)}{\Gamma(c_1)}\right)^m \exp\left\{\left(\sum_{i=1}^{m} \log \beta_i^1 - \lambda\right) c_1\right\},$$

$$f(d_1|\beta_i^1, i = 1, \ldots, m, c_1) \quad\propto\quad \left(\frac{\Gamma(c_1 + d_1)}{\Gamma(d_1)}\right)^m \exp\left\{\left(\sum_{i=1}^{m} \log(1 - \beta_i^1) - \lambda\right) d_1\right\},$$

$$f(c_0|\beta_i^0, i = 1, \ldots, m, d_0) \quad\propto\quad \left(\frac{\Gamma(c_0 + d_0)}{\Gamma(c_0)}\right)^m \exp\left\{\left(\sum_{i=1}^{m} \log \beta_i^0 - \lambda\right) c_0\right\},$$

$$f(d_0|\beta_i^0, i = 1, \ldots, m, c_0) \quad\propto\quad \left(\frac{\Gamma(c_0 + d_0)}{\Gamma(d_0)}\right)^m \exp\left\{\left(\sum_{i=1}^{m} \log(1 - \beta_i^0) - \lambda\right) d_0\right\}.$$

Nyní použijeme *Metropolisův algoritmus* (viz [6]) pro získání hodnot z požadovaných rozdělení. Gibbsův algoritmus nelze použít vzhledem k nestandardním hustotám jednotlivých parametrů. Pro parametry c_1, d_1, c_0 a d_0 získáme kandidátní hodnotu pro dalšího bod v rámci Metropolisova řetězce pomocí normálního rozdělení se střední hodnotou v aktuálním bodě a s dostatečně velkým rozptylem. Pro parametry β_i^1 a β_i^0 použijeme pro pohyb v řetězci rovnoměrné rozdělení. Potřebnou teorii k MCMC metodám je možné nalézt v [33].

2.16.2 Alternativní hierarchické modely

Volba rozdělení parametru $\boldsymbol{\beta}_i$ jako beta rozdělení z kapitoly 2.16.1 není jedinou možnou. V [20], kap. 1, je popsáno použití beta-binomického modelu na datech skládajících se z volebních obvodů, kde zkoumáme závislost převážně politického názoru, tj. volby parlamentní strany, na barvě pleti. Nicméně v epidemiologii se můžeme setkat například s Poissonovým rozdělením v případě zkoumání vzácných jevů (viz [34]). Předpokládáme, že pozorujeme vzácnou neinfekční nemoc, jejíž výskyt modelujeme jako

$$n_i^j|\mu_j, \delta_i \sim \mathrm{Po}(x_i^j \exp(\mu_j + \delta_i)), \qquad j = 0, 1, \qquad i = 1, \ldots, m,$$

kde $x_i^0 = n_i - x_{i\bullet}$ a $x_i^1 = x_{i\bullet}$ (viz tabulka 2.2). x_i^1 značí počet osob, jež byly v daném i-tém souboru vystaveny nemoci. Veličina n_i^j odpovídá značení dle tabulky 2.1 a pro $j = 1$ značí počet nemocných, kteří byli vystaveni nemoci v i-tém souboru. Parametry modelu μ_j a δ_i jsou neznámé. Parametru $\nu = \mu_1 - \mu_0$ říkáme *log-relativní riziko*. Předpokládáme, že ν je konstantní pro všechny soubory. S tímto parametrem se můžeme setkat také například v [21]. Agregací získáme

$$\bar{Y}_{i\bullet}|\mu_0, \nu, \delta_i \sim \mathrm{Po}(n_i \exp(\mu_0 + \delta_i)\{(1 - \bar{x}_{i\bullet}) + \bar{x}_{i\bullet} \exp(\nu)\}), \qquad i = 1, \ldots, m,$$

kde pro upřesnění $\bar{x}_{i\bullet}$ značí podíl osob, jež byly vystaveny nemoci.

Další možné přístupy nalezneme v [17]. Nechť platí předpoklad o nekorelovanosti $\boldsymbol{\beta}_i$ a x_i dle (2.23) a (2.24). Potom modelujeme veličinu $\boldsymbol{\beta}_i^* = (\beta_i^{1*}, \beta_i^{0*})' = (\mathrm{logit}(\beta_i^1), \mathrm{logit}(\beta_i^0))'$ jako

$$\boldsymbol{\beta}_i^*|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

kde $\boldsymbol{\mu}$ je vektor populačních průměrů a $\boldsymbol{\Sigma}$ je pozitivně definitní rozptylová matice. Parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ můžeme dále modelovat pomocí

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathrm{N}\left(\boldsymbol{\mu}_0, \frac{\boldsymbol{\Sigma})}{\tau_0^2}\right), \qquad \boldsymbol{\Sigma} \sim \mathrm{InvWish}(\nu_0, S_0^{-1}),$$

kde $\boldsymbol{\mu}_0$ je vektor apriorní střední hodnoty, τ_0 je skalár, ν_0 je apriorní počet stupňů volnosti inverzního Wishartova rozdělení, které je mnohorozměrnou obdobou chí-kvadrát rozdělení. Dále pak S_0 je pozitivně definitní apriorní matice. Tyto parametry nastavíme podle dostupných informací. Pokud žádné dostupné informace nemáme, doporučuje se volit $\boldsymbol{\mu}_0 = \mathbf{0}$, $S_0 = 10I_2$, kde I_2 je jednotková matice dimenze 2. Dále pak $\tau_0 = 2$ a $\nu_0 = 4$.

Toto rozdělení je možné nalézt i v [31], přehledněji pak v [35] a [36]. Další možností volby je například apriorní Dirichletův proces, který je použit v [17].

3. Aplikace

3.1 Simulace

Než přikročíme k aplikaci na reálných datech, zkusíme ověřit kvalitu odhadů jednotlivých popsaných modelů na třech typech simulovaných dat, kde budeme mít možnost porovnat pozorované vychýlení odhadů parametrů β_i^1 a β_i^0 a odmocninovou střední čtvercovou chybu RMSE.

V této kapitole budeme používat zjednodušené značení y_i a x_i.

Použijeme reálná data analyzovaná Kingem [18], kap. 10. Tato data jsou součástí knihovny EI[1] v softwaru R [29]. Data obsahují pozorované podíly registrovaných voličů, podíly občanů tmavé pleti a počty jedinců v jednotlivých správních obvodech (souborech) za státy Florida, Louisiana, Severní Karolína a Jižní Karolína. Celkem máme 268 pozorování. Z těchto dat použijeme pouze podíly občanů tmavé pleti, které si označíme jako x_i a počty jedinců n_i.

K simulaci různých nastavení pravděpodobností β_i^1 a β_i^0 pro jednotlivé správní obvody použijeme generátor normálního rozdělení.

