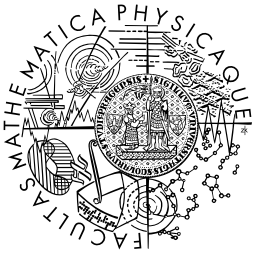


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Martin Splítek

Statistická inference v modelech s proměnlivými koeficienty

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	RNDr. Maciak Matúš, Ph.D.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie

Praha 2018

Prohlašuji, že jsem tuto diplomovou práci vypracoval 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 má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 4.1.2018 Martin Splítek

Název práce: Statistická inference

v modelech s proměnlivými koeficienty

Autor: Bc. Martin Splítek

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Maciak Matúš, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá modely s proměnlivými koeficienty se zaměřením na statistickou inferenci. Hlavní myšlenkou těchto modelů je použití regresních koeficientů, měnících se v závislosti na nějakém modifikátoru vlivu, namísto konstantních koeficientů klasické lineární regrese. Nejprve si definujeme tyto modely a jejich odhadové procedury, kterých bylo doposud publikováno několik variant. K odhadu se používá lokální regrese nebo různé druhy splajnů – vyhlazovací, polynomiální či penalizované. Od metody odhadu se následně odvíjí i daná statistická inference, ke které uvedeme odvozené vychýlení, rozptyl, asymptotickou normalitu, konfidenční pásma a testování hypotéz. Hlavním cílem naší práce je kompaktně shrnout vybrané metody a jejich inferenci. Na závěr je navržena proceduru pro výběr proměnných.

Klíčová slova: modely s proměnlivými koeficienty, odhad, inference

Title: Statistical inference in varying coefficient models

Author: Bc. Martin Splítek

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Maciak Matúš, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis discusses varying coefficient models with focus on statistical inference. The main idea of these models is the use of regression coefficients varying over some effect modifier instead of constant coefficients of classical linear regression. First, we define these models and their estimation procedures, which have been published in several variants to date. Local regression or different spline types - smoothing, polynomial or penalized, can be used to estimate these models. From the estimation method, we also derive the given statistical inference, to which we refer deduced bias, variance, asymptotic normality, confidence bands, and hypothesis testing. The main aim of our work is to summarize the selected methods and their inference. Finally, a procedure for variable selection is proposed.

Keywords: varying coefficient models, estimation, inference

Rád bych poděkoval RNDr. Matúši Maciakovi, Ph.D. za cenné rady, věcné připomínky a vstřícnost při konzultacích a vypracování diplomové práce. V neposlední řadě také mé rodině za jejich neochvějnou podporu.

Obsah

Úvod

Přechod do 21. století sebou přinesl mnoho změn. Mezi hlavní z nich zajisté můžeme počítat vzestup osobních počítačů a informačních technologií, který se nyní rozvíjí přímo exponenciálním tempem. Všechny tyto prostředky, zejména jejich spojení pomocí internetu, v současné době generují nepřeberné množství dat. Právě tato data se pomalu, ale jistě, stávají komoditou budoucnosti. A zde na scénu přicházíme my, matematici, statistici či ekonometři v neustávajícím pokusu z těchto dat vytěžit relevantní informace a být schopni do jisté míry předpovídat budoucnost. Analytici finančních trhů odhadují budoucí vývoj úrokových sazeb na základě dosavadního vývoje, statistici v bankovním sektoru modelují pravděpodobnosti, že si klient pořídí úvěrový produkt, mediální společnosti se snaží předpovídat sledovanost daného kanálu v průběhu dne a takto bychom mohli vyjmenovat nespočet dalších možných uplatnění matematiků v soukromém sektoru. Způsobů, jak přistupovat k matematickému modelování, ať už volba podkladového modelu či metoda jeho odhadu, existuje velké množství, avšak v současné době se v praxi asi nejvíce setkáme se zobecněnou lineární regresí. Základní lineární regrese a logistická regrese pro odhad binární odezvy nyní tvoří praktickou páteř matematického modelování. Od jejich publikace v (Nelder a Wedderburn, 1972) ale již uplynulo 45 let a neúprosný vývoj kupředu si žádá nové a pokročilejší metody. Společnosti po celém světě zkoušejí nové přístupy k modelování. Obzvlášť metody takzvaného strojového učení nabývají na popularitě. Mezi nejznámější z těchto metod se řadí například neuronové sítě či rozhodovací stromy a jejich pokročilá verze náhodné lesy. V praxi se ale ukazuje, že tyto alternativní metody trpí jedním velice zásadním problémem – ve většině případů totiž dávají horší či srovnatelné výsledky než zobecněná lineární regrese, a to za cenu interpretovatelnosti. Vzpomeňme si, jak snadné je interpretovat vliv regresoru na odezvu pomocí jeho odhadnutého koeficientu v klasické lineární regresi. Metody strojového učení takovouto vlastností bohužel nedisponují a často se o nich hovoří jako o tzv. "black boxech", kdy ze vstupu vytvoří výstup, ale způsob jakým toho dosáhly není zcela zřejmý. Po možném nástupci zobecněné lineární regrese tedy požadujeme dvě vlastnosti – alespoň stejně snadnou interpretovatelnost výsledného modelu a přesnost odhadu. Dosavadní alternativní metody bohužel obou těchto vlastností nedosahují a hledání proto pokračuje.

V naší práci čtenáře seznámíme s poměrně novým a prozatím nepříliš známým typem matematických modelů – modely s proměnlivými koeficienty, poprvé představených v (Hastie a Tibshirani, 1993). Právě o nich si totiž myslíme, že v blízké budoucnosti mají potenciál stát se nástupcem modelů zobecněné lineární regrese. Tyto modely stojí na pomezí zobecněné lineární regrese a neparametrických modelů, snaží se kombinovat to nejlepší z obou. Jejich hlavní myšlenka spočívá v předpokladu, že regresní koeficienty nejsou konstantami, nýbrž hladkými funkcemi závislými na jiných regresorech, označovaných jako modifikátory vlivu. K odhadu těchto koeficientů ale již není vhodné použít klasický parametrický odhad, a proto se využívají silnější neparametrické metody odhadu. Výsledné koeficienty ve formě hladkých křivek se interpretují stejně snadno jako v případě lineární regrese. Navíc na nich ještě můžeme pozorovat dodatečnou informaci o změně parametru v závislosti na modifikátoru. Za zmínku stojí hlavně

modely, kde je modifikátorem vlivu čas. Při dlouhodobých pozorováních je totiž nepravděpodobné, že by vlivy regresorů zůstávaly s časem konstantní (Fan a Zhang, 2008). Neparametrický odhad by taktéž měl být přesnější nebo alespoň stejně přesný, než odhad parametrický, a proto tyto modely mají zmíněný potenciál nahradit zobecněnou lineární regresi. Tento způsob odhadu však s sebou nese i svá úskalí. Při velkém počtu regresorů nejsou tyto metody kvůli výpočetní složitosti schopny dojít k výsledkům. Tento fenomén se označuje jako kletba dimenzionality. Díky omezení se na hladké křivky se modelům s proměnlivými koeficienty podařilo tomuto fenoménu vyhnout, avšak výpočetní složitost těchto modelů je stále velmi vysoká.

Modely s proměnlivými koeficienty nám tedy oproti zobecněné lineární regresi do modelu přinášejí dodatečnou informaci o závislosti regresoru na nějakém modifikátoru vlivu, díky které by výsledný model měl věrněji odrážet skutečnost. To však tvoří pouze špičku ledovce, neboť hledání takovýchto modifikátorů vlivu může odhalit mnoho různých závislostí ve vstupních datech, které na první pohled nejsou zcela zřejmé a zůstaly by jinak skryty.

Od svého vzniku prošly modely s proměnlivými koeficienty díky velkému množství akademických publikací značným vývojem. V této práci se pokusíme přehledně shrnout dosavadní teoretické poznatky a metodologii modelů s proměnlivými koeficienty. Zaměříme se převážně na statistickou inferenci v těchto modelech. Ta je ale úzce spjata se zvolenou metodou odhadu. Ty zde proto také stručně popíšeme.

V první kapitole si definujeme základní Gaussovský tvar modelu s proměnlivými koeficienty zvlášť pro standardní data s nezávislými pozorováními a podíváme se na jeho rozšíření na množinu exponenciálních rozdělení. Pro demonstraci široké škály možností vývoje a využití těchto modelů čtenáře zběžně seznámíme i s několika speciálními případy těchto modelů. Dále si definujeme i model s proměnlivými koeficienty pro longitudinální data.

V druhé kapitole si nastíníme metody odhadu, které se dělí na dvě hlavní skupiny – odhad pomocí splajnů a odhad pomocí lokální regrese. Splajnové metody se pak dále dělí dle použitých dat a dle použitého druhu splajnů. Hlavním rozdílem mezi nimi je forma penalizace hladkosti a volba uzlů.

Hlavní část diplomové práce, statistická inference v modelech s proměnlivými koeficienty, se nachází ve třetí kapitole. V té se budeme věnovat odhadům vychýlení, rozptylu a kovarianční struktury, asymptotickým vlastnostem, konfidenčním intervalům a testování hypotéz o proměnlivosti koeficientů.

Ve čtvrté kapitole navrhneme vlastní proceduru pro výběr proměnných operující na bázi "forward selekce"a s modifikacemi pro použití na modelech s proměnlivými koeficienty.

V páté kapitole se zaměříme na praktické využití modelů s proměnlivými koeficienty. Zmíníme několik použití těchto modelů na reálných problémech a rovněž v současné době dostupný software, ve kterém lze modely s proměnlivými koeficienty odhadovat. Dále otestujeme empirické vlastnosti dvou testů konstantnosti proměnlivého koeficientu pomocí simulací. Naše poznatky pak aplikujeme na příkladu reálných dat.

1. Modely s proměnlivými koeficienty

Jak již bylo řečeno, modely s proměnlivými koeficienty kombinují přednosti neparametrické regrese a snadné interpretovatelnosti. Stejně jako v lineární regresi do modelu vstupují parametry v lineárním tvaru. Výrazná změna nastává až v jejich podobě, kde již nejsou konstantami, nýbrž hladkými funkcemi závislými na nějakém modifikátoru vlivu.

Mějme náhodnou veličinu závislé proměnné Y, náhodný vektor regresorů $\boldsymbol{X} = (X_1,...,X_k)^T$ a náhodný vektor modifikátorů vlivu $\boldsymbol{U} = (U_1,...,U_k)^T$. Označme si $\boldsymbol{Y} = (Y_1,...,Y_n)^T$ jako vektor závislé proměnné.

Předpokládejme, že máme k dispozici data složená z n pozorování nezávislých a stejně rozdělených náhodných vektorů $(Y_i, \boldsymbol{X}_i^T, \boldsymbol{U}_i^T)^T$ s rozdělením náhodného vektoru $(Y, \boldsymbol{X}^T, \boldsymbol{U}^T)^T$. Takováto data budeme dále označovat jako standardní data s nezávislými pozorováními.

Dále si označme $\boldsymbol{X}_i = (X_{i,1},...,X_{i,k})^T, i = 1,...,n$ jako vektor regresorů pro i-té pozorování a $\boldsymbol{X}^j = (X_{1,j},...,X_{n,j})^T$ jako vektor j-tého regresoru. Analogicky označíme i $\boldsymbol{U}_i = (U_{i,1},...,U_{i,k})^T, i = 1,...,n$ jako vektor modifikátorů vlivu pro i-té pozorování a $\boldsymbol{U}^j = (U_{1,j},...,U_{n,j})^T$ jako vektor j-tého modifikátoru vlivu.

Definujme si ještě matici regresorů jako

$$\mathcal{X} = \begin{pmatrix} X_{1,1} & \cdots & X_{1,k} \\ \vdots & \vdots & \vdots \\ X_{n,1} & \cdots & X_{n,k} \end{pmatrix} = (\boldsymbol{X}^1,...,\boldsymbol{X}^k),$$

a matici modifikátorů vlivu následovně

$$\mathcal{U} = \begin{pmatrix} U_{1,1} & \cdots & U_{1,k} \\ \vdots & \vdots & \vdots \\ U_{n,1} & \cdots & U_{n,k} \end{pmatrix} = (\boldsymbol{U}^1,...,\boldsymbol{U}^k).$$

Základní tvar modelu pro standardní data o n nezávislých pozorováních je pak definován jako

$$Y_i = \sum_{j=1}^{k} \beta_j(U_{i,j}) X_{i,j} + \varepsilon_i. \tag{1.1}$$

Náhodné chyby $\varepsilon_i, i = 1,...,n$ jsou nezávislé a normálně rozdělené se střední hodnotou $E(\varepsilon_i \mid \boldsymbol{X}_i, \boldsymbol{U}_i) = 0$ a rozptylem $var(\varepsilon_i \mid \boldsymbol{X}_i, \boldsymbol{U}_i) = \sigma^2$. Koeficienty $\beta_j(\cdot)$ jsou definovány jako neznámé hladké funkce.

Základní definici modelu s normálně rozdělenými chybami je možno zobecnit na množinu exponenciálních rozdělení pomocí lineárního prediktoru a přenosové funkce jako

$$\eta = \sum_{j=1}^{k} \beta_j(\boldsymbol{U}^j)\boldsymbol{X}^j, \tag{1.2}$$

kde η je zmíněný lineární prediktor a střední hodnotu $\mu = (EY \mid X,U)$ vyjádříme pomocí přenosové funkce g jako $\eta = g(\mu)$ (Hastie a Tibshirani, 1993).

Na první pohled z definice (1.1) nemusí být zřejmé, jak širokou oblast tyto modely pokrývají. Taktéž zmíněná souvislost mezi modely s proměnlivými koeficienty a modely lineární regrese spolu s aditivními modely nám může unikat. Podívejme se proto na několik speciálních případů modelu s proměnlivými koeficienty.

- Pokud vezmeme všechny koeficienty $\beta_j(U_j)$ jako konstanty β_j, pak dostáváme klasický model lineární regrese či zobecněné lineární regrese. Lineární regrese je tedy speciálním případem modelů s proměnlivými koeficienty.

- Když bychom použili regresory $X_j = c$, kde c je nějaká konstanta (bez újmy na obecnosti kupříkladu $c = 1$), pak nám v rovnici zbývají pouze členy $\beta_j(U_j)$. Ty jsou definovány jako jakési hladké funkce závislé na U_j a tím pádem takovýto model je aditivním modelem s regresory $(U_1,...,U_k)$. Při použití regresorů $(X_1,...,X_k)$ jako modifikátorů vlivu pak dostaneme aditivní model pro původní regresory. Aditivní modely tedy taktéž můžeme považovat za speciální případ modelů s proměnlivými koeficienty.

- Pokud si koeficienty $\beta_j(U_j)$ definujeme jako $\beta_j U_j$, pak vytvoříme interakci lineárního regresního modelu ve tvaru $\beta_j U_j X_j$.

- Předpokládejme jednoduchý model pouze s jediným regresorem X. Tento regresor použijme i jako modifikátor vlivu. Dostáváme pak rovnici modelu $Y_i = \beta(X_i)X_i + \varepsilon_i$. Takovýto model je jak speciálním případem modelu s proměnlivými koeficienty, tak i modelem vyhlazovací či neparametrické regrese.

- Taktéž je třeba se více zamyslet nad tvary modifikátorů vlivu $U_j, j = 1,...,k$. Modifikátor vlivu si můžeme definovat nejen jako skalár, ale i jako vektor. Volba vektorového modifikátoru vlivu se nabízí kupříkladu pro zeměpisné nebo GPS souřadnice, když očekáváme změny vlivu parametrů v závislosti na lokaci. Ve zbytku práce budeme uvažovat pouze skalární modifikátory vlivu.

- V definici modelu s proměnlivými koeficienty (1.1) je ke každému regresoru $X_j, j = 1,...,k$ přiřazen vlastní modifikátor vlivu U_j. Všechny tyto modifikátory vlivu ale mohou být jedna a ta samá proměnná. V další kapitole nastíníme několik různých metod odhadů modelů s proměnlivými koeficienty. Některé z nich jsou navrženy pro různý modifikátor vlivu pro každý regresor, ale některé používají pouze jediný modifikátor vlivu, stejný pro všechny regresory.

Dále pouze pro zajímavost uvedeme i dva pokročilé speciální typy modelů s proměnlivými koeficienty, kterými se už ale v práci dále nebudeme zabývat.

- (Hastie a Tibshirani, 1993) navrhli Bayessovský tvar modelu s časově proměnlivými koeficienty $Y_t = X_t\beta(t) + \varepsilon_t$ pro pozorování v časech $t = 1,...,n$, odezvu Y_t a jediný regresor X_t. U koeficientu $\beta(t)$ předpokládáme apriorní informaci a model pak vypadá následovně:

$$\begin{aligned} Y_t &= X_t\beta(t) + \nu_t, & \nu_t &\sim N(0,V_t), \\ \beta(t) &= G_t\beta(t-1) + t\omega_t, & \omega_t &\sim N(0,W_t). \end{aligned} \tag{1.3}$$

  První rovnice se nazývá rovnice pozorování a druhá evoluční. Regresní parametr β s časovým modifikátorem vlivu t je zde definován jako Markovův proces.

- (Fan a kol., 2003) představili zcela novou třídu modelů s proměnlivými koeficienty – adaptivní modely. Ty zde zmíníme proto, aby si čtenář udělal představu o obrovském potenciálu těchto modelů. Jak jsme již zmínili, velký přínos modelů s proměnlivými koeficienty vidíme v hledání skrytých závislostí mezi trojicí odezva, regresor a modifikátor vlivu. Při velkém množství vstupních proměnných před modelářem stojí zásadní problém výběru modifikátoru či modifikátorů vlivu. Proměnné jako čas či věk se nabízejí jako zřejmé, avšak pro nalezení oněch ne zcela zřejmých závislostí je třeba testovat různé modely. To ale při velkém počtu regresorů není prakticky možné. Analogicky ani v lineární regresi nelze při počtu prediktorů v rámci stovek a tisíců ručně přidávat či odebírat proměnné z modelu. Proto jsou používány různé výběrové algoritmy, kupříkladu forward či backward selekce. Tyto metody bohužel pro modely s proměnlivými koeficienty doposud nejsou k dispozici, ale jako krok tímto směrem vidíme právě adaptivní modely, které jako modifikátor vlivu dosadí kombinaci všech regresorů, z nichž je pak podle hodnot odhadnutých koeficientů α možno vybrat ty relevantní, tj. ty s nenulovou hodnotou. Jak jsme již naznačili, autoři podstatně rozšířili modelovací schopnosti těchto modelů nahrazením modifikátoru vlivu lineární kombinací všech modifikátorů vlivu. Jejich model je ve tvaru:

$$Y_i = \sum_{j=1}^{k} \beta_j(\alpha^T \boldsymbol{U}_i)X_{i,j} + \varepsilon_i, \tag{1.4}$$

  kde se nám oproti základnímu tvaru modelu s proměnlivými koeficienty v rovnici (1.1) namísto modifikátoru vlivu j-tého regresoru $U_j, j = 1,...,k$ objevuje kompozitní modifikátor vlivu $\alpha^T\boldsymbol{U}_i$. Koeficient $\alpha \in \mathbb{R}^k$ značí vlivy jednotlivých modifikátorů na výsledný modifikátor vlivu $\alpha^T\boldsymbol{U}_i$.

  Všimněme si, že koeficient α nezávisí na pořadí regresoru $j = 1,...,k$. Tím pádem všechny regresory mají modifikátor vlivu kalkulovaný stejným způsobem. Pokud pomineme výpočetní náročnost odhadu takovéhoto typu modelu s proměnlivými koeficienty, pak by bylo možno si do budoucna představit rozvoj těchto modelů přidáním odhadu α_j pro každý vektor regresoru $\boldsymbol{X}^j, j = 1,...,k$ zvlášť.

Tím jsme popsali tvar modelu s proměnlivými koeficienty pro standardní data s nezávislými pozorováními a některé jeho speciální případy. Často se ale můžeme setkat i s jiným druhem dat. Řeč je o datech longitudinálních, tedy datech, kde pro jednotlivé subjekty máme k dispozici vícero různých pozorování. Ty reprezentují spojení regresní analýzy a analýzy časových řad. S takovýmto typem dat se asi např. setkáme v lékařství, kdy pro jednotlivé pacienty disponujeme záznamy z různých časů. Modely s proměnlivými koeficienty lze aplikovat i na tyto data. Metody jejich odhadu a inference byly popsány v (Huang a kol., 2004).

Budeme uvažovat časová longitudinální data a zavedeme si následující značení. Mějme k dispozici n pozorovaných nezávislých subjektů, přičemž ke každému z nich ještě máme n_i pozorování v časech $t_{i,l}, l = 1,...,n_i$. Závislou proměnnou pro i-tý subjekt a jeho l-té pozorování označíme $Y_{i,l} = Y_i(t_{i,l})$ a vektor pozorování závislé proměnné pro i-tý subjekt jako $\boldsymbol{Y}_i = (Y_{i,1},...,Y_{i,n_i})^T$. Dále si definujme $X_{i,l,j}, j = 1,...,k$ jako j-tý regresor u l-tého pozorování i-tého subjektu a vektor $\boldsymbol{X}_{i,l} = (X_{i,l,1},...,X_{i,l,k})^T = \boldsymbol{X}_i(t_{i,l})$. Pozorování regresorů si označme jako $\mathcal{D} = \{(\boldsymbol{X}_i(t_{i,l}), t_{i,l}), i = 1,...,n, l = 1,...,n_i\}$.

Vzhledem k povaze těchto dat budeme používat pouze jediný modifikátor vlivu, a to čas l-tého pozorování i-tého subjektu $t_{i,l}$. Základní tvar modelu s proměnlivými koeficienty a longitudinálními daty je pak definován následovně:

$$Y_{i,l} = \sum_{j=1}^{k} \beta_j(t_{i,l}) X_{i,l,j} + \varepsilon_{i,l}, \tag{1.5}$$

kde $\beta_j(t_{i,l})$ značí j-tý koeficient v čase $t_{i,l}$ a $\varepsilon_{i,l}$ náhodnou chybu s rozdělením $N(0, \sigma^2(t_{i,l}))$. Vektory chyb i-tých subjektů $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1},...,\varepsilon_{i,n_i})^T, i = 1,...,n$ jsou nezávislé a stejně rozdělené. Jednotlivé chyby pro i-tý subjekt $\varepsilon_{i,l}, l = 1,...,n_i$ však již nezávislé být nemusí, a proto jsou jak odhad, tak statistická inference pro tyto modely složitější, než u modelu s proměnlivými koeficienty pro standardní data definovaného rovnicí (1.1).

2. Metody odhadů

Modely s proměnlivými koeficienty se oproti klasickým zobecněným lineárním modelům liší ve dvou bodech. Jedná se o proměnlivou strukturu parametrů v závislosti na nějakém modifikátoru vlivu a jimi podmíněné složitější metody odhadu. Právě neparametrický přístup poskytuje modelům s proměnlivými koeficienty flexibilitu. Jak již ovšem bylo řečeno, nespočetné možnosti odhadnuté funkce v neparametrických metodách přináší svá vlastní úskalí. Kvůli kletbě dimenzionality je třeba množinu funkcí použitou k odhadu dostatečně omezit. (Hastie a Tibshirani, 1993) popisují několik možností takovéto restrikce. Jako nejjednodušší příklad uvádějí definovat si nějakou parametrickou bázi, kupříkladu polynomy či trigonometrické funkce. Tato metoda ovšem poskytuje příliš málo flexibility. Alternativou by mohlo být použití regresních splajnů s pevně stanovenými uzly. U této metody ale nesprávná volba umístění uzlů vede k vychýleným výsledkům. Dále je možné omezit se kupříkladu na Fourierovy řady či použít nějakou obecnější neparametrickou restrikci jako hladké funkce. Odhad takovýchto obecných neparametrických funkcí je pak taktéž možno aplikovat různými způsoby, ať už přes lokální jádrové metody, penalizaci nebo stochastické Bayesovy formulace.

Způsobů odhadu koeficientů v modelech s proměnlivými koeficienty bylo v literatuře prozatím publikováno pouze několik. Ty se dají rozdělit na dva různé přístupy – lokálně regresní metody a odhady pomocí splajnů. Je jisté, že s rozvojem v oblasti neparametrické regrese se budou vyvíjet i modely s proměnlivými koeficienty, ať už publikací nových metod odhadu či hlubší statistickou inferencí. V dnešní době jsou ale v literatuře důkladně popsány pouze tyto dva způsoby odhadu. V následujících podkapitolách si stručně nastíníme tyto metody, jelikož se od nich bude odvíjet celá konstrukce statistické inference.

V první kapitole jsme definovali jak základní gaussovský tvar modelů s proměnlivými koeficienty, tak i tvar zobecněný. V této kapitole se budeme věnovat pouze odhadu gaussovské varianty. Odhady zobecněného tvaru se konstruují obdobně, ale s pomocí maximální věrohodnosti a aplikace algoritmu typu Newton-Raphson. Odhady pomocí vyhlazovacích splajnů (tzv. smoothing spline), penalizovaných splajnů a pomocí lokální regrese byly autory definovány pro základní Gaussovský tvar modelů s proměnlivými koeficienty a standardními daty (viz. (1.1)). Naproti tomu odhad pomocí polynomiálních splajnů byl zkonstruován speciálně pro použití longitudinálních dat a tak ho také uvedeme. Pro tento přístup je třeba brát definici (1.5) a její příslušné značení, které se od výše zmíněného základního trochu liší.

2.1 Odhad pomocí splajnů

V této podkapitole se seznámíme se třemi různými způsoby odhadu modelu s proměnlivými koeficienty pomocí splajnů – vyhlazovacími splajny, polynomiálními splajny a penalizovanými splajny. Splajn řádu d s uzly $\xi_0 \leq \xi_1 \leq ... \leq \xi_{M+1}$ je hladká křivka tvořená na každém intervalu $[\xi_g,\xi_{g+1}], g = 0,...,M$ polynomem řádu d. Hladkost křivky znamená, že má na celém svém definičním oboru $[\xi_0, \xi_{M+1}]$ spojité derivace až do řádu $d-1$. Dvojice $\{\xi_0,\xi_{M+1}\}$ je často označována jako vnější uzly ohraničující definiční obor splajnu a $\xi_1,...,\xi_M$ jako uzly vnitřní.

Odhadnuté funkcionální regresní koeficienty v těchto metodách budou pro modely s proměnlivými koeficienty vždy splajny. V čem se tedy tyto metody liší? Rozdíl není v typu křivky, ale ve volbě uzlů ξ_1,...,ξ_M a penalizaci nehladkosti.

- Vyhlazovací splajny definujeme jako splajny, kde si za uzly zvolíme všechny unikátní pozorované hodnoty proměnné, kterou dosazujeme na pomyslnou osu x. V případě modelů s proměnlivými koeficienty tedy půjde o unikátní hodnoty pozorovaných hodnot příslušného modifikátoru vlivu U_j. Zároveň je v rámci odhadu aplikován penalizační parametr λ definující míru vyhlazení odhadnuté křivky.

- Polynomiální splajny bývají často označovány jako regresní splajny. Na rozdíl od vyhlazovacích splajnů si u nich můžeme sami zvolit řád splajnu a při jejich odhadu nedochází k penalizaci nehladkosti. Taktéž výběr vnitřních uzlů probíhá jiným způsobem. Nejprve se kupříkladu pomocí tzv. cross-validace vybere počet vnitřních uzlů $K < n$. Ty jsou následně vybrány buď jako odpovídající kvantily náhodné veličiny U nebo ekvidistantně, tj. rozdělením definičního oboru na $K+1$ stejně velkých intervalů s uzly v hraničních bodech, případně jinými vhodnými způsoby.

- Penalizované splajny kombinují přístup vyhlazovacích a polynomiálních splajnů. Volba uzlů a řádu probíhá shodně s polynomiálními splajny a navíc se při odhadu aplikuje penalizační člen nehladkosti.

U publikovaných metod odhadů modelů s proměnlivými koeficienty všechny tyto tři přístupy ke konstrukci splajnů používají B-splajnovou bázi. Jejími argumenty jsou řád splajnu a množina uzlů (vnitřních i vnějších). B-splajnová báze pak tvoří bázi lineárního prostoru všech splajnových funkcí daného řádu a uzlů. Konstruuje se rekurzivně pomocí de Boorovy formule.

Mějme definován řád splajnu d a uzly $\xi_0 \leq \xi_1 \leq ... \leq \xi_{M+1}$. Vnější uzly d-krát zreplikujeme a dostaneme posloupnost uzlů $\xi_{-d} = ... = \xi_0 \leq \xi_1 ... \leq \xi_{M+1} = ... = \xi_{M+d+1}$. Označme si $B_{v,p}(\xi)$ jako v-tou B-splajnovou bazickou funkci řádu p v bodě $\xi \in [\xi_0, \xi_{M+1}]$. Ta je definována následovně

$$B_{v,0}(\xi) = \begin{cases} 1 & \text{pro } \xi_v \leq \xi < \xi_{v+1}, \\ 0 & \text{jinak,} \end{cases}$$

$$B_{v,p}(\xi) = \frac{\xi - \xi_v}{\xi_{v+p} - \xi_v} B_{v,p-1}(\xi) + \frac{\xi_{v+p+1} - \xi}{\xi_{v+p+1} - \xi_{v+1}} B_{v+1,p-1}(\xi),$$

pokud je nějaký z jmenovatelů nulový, pak je danému zlomku přiřazena nulová hodnota. Pomocí této rekurzivní formule zkonstruujeme bazické funkce požadovaného řádu d a ty tvoří B-splajnovou bázi vybraného splajnu. Existují i jiné způsoby konstrukce báze prostoru splajnových funkcí, avšak B-splajnová báze se těší největší popularitě kvůli její výpočetní stabilitě. Většina statistických softwarů dnes disponuje funkcí konstruující B-splajnovou bázi.