- **Simulace I.** Generujeme vektory $(\beta_i^1, \beta_i^0)'$ jako nezávislé realizace z dvourozměrného normálního rozdělení $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde

$$\boldsymbol{\mu} = \begin{pmatrix} 0{,}7 \\ 0{,}3 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 0{,}020 & 0{,}005 \\ 0{,}005 & 0{,}020 \end{pmatrix},$$

  přičemž hodnoty mimo interval $[0, 1]^2$ upravíme na nejbližší hodnotu tohoto intervalu[2]. Pozorované průměry β_i^1 a β_i^0 nejsou v tomto případě vzdáleny od uvažované střední hodnoty.

- **Simulace II.** Generujeme β_i^1 a β_i^0 jako nezávislé realizace ze směsi dvou dvourozměrných normálních rozdělení s pravděpodobností 0,4, že realizace pochází z rozdělení z předchozí simulace a s pravděpodobností 0,6, že realizace pochází z dvourozměrného normálního rozdělení s vektorem středních hodnot $(0{,}2, 0{,}8)'$, rozptyly $(0{,}015, 0{,}010)'$ a nulovou kovariancí. Pozorované průměry β_i^1 a β_i^0 vychází 0,4, resp. 0,6.

- **Simulace III.** Generujeme β_i^0 a β_i^1 jako realizace pocházející z dvourozměrného normálního rozdělení s vektorem středních hodnot

$$(0{,}7 + 0{,}8(x_i - \bar{x}), 0{,}3 + 0{,}5(x_i - \bar{x}))'$$

  a rozptylovou maticí stejnou jako v simulaci I. Pozorované průměry β_i^1 a β_i^0 vychází 0,6, resp. 0,4. Oproti předchozím případům je zde značná korelace mezi β_i^1 a x_i, resp. β_i^0 a x_i, konkrétně $(0{,}7, 0{,}3)'$.

[1] URL: `http://gking.harvard.edu/eiR`

[2] Alternativním postupem k získání parametrů β_i^1 a β_i^0 by bylo generování $(\beta_i^{1*}, \beta_i^{0*})'$ z dvourozměrného normálního rozdělení s neomezenými parametry, na tyto hodnoty následně aplikujeme logistickou funkci.

Z takto získaných pravděpodobností volební účasti pro občany tmavé pleti a pro ostatní získáme pomocí vztahu (2.3) veličinu y_i. Dále budeme považovat y_i, x_i a n_i za známé a budeme odhadovat parametry β_i^1 a β_i^0. Označme si jejich odhady jako b_i^1 a b_i^0. Pro jednotlivé odhady definujeme pozorované vychýlení jako $\text{Bias}^j = \frac{1}{n}\sum_{i=1}^{n}(b_i^j - \beta_i^j)$, $j = 0, 1$, kde $n = 268$ je počet pozorování a β_i^j jsou námi nasimulované hodnoty. Obdobně pak definujeme RMSE jako $\text{RMSE}^j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_i^j - \beta_i^j)^2}$, $j = 0, 1$.

Obrázek 3.1 znázorňuje pozorované hodnoty y_i a x_i pro simulace I-III. Pevnými čarami je navíc zobrazeno prvních 20 regresních přímek, které jsou platné pro jednotlivé správní obvody. Na první pohled vidíme, že simulaci II netvoří jedna sada pozorování. Lépe je tento problém vidět na obrázku 3.2. Zde jsou patrné dvě oblasti s vyšší hustotou průsečíku tomografických linek, danými ze vztahu (2.3). Jak King [18] udává, lze na základě hustoty průsečíku a vhodného jádrového odhadu vytvořit odhad sdružené hustoty odhadovaných parametrů β_i^1 a β_i^0.

Z obr. 3.2 lze krom zjevného porušení předpokladu o unimodálním rozdělení pro Kingův model vyčíst i případné porušení předpokladu o nezávislosti β_i^1 a x_i, resp. β_i^0 a x_i, které je v případné modifikované verzi velice důležité zejména pro Goodmanův model. Pokud je předpoklad nezávislosti porušen, pak oblast, kde se tomografické přímky budou nejvíce protínat, bude ležet mimo interval $[0, 1]^2$.

Pro jednotlivé simulace odhadneme parametry β_i^1 a β_i^0 pomocí Goodmanova modelu z kap. 2.2.1. Dále pak použijeme prostředky intervalů z kap. 2.7. Z následující kapitoly 2.8 použijeme základní neighbourhood model a jeho zobecněnou variantu, theta model s parametrem $\alpha = 0,05$. Dále použijeme Kingův model, kap. 2.10, který je k dispozici v rámci knihovny EI s výchozím nastavením iniciálních hodnot a s nastavením 100 000 iterací, přičemž prvních 50 000 odstraníme. Posledním uvažovaným modelem je beta-binomický model, kap. 2.16.1, zprogramovaný v knihovně EIPACK, opět necháme iniciální hodnoty parametrů, pouze nastavíme 100 000[3] iterací, přičemž prvních 50 000 opět odstraníme. Časová náročnost je při tomto počtu iterací v řádu minut.

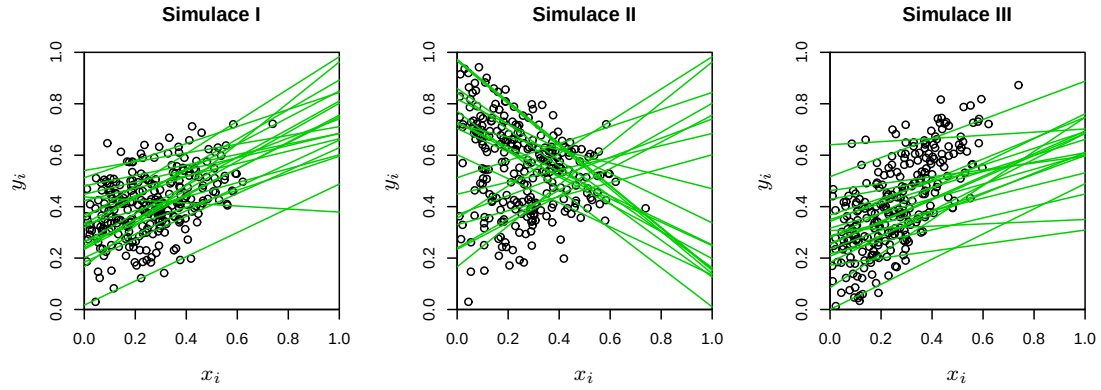
Výsledky pozorovaných vychýlení odhadů parametrů β_i^1 a β_i^0 pro jednotlivé představené modely jsou sepsány přehledně v tabulce 3.1, ze které je patrné, že pro Simulaci I, tj. v situaci, kdy není porušen žádný předpoklad modelu, nám dávají parametrické odhady dobré výsledky. Metoda hranic je pro tato data nepoužitelná a to pro svůj příliš široký interval. Neighbourhood model i jeho zobecnění theta model nedávají dobré výsledky, je to však v důsledku porušení předpokladu o nezávislosti y_i na konkrétní hodnotě x_i.

Z tab. 3.2 dostáváme informaci o velikosti chyby pro jednotlivé správní obvody. Odhady celosouborových parametrů jsou v případě simulace II blíže pravdě, ale pro odhady parametrů v rámci jednotlivých souborů je situace oproti simulaci I opačná. To je dáno tím, že data jsou tvořena pomocí směsi dvou modelů, takže v průměru jsou odhady přesné, ale pro jednotlivá pozorování toto neplatí. Velice dobře si pro tento typ dat vedou neighbourhood model a theta model, při bližním pohledu na prostřední obrázek 3.1, je patrné, že y_i nezávisí na x_i, což je předpoklad těchto dvou modelů.

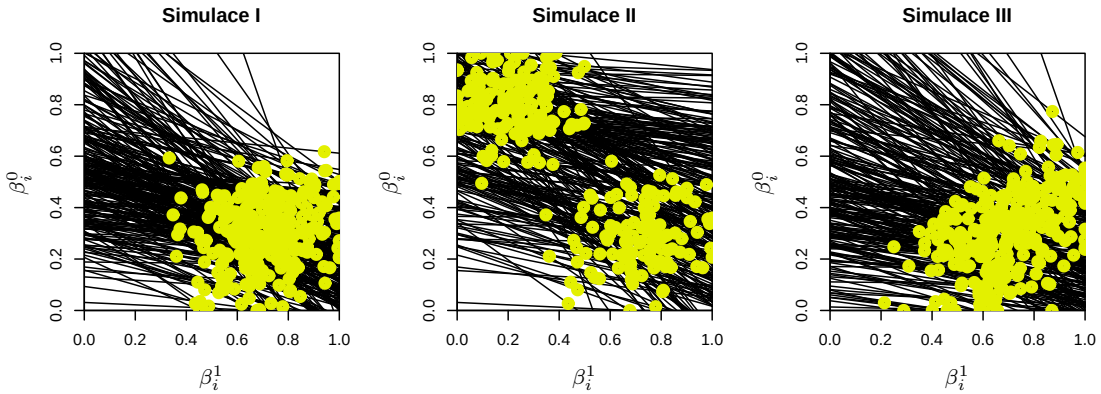
[3]Při desetinovém počtu iterací beta-binomický model nedává zdaleka tak dobré výsledky. Kingův model dává i při nižším počtu iterací podobné výsledky.

Simulace III představuje nejzávažnější porušení předpokladů většiny modelů. Z tab 3.1 vyplývá, že Goodmanův model naprosto selhává. Nejhůře překvapivě nedopadá neighbourhood model, který používá jiný předpoklad, který je při pohledu na 3.1 porušen. Metoda hranic se jeví totožně napříč různými nastaveními, nicméně nelze říci, že by odhady touto metodou byly přesné. Sofistikované modely Kingův a beta-binomický model podávají nejuspokojivější výsledky z množiny námi testovaných způsobů odhadů souborových pravděpodobností β^1 a β^0.

Nyní provedeme opakování generování dat pro jednotlivé simulace I-III. Celkem tak dostaneme 10 opakování. Výsledky pozorovaných vychýlení pro jednotlivé metody jsou zobrazeny na obrázcích 3.3 a 3.4. Je patrné, že pozorované vychýlení můžeme pro jednotlivé modely považovat za stabilní. Pro vývoj RMSE jsou zde obr. 3.5 a 3.6. Jednotlivé modely i zde prokazují stabilní chování.



Obrázek 3.1: Bodový diagram simulací I, II, III. Kroužky jsou znázorněny pozorované hodnoty y_i a x_i. Pevnými čarami jsou znázorněny regresní přímky prvních dvaceti souborů.



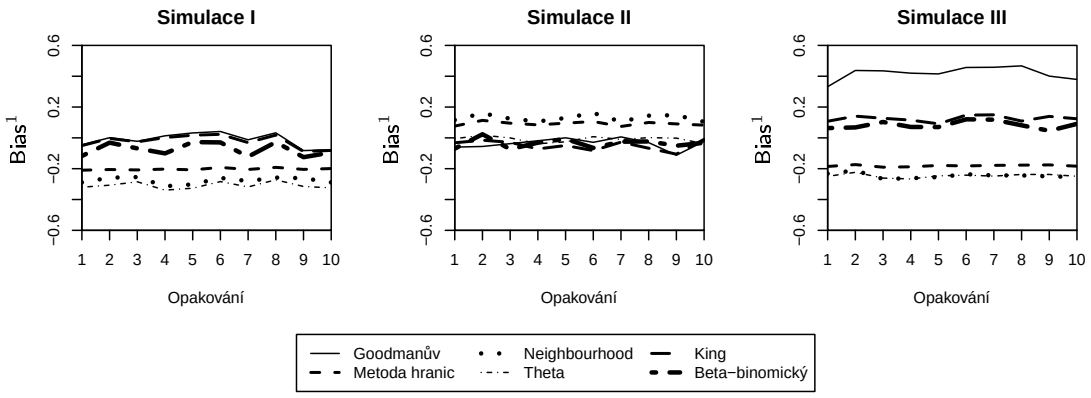
Obrázek 3.2: Tomogram simulací I, II, III. Pevné čáry určují deterministický vztah parametrů β_i^1 a β_i^0 určený rovnicí (2.3). Tečkami jsou označeny skutečné hodnoty $(\beta_i^1, \beta_i^0)'$.

Tabulka 3.1: Pozorované vychýlení odhadů parametrů β_i^1 a β_i^0 pro simulace I-III a jednotlivé modely.