2.1.1 Model se standardními daty

V této sekci se zaměříme na metody odhadů modelů s proměnlivými koeficienty pomocí splajnů na standardních datech s nezávislými pozorováními de-

finovaných v rovnici (1.1). Každý regresor $X_j, j = 1,...,k$ má přiřazený vlastní modifikátor vlivu U_j.

Vyhlazovací splajny

První odhadová procedura modelů s proměnlivými koeficienty byla navržena již v (Hastie a Tibshirani, 1993) a to odhad pomocí vyhlazovacího splajnu přes metodu penalizovaných nejmenších čtverců.

Budeme uvažovat model v klasickém tvaru

$$Y_i = \sum_{j=1}^{k} \beta_j(U_{i,j}) X_{i,j} + \varepsilon_i, \qquad i = 1,...,n, \tag{2.1}$$

kde $Y_1,...,Y_n$ značí pozorování závislé proměnné Y, $X_{i,j}$ i-té pozorování j-tého regresoru a $U_{i,j}$ ekvivalentně definovaný modifikátor vlivu. Pro odhad funkcí β_j se použije metoda penalizovaných nejmenších čtverců a budeme minimalizovat

$$\min_{\beta_j \in \mathbb{C}^2, j=1,...,k} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} X_{i,j} \beta_j(U_{i,j}) \right)^2 + \sum_{j=1}^{k} \lambda_j \int \beta_j''(u)^2 du, \tag{2.2}$$

kde $\mathbb{C}^2$ značí funkce se spojitými derivacemi do řádu 2. Parametry λ_j kontrolují míru vyhlazení a to tím způsobem, že vyšší hodnota implikuje hladší výslednou křivku. Před minimalizací je zapotřebí nejprve zvolit jejich hodnoty. Ty můžeme získat metodami cross-validace či zobecněné cross-validace, nebo pomocí volby požadovaných stupňů volnosti dané funkce a následné volby parametrů λ_j, které nám dají požadované výsledky.

Taktéž je třeba si uvědomit, že tato minimalizace probíhá přes nějaké funkce $\beta_1,..., \beta_k$ a tím pádem před námi stojí nekonečně dimenzionální problém. Takovou úlohu neumíme efektivně vypočítat a je proto třeba tento problém řešit jiným způsobem. Reprezentujme si funkce $\beta_j, j = 1,...,k$ pomocí kubických splajnů a jejich tvar pak lze pomocí B-splajnové báze přepsat do tvaru

$$\beta_j(U_{i,j}) = \sum_{s=1}^{S_j} \gamma_{s,j} B_{s,d}^{(j)}(U_{i,j}). \tag{2.3}$$

B-splajnová báze je definována jednoznačně stupněm splajnu a jeho uzly. Vzhledem ke tvaru minimalizační úlohy (2.2) bude takto konstruovaný splajn kubický, tedy $d = 3$. Horní index j v zápisu B-splajnové báze $B_{s,d}^{(j)}(U_{i,j})$ znamená, že pro její konstrukci používáme množinu uzlů pro j-tý koeficient. Jako uzly bereme pozorování modifikátoru vlivu $\boldsymbol{U}_j$. B-splajnová báze již je konečná a budeme řešit konečně dimenzionální problém odhadu parametrů $\gamma_{s,j}, j = 1,...,k, s = 1,...,S_j$.

Pokud jako $\boldsymbol{\beta}_j$ označíme $(\beta_j(U_{1,j}),..., \beta_j(U_{n,j}))^T$ a bazickou matici $\mathcal{B}_j$ s is-tým členem definovaným jako $B_{s,d}^{(j)}(U_{i,j})$, pak lze rovnici (2.3) zapsat maticově jako

$$\boldsymbol{\beta}_j = \mathcal{B}_j \boldsymbol{\gamma}_j, \tag{2.4}$$

kde $\boldsymbol{\gamma}_j = (\gamma_{1,j},..., \gamma_{S_j,j})^T$. Minimalizační problém přes koeficienty $\boldsymbol{\gamma}_j$ pak lze zapsat v maticovém tvaru následovně

$$\min_{\boldsymbol{\gamma}_j \in \mathbb{R}^{S_j}, j=1,...,k} \left\| \boldsymbol{Y} - \sum_{j=1}^{k} \mathbf{D}_j \mathcal{B}_j \boldsymbol{\gamma}_j \right\|^2 + \sum_{j=1}^{k} \lambda_j \parallel \boldsymbol{\gamma}_j \parallel^2_{\Omega_j}, \tag{2.5}$$

kde $\boldsymbol{Y} = (Y_1,...,Y_n)^T$ je vektor pozorovaných hodnot závislé proměnné, $\mathbf{D}_j$ diagonální matice s n pozorovanými hodnotami regresoru X_j, $\parallel \boldsymbol{\gamma}_j \parallel^2_{\Omega_j} = \boldsymbol{\gamma}_j^T \boldsymbol{\Omega}_j \boldsymbol{\gamma}_j$ a $\boldsymbol{\Omega}_j$ značí matici s ik-tým elementem definovaným jako $\int_0^\infty B_{i,d}^{(j)''}(u) B_{k,d}^{(j)''}(u) du$. Výraz $\| \cdot \|$ zde značí L_2-normu.

Úlohy (2.2) a (2.5) nejsou ekvivalentní, avšak bylo ukázáno, že při reprezentaci pomocí vyhlazovacího splajnu (řád 3 a uzly jako unikátní pozorované hodnoty modifikátorů vlivu $U_j, j = 1,...,k$) tyto dvě úlohy dávají stejná řešení vzhledem k datům. Ze získaných odhadů $\hat{\boldsymbol{\gamma}}_j, j = 1,...,k$ z minimalizace (2.5) pak tím pádem můžeme dopočítat odhady $\hat{\beta}_j$ přes kupříkladu tzv. backfittingové procedury.

Penalizované splajny

Nevýhodou odhadu pomocí polynomiálních splajnů je nutnost správné volby počtu a lokace uzlů. Jedná se o komplexní problém nelineární optimalizace a při špatné volbě těchto parametrů dochází k nepřesnostem v odhadech. (Marx, 2010) proto navrhli metodu penalizovaných splajnů (P-splajny), která vyhlazuje křivky odhadnutých parametrů ve dvou krocích. Nejprve využije bohaté regresní báze k záměrnému přefitování vektoru hladkých parametrů se skromným počtem stejně od sebe vzdálených B-splajnů. V kroku druhém probíhá vyhlazování přes penalizaci diference mezi sousedícími koeficienty B-splajnů, čímž nám vzniknou právě hledané P-splajny.

Jako v předchozích metodách se funkce $\beta_j, j = 1,...,k$ vyjadřují pomocí B-splajnové báze přes parametry $\boldsymbol{\gamma}_j$ a cílem je minimalizovat výraz

$$\min_{\boldsymbol{\gamma}_j \in \mathbb{R}^{S_j}, j=1,...,k} \left\| \boldsymbol{Y} - \sum_{j=1}^{k} \mathbf{D}_j \mathcal{B}_j \boldsymbol{\gamma}_j \right\|^2 + \sum_{j=1}^{k} \lambda_j \| \Delta_d \boldsymbol{\gamma}_j \|^2, \tag{2.6}$$

kde $\boldsymbol{Y}$ je vektor pozorované závislé proměnné, $\mathbf{D}_j$ diagonální matice s n pozorovanými hodnotami regresoru X_j a $\mathcal{B}_j$ matice B-splajnové báze. $\Delta_z \boldsymbol{\gamma}_j$ značí z-tou diferenci $\boldsymbol{\gamma}_j$, kde první diference $\Delta_1 \boldsymbol{\gamma}_j$ je definována jako $\boldsymbol{\gamma}_{j+1} - \boldsymbol{\gamma}_j$, druhá diference $\Delta_2 \boldsymbol{\gamma}_j$ jako $(\Delta_1 \boldsymbol{\gamma}_{j+1} - \Delta_1 \boldsymbol{\gamma}_j)$ a analogickým způsobem pro další stupně. Definujme si matici $\mathbf{V} = \mathbf{D}_j \mathcal{B}_j$ a vektor $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T,..., \boldsymbol{\gamma}_k^T)^T$. Dále použijeme také matici $\mathbf{M}_z$, která konstruuje z-té diference $\boldsymbol{\gamma}$ jako $\mathbf{M}_z \boldsymbol{\gamma} = \Delta_z \boldsymbol{\gamma}$. Pak můžeme odhad parametrů $\hat{\boldsymbol{\gamma}}$ pro nějakou $\boldsymbol{\lambda} = (\lambda_1,..., \lambda_k)^T$ zapsat jako

$$\hat{\boldsymbol{\gamma}} = (\mathbf{V}^T \mathbf{V} + \lambda \mathbf{M}_z^T \mathbf{M}_z)^{-1} \mathbf{V}^T \boldsymbol{Y}. \tag{2.7}$$

I v této metodě je nutné předem zvolit parametry λ_j pro každou hledanou křivku parametru β_j, čehož je dosaženo použitím cross-validace nebo AIC kritéria. Dále je možno volit i proměnnou z vyjadřující stupeň diference. (Marx, 2010) doporučují volbu alespoň druhého stupně diference a bázi kvadratického nebo kubického B-splajnu.

2.1.2 Model s longitudinálními daty

V této podkapitole se seznámíme ještě s jedním způsobem odhadu modelů s proměnlivými koeficienty založeným na splajnových metodách – odhadem pomocí polynomiálního splajnu. Ten je však navržen speciálně pro použití na longitudinálních datech a proto se od výše popsaných metod liší.

Polynomiální splajny

Další metoda odhadu modelů s proměnlivými koeficienty pomocí splajnů je založena na polynomiálních splajnech a byla navržena v (Huang a kol., 2004). Polynomiální splajny jsou křivky tvořené po částech polynomy, které na sebe hladce navazují na množině vnitřních bodů, kterým se říká uzly. Polynomiální splajn stupně d s uzly $\xi_0 < \xi_1 < ... < \xi_{M+1}$, kde ξ_0 a ξ_{M+1} jsou koncové body intervalu uzlů, je na každém z intervalů $[\xi_m,\xi_{m+1}), 0 \leq m \leq M-1$ a $[\xi_M,\xi_{M+1}]$ polynomem stupně d a má všude spojitou derivaci do řádu $d-1$. Reprezentace koeficientů $\beta_j, j = 1,...,k$ pomocí těchto splajnů můžeme zapsat jako

$$\beta_j(t_{i,l}) = \sum_{s=1}^{S_j} \gamma_{s,j} B_{s,d}^{(j)}(t_{i,l}), \qquad j = 1,...,k, \tag{2.8}$$

kde pro každé $j = 1,...,k, B_{s,d}^{(j)}(.), s = 1,...,S_j$ je opět B-splajnová báze stupně d na intervalu $[\xi_0,\xi_{M+1}]$ s fixním počtem uzlů S_j a jejich pořadím. Všechny odhadované hodnoty času t tedy musí ležet na intervalu $[\xi_0,\xi_{M+1}]$.

Model s proměnlivými koeficienty pro longitudinální data nyní můžeme zapsat ve tvaru

$$Y_{i,j} = \sum_{j=1}^{k} \sum_{s=1}^{S_j} X_{i,l,j} B_{s,d}^{(j)}(t_{i,j}) \gamma_{s,j} + \varepsilon_{i,j}. \tag{2.9}$$

Definujme si následující výrazy:

$$\boldsymbol{B}(t) = \begin{pmatrix} B_{1,d}^{(1)}(t) & ... & B_{S_1,d}^{(1)}(t) & 0 & ... & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & 0 & 0 & ... & 0 & B_{1,d}^{(k)}(t) & ... & B_{S_k,d}^{(k)}(t) \end{pmatrix},$$

a pro $i = 1,...,n$, $j = 1,...,k$ a $l = 1,...,n_i$

$$\begin{aligned} \boldsymbol{U}_{i,l}^T &= \boldsymbol{X}_i^T(t_{i,l}) \boldsymbol{B}(t_{i,l}), \\ \boldsymbol{U}_i &= (\boldsymbol{U}_{i,1},...,\boldsymbol{U}_{i,n_i})^T, \\ \boldsymbol{W}_i &= diag(\omega_i,...,\omega_i), \end{aligned}$$

kde váhy ω_i mohou nabývat hodnoty 1 pro přiřazení stejné váhy každému pozorování nebo hodnoty $1/n_i$ pro přiřazení stejné váhy každému pozorovanému subjektu. Matice $\boldsymbol{W}_i$ má dimenzi k.

Odhady $\gamma_{s,j}, j = 1,...,k, s = 1,...,S_j$ z rovnice (2.9) vypočteme minimalizací výrazu

$$\min_{\gamma_j \in \mathbb{R}^{S_j}, j=1,...,k} \sum_{i=1}^{n} \omega_i \sum_{l=1}^{n_i} \left(Y_{i,j} - \sum_{j=1}^{k} \sum_{s=1}^{S_j} X_{i,l,j} B_{s,d}^{(j)}(t_{i,l}) \gamma_{s,j} \right)^2. \tag{2.10}$$

Rovnici (2.10) pak můžeme pomocí výše uvedených výrazů zapsat maticově jako

$$\min_{\gamma_j \in \mathbb{R}^{S_j}, j=1,...,k} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \boldsymbol{U}_i \boldsymbol{\gamma})^T \boldsymbol{W}_i (\boldsymbol{Y}_i - \boldsymbol{U}_i \boldsymbol{\gamma}), \tag{2.11}$$

díky které můžeme vypočítat unikátní odhad $\hat{\boldsymbol{\gamma}}$ ve tvaru

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{n} \boldsymbol{U}_i^T \boldsymbol{W}_i \boldsymbol{U}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{U}_i^T \boldsymbol{W}_i \boldsymbol{Y}_i, \tag{2.12}$$

a odhad hledaného parametru $\hat{\beta}_j(t), j = 1,...,k$ v nějakém čase t má tvar

$$\hat{\beta}_j(t) = \sum_{s=1}^{S_j} \hat{\gamma}_{s,j} B_{s,d}^{(j)}(t).$$

V této metodě je třeba předem zvolit počty uzlů pro jednotlivé splajnové funkce $\beta_j, j = 1,...,k$ a případně i jejich umístění (pokud bychom se rozhodli nepoužít ekvidistantní uzly či uzly v příslušných kvantilech daného modifikátoru vlivu). K tomu je možno použít cross-validaci nebo informační kritéria jako AIC či BIC. (Huang a kol., 2004) ukázali, že nejoptimálněji se jeví volba počtu ekvidistantních uzlů podle kritéria AIC.

2.2 Odhad pomocí lokální regrese

Všechny doposud popsané metody používali pro odhad parametrických křivek splajny. Jako kvalitní alternativa se jeví použití lokální regrese, která operuje pouze s lokálními daty a ve výsledku je považována za méně náročnou ve smyslu výpočetní složitosti. Lokálnost zde spočívá v principu, že odhad funkce β_j konstruujeme vždy pouze pro určité okolí zvoleného bodu u_0 z oboru hodnot modifikátoru vlivu U namísto konstrukce celé funkce jako v předchozích metodách odhadu pomocí splajnů. Blíže tuto metodu vysvětlíme u minimalizační funkce jednokrokového odhadu.

Dalším rozdílem oproti odhadu pomocí splajnů je počet různých modifikátorů vlivu. Vzpoměnme si, že u odhadu pomocí splajnů byl každému regresoru $X_j, j = 1,...,k$ definován vlastní modifikátor vlivu U_j. U odhadu pomocí lokální regrese však používáme pouze jediný modifikátor vlivu U. Při použití vlastního modifikátoru vlivu ke každému regresoru bychom totiž museli zvolit k-rozměrný bod u_0 pro konstrukci globálního odhadu požadovaných křivek $\beta_j(U)$, což by ale vyžadovalo více pozorování a výpočetní náročnost takového úkolu by pravděpodobně byla neúnosně vysoká.

Tato metoda byla navržena pouze pro standardní data s nezávislými pozorováními.

K aplikaci zmíněné lokálnosti se budou používat jádrové funkce (tzv. kernely). Ty jsou definovány jako symetrické funkce K takové, že $K(x) \geq 0$ splňuje podmínky $\int_{\mathbb{R}} K(x)dx = 1$, $\int_{\mathbb{R}} xK(x)dx = 0$ a $\int_{\mathbb{R}} x^2K(x)dx > 0$. Nejčastěji se používá Epanechnikova jádrová funkce definovaná ve tvaru

$$K(x) = \frac{3}{4}(1 - x^2)I_{|x|\leq 1}(x), \text{ pro } x \in \mathbb{R}.$$

Jádrových funkcí existuje velké množství a v této metodě je možné zvolit si jakoukoliv z nich, avšak výsledné odhady by se při použití různých jádrových funkcí neměly výrazně lišit. Naopak zásadní roli v odhadu hraje vhodná volba šířky pásma, kterou popíšeme později.

Jednokrokový odhad

Fan a Zhang (1999) navrhli jednokrokovou metodu odhadu modelu s proměnlivými koeficienty pomocí lokální regrese. Předpokládejme, že máme k dispozici standardní data s nezávislými pozorováními $\{(Y_i, X_{i,1},..., X_{i,k}, U_i)\}_{i=1}^n$ a model s proměnlivými koeficienty s jediným modifikátorem vlivu U. Proměnlivé koeficienty β_j budeme odhadovat lokální lineární regresí a pro každý bod u_0 si tyto funkce lokálně aproximujeme useknutou Taylorovou řadou jako

$$\beta_j(U) \approx a_j + b_j(U - u_0). \tag{2.13}$$

Následně se minimalizuje výraz níže

$$\min_{a_j,b_j\in\mathbb{R},j=1,...,k} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} \big[a_j + b_j(U_i - u_0)\big] X_{ij} \right)^2 K_h(U_i - u_0), \tag{2.14}$$

pro zvolené jádro K s šířkou pásma h a výrazem $K_h(\cdot)$ definovaným jako $K(\cdot/h)/h$.

Řešení této rovnice nám ale dá odhadnuté koeficienty křivek $\hat{a}, \hat{b}$ pouze v okolí bodu u_0. Logicky je proto nutné tento postup opakovat i pro jiné body u_0, abychom dostali odhady po celé délce křivky. Máme k dispozici n pozorovaných hodnot modifikátoru vlivu U_i a proto zkonstruujeme n lokálních odhadů pro zvolené body $u_0 = U_i, i = 1,..., n$ pomocí minimalizace rovnicí (2.14). Bodové odhady $\hat{\beta}_j(U_i), i = 1,..., n, j = 1,..., k$ jsou pak rovny odhadům $\hat{a}_j(U_i), i = 1,..., n, j = 1,..., k$. Tímto způsobem z hledaných křivek β_j dostaneme jejich odhady v bodech $U_i, i = 1,..., n$ a při dostatečně hustém pokrytí domény U i výslednou křivku.

Tato metoda tedy namísto jedné velké a složité minimalizace používá mnoho jednoduchých minimalizací a právě v tom spočívá její menší výpočetní složitost při složitějších úlohách. Taktéž se zde naskýtá prostor pro vhodnější volbu množiny bodů $U_i, i = 1,..., n$ na kterých provádíme lokální odhad. Kupříkladu je možné vzít si pouze unikátní hodnoty $U_i, i = 1,..., n$ abychom neprováděli stejnou operaci vícekrát nebo si zvolit jen část z této množiny bodů dle nějakého kritéria.

Odhady $\hat{\boldsymbol{\beta}}(u_0) = (\hat{\beta}_1(u_0),..., \hat{\beta}_k(u_0))^T$ lze z rovnice (2.14) maticově vyjádřit jako

$$\hat{\boldsymbol{\beta}}(u_0) = (\boldsymbol{I}_k, \boldsymbol{0}_k)(\boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^h \boldsymbol{\Gamma}_{u_0})^{-1} \boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^h \boldsymbol{Y}, \tag{2.15}$$

kde $\boldsymbol{I}_k$ je jednotková matice velikosti k, $\mathbf{0}_k$ matice taktéž velikosti k, která má každý člen nulový a

$$\begin{aligned} \mathbf{U}_{u_0} &= diag(U_1 - u_0,...,U_n - u_0), \\ \boldsymbol{\Gamma}_{u_0} &= (\mathcal{X}, \mathbf{U}_{u_0}\mathcal{X}), \\ \boldsymbol{W}_{u_0}^h &= diag(K_h(U_1 - u_0),...,K_h(U_n - u_0)). \end{aligned}$$

Obrovská nevýhoda této metody ale spočívá v předpokladu, že všechny funkce parametrů β_j disponují stejnou mírou hladkosti, která je algoritmem aplikována šířkou pásma h a useknutou Taylorovou řadu, jíž aproximujeme funkce $\beta_j(U)$. Hodnotu šířky pásma h volíme pomocí cross-validace.

Dvoukrokový odhad

Fan a Zhang (1999) kromě jednokrokového odhadu navrhli i odhad dvoukrokový, který řeší problémy s různými stupni hladkosti. Autoři ukázali, že předchozí metoda nedokáže optimálně odhadnout hladší komponenty, nezávisle na volbě šířky pásma. V prvním kroku tento algoritmus zvolí malou šířku pásma a najde odhady stejným způsobem jako v jednokrokovém odhadu. V kroku druhém jsou tyto odhady použity u méně hladkých parametrů a u zbylých (hladších) parametrů dochází k dalšímu vyhlazení pomocí větší šířky pásma a aproximací useknutou Taylorovou řadou většího stupně.

Nechť $\boldsymbol{\beta} = (\beta_1,...,\beta_k)^T$. Bez újmy na obecnosti předpokládejme, že β_k je hladší než $\beta_j, j = 1,...,k-1$, které mají stejný stupeň hladkosti. Hladkost zde můžeme vyjádřit pomocí spojitosti derivací. Předpokládejme, že β_k má spojitou čtvrtou derivaci a zbylé β_j pouze druhou. Model s proměnlivými koeficienty pak zapíšeme ve tvaru

$$Y_i = \sum_{j=1}^{k-1} \beta_j(U_i)X_{i,j} + \beta_k(U_i)X_{i,k} + \varepsilon_i, \qquad i = 1,...,n. \tag{2.16}$$

V prvním kroku dojde k odhadu jednokrokovou metodou a odhady $\hat{\beta}_j$ jsou ve tvaru z rovnice (2.15). V modelu (2.16) nahradíme méně hladké parametry $\beta_j, j = 1,...,k-1$ těmito odhady a dostáváme výraz

$$Y_i - \sum_{j=1}^{k-1} \hat{\beta}_j(U_i)X_{i,j} = \beta_k(U_i)X_{i,k} + \varepsilon_i, \qquad i = 1,...,n. \tag{2.17}$$

V prvním kroku byly parametry β_j aproximovány useknutou Taylorovou řadou. My se ale teď zabýváme pouze parametrem β_k, který je hladší a proto ho můžeme odhadnout Taylorovou řadou třetího stupně

$$\beta_k(U) \approx \sum_{m=0}^{3} \frac{\beta_k^{(m)}(u_0)(U - u_0)^m}{m!},$$

a tento výraz pak minimalizujeme s větší šířkou pásma h_1

$$\min_{a_{k,0},..,a_{k,3}\in\mathbb{R}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k-1} \hat{\beta}_j(U_i)X_{i,j} - X_{i,k} \sum_{m=0}^{3} a_{k,m}(U_i - u_0)^m \right)^2 K_{h_1}(U_i - u_0). \tag{2.18}$$

Výsledný odhad $\hat{\beta}_k$ koresponduje s odhadem minimalizovaného parametru $\hat{a}_{k,0}$ a maticově má tvar

$$\hat{\beta}_k(u_0) = \boldsymbol{e}_{1,4}^T(\boldsymbol{G}_{\boldsymbol{u_0}}^T \boldsymbol{W}_{u_0}^{h_1} \boldsymbol{G}_{\boldsymbol{u_0}})^{-1} \boldsymbol{G}_{\boldsymbol{u_0}}^T \boldsymbol{W}_{u_0}^{h_1} \tilde{\boldsymbol{Y}}, \tag{2.19}$$

kde $\tilde{\boldsymbol{Y}} = (\tilde{Y}_1,...,\tilde{Y}_n)$, $\boldsymbol{e}_{1,4}$ je 4×1 vektor s jedničkou v prvním řádku a nulami jinde a

$$\begin{aligned}
\tilde{Y}_i &= Y_i - \sum_{j=1}^{k-1} \hat{\beta}_j(U_i) X_{i,j}, \\
\boldsymbol{W}_{u_0}^{h_1} &= diag(K_{h_1}(U_1 - u_0),..., K_{h_1}(U_n - u_0)), \\
\boldsymbol{G}_{u_0} &= diag(X_{1,k},..., X_{n,k}) \boldsymbol{Q}_{u_0},
\end{aligned}$$

$$\boldsymbol{Q}_{u_0} = \begin{pmatrix} 1 & U_1 - u_0 & (U_1 - u_0)^2 & (U_1 - u_0)^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & U_n - u_0 & (U_n - u_0)^2 & (U_n - u_0)^3 \end{pmatrix}.$$

Alternativou k tomuto dodatečnému vyhlazení koeficientu β_k může být nahrazení druhého kroku jednoduchým vyhlazením $\hat{\beta}_k$ z prvního kroku pomocí lokální kubické regrese se šířkou pásma h_1. Fan a Zhang (1999) ukázali, že dvoukroková metoda vždy přináší lepší výsledky než procedura jednokroková. I v případě, že všechny parametry mají stejnou hladkost, dosahuje dvoukroková metoda stejně kvalitních výsledků jako metoda jednokroková.

3. Statistická inference

V předchozí kapitole jsme si nastínili dva přístupy, kterými lze modely s proměnlivými koeficienty konstruovat. Ač algoritmy pro odhad jsou podrobně popsány u všech, jejich inference je následně zkoumána pouze u dvou metod – odhadu pomocí polynomiálních splajnů (Huang a kol., 2004) a odhadu pomocí lokální regrese (Fan a Zhang, 2008). Tyto metody se od sebe liší nejenom v přístupu k odhadu křivek $\beta_j, j = 1,...,k$, ale i v použitých datech. (Huang a kol., 2004) svoji práci aplikovali na longitudinální data. Z toho důvodu tuto kapitolu rozdělíme na dvě části. V první se budeme věnovat statistické inferenci pro model se standardními daty odhadnutý pomocí lokální regrese. V druhé části se pak zaměříme pouze na model s longitudinálními daty odhadnutý pomocí polynomiálních splajnů. V obou těchto částech se podíváme na odhad vychýlení, rozptylu a kovarianční struktury, asymptotickou teorii a z nich vycházející konstrukci konfidenčních intervalů a pásem a testování hypotéz. Nakonec popíšeme i několik způsobů testování hypotéz pro zbylé metody odhadu.

3.1 Model se standardními daty

V této podkapitole se zaměříme na inferenci pro odhad pomocí lokální regrese na standardních datech s nezávislými pozorováními 2.2.

3.1.1 Vychýlení a rozptyl

V případě odhadu modelu s proměnlivými koeficienty pomocí lokální regrese je odhad vychýlení a rozptylu úzce spjat s volbou šířky pásma h, kde její optimální hodnota minimalizuje střední čtvercovou chybu MSE, pro jejíž výpočet je třeba nejprve získat právě odhady vychýlení a rozptylu.

V této podkapitole uvedeme vzorce vychýlení a rozptylu pouze pro jednokrokový odhad. Pro odhad dvoukrokový by se odvodily obdobně, avšak s použitím vyššího řádu Taylorova rozvoje funkcí $\beta_j, j = 1,...,k$.

Z rovnice (2.15) vidíme, že vychýlení je rovno

$$bias(\hat{\boldsymbol{\beta}}(u_0)|\mathcal{X},\mathcal{U}) = (\boldsymbol{I}_k, \boldsymbol{0}_k)(\boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^h \boldsymbol{\Gamma}_{u_0})^{-1} \boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^h \boldsymbol{d}, \tag{3.1}$$

kde

$$\boldsymbol{d} = (d_1,...,d_n)^T, \quad d_i = \sum_{j=1}^{k} \Big(\beta_j(U_i) - \big(a_j + b_j(U_i - u_0)\big) \Big) X_{i,j}.$$

Vzhledem k původní aproximaci proměnlivého koeficientu β_j useknutou Taylorovou řadou

$$\beta_j(U) \approx a_j + b_j(U - u_0),$$

nyní můžeme výše definovaný člen $\boldsymbol{d}$ aproximovat jako $\boldsymbol{\tau}$, což je sloupcový vektor délky n s i-tým členem definovaným jako

$$\tau_i = \sum_{j=1}^{k} \left(\frac{1}{2}\beta_j^{(2)}(u_0)(U_i - u_0)^2 + \frac{1}{3}\beta_j^{(3)}(u_0)(U_i - u_0)^3 \right) X_{i,j}.$$

Pro vychýlení tedy dostáváme výraz

$$bias(\hat{\boldsymbol{\beta}}(u_0)|\mathcal{X},\mathcal{U}) \approx (\boldsymbol{I}_k, \boldsymbol{0}_k)(\boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^h \boldsymbol{\Gamma}_{u_0})^{-1} \boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^h \boldsymbol{\tau}. \tag{3.2}$$

Odhad podmíněného vychýlení $\hat{\boldsymbol{\beta}}(u_0)$ na $(\mathcal{X},\mathcal{U})$ pak získáme dosazením $\hat{\boldsymbol{\tau}}$ za $\boldsymbol{\tau}$ v rovnici (3.2), kde v $\hat{\boldsymbol{\tau}}$ je $\beta_j^{(k)}(u_0)$ nahrazena odhadem $\hat{\beta}_j^{(k)}(u_0), k = 2{,}3$, který se dá získat lokálním kubickým vyrovnáváním s pilotní šířkou pásma h_*.