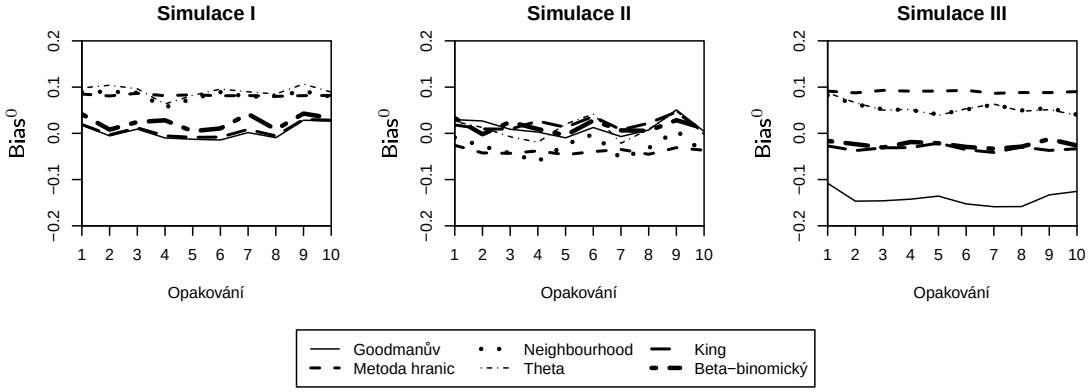
Model	Simulace I		Simulace II		Simulace III	
	Bias^1	Bias^0	Bias^1	Bias^0	Bias^1	Bias^0
Goodmanův	−0,045	0,018	0,060	−0,030	0,332	−0,108
Metoda hranic	−0,210	0,085	0,077	−0,026	−0,185	0,091
Neighbourhood	−0,289	0,089	0,116	−0,009	−0,232	0,083
Theta	−0,320	0,098	0,005	−0,026	−0,250	0,088
Kingův	−0,051	0,019	−0,023	0,015	0,105	−0,026
Beta-binomický	−0,032	0,011	−0,055	0,028	0,047	−0,013

Tabulka 3.2: Odmocninová střední čtvercová chyba RMSE odhadů parametrů β_i^1 a β_i^0 pro simulace I-III a jednotlivé modely.

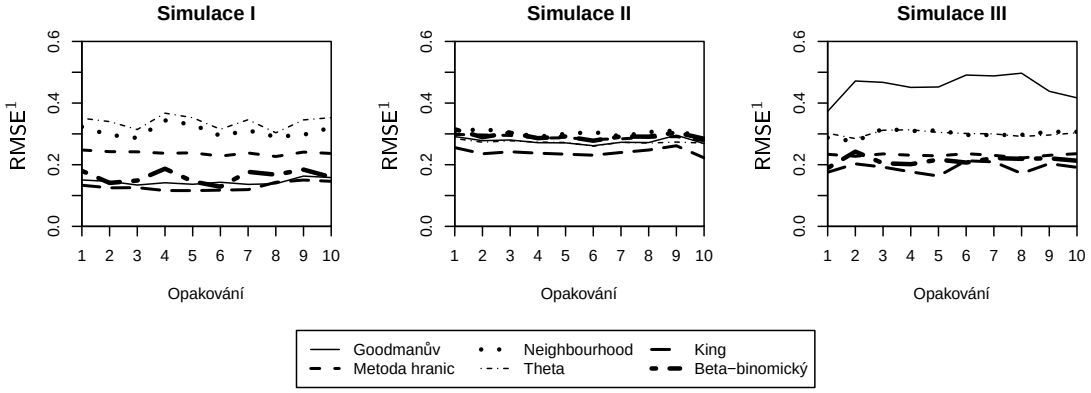
Model	Simulace I		Simulace II		Simulace III	
	RMSE^1	RMSE^0	RMSE^1	RMSE^0	RMSE^1	RMSE^0
Goodmanův	0,151	0,133	0,291	0,282	0,373	0,190
Metoda hranic	0,248	0,121	0,298	0,144	0,234	0,130
Neighbourhood	0,323	0,160	0,308	0,281	0,288	0,177
Theta	0,351	0,165	0,285	0,282	0,302	0,179
Kingův	0,134	0,060	0,254	0,140	0,172	0,056
Beta-binomický	0,131	0,060	0,309	0,151	0,205	0,072



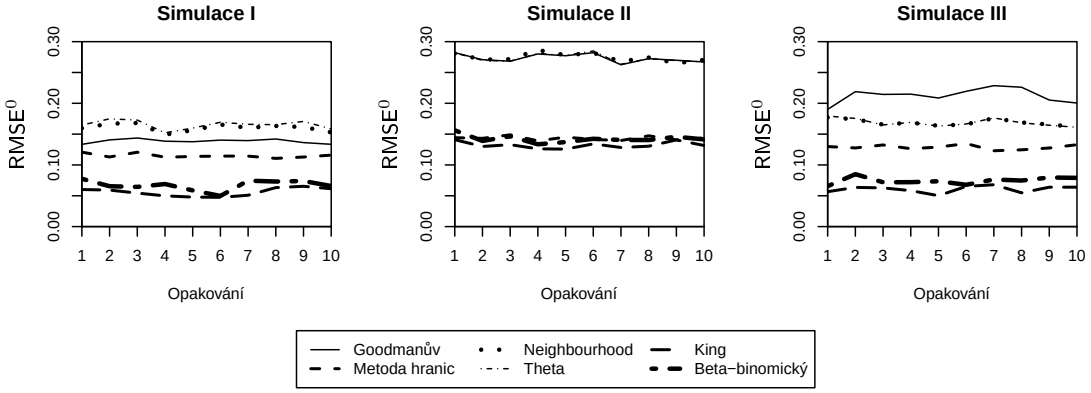
Obrázek 3.3: Pozorované vychýlení odhadů parametru β_i^1 při opakování simulací pro simulace I-III a jednotlivé modely.



Obrázek 3.4: Pozorované vychýlení odhadů parametru β_i^0 při opakování simulací pro simulace I-III a jednotlivé modely.



Obrázek 3.5: Odmocninová střední čtvercová chyba odhadů parametru β_i^1 při opakování simulací pro simulace I-III.



Obrázek 3.6: Odmocninová střední čtvercová chyba odhadů parametru β_i^0 při opakování simulací pro simulace I-III a jednotlivé modely.

3.2 Aplikace na reálných datech

Studii ekologické regrese na reálných datech provedeme na vztahu barvy pleti a negramotnosti. Data byla původně publikována již v [28]. Robinson zkoumal na těchto datech vztah mezi korelačním koeficientem pro neagregovaná data a pro agregovaná data. Analýzou dat došel k závěru, že právě toto je případ, kdy je patrně porušena základní podmínka ekologické regrese a to nezávislost parametrů β_i^1 a β_i^0 na vysvětlující proměnné. King [18] rozšířil původních 48 pozorování na 1 040 a podrobně data zkoumal a testoval na nich přesnost svého modelu, zde označeného jako Kingův model. Data[4] obsahují krom agregovaných proměnných i skutečné hodnoty parametrů β_i^1 a β_i^0. Můžeme tak porovnat pozorované vychýlení modelu, stejně jako v předcházejícím příkladu. Vzhledem k většímu množství dat použijeme u iteračních metod 20 00 iterací, přičemž prvních 10 000 zahodíme. Tab. 3.3 sumarizuje pozorované odchylky jednotlivých odhadů od skutečných hodnot. Stejně jako v teoretickém příkladu i zde, dává nejpřesnější odhady Kingův a beta-binomický model. Obr. (3.7) zobrazuje bodový diagram jednotlivých pozorování spolu se standardním odhadem metodou nejmenších čtverců, který je ekvivalentní Goodmanově modelu. Obr. (3.8) nám poskytuje tu samou informaci vzhledem k parametrům zájmu β_i^1 a β_i^0. Z obrázku je patrné solidní chování dat vůči předpokladům Goodmanova modelu, jednotlivé čáry nemají nejvyšší hustotu mimo $[0,1]^2$.