Z rovnice (2.15) zřejmě vidíme, že matice podmíněného rozptylu je rovna

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}(u_0)|\mathcal{X},\mathcal{U}) = (\boldsymbol{I}_k, \boldsymbol{0}_k)(\boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^h \boldsymbol{\Gamma}_{u_0})^{-1} (\boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^{h\ \otimes 2} \boldsymbol{\Gamma}_{u_0}) \\ \times (\boldsymbol{\Gamma}_{u_0}^T \boldsymbol{W}_{u_0}^h \boldsymbol{\Gamma}_{u_0})^{-1} (\boldsymbol{I}_k, \boldsymbol{0}_k)^T \sigma^2(u_0), \end{aligned} \tag{3.3}$$

a pro odhad tohoto rozptylu pak stačí nahradit $\hat{\sigma}^2(u_0)$ za $\sigma^2(u_0)$. Odhad $\hat{\sigma}^2(u_0)$ dostaneme jako vedlejší produkt lokálního kubického vyrovnávání $\boldsymbol{\beta}^{(k)}(u_0), k = 2{,}3$ s pilotní šířkou pásma h_* jako:

$$\hat{\sigma}^2(u_0) = \frac{\boldsymbol{Y}^T \{ \boldsymbol{W}_{u_0}^{h_*} - \boldsymbol{W}_{u_0}^{h_*} \boldsymbol{\Gamma}_{u_0}^* (\boldsymbol{\Gamma}_{u_0}^{*T} \boldsymbol{W}_{u_0}^{h_*} \boldsymbol{\Gamma}_{u_0}^*)^{-1} \boldsymbol{\Gamma}_{u_0}^{*T} \boldsymbol{W}_{u_0}^{h_*} \} \boldsymbol{Y}}{tr\{ \boldsymbol{W}_{u_0}^{h_*} - (\boldsymbol{\Gamma}_{u_0}^{*T} \boldsymbol{W}_{u_0}^{h_*} \boldsymbol{\Gamma}_{u_0}^*)^{-1} (\boldsymbol{\Gamma}_{u_0}^{*T} \boldsymbol{W}_{u_0}^{h_*\ \otimes 2} \boldsymbol{\Gamma}_{u_0}^*) \}}, \tag{3.4}$$

kde $\boldsymbol{\Gamma}_{u_0}^* = (\mathcal{X}, \boldsymbol{U}_{u_0}\mathcal{X}, \boldsymbol{U}_{u_0}^2\mathcal{X}, \boldsymbol{U}_{u_0}^3\mathcal{X})$.

3.1.2 Asymptotické vlastnosti

Konstrukce asymptotické teorie zde probíhá obdobně jako u lineární regrese. Nejprve je definováno několik technických podmínek, díky kterým jsou následně vyřčeny věty o asymptotickém chování vychýlení a rozptylu. Tyto podmínky zde nebudeme přepisovat, ale odkážeme čtenáře na (Fan a Zhang, 1999). Převážně se jedná o podmínky spojitosti a derivovatelnosti funkcí $\beta_j, j = 1,...,k$ a existence středních hodnot z regresorů $\boldsymbol{X}$. Autoři vypracovali asymptotické vlastnosti pro obě své metody odhadu – jednokrokovou i dvoukrokovou. Pro formulaci následujících vět je ale nejprve třeba definovat několik výrazů:

$$\begin{aligned} \mu_p &= \int_{\mathbb{R}} t^p K(t) dt, \\ \nu_p &= \int_{\mathbb{R}} t^p K^2(t) dt, \end{aligned}$$

dále položme $r_{s,j} = r_{s,j}(u_0) = E(\boldsymbol{X}^s \boldsymbol{X}^j \mid U = u_0)$ pro $s,j = 1,...,k$ a definujme

$$\begin{aligned} &\boldsymbol{\Psi} = diag(\sigma^2(U_1),...,\sigma^2(U_n)), \\ &\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j(u_0) = (r_{1,j}(u_0),...,r_{k-1,j}(u_0))^T, \\ &\boldsymbol{\Omega}_j = \boldsymbol{\Omega}_j(u_0) = E[(\boldsymbol{X}^1,...,\boldsymbol{X}^j)^T(\boldsymbol{X}^1,...,\boldsymbol{X}^1) \mid U = u_0]. \end{aligned}$$

V dalším textu se setkáme s několika různými šířkami pásma. Parametr h_0 značí začáteční šířku pásma, h_1 šířku pro jednokrokový odhad a h_2 šířku pásma pro dvoukrokový odhad. Pro jednokrokový odhad k zápisu koeficientu přidáme index JK (pro dvoukrokový odhad pak analogicky DK) a dostáváme následující asymptotické vlastnosti:

Věta 1. *(Vychýlení a rozptyl, jednokrokový odhad)*
Předpokládejme, že podmínky z (Fan a Zhang, 1999) jsou splněny. Pokud $h_1 \to 0$ *tak, že* $nh_1 \to \infty$*, pak asymptotické podmíněné vychýlení* $\hat{\beta}_{k,JK}(u_0)$ *je rovno*

$$bias(\hat{\beta}_{k,JK}(u_0) \mid \mathcal{X},\mathcal{U}) = -\frac{h_1^2\mu_2}{2r_{k,k}}\sum_{j=1}^{k-1} r_{k,j}\beta_j''(u_0) + o_p(h_1^2),$$

a asymptotický podmíněný rozptyl $\hat{\beta}_{k,JK}(u_0)$ *je*

$$var(\hat{\beta}_{k,JK}(u_0) \mid \mathcal{X},\mathcal{U}) = \frac{\sigma^2(u_0)(\lambda_2 r_{k,k} + \lambda_3 \boldsymbol{\alpha}_k^T \boldsymbol{\Omega}_{k-1}^{-1}\boldsymbol{\alpha}_k)}{nh_1 f(u_0)\lambda_1 r_{k,k}(r_{k,k} - \boldsymbol{\alpha}_k^T\boldsymbol{\Omega}_{k-1}^{-1}\boldsymbol{\alpha}_k)}(1 + o_p(1)),$$

kde $\lambda_1 = (\mu_4 - \mu_2^2), \lambda_2 = \nu_0\mu_4^2 - 2\nu_2\mu_2\mu_4 + \mu_2^2\nu_4$ *a* $\lambda_3 = 2\mu_2\nu_2\mu_4 - 2\nu_0\mu_2^2\mu_4 - \mu_2^2\nu_4 + \nu_0\mu_2^4$.

Všimněme si několika věcí. Při volbě šířky pásma $h_1 = O(n^{-1/5})$ dosáhne podmíněná MSE jednokrokového odhadu $\hat{\beta}_{k,JK}(u_0)$ míry $O_P(n^{-4/5})$. Vyjádření vychýlení výše jasně ukazuje, že aproximační chyby funkcí $\beta_1,...,\beta_{k-1}$ se přenesou do vychýlení β_k. Tím pádem jednokrokový odhad β_k nabývá nezanedbatelné aproximační chyby a není optimální.

Následující část se již bude týkat asymptotického chování dvoukrokového odhadu.

Věta 2. *(Vychýlení a rozptyl, dvoukrokový odhad)*
Předpokládejme, že podmínky z (Fan a Zhang, 1999) jsou splněny. Pak lze asymptotické podmíněné vychýlení $\hat{\beta}_{k,DK}(u_0)$ *vyjádřit jako*

$$bias(\hat{\beta}_{k,DK}(u_0) \mid \mathcal{X},\mathcal{U}) = \frac{1}{4!}\frac{\mu_4^2 - \mu_6\mu_2}{\mu_4 - \mu_2^2}\beta_k^{(4)}(u_0)h_2^4 - \frac{\mu_2 h_0^2}{2r_{k,k}}\sum_{j=1}^{k-1}\beta_j''(u_0)r_{k,j} + o_P(h_2^4 + h_0^2),$$

a asymptotický podmíněný rozptyl $\hat{\beta}_{p,DK}(u_0)$ *jako*

$$var(\hat{\beta}_{k,DK}(u_0) \mid \mathcal{X},\mathcal{U}) = \frac{(\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4)\sigma^2(u_0)}{nh_2 f(u_0)(\mu_4 - \mu_2^2)^2}\boldsymbol{e}_{k,k}^T\boldsymbol{\Omega}_k^{-1}\boldsymbol{e}_{k,k}(1 + o_P(1)).$$

Asymptotický rozptyl dvoukrokového odhadu je nezávislý na původní šířce pásma h_0, pokud platí, že $nh_0^\gamma \to \infty$. Tím pádem původní šířka pásma by se měla volit co nejmenší aby stále ještě splňoval tuto podmínku. Dále také lze ukázat, že pokud zvolíme optimální šířku pásma h_2 řádu $n^{-1/9}$, tak podmíněný MSE dvoukrokového odhadu dosáhne optimální míry konvergence $O_P(n^{-8/9})$.

Ve větách výše byly odvozeny asymptotické vychýlení a rozptyl $\hat{\beta}_k(u_0)$ pro jednokrokový i dvoukrokový odhad. Jelikož odhadujeme pomocí lokální regrese, tak $\hat{\beta}_k(u_0)$ je lineární odhad $\beta_k(u_0)$ a má proto asymptotické normální rozdělení se střední hodnotou rovnou $(\beta_k(u_0)|\mathcal{X},\mathcal{U})$ a rozptylem $var(\hat{\beta}_k(u_0)|\mathcal{X},\mathcal{U})$.

3.1.3 Konfidenční pásma

Ukázalo se, že konfidenční intervaly pro funkcionální koeficienty modelů s proměnlivými koeficienty nepředstavují validní přístup k danému problému. Vezměme si neznámou funkci $g(\cdot)$, jejíž $1-\alpha$ konfidenční interval $(g_1(\cdot), g_2(\cdot))$ nám zaručí pouze fakt, že $P(\hat{g}_1(u) \leq g(u) \leq \hat{g}_2(u)) = 1-\alpha$ pro všechna u. To ale neznamená, že by platilo i $P(\hat{g}_1(u) \leq g(u) \leq \hat{g}_2(u), \forall u) = 1-\alpha$.

K praktickému odhadu konfidenčních mezí se tedy používají konfidenční pásma. K jejich odhadu je třeba nejprve zjistit distribuci maximálního rozdílu mezi odhadnutou funkcí $\hat{\beta}_j$ a skutečnou funkcí β_j. Teoretická zjištění potřebná k tomuto účelu byla podrobně popsána v (Fan a Zhang, 2000). Autoři se zde bez újmy na obecnosti zaměřili na konfidenční intervaly na intervalu [0,1]. Opět si definovali několik technických podmínek, které zde nebudeme uvádět. V této práci došli k velice důležité větě:

Věta 3. *(Distribuce maximálního rozdílu mezi odhadnutým a skutečným funkcionálním koeficientem) Za platnosti podmínek uvedených v (Fan a Zhang, 2000) dostáváme*

$$P\left[(-2logh)^{1/2}\left(\sup_{U\in[0,1]} \frac{\mid \hat{\beta}_j(u_0) - \beta_j(u_0) - \widehat{bias}(\hat{\beta}_j(u_0)|\mathcal{X},\mathcal{U}) \mid}{(\widehat{var}(\hat{\beta}_j(u_0)|\mathcal{X},\mathcal{U}))^{1/2}} - d_{v,n}\right) < x\right] \xrightarrow{d} e^{-2e^{-x}},$$

pro $j = 1,...,k$ *a kde*

$$d_{v,n} = (-2logh)^{1/2} + \frac{1}{(-2logh)^{1/2}} log\left(\frac{1}{4\nu_0\pi}\int (K^{'}(t))^2 dt\right),$$
$$\nu_0 = \int K^2(t)dt.$$

Dle této věty lze konfidenční pásma pro $\beta_j(u_0), j = 1,...,k$ jednoduše zkonstruovat jako

$$\hat{\beta}_j(u_0) - \widehat{bias}(\hat{\beta}_j(u_0)|\mathcal{X},\mathcal{U}) \pm \Delta_{j,\alpha}(u_0),$$

kde

$$\Delta_{j,\alpha}(u_0) = \left(d_{v,n} + \big[log2 - log\big(-log(1-\alpha)\big)\big](-2logh)^{1/2}\right) \times \left(\widehat{var}(\hat{\beta}_j(u_0)|\mathcal{X},\mathcal{U})\right)^{1/2}.$$

Odhady $\widehat{bias}(\hat{\beta}_j(u_0)|\mathcal{X},\mathcal{U})$ a $\widehat{var}(\hat{\beta}_j(u_0)|\mathcal{X},\mathcal{U})$ byly popsány v sekci (3.1.1). (Fan a Zhang, 2000) taktéž na simulacích ukázali, že takto konstruovaná konfidenční pásma fungují velice dobře.

3.1.4 Testování hypotéz

U odhadnutých funkcionálních koeficientů nás bude zajímat především informace o tom, zda-li jsou skutečně proměnlivé dle svého modifikátoru vlivu nebo zda-li nabývají konstantní hodnoty a odhadujeme je proto zbytečné odhadovat

je neparametricky. Právě k tomuto účelu použijeme principy testování hypotéz a test s hypotézami

$$\begin{aligned} H_0 &: \beta_j(U) = c_j \\ H_1 &: \beta_j(U) \neq c_j, \end{aligned} \tag{3.5}$$

kde c_j značí jakousi konstantu. Pro odhady $\beta_j, j = 1,...,k$ kalkulované pomocí lokální regrese byla ukázána jejich asymptotická normalita. Při dostatečném počtu pozorování je proto možné pro výše uvedený test konstruovat klasickou t-testovou statistiku z lineární regrese založenou právě na znalosti normálního rozdělení odhadnutého koeficientu. Takovýto postup je ale aplikován pouze na jediný bod výsledné křivky. Postupuje se tedy tak, že pro každý v datech pozorovaný modifikátor vlivu zkonstruujeme a vyhodnotíme tento test. Pokud pro nějaký z těchto bodů nebo nějakou jejich část zamítáme nulovou hypotézu, že funkce odhadnutého koeficientu v daném bodě je rovna zvolené konstantě, pak lze tvrdit, že odhadnutá křivka není konstantní. Samozřejmě je třeba vhodně zvolit testovanou konstantu, kupříkladu jako průměr všech odhadnutých hodnot koeficientu β_j, tj. odhady pro pozorovaný modifikátor vlivu $U_i, i = 1,...,n$. Taktéž je třeba vhodným způsobem upravit výsledné p-hodnoty, jelikož se jedná o problém vícenásobného testování.