Tabulka 3.3: Pozorované vychýlení, RMSE a MAE odhadů parametrů β_i^1 a β_i^0 pro jednotlivé modely.

Model	Bias		RMSE		MAE	
	b^1	b^0	b^1	b^0	b^1	b^0
Goodmanův	−0,072	0,016	0,128	0,059	0,104	0,040
Metoda hranic	0,022	−0,088	0,140	0,165	0,120	0,100
Neighbourhood	0,141	−0,093	0,151	0,128	0,141	0,093
Theta	0,080	−0,021	0,132	0,061	0,103	0,048
Kingův	−0,063	0,013	0,093	0,031	0,070	0,023
Beta-binomický	−0,054	0,009	0,091	0,030	0,066	0,023

Pozn: Střední absolutní chyba MAE je dána vztahem $\sum_{i=1}^{n} |b_i^j - \beta_i^j|/n$.

Obrázky 3.9 a 3.10 nám dávají informaci o účinnosti metody hranic pro daný soubor dat. Na první pohled je patrné, že metoda hranic dokáže zmenšit možný interval β_i^0 na třetinu původní hodnoty a to pro více jak polovinu pozorování. Jednotlivé čáry na obr. 3.9 odpovídají postavení dvojice $(x_i, y_i)'$ na obr. 3.10.

Pro Kingův model získáváme hned několik možných výsledných zobrazení. V první řadě to je obr. 3.11, který zobrazuje odhadnutou aposteriorní hustotu pro celosouborové pravděpodobnosti β^1 a β^0 pomocí Kingova modelu. Dále tu máme 3.12, který nám zobrazuje odhady parametrů zájmu pro jednotlivé soubory. Rozdíl mezi Goodmanovým modelem a přístupem, který zvolil King, je vidět na pásech spolehlivosti kolem regresní přímky (viz obr. 3.7 a 3.13).

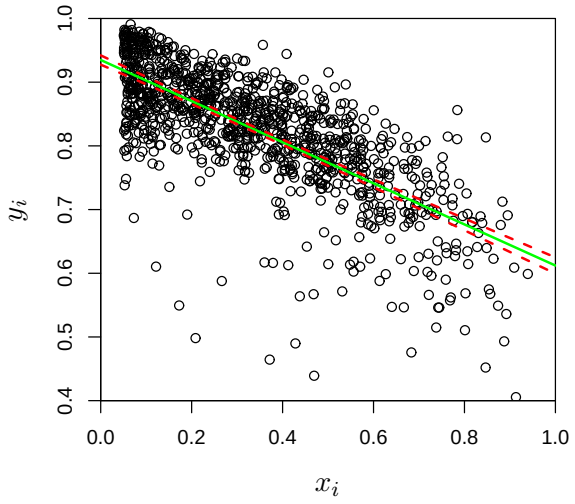
Obrázky 3.14 a 3.15 popisují odhady pomocí beta-binomického modelu. Rozložení jednotlivých odhadů $(\beta_i^1, \beta_i^0)'$ se do značné míry liší od odhadů Kingova

[4]Data *census* z knihovny ECO. URL: `http://imai.princeton.edu/software/eco.html`

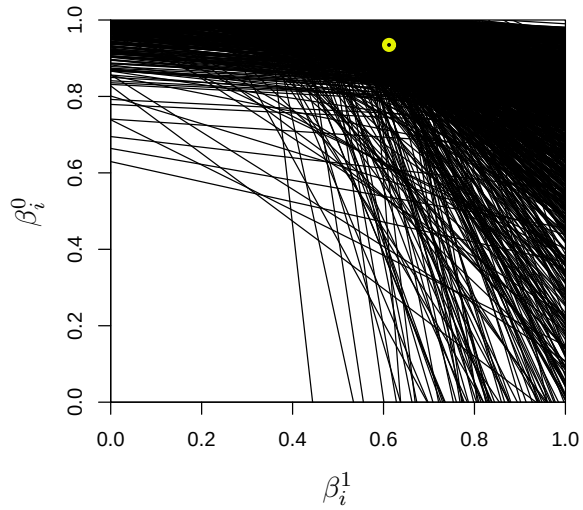
modelu na obr. 3.12, nicméně jak ukazuje tab. 3.3, nemá to negativní vliv na pozorované vychýlení a tento model na data sedí nejlépe ze všech testovaných. Je dobré si i přesto povšimnout na obr. 3.15, že hustota β^1_{50} dostatečně nepokrývá skutečnou hodnotu.

Skutečné rozložení parametrů zájmu jsou až na pár odlehlých pozorování koncentrované kolem hodnoty $(0{,}7, 0{,}9)'$ (viz obr. 3.16), což odpovídá odhadům celosouborových parametrů β^1 a β^0 pro Kingův i beta-binomický model.

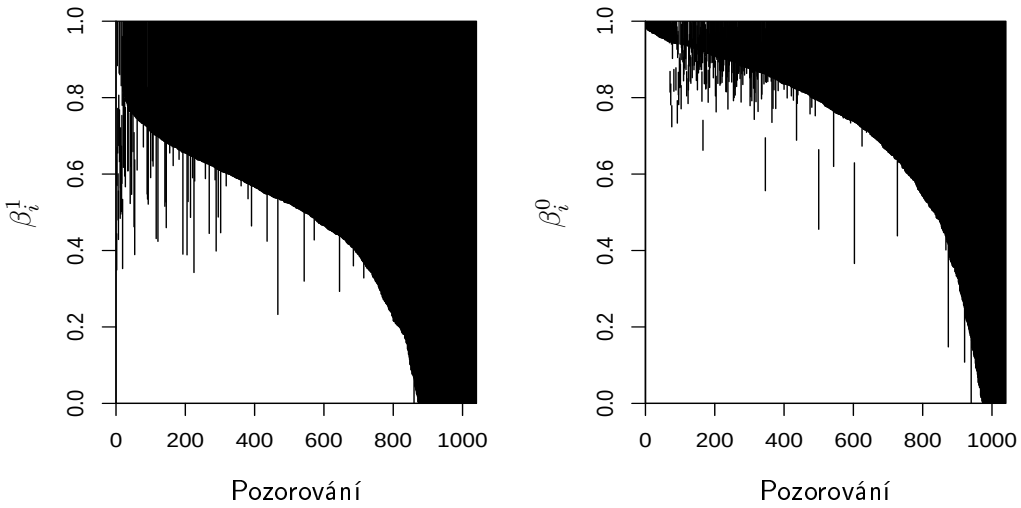
Na úplný závěr poznamenejme, že dobré výsledky hierarchických modelů v tomto příkladě neznamenají obecně dobré chování pro jiná data, na druhou stranu, jsou méně závislé na porušení předpokladu chování pravděpodobností β^1_i a β^0_i vůči vysvětlující proměnné x_i.