Pokud bychom však raději přesnější výsledky nebo disponujeme příliš málo pozorováními pro splnění asymptotické normality, pak lze použít i jiné složitější metody. (Fan a Zhang, 2000) navrhli vlastní verzi testování hypotéz pro jimi zkoumané modely s proměnlivými koeficienty odhadnuté pomocí lokální regrese. Jejich metoda je založena na distribuci maximálního rozdílu mezi odhadnutou funkcí a skutečnou funkcí, jež byla formulována ve větě 3 pro konfidenční pásma. Tato metoda umožňuje dvě různá použití. Za prvé je možné testovat hypotézu

$$\begin{aligned} H_0 &: \beta_j(U) = \beta_0(U) \\ H_1 &: \beta_j(U) \neq \beta_0(U), \end{aligned}$$

pro nějakou pevně stanovenou funkci $\beta_0(\cdot)$. Přirozeně se naskýtá otázka, zda-li nějaká takováto funkce spadá do konfidenčního pásma, což je ekvivalentní s použitím testové statistiky

$$T = (-2logh)^{1/2}\left(\left\| \left(\widehat{var}(\hat{\beta}_j(U)|\mathcal{X},\mathcal{U})\right)^{-1/2}\left(\hat{\beta}_j(U) - \beta_0(U) - \widehat{bias}(\hat{\beta}_j(U)|\mathcal{X},\mathcal{U})\right)\right\| - d_{v,n}\right),$$

kde zamítáme nulovou hypotézu, pokud tato testová statistika překročí hodnotu $c_\alpha = -log(-0.5log\alpha)$.

Nejdůležitější metoda navržená v (Fan a Zhang, 2000) se zabývá otázkou, zda-li odhadnutý funkcionální koeficient není konstantní. Test je formulován jako

$$\begin{aligned} H_0 &: \beta_k(U) = c \\ H_1 &: \beta_k(U) \neq c. \end{aligned}$$

Hlavním problémem je vhodná volba hodnoty c. Nejprve je třeba odhadnout, jaké velikosti by testovaná konstanta měla za platnosti nulové hypotézy nabývat. Za platnosti nulové hypotézy je model definován jako

$$\boldsymbol{Y} = \sum_{j=1}^{k-1} \beta_j(\boldsymbol{U})\boldsymbol{X}^j + c\boldsymbol{X}^k + \varepsilon.$$

Odhad $\hat{c}$ vypočteme dvoukrokově. Nejprve zanedbáme fakt, že by c měla být konstanta a namísto toho k ní přistupujeme jako k nějaké funkci $\beta_k(U)$. Lokální regresí dostaneme odhady $\hat{\beta}_k(U_i), i = 1,...,n$. Každý z těchto odhadů je taktéž i odhadem neznámé konstanty c za platnosti nulové hypotézy. Ve druhém kroku pak už jen tyto odhady zprůměrujeme a dostaneme

$$\hat{c} = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_k(U_i) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{e}_{l,4}^T(\boldsymbol{G}_{U_i}^T\boldsymbol{W}_{U_i}^h\boldsymbol{G}_{U_i})^{-1}\boldsymbol{G}_{U_i}^T\boldsymbol{W}_{U_i}^h\boldsymbol{Y},$$

kde $l = 4(k-1)+1$. Pro zjednodušení se zde předpokládá, že modifikátor vlivu U se nachází na intervalu [0,1], kde je spojitý nezáporný. Autoři formulovali větu o asymptotickém chování vychýlení a rozptylu odhadnuté konstanty $\hat{c}$, díky níž a díky větě o distribuci maximálního rozdílu mezi odhadnutou funkcí a skutečnou funkcí pak lze pro tento test formulovat testovou statistiku ve tvaru

$$T = (-2logh)^{1/2}\left(\left\| \left(\widehat{var}(\hat{\beta}_k(U)|\mathcal{X},\mathcal{U})\right)^{-1/2}\left(\hat{\beta}_k(U) - \hat{c} - \widehat{bias}(\hat{\beta}_k(U)|\mathcal{X},\mathcal{U})\right)\right\| - d_{v,n}\right),$$

a zamítat nulovou hypotézu pro velké hodnoty této statistiky, tedy při překročení kritické hodnoty $c_\alpha = -log(-0.5log\alpha)$ na hladině α.

3.2 Model s longitudinálními daty

Další podkapitola se věnuje inferenci pro odhad pomocí polynomiálních splajnů na longitudinálních datech.

3.2.1 Rozptyl a kovarianční struktura

Oproti odhadu rozptylu a kovarianční struktury u modelu se standardními daty je tato operace u longitudinálních dat poněkud komplikovanější. Důvodem je závislost mezi různými pozorováními pro jednotlivé subjekty. Matici rozptylu odhadu $\hat{\boldsymbol{\gamma}}$ podmíněnou na množině $\mathcal{D}$ lze zapsat jako

$$var(\hat{\boldsymbol{\gamma}}|\mathcal{D}) = \left(\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{W}_i\boldsymbol{U}_i\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{W}_i\boldsymbol{V}_i\boldsymbol{W}_i\boldsymbol{U}_i\right)\left(\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{W}_i\boldsymbol{U}_i\right)^{-1}, \quad (3.6)$$

kde $\boldsymbol{V}_i = var(\boldsymbol{Y}_i) = C_\varepsilon(t_{i,j}, t_{i,j'})$ a $C_\varepsilon(t,s)$ je kovariance $\varepsilon(t)$. Pro finální odhady parametrů $\hat{\boldsymbol{\beta}}$ je třeba odhady koeficientů $\hat{\boldsymbol{\gamma}}$ aplikovat na použitou B-splajnovou bázi a matice rozptylu $\hat{\boldsymbol{\beta}}(t)$ podmíněný množinou $\mathcal{D}$ má podobu

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}(t)|\mathcal{D}) = \boldsymbol{B}(t)&\left(\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{W}_i\boldsymbol{U}_i\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{W}_i\boldsymbol{V}_i\boldsymbol{W}_i\boldsymbol{U}_i\right) \\ &\times\left(\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{W}_i\boldsymbol{U}_i\right)^{-1}\boldsymbol{B}^T(t). \end{aligned} \quad (3.7)$$

Pokud bychom chtěli získat rozptyl pouze pro jednotlivý parametr $\hat{\beta}_l(t)$, tak stačí definovat si $\boldsymbol{e}_{j+1}$ jako $(k+1)$-dimenzionální nulový vektor s $(j+1)$-ním členem rovným jedné. Pak podmíněný rozptyl $\hat{\beta}_l(t)$ na $\mathcal{D}$ je

$$var(\hat{\beta}_j(t)|\mathcal{D}) = \boldsymbol{e}_{j+1}^T var(\hat{\boldsymbol{\beta}}(t)|\mathcal{D})\boldsymbol{e}_{j+1}, \qquad j = 1,...,k. \tag{3.8}$$

Všimněme si, že ze všech použitých proměnných neznáme podobu matice $\boldsymbol{V}_i = C_\varepsilon(t_{i,l}, t_{i,l'})$. K té je zapotřebí určit kovarianční strukturu procesu chyb $\varepsilon(t)$. To je ale kvůli longitudinální podobě vstupních dat podstatně složitější než při použití standardních dat s nezávislými pozorováními. Na druhou stranu to ale podstatně rozšiřuje možnosti použití modelu s proměnlivými koeficienty s odhadem pomocí polynomiálních splajnů, jelikož v poslední době se longitudinální data začínají využívat mnohem častěji a to nejen v lékařských pozorováních.

(Huang a kol., 2004) navrhli metodu založenou na odhadu kovarianční funkce pomocí splajnů, kde funkci $C_\varepsilon(t,s)$ aproximovali jako tensorový produkt splajnů na $[\xi_0, \xi_{M+1}] \times [\xi_0, \xi_{M+1}]$

$$C_\varepsilon(t,s) \approx \sum_a \sum_b u_{a,b} B_{a,d}^\xi(t) B_{b,d}^\xi(s), \qquad t,s \in [\xi_0, \xi_{M+1}], t \neq s, \tag{3.9}$$

kde $B_{a,d}^\xi(t)$ je a -tá B-splajnová báze stupně d na množině uzlů $\xi_0,...,\xi_{M+1}$. Zřejmě platí, že $E(\varepsilon(t_{i,l})\varepsilon(t_{i,l'}) = C_\varepsilon(t_{i,l},t_{i,l'})$ pro $l \neq l'$ a $C_\varepsilon(t,s) = C_\varepsilon(s,t)$. Při pozorovaných $\{\varepsilon_i(t_{i,l}), i = 1,...,n, l = 1,...,n_i\}$ lze $C_\varepsilon(s,t), s \neq t$ odhadnout přes minimalizaci výše uvedených parametrů $\{u_{a,b} : u_{a,b} = u_{b,a}\}$ pomocí výrazu

$$\min_{u_{a,b}\in\mathbb{R}:u_{a,b}=u_{b,a}} \sum_{i=1}^{n} \sum_{l,l'=1,l<l'}^{n_i} \left(\varepsilon_i(t_{i,l})\varepsilon_i(t_{i,l'}) - \sum_a \sum_b u_{a,b} B_{a,d}^\xi(t_{i,l}) B_{b,d}^\xi(t_{i,l'}) \right)^2. \tag{3.10}$$

Skutečná $\varepsilon_i(t_{i,l})$ však nejsou pozorována a proto je třeba místo nich dosazovat jejich odhady – rezidua $\hat{\varepsilon}_i(t_{i,l}) = Y_{i,l} - \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}(t_{i,l})$. Výsledný odhad kovarianční funkce pro $t \neq s$ pak snadno dostaneme dosazením odhadnutých parametrů $\hat{u}_{k,l}$ do vzorce (3.9). Odhad rozptylu $\sigma_\varepsilon^2(t) = C_\varepsilon(t,t)$ dostaneme obdobně aproximací $C_\varepsilon(t,t) \approx \sum_a v_a B_{a,d}^\xi(t)$, kde odhad $\hat{v}_k$ opět získáme minimalizací

$$\min_{v_a\in\mathbb{R}} \sum_{i=1}^{n} \sum_{l=1}^{n_i} \left(\hat{\varepsilon}_i^2(t_{i,j}) - \sum_a v_a B_{a,d}^\xi(t_{i,l}) \right)^2. \tag{3.11}$$

Tímto způsobem je odhadnuta celá rozptylová matice. Autoři se dále zamýšlejí i nad kvalitou tohoto odhadu, který zřejmě závisí na volbě splajnového prostoru, tedy volbě umístění uzlů. Opět zde narážíme na problém výpočetní složitosti, kdy volba počtu stejně vzdálených uzlů pomocí cross-validace vyžaduje velké množství výpočetního výkony. Autoři podle svých zkušeností proto doporučují vybrat počet uzlů ručně v rozmezí 5 až 10.

3.2.2 Asymptotické vlastnosti

V této části se blíže podíváme na asymptotické vlastnosti odhadnutých funkcí $\hat{\boldsymbol{\beta}}$. Asymptotiku pro longitudinální data zde zkoumáme pro množinu subjektů

$i = 1,...,n$, kde jejich počet n jde k nekonečnu. Počet pozorování u i-tého subjektu označený jako n_i může nebo nemusí jít k nekonečnu. Pro každý subjekt je rovněž použita váha $w_i = 1/n_i$.

(Huang a kol., 2004) si nejprve definovali několik technických podmínek. Opět je zde nebudeme uvádět a čtenáře odkážeme na citovaný článek. Abychom zde pouze nepřepisovali tento článek, tak v plném znění uvedeme pouze několik hlavních vět a zbylé poznatky jen popíšeme.

Autoři ukázali, že výsledné odhady $\hat{\beta}_j, j = 1,...,k$ jsou definovány jednoznačně a s pravděpodobností jdoucí k jedné. Tyto odhady jsou taktéž konzistentní. Zavádí se zde nová proměnná $\tilde{\beta}_j(t) = E[\hat{\beta}_j(t)|\mathcal{D}]$ jako střední hodnota $\hat{\beta}_j(t)$ podmíněná na $\mathcal{D}$

V další větě je vyčíslena míra konvergence $\| \hat{\beta}_j - \beta_j \|^2$, $\| \tilde{\beta}_j - \beta_j \|$ a $\| \hat{\beta}_j - \tilde{\beta}_j \|^2$. Tato věta implikuje, že rozsah vychýlení je omezen v pravděpodobnosti nejlepší aproximační mírou dosažitelnou v prostorech splajnových funkcí. Když je počet pozorování pro každý subjekt omezen, tedy $n_i \leq C, 1 \leq i \leq n$ pro nějakou konstantu C, pak míra konvergence $\| \tilde{\beta}_j - \beta_j \|^2$ je omezena na $O_P(K_n/n + \rho_n^2)$, což je stejná míra jako u standardních dat s nezávislými pozorováními.

Pokud zvolíme $K_n \sim (n^{-2}) \sum_i n_i^{-1})^{-1/5}$ a počet pozorování pro každý subjekt je omezen, pak lze dosáhnout $\| \hat{\beta}_j - \beta_j \|^2 = O_P(n^{-4/5})$, což je taktéž stejná optimální míra jako u dat s nezávislými pozorováními. Když dále omezíme počet pozorování pro každý subjekt n_i fixní konstantou, pak nám pro dosažení $n^{-4/5}$ míry stačí $K_n \sim n^{1/5}$.

Věta 4. *(Vychýlení)*
Předpokládejme, že podmínky z (Huang a kol., 2004) jsou splněny a $\frac{K_n log K_n}{n} \to 0$, *pro* $n \to \infty$. *Pak* $\sup_{t \in [\xi_0, \xi_{M+1}]} \left| \tilde{\beta}_j(t) - \beta_j(t) \right| = O_P(\rho_n), j = 1,...,k.$

Tato věta nám dává postačující podmínku k tomu, aby vychýlení bylo zanedbatelné vzhledem k rozptylu.

Věta 5. *(Asymptotická normalita)*
Předpokládejme, že podmínky z (Huang a kol., 2004) jsou splněny.
Pokud $\frac{K_n log K_n}{n} \to 0$, *pro* $n \to \infty$ *a* $\frac{K_n \max_i n_i}{n} \to 0$, *pro* $n \to \infty$, *pak*

$$\frac{\hat{\boldsymbol{\beta}}(t) - \tilde{\boldsymbol{\beta}}(t)}{\sqrt{var(\hat{\boldsymbol{\beta}}(t))}} \xrightarrow{d} N(0,I),$$

v distribuci, kde $\tilde{\boldsymbol{\beta}}(t) = (\tilde{\beta}_1(t),...,\tilde{\beta}_k(t))^T$. *Speciálně pak pro* $j = 1,...,k$ *platí, že*

$$\frac{\hat{\beta}_j(t) - \tilde{\beta}_j(t)}{\sqrt{var(\hat{\beta}_j(t))}} \xrightarrow{d} N(0,1).$$

O $\tilde{\beta}_j(t)$ lze smýšlet jako o odhadnutelné části $\beta_j(t)$ a větu o asymptotické normalitě v další části použijeme ke konstrukci konfidenčních intervalů pro $\tilde{\beta}_j(t)$ a potažmo i $\beta_j(t)$.

Výše uvedené věty lze aplikovat i na data s deterministicky stanovenými časy $t_{i,l}$. V tom případě je ale třeba vhodným způsobem upravit jednu z technických podmínek.

3.2.3 Konfidenční intervaly a pásma

V této části využijeme asymptotické vlastnosti modelů s proměnlivými koeficienty, odhadnutými pomocí polynomiálních splajnů, ke konstrukci bodových konfidenčních intervalů pro jednotlivé proměnlivé koeficienty $\hat{\beta}_j, j = 1,...,k$. Vzhledem ke křivkovému charakteru odhadnutých parametrů ale takováto metoda není zcela validní a proto je užitečnější konstruovat spíše simultánní konfidenční pásma, která zde také popíšeme.