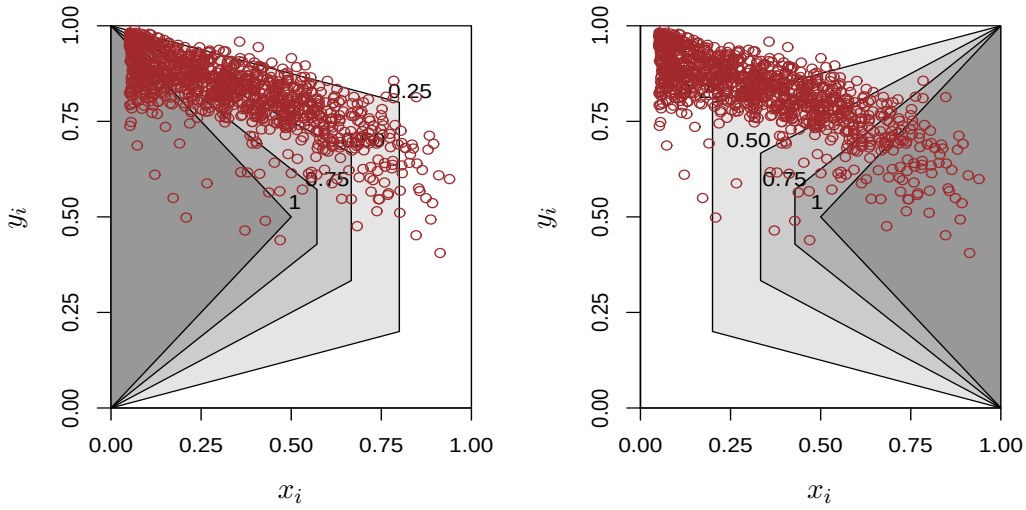
Obrázek 3.7: Bodový diagram pozorovaných hodnot a Goodmanův odhad s 95% pásem spolehlivosti.



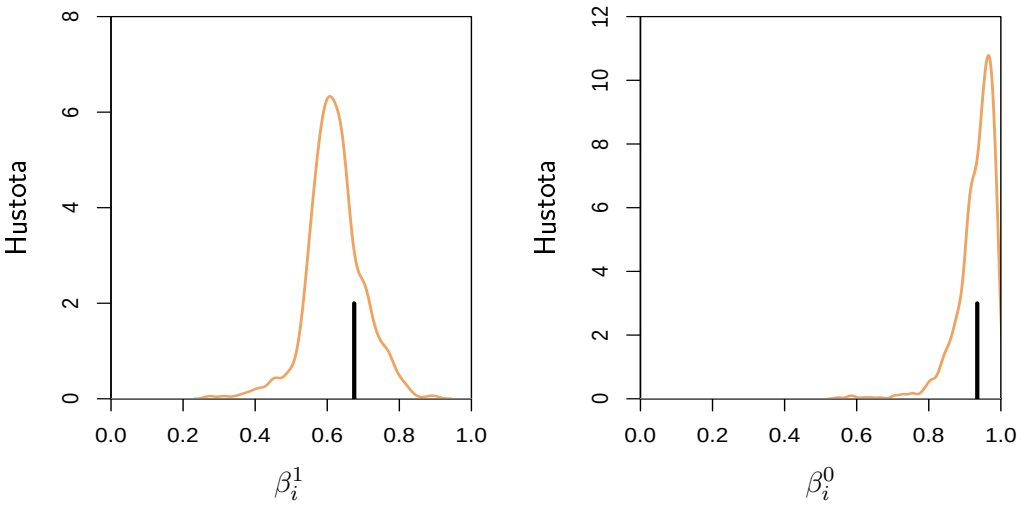
Obrázek 3.8: Tomogram možných nastavení odhadovaných parametrů β_i^1 a β_i^0. Bodový odhad pomocí Goodmanova modelu je zobrazen kolečkem.



Obrázek 3.9: Horní a dolní hranice určující množinu možných hodnot parametrů β_i^1 a β_i^0. Pozorování jsou seřazena podle délky intervalu.

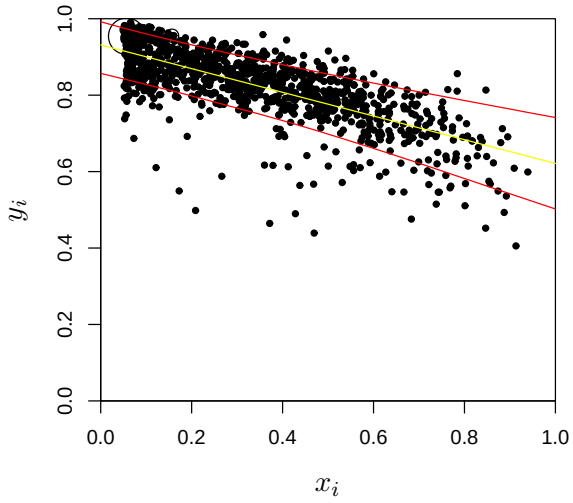


Obrázek 3.10: Alternativní pohled k určení délky intervalů pomocí metody hranic. Obrázek vlevo udává délky intervalů parametru β_i^1. Obrázek vpravo nám dává totéž pro parametr β_i^0.

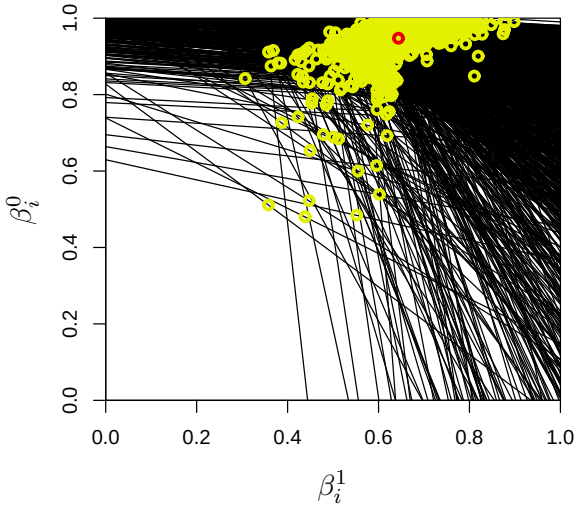


Obrázek 3.11: Kingovy odhady aposteriorních marginálních hustot useknutého normálního rozdělení. Černou úsečkou je znázorněna skutečná hodnota odhadovaného parametru.

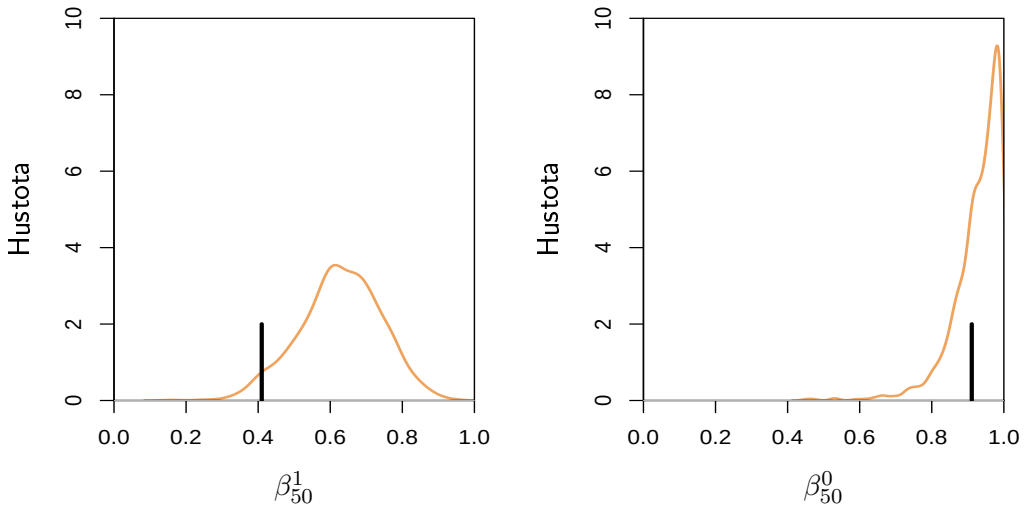
Obrázek 3.12: Bodové odhady $(\beta_i^1, \beta_i^0)'$ pro jednotlivá pozorování pomocí Kingova modelu (světlé body). Bodový odhad celosouborového vektoru $(\beta_i^1, \beta_i^0)'$ (tmavý bod).



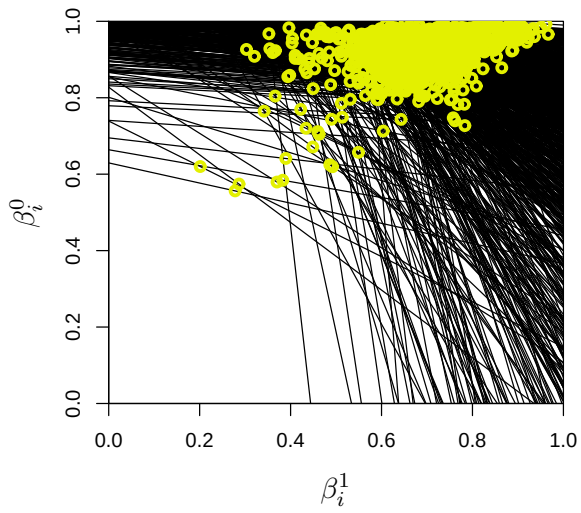
Obrázek 3.13: Bodový diagram pozorovaných hodnot a Kingův odhad regresního modelu s 80% pásem spolehlivosti.



Obrázek 3.14: Bodové odhady $(\beta_i^1, \beta_i^0)'$ pro jednotlivá pozorování pomocí beta-binomického modelu (světlé body). Bodový odhad celosouborového vektoru $(\beta_i^1, \beta_i^0)'$ (tmavý bod).



Obrázek 3.15: Odhady aposteriorních marginálních hustot β_{50}^1 a β_{50}^0 určené pomocí beta-binomického modelu pro padesáté pozorování. Černou úsečkou je znázorněna skutečná hodnota odhadovaného parametru β_{50}^1, resp. β_{50}^0.



Obrázek 3.16: Skutečné hodnoty $(\beta_i^1, \beta_i^0)'$.

Závěr

Ekologická regrese je statistická metoda, kdy používáme regresi na agregovaná data (typicky se jedná o průměry v rámci různých geografických oblastí) a odhady takto získané interpretujeme jako vztahy na úrovni neagregovaných jednotek, jednotlivců. Po celou dobu, co se ekologická regrese používá, je znám fakt, že samotná regrese závisí na různých předpokladech, které jsou v praxi netestovatelné díky ztrátě informace v důsledku agregace. Cílem této práce bylo představit jednotlivé přístupy k analýze těchto agregovaných dat, definovat modely a specifikovat podmínky, za kterých tyto modely dávají rozumné odhady.

V teoretické části jsme rozkryli vznik vychýlení v důsledku agregace dat a definovali si postupy, jak se tohoto vychýlení odhalit a případně jak ho také odhadnout. Zavedli jsme několik modelů, které pracují s ekologickými daty. Za zmínku stojí zejména Goodmanův model, jakožto základní přístup k odhadům neznámých parametrů v ekologické regresi. Dále jsme si představili jeho zobecnění a rozšíření a další pomocné vysvětlující proměnné. Prozkoumali jsme různé alternativy ke Goodmanově modelu, jako je například sousedský model a metoda hranic. Představili jsme hierarchické modely využívající bayesovský přístup k chování neznámých parametrů a využívajících metody maximální věrohodnosti pro určení dodatečných parametrů modelu a MCMC metody pro odhady parametrů zájmu.

Jednotlivé modely jsme následně v praktické části použili pro získání odhadů při různých strukturách dat. Navzájem porovnali jejich efektivitu a stabilitu. Veškeré znalosti jsme posléze použili pří odhadovaní parametrů chování jednotlivců v rámci agregovaných jednotek, tak k odhadnutí parametrů určující chování pro celou populaci.

Další možností, jak rozšířit tuto práci, je zobecnění modelu binárních proměnných na multinomické proměnné, případně zaměření se na analýzu dat v oblasti epidemiologie.

Seznam použité literatury

[1] Achen, Ch. H. a Shively, P. W. *Cross-level Inference*. Chicago: University of Chicago Press, 1995. ISBN 0-226-00220-9.

[2] Anděl, J. *Základy matematické statistiky*. 2. vydání. Praha: Matfyzpress, 2011. ISBN 80-7378-001-1.

[3] Ansolabehere, S. a Rivers, D. *Bias in Ecological Regression*. Presented at the Midwest Political Science Association Meetings, Chicago, 1994.

[4] Blakely, T. A. a Woodward, A. J. Ecological effects in multi-level studies. *J Epidemiol Community Health*, 2000, roč. 54, s. 367–374.

[5] Breslow, N. E. a Day, N. E. Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC Scientific Publications*, 1980, roč. 32, s. 5–338.