Za platnosti technických podmínek a z nich vyvozených asymptotických vět víme, že pro $j = 1,...,k$ a $t \in [\xi_0, \xi_{M+1}]$ platí následující konvergence

$$\big(var(\hat{\beta}_j(t))\big)^{-1/2}\big(\hat{\beta}_j(t) - E[\hat{\beta}_j(t)]\big) \xrightarrow{d} N(0,1), \text{ pro } n \to \infty, \tag{3.12}$$

kde rozptyl a střední hodnota $\hat{\beta}_j(t)$ je opět podmíněna na $\mathcal{D}$. Pokud nalezneme odhad $\widehat{var}(\hat{\beta}_j(t))$ takový, že $\widehat{var}(\hat{\beta}_j(t))/var(\hat{\beta}_j(t)) \xrightarrow{p} 1$ pro $n \to \infty$, pak (3.12) platí i pro tento odhad a $(1-\alpha)$ asymptotický interval $E[\hat{\beta}_j(t)]$ je stanoven jako rozmezí

$$\hat{\beta}_j(t) \pm z_{\alpha/2}\big(\widehat{var}(\hat{\beta}_j(t))\big)^{1/2}, \tag{3.13}$$

kde $z_{\alpha/2}$ je $(1-\alpha/2)$-tý kvantil normálního rozdělení. V předchozí části si vzpomeňme na diskuzi o asymptotické zanedbatelnosti vychýlení $E[\hat{\beta}_j(t)] - \beta_j(t)$. Pokud tomu tak je, pak vzorec v (3.13) definuje $(1-\alpha)$ asymptotický konfidenční interval i pro $\beta_j(t)$.

Konstrukce simultánních konfidenčních pásem pro $E[\hat{\beta}_j(t)]$ a $\beta_j(t)$ na nějakém intervalu $[a,b] \in [\xi_0, \xi_{M+1}]$ rozšiřuje výše popsané postupy pro konstrukci konfidenčních intervalů. Interval $[a,b]$ rozdělíme na $M+1$ ekvidistantních úseků pomocí uzlů $a = \xi_0 < ... < \xi_{M+1} = b$ a pro každý z nich vypočteme meze $(1-\alpha)$ simultánních konfidenčních intervalů $(l_{j,\alpha}(\xi_r), u_{j,\alpha}(\xi_r))$ pro $E[\hat{\beta}_j(\xi)_r]$ takových, že $lim_{n\to\infty} P[l_{j,\alpha}(\xi_r)) \leq E[\hat{\beta}_j(\xi)_r] \leq u_{j,\alpha}(\xi_r), r = 1,..., M+1] \geq 1-\alpha$. Jednoduchý přístup založený na Bonferroniho postupu navrhuje zvolit $(l_{j,\alpha}(\xi_r), u_{j,\alpha}(\xi_r))$ jako

$$\hat{\beta}_j(\xi_r) \pm z_{\alpha/(2(M+1))}\big(\widehat{var}(\hat{\beta}_j(\xi_r))\big)^{1/2}. \tag{3.14}$$

Střední hodnotu $E[\hat{\beta}_j(t)]$ budeme konstruovat jako lineární interpolaci $E^{(I)}[\hat{\beta}_j(t)]$ mezi $E[\hat{\beta}_j(\xi_r)]$ a $E[\hat{\beta}_j(\xi_{r+1})]$ pro $\xi_r \leq t \leq \xi_{r+1}$ dle vzorce

$$E^{(I)}[\hat{\beta}_j(t)] = M\left(\frac{\xi_{r+1} - t}{b-a}\right)E[\hat{\beta}_j(\xi_r)] + M\left(\frac{t-\xi_r}{b-a}\right)E[\hat{\beta}_j(\xi_{r+1})].$$

Stejným způsobem se pak lineárně interpolují i krajní meze $(1-\alpha)$ konfidenčního intervalu pro $E^{(I)}[\hat{\beta}_j(t)]$. Pro konstrukci konfidenčních pásem předpokládáme splnění jedné z podmínek

$$sup_{t\in[a,b]}\big|\big(E[\hat{\beta}_j(t)]\big)'\big| \leq c_1 \text{ pro známou konstantu } c_1 > 0, \tag{3.15}$$

$$sup_{t\in[a,b]}\big|\big(E[\hat{\beta}_j(t)]\big)''\big| \leq c_2 \text{ pro známou konstantu } c_2 > 0. \tag{3.16}$$

Kalkulací pomocí Taylorova rozvoje lze dojít k následujícím výsledkům

$$|E[\hat{\beta}_j(t)] - E^{(I)}[\hat{\beta}_j(t)]| = \begin{cases} 2c_1 M(\frac{(\xi_{r+1}-t)(t-\xi_r)}{b-a}) & \text{za platnosti (3.15),} \\ \frac{1}{2}c_2 M(\xi_{r+1}-t)(t-\xi_r) & \text{za platnosti (3.16).} \end{cases}$$

Pomocí těchto členů pak upravíme interpolované meze $(1-\alpha)$ konfidenčních intervalů a získáme tak odhadnutá $(1-\alpha)$ konfidenční pásma pro $E[\hat{\beta}_j(t)]$ za platnosti (3.15) nebo (3.16) jako

$$\left(l_{j,\alpha}^{(I)}(t) - 2c_1 M\left(\frac{(\xi_{r+1}-t)(t-\xi_r)}{b-a}\right), u_{j,\alpha}^{(I)}(t) + 2c_1 M\left(\frac{(\xi_{r+1}-t)(t-\xi_r)}{b-a}\right)\right), \tag{3.17}$$

$$\left(l_{j,\alpha}^{(I)}(t) - \frac{1}{2}c_2 M(\xi_{r+1}-t)(t-\xi_r), u_{j,\alpha}^{(I)}(t) + \frac{1}{2}c_2 M(\xi_{r+1}-t)(t-\xi_r)\right). \tag{3.18}$$

Pokud je vychýlení $E[\hat{\beta}_j(t)] - \beta_j(t)$ asymptoticky zanedbatelné a jedna z podmínek (3.15), (3.16) je splněna, pak výsledky výše tvoří odpovídající asymptotické konfidenční pásmo i pro $\beta_j(t), j = 1,...,k$.

3.2.4 Testování hypotéz

Stejně jako u testování hypotéz pro model se standardními daty pomocí lokální regrese nás bude zajímat, jestli je odhadnutý koeficient proměnlivý nebo konstantní. Takový test můžeme formulovat jako

$$\begin{aligned} H_0 &: \beta_j(t) = c_j \\ H_1 &: \beta_j(t) \neq c_j, \end{aligned} \tag{3.19}$$

kde c_j značí jakousi konstantu. Pro odhady $\beta_j, j = 1,...,k$ kalkulované pomocí polynomiálních splajnů byla taktéž ukázána jejich asymptotická normalita a lze proto pro tento test konstruovat t-testovou statistiku obdobně jako v lineární regresi. Samotné vyhodnocení testu pak probíhá stejně jako u testování hypotéz pro lokální regresi v předchozí podkapitole, tj. testujeme vždy pro jednotlivý bod a tuto operaci provedeme pro křivku ve všech pozorovaných bodech časového modifikátoru vlivu t. Jedná se opět o problém vícenásobného testování a to je třeba vhodným způsobem ošetřit. Nejvhodnější volba porovnávané konstanty je opět průměr křivky.

(Huang a kol., 2002) taktéž vyvinuli vlastní test pro ověření proměnlivosti koeficientů založený na tzv. bootstrappingu, kdy není potřeba apriorní znalost rozdělení odhadnutých koeficientů. Jejich test se namísto na jednotlivé funkce $\beta_j, j = 1,...,k$ zaměřuje na celý model. Testuje se hypotéza

$$\begin{aligned} &H_0 : \beta_j(t) = c_j, \qquad j = 2,...,k, t \in [\xi_0, \xi_{M+1}] \\ &H_1 : \text{nějaký z funkcionálních koeficientů je časově proměnlivý,} \end{aligned}$$

kde $c_j, j = 2,...,k$ jsou nějaké konstanty. Regresor X_1 budeme uvažovat jako intercept. Za platnosti nulové hypotézy se tedy pouze tento intercept s časem

mění, zatímco ostatní funkcionální koeficienty jsou konstantní. Vážený reziduální součet čtverců za platnosti nulové hypotézy pak má tvar

$$RSS_0 = \sum_{i=1}^{n}\sum_{l=1}^{n_i} w_i\left(Y_{i,l} - \sum_{s=1}^{S_1} X_{i,l,1}B^1_{s,d}(t_{i,l})\hat{\gamma}_{s,1} - \sum_{j=2}^{k} X_{i,l,j}\hat{c}_j\right)^2,$$

kde odhady $\hat{c}_j$ a $\hat{\gamma}_{s,1}$ minimalizují daný vážený reziduální součet čtverců. Za platnosti obecné alternativy, že všechny funkce koeficientů se mohou s časem lišit má takovýto vážený součet čtverců reziduí tvar

$$RSS_1 = \sum_{i=1}^{n}\sum_{l=1}^{n_i} w_i\left(Y_{i,l} - \sum_{j=1}^{k}\sum_{s=1}^{S_j} X_{i,l,j}B^j_{s,d}(t_{i,l})\hat{\gamma}_{s,j}\right)^2.$$

Dále si definujme testovou statistiku $T = (RSS_0 - RSS_1)/RSS_1$. Autoři formulovali větu o této testové statistice a její kritické hodnotě, která ukazuje, že testovou statistiku budeme zamítat při překročení odpovídající kritické hodnoty. K tomu je ale třeba použít bootstrap algoritmus s převzorkováním pozorování, abychom zjistili distribuci statistiky T za platnosti nulové hypotézy. Jako $\hat{\varepsilon}_{i,l}$ budeme používat odhady reziduí z modelu za platnosti alternativní hypotézy, tedy z klasického modelu se všemi proměnlivými koeficienty. Dále si definujme množinu $\{Y^p_{i,l} = i = 1,...,n, l = 1,...,n_i\}$ jako

$$Y^p_{i,l} = \sum_{s=1}^{S_1} X_{i,l,1}B^1_{s,d}(t_{i,l})\hat{\gamma}_{s,1} + \sum_{j=2}^{k}\sum_{s=1}^{S_j} X_{i,l,j}\hat{c}_j + \hat{\varepsilon}_{i,l},$$

značící jakousi pseudo-odezvu za platnosti nulové hypotézy. Následující algoritmus zjistí rozdělení testové statistiky a hledané p-hodnoty

(1. krok) Nahradíme n vzorků pozorováními z množiny $\{(Y^p_{i,l}, \boldsymbol{X}_i(t_{i,l}),t_{i,l}), i = 1,...,n, l = 1,...,n_i\}$ a získáme vzorek pro bootstrap $\{(Y^{p*}_{i,l},\boldsymbol{X}^*_i(t^*_{i,l}),t^*_{i,l}), i = 1,...,n, l = 1,...,n^*_i\}$

(2. krok) Předchozí krok zopakujeme B-krát

(3. krok) Z každého vzorku vypočteme testovou statistiku T^*. Z B nezávislých vzorků pak odvodíme empirické rozdělení T^*.

(4. krok) Nulovou hypotézu H_0 zamítáme na hladině α, pokud je hodnota testové statistiky T větší nebo rovna $(100(1-\alpha))$-tému percentilu empirického rozdělení T^*. P-hodnota testu je rovna empirické pravděpodobnosti že $\{T^* \geq T\}$.

Tuto proceduru lze upravit, když bychom kupříkladu chtěli testovat, zda-li nějaká podmnožina koeficientů není konstantní. Procedura byla taktéž popsána speciálně pro model s longitudinálními daty, ale lze ji v příslušně přeformulovaném tvaru aplikovat i na jiné metody odhadů jako odhad pomocí vyhlazovacího splajnu.

3.3 Další metody

U odhadu pomocí vyhlazovacích a penalizovaných splajnů bohužel nebyla důkladněji vypracována jejich inference a neznáme proto vychýlení či rozptyl jejich odhadnutých koeficientů, nevíme zda-li u nich za určitých podmínek platí asymptotická normalita a neumíme pro ně zkonstruovat konfidenční pásma. Avšak

existují pro ně jisté alternativní testy, díky kterým můžeme určit, zda-li je daný koeficient proměnlivý nebo konstantní. Formulujme si takovýto test jako

$$\begin{aligned} H_0 &: \beta_j(U_j) = c_j \\ H_1 &: \beta_j(U_j) \neq c_j, \end{aligned} \tag{3.20}$$

kde c_j značí jakousi konstantu. (Hastie a Tibshirani, 1993) navrhli metodu založenou na přibližných stupních volnosti. Tato metoda je použitelná pro vyhlazené členy, které monotónně závisí na nějakém vyhlazovacím parametru, což platí jak pro vyhlazování spajny, tak pro odhad lokální regresí přes jádrové funkce. Vezměme si jednoduchý hladký odhad $\hat{\boldsymbol{Y}} = \boldsymbol{S}\boldsymbol{Y}$. Autoři ukázali, že přibližné stupně volnosti pro takovýto odhad jsou $\nu = tr(2\boldsymbol{S} - \boldsymbol{S}\boldsymbol{S}^T)$ a distribuce následné testové statistiky přibližně $F_{\nu,n-\nu}$. Odhad pomocí vyhlazovacích splajnů navržený taktéž v (Hastie a Tibshirani, 1993) má vyhlazovací člen $\boldsymbol{S}$, který tvoří odhad $\hat{\beta}_j(U)X_j$, podobu $\boldsymbol{D}_j\mathcal{B}_j(\mathcal{B}_j^T\boldsymbol{D}_j^2\mathcal{B}_j + \lambda\boldsymbol{\Omega}_j)^{-1}\mathcal{B}_j\boldsymbol{D}_j$. Vidíme, že $\nu = tr(2\boldsymbol{S}' - \boldsymbol{S}'\boldsymbol{S}'^T)$ kde $\boldsymbol{S}'$ je vážený vyhlazovací člen kubického splajnu použitý v algoritmu odhadu modelů s proměnlivými koeficienty pomocí vyhlazovacího splajnu. Jako velice užitečný se ukázal fakt, že stupně volnosti členu $\hat{\beta}_j(U)X_j$ jsou stejné jako stupně volnosti odpovídajícího vyhlazovacího členu. Stačí nám tedy vypočítat ν. Ukazuje se ale, že na rozdíl od $tr(\boldsymbol{S}')$ je $tr(\boldsymbol{S}'\boldsymbol{S}'^T)$ těžko spočitatelná. Autoři proto vyvinuli její aproximaci

$$tr(2\boldsymbol{S}' - \boldsymbol{S}'\boldsymbol{S}'^T) \approx 1.25tr(\boldsymbol{S}) - 0.5.$$

Pomocí této aproximaci pak lze spočítat potřebné přibližné stupně volnosti a k nim odpovídající F rozdělení, s jehož znalostí pak lze testovat požadovanou hypotézu.