[6] Chib, S. a Greenberg, E. Understanding the Metropolis-Hastings Algorithm. *The American Statistican*, 1995, roč. 49, č. 4, s. 327–335.

[7] Cipra, T. *Finanční ekonometrie*. 1. vyd. Praha: Ekopress, 2008. ISBN 978-80-86929-43-9.

[8] Duncan, D. O. *Statistical Geography: Problems in Analyzing Areal Data*. Westport, Conn.: Free Press, 1961.

[9] Firebaugh, G. A Rule for Inferring Individual-Level Relationships from Aggregate Data. *American Sociological Review*, 1978, roč. 43, č. 4, s. 557–572.

[10] Freedman, D. A. Klein, S. P., Sacks, J., Smyth, Ch. A. a Everett, Ch. G. Ecological regression and voting rights. *Evaluation Review*, 1991, roč. 15, č. 6, s. 673–711.

[11] Gamerman, D. *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*.1st ed. London: Chapman & Hall/CRC, 1997. ISBN 0-412-81820-5.

[12] Gelman, A., Park, D. K., Ansolabehere, S., Price, P. N. a Minnite, L. C. Models, Assumptions and Model Checking in Ecologicl Regressions. *Journal of Royal Statistical Society. Series A (Statistics in Society)*, 2001, roč. 164, č. 1, s. 101–118.

[13] Goodman, L. A. Ecological Regressions and the Behavior of Individuals. *American Sociological Review*, 1953, roč. 18, s. 663–666.

[14] Goodman, L. A. Some Alternatives to Ecological Correlation. *American Journal of Sociology*, 1959, roč. 64, č. 6, s. 610–625.

[15] Hušková, M. *Bayesovské metody*. 1. vyd. Praha: Státní pedagogické nakladatelství, 1985.

[16] Chambers, R. L. a Steel, D. G. Simple methods for ecological inference in 2x2 tables. *Journal of Royal Statistical Society. Series A (Statistics in Society)*, 2001, roč. 164, č. 1, s. 175–192.

[17] Imai, K., Lu, Y. a Strauss, A. Bayesian and Likelihood Inference for 2 x 2 Ecological Tables: An Incomplete Data Approach. *Political Analysis*, 2008, roč. 16, č. 1, s. 41–69.

[18] King, G. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press, 1997. ISBN 0-691-01240-7.

[19] King, G., Rosen, O. a Tanner, M. A. Binomial-Beta Hierarchical Models for Ecological Inference. *Sociological Methods & Research*, 1999, roč. 28, č. 1, s. 61–90.

[20] King, G., Rosen, O. a Tanner, M. A. *Ecological inference : new methodological strategies*. New York: Cambridge University Press, 2004. ISBN 0-521-54280-4.

[21] Kupper, L. L., Karon, J. M., Kleinbaum, D. G., Morgenstern, H. a Lewis, D. K. Matching in Epidemiologic Studies: Validity and Efficiency Considerations. *Biometrics*, 1981, roč. 37, č. 2, s. 271–291.

[22] Manjunath, B. G. a Wilhelm, S. Moments Calculation For the Doubly Truncated Multivariate Normal Density. *SSRN Electronic Journal*, 2012, DOI:10.2139/ssrn.1472153.

[23] Morgenstern, H. Uses of ecologic analysis in epidemiologic research. *American Journal Public Health*, 1982, roč. 72, s. 1336–1344.

[24] Morgenstern, H. a Thomas D. Principles of study design in environmental epidemiology. *Environ Health Perspect.*, 1993, roč. 101, s. 23-–38.

[25] Papalia, R. B. Incorporating Spatial Structures in Ecological Inference: An Information Theory Approach. *Entropy.*, 2010, roč. 12, s. 2171-–2185.

[26] Piantadosi, S., Byar, D. P. a Green, S. B. The ecological fallacy. *American Journal of Epidemiology*, 1988, roč. 127, č. 5, s. 893–904.

[27] Raudenbush, S. W. a Bryk, A. S. *Hierarchical linear models: applications and data analysis methods*. 2nd ed. Thousand Oaks: Sage Publications, c2002, xxiv, 485 s.

[28] Robinson, W. S. Ecological correlations and behaviour of individuals. *American Sociological Review*, 1950, roč. 15, č. 3, s. 351–357.

[29] R Core Team (2015). *R: A language and environment for statistical computing*. [počítačový program]. R Foundation for Statistical Computing.`http://www.R-project.org/` [verze 3.2.0].

[30] Schuessler, A. A. Ecological inference. *Proc Natl Acad Sci U S A*, 1999, roč. 96, s. 10578–10581.

[31] Subramanian, S. V., Jones, K., Kaddour, A. a Krieger, N. Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology*, 2009, roč. 38, s. 342–360.

[32] Tam Cho, W. K. Latent Groups and Cross-Level Invefences. *Electoral Studies*, 2001, roč. 20, s. 243–269.

[33] Tierney, L. Markov chains for exploring posterior distribution. *The Annals of Statistics*, 1994, roč. 22, č. 4, s. 1701–1762.

[34] Wakefield, J. Ecological inference for 2 x 2 tables. *Journal of Royal Statistical Society. Series A (Statistics in Society)*, 2004, roč. 167, č. 3, s. 385–445.

[35] Wakefield, J. Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology*, 2009, roč. 38, s. 330–336.

[36] Wakefield, J, Haneuse, S., Dobra, A. a Teeplec, E. Bayes Computation for Ecological Inference. *Statistics in Medicine*, 2011, roč. 30, č. 12, s. 1381-1396.

[37] Zvára, K. *Regrese*. Praha: Matfyzpress, 2008. ISBN 978-80-7378-041-8.

Seznam obrázků

Seznam tabulek

Přílohy

Příloha 1. Součástí práce je přiložené CD, které obsahuje všechny zdrojové kódy z programu R pro jednotlivé příklady a simulace. CD obsahuje soubory:

- *data_census.RData* je soubor podkladových dat programu *R* požívaných v podkapitole 3.2.
- *data_matproii.RData* je soubor podkladových dat programu *R* požívaných v podkapitole 3.1.
- *kap_2_priklad.R* je zdrojový soubor z *R* pro příklad z podkapitoly 2.
- *kap_2_2_hranice.R* je zdrojový soubor z *R* pro tvorbu obrázků z podkapitoly 2.7.
- *kap_3_1_simulace.R* je zdrojový soubor z *R* pro simulaci dat a jejich odhadování z podkapitoly 3.1.
- *kap_3_1_simulace_opakovani.R* je zdrojový soubor z *R* pro testovaní stability v podkapitole 3.1.
- *kap_3_2_realna_data.R* je zdrojový soubor z *R* pro analýzu dat z podkapitoly 3.2.
- *pouzite_balicky.R* je zdrojový soubor z *R* s přehledem použitých knihoven.