V celé práci jsme se zaměřili pouze na standardní Gaussovský tvar modelů s proměnlivými koeficienty. Pro zobecněný tvar byl ale taktéž vyvinut test, který by zde stálo za to zmínit. (Fan a kol., 2001) vyvinuli zobecněný test poměru maximální věrohodnosti a (Cai a kol., 2000) ho aplikovali na testování hypotéz v zobecněných modelech s proměnlivými koeficienty. Jedná se o test

$$\begin{aligned} H_0 &: \beta_j(U) = c_j, \qquad j = 1,...,k \\ H_1 &: \text{nějaký z funkcionálních koeficientů je časově proměnlivý,} \end{aligned}$$

a testovou statistikou ve tvaru

$$T = 2(l(H_1) - l(H_0)),$$

kde $l(H_0)$ a $l(H_1)$ jsou logaritmické věrohodnostní funkce zkonstruované za platnosti nulové a alternativní hypotézy. Takovýto test se dá relativně snadno implementovat a je dostatečně silný. Testová statistika má za platnosti nulové hypotézy asymptotické normální rozdělení a nezávisí na hodnotách $c_j, j = 1,...,k$. Tento jev je označován jako Wilksův fenomén a je podrobněji popsán v (Fan a kol., 2001). Při menším počtu pozorování je autory doporučeno raději použít bootstrap metodu za platnosti nulové hypotézy pro zjištění kritické hodnoty testové statistiky.

4. Výběr proměnných

V předchozích kapitolách jsme popsali jak modely s proměnlivými koeficienty vypadají, jakým způsobem je odhadnout a pro některé metody pak i jak se tyto odhady chovají z hlediska statistické inference. Popsali jsme procedury, jimiž lze otestovat, zda-li jsou všechny, nebo pouze některé, koeficienty konstantní či nikoliv. Dalo by se říci, že některé metody odhadů modelů s proměnlivými koeficienty jsou v současné době více než dostatečně vypracované pro praktické použití, až na absenci automatických procedur pro výběr regresorů. V následující kapitole dokonce zmíníme i několik příkladů, kdy autoři citovaných článků otestovali svoje poznatky na praktických aplikacích. V úvodní kapitole jsme rozebírali výhody a nevýhody modelů s proměnlivými koeficienty v porovnání s klasickou lineární regresí a konstatovali jsme naše očekávání, že modely s proměnlivými koeficienty nahradí lineární regresi na špici pomyslného žebříčku nejpoužívanějších metod matematického modelování. Avšak i přes všechna tato pozitiva se v současné době s praktickým použitím modelů s proměnlivými koeficienty setkáme v podstatě pouze v akademické sféře. Je otázkou, proč je takovýto revoluční nástroj stále přehlížen ve vysoce kompetitivním tržním prostředí, kde každý neustále hledá nějakou výhodu oproti svým oponentům.

Dle našeho názoru je tomu tak ze tří hlavních důvodů. Za prvé celá teorie modelů s proměnlivými koeficienty je stále dosti neucelená. Někteří autoři pracují se standardními daty, jiní s longitudinálními. Někteří popisují odhad pomocí vyhlazovacích splajnů, jiní zase pomocí lokální regrese. Když si čtenář prostuduje několik základních článků o modelech s proměnlivými koeficienty, tak bude nejspíše zmaten tím, že každý z těchto článků pojednává o něčem jiném. Právě toto větvení se do různých směrů často vyvolává zmatky a evokuje neucelenost celého tématu modelů s proměnlivými koeficienty. A tato neucelenost má nejspíše za následek fakt, že téma modelů s proměnlivými koeficienty není na většině vysokých škol vyučováno. Ve výsledku pak o těchto modelech mají přehled pouze akademici specializující se na statistiku a odborná veřejnost, která se s tímto tématem někde setkala. Prvotním problémem tedy je informovanost o modelech s proměnlivými koeficienty. Jako dobrý začátek se jeví jakýmsi způsobem spojit všechny, či alespoň ty nejvíce rozpracované popsané směry do unifikovaného celku. Netroufáme si tvrdit, že naše práce tohoto cíle dosáhla, avšak snažili jsme vzít si základy metod odhadů a statistickou inferenci ze všech různých směrů a zcelit je do přehledného celku.

Druhým důvodem je výpočetní složitost. V praxi je třeba modelovat nad daty o desetitisících pozorování a stovkách až tisících proměnných. V takovýchto případech již je třeba využívat výkonu serverů a i tak je výpočetní čas stále příliš dlouhý. Nabízejí se dva způsoby, jak překonat tuto překážku. Buď můžeme počkat několik let, až nám rozvoj informačních technologií dovolí počítat i takto složité problémy v krátkých časech. Nebo je třeba zapracovat na metodách odhadů. Výzkumníci by v tomto případě museli přijít s novými přístupy, algoritmy či heuristikami jak výpočet urychlit. Kterou z těchto cest se svět bude ubírat nám ukáže až čas.

Poslední důvod a pojednání této kapitoly je výběr proměnných. Při praktickém využití je třeba pracovat se stovkami až tisíci prediktory. Není možné,

nebo přesněji řečeno není přípustné, vybírat proměnné do modelu ručně jednu po druhé. Nejprve je vhodné jakýmsi způsobem zredukovat počet proměnných pouze na ty relevantní a až pak přichází na řadu ruční modelování. V lineární regresi tento účel plní například selekce proměnných typu forward či stepwise. V modelech s proměnlivými koeficienty takovýto algoritmus prozatím nebyl ustanoven. V této podkapitole se pokusíme nastínit jak by jakási obdoba forward selekce proměnných mohla vypadat pro modely s proměnlivými koeficienty.

Nejprve se zamysleme nad situací, kdy nějaký proměnlivý koeficient vyjde konstantní či téměř konstantní. Nebylo by pak vhodnější ho odhadovat rovnou jako konstantní, čímž si ušetříme výpočetní výkon. Naše práce pojednává na téma modelů s proměnlivými koeficienty, ale v literatuře se často setkáme s pojmem model se semi-proměnnými koeficienty. Jedná se o model s proměnlivými koeficienty, kde část regresorů má koeficienty konstantní, tedy jsou definovány jako v lineární regresi. Model má tvar

$$Y_i = \sum_{j=1}^{m} \beta_j(U_{i,j}) X_{i,j} + \sum_{l=m+1}^{k} \beta_l X_{i,l} + \varepsilon_i$$

kdy prvním m koeficientů je proměnlivých a zbylých $k - m$ konstantních. (Fan a Zhang, 2008) navrhli metodu odhadu takovýchto modelů pomocí lokální regrese. Princip je relativně jednoduchý. Nejprve je celý model odhadnut jako model s proměnlivými koeficienty, tedy i koeficienty dané jako konstantní odhadneme proměnlivě. Následně hodnoty $\hat{\beta}_j(U_{i,l}), j = m+1,...,k$ zprůměrujeme přes $U_i, i = 1,...,n$ a tím získáme odhady $\hat{\beta}_j, j = m+1,...,k$. Dále pak v druhém kroku použijeme tyto odhady a odhadneme i zbylé proměnlivé koeficienty.

Náš algoritmus navrhujeme pro tyto modely ze dvou důvodů. Za prvé pokud je daný koeficient konstantní, pak dává smysl k němu tak při odhadu i přistupovat a ušetřit si výpočetní složitost neparametrického odhadu. Za druhé se nám zjednoduší testování hypotéz. Pokud nám v prvotním testu vyjde koeficient jako konstantní, pak ho tak můžeme i odhadnout a při testování jeho relevance použít klasický t-test z lineární regrese.

Při rozhodování se o přidání či nepřidání proměnné do modelu se před námi nachází tři otázky. Přidat proměnnou s proměnlivým parametrem? Přidat proměnnou s konstantním parametrem jako v lineární regresi? A nebo proměnnou do modelu nepřidat? Náš algoritmus bude fungovat na stejném principu jako forward výběr proměnných. V každém kroku do modelu přidáme další proměnnou a pomocí nějakých statistických testů rozhodneme, zda-li její koeficient bude proměnlivý, konstantní či ji nepřidáme. Na to ovšem budeme potřebovat až dva testy, první zda-li je odhadnutý koeficient proměnlivý a pokud ne tak ho odhadnout jako konstantní a otestovat, zda-li není nulový. To už je ale problém vícenásobného testování a nestačí proto pouze zvlášť provést tyto testy, ale je třeba aplikovat nějakou korekci aby nám výsledné p-hodnoty odpovídaly. K tomu navrhujeme použít Benjamini-Hochbergovu korekci, která je definována následovně

Věta 6. *(Benjamini-Hochbergova korekce)*

Mějme m nezávislých testovacích procedur s nulovými hypotézami H_1,...,H_m a k nim odpovídající p-hodnoty P_1,...,P_m. Ty seřadíme od nejmenší po největší p-hodnotu a označíme je $P_{(1)}$,...,$P_{(m)}$. Korekce pro stupeň α má následující dva kroky

1) Pro požadovaný stupeň α najdeme nejvyšší index k splňující $P_{(k)} \leq \frac{k}{m}\alpha$

2) Zamítáme nulové hypotézy pro testovací procedury $H_{(i)}, i = 1,...,k$

Tímto způsobem vykompenzujeme chyby mnohonásobného testovaní a závěry těchto testů by měly být relativně přesné. V našem případě budeme používat pouze dvě testovací procedury a proto $m = 2$.

Dále je třeba se zamyslet nad složitostí. Odhadnout samotný model s proměnlivými koeficienty je dost výpočetně náročné a tento algoritmus v každém kroku odhaduje model jednou až dvakrát (pokud koeficient u nově přidaného regresoru vyjde jako proměnlivý pak pouze jednou, pokud ne, tak je třeba model odhadnout znovu, akorát s tímto regresorem definovaným s konstantním koeficientem). Je tedy třeba co nejvíce snížit výpočetní složitost a do jisté přijatelné míry i za cenu korektnosti a přesnosti.

Prvně navrhujeme zafixovat stav regresoru tak, jak byl v kroku kdy byl do modelu přidán klasifikován. Pro zjednodušení zde budeme používat několik vlastních pojmů. Pod pojmem proměnlivý regresor budeme myslet takový regresor, jehož odhadnutý koeficient v modelu je funkce, tedy proměnlivý koeficient závislý na nějakém modifikátoru vlivu. Jako konstantní regresor pak myslíme takový regresor, jehož odhadnutý koeficient je konstantní, čili klasický regresor z lineární regrese. Pokud nově přidaný regresor otestujeme a vyjde nám jako proměnlivý, pak už po celý zbytek chodu algoritmu bude do modelu vstupovat s proměnlivým koeficientem. Stejně tak pokud proměnlivost zamítneme a regresor vyjde konstantní, ale nenulový. Nemůžeme vyloučit, že s příchodem dalších regresorů se nějaký dříve přidaný nestane z proměnlivého konstantní či zanedbatelný, avšak to je cena za menší výpočetní složitost. Abychom zmenšili pravděpodobnost, že takováto chyba nám při algoritmu nastane, tak hodláme test o proměnlivosti regresoru položit na vysoké hladině $1 - \alpha$.

Taktéž by bylo vhodné předem stanovit pořadí, ve kterém budeme regresory do modelu algoritmem přidávat. Navrhujeme nejprve kupříkladu odhadnout klasický model lineární regrese a seřadit proměnné dle vysvětlující síly, tedy třeba dle chí-kvadrátu jež klasicky vidíme při výstupu. Důvodem proč je výpočetní složitost. Kdybychom si vzali opačný příklad a proměnné seřadili od největší relevantnosti, tak čím dříve proměnná do modelu vstoupí, tím větší má pravděpodobnost, že v něm i zůstane. Tím pádem pro proměnné na konci řady budeme muset v každém kroku znovu a znovu odhadovat model s většinou proměnných které do něj ve finále patří. Při postupu od nejméně relevantních proměnných ale ze začátku většinu těchto proměnných zamítneme a ve výsledku ušetříme výpočetní výkon.

Tím jsme nastínili náš algoritmus a důvody jeho podoby a nyní jej pojďme přesně formulovat:

Mějme data o $i = 1,...,n$ nezávislých pozorováních s $j = 1,...,k$ regresory $X_{i,j}$ a odezvou Y_i.
(0) Odhadneme model lineární regrese pro odezvu Y a pro regresory $X_j, j = 1,...,k$. Regresory pak seřadíme od nejmenší relevantnosti po největší dle chíkvadrátu a označíme je jako $X_{(1)},...,X_{(k)}$. Položme $p = 0$.
(1) Položme $p = p+1$. Definujme si množinu regresorů $X = \{X_{(1)},...,X_{(p)}\}$. Ke každému regresoru si definujeme i jeho stav, tedy stav (proměnlivý/konstantní/nevstupuje) v jakém byl do modelu klasifikován a případně stav neznámý pokud se v tomto kroku v množině regresorů ocitnul poprvé.
(2) Proměnné $X_{(j)}, j = 1,...,p-1$ do modelu vstupují v takovém stavu (proměnlivý/konstantní/nevstupuje), v jakém byly dříve klasifikovány. Proměnná $X_{(p)}$ je v modelu definována jako proměnlivá. Model odhadneme. Otestujeme, zdali je proměnná $X_{(p)}$ proměnlivá (pomocí nějakého vhodného testu popsaného v podkapitole o testování hypotéz). Taktéž na výslednou p-hodnotu aplikujeme Benjamini-Hochbergovu korekci. Pokud zamítáme nulovou hypotézu konstantnosti koeficientu $\beta_{(p)}$, pak přeskočíme následující kroky a iterujeme krokem (1). Pokud nulovou hypotézu nezamítáme, pak pokračujeme krokem (3).
(3) Model je definován následovně. Proměnné $X_{(j)}, j = 1,...,p-1$ do modelu vstupují v takovém stavu, v jakém byly dříve klasifikovány. Proměnná $X_{(p)}$ je v modelu definována jako konstantní. Model odhadneme a klasickým t-testem otestujeme, zda-li je koeficient regresoru $X_{(p)}$ konstantní. Na výslednou p-hodnotu aplikujeme Benjamini-Hochbergovu korekci. Pokud zamítáme nulovou hypotézu, že $\beta_{(p)} = 0$, pak regresor do modelu vstupuje jako konstantní. Pokud nulovou hypotézu nezamítáme, tak regresor do modelu nevstupuje. Iterujeme krokem (1).

Tímto způsobem po poslední iteraci získáme stavy ke každému regresoru a můžeme přejít k ručnímu modelování.

Tento algoritmus byl sepsán dosti obecně a nabízí mnoho příležitostí pro modifikaci. Prvotní otázkou je metoda odhadu modelu s proměnlivými koeficienty. Vzhledem k tomu, že zde navrhujeme používat modely se semi-proměnlivými koeficienty, jejichž odhad byl popsán v (Fan a Zhang, 2008), tak se metoda odhadu pomocí lokální regrese jeví jako nejvhodnější. Avšak i splajnové metody by se daly modifikovat pro tento typ modelů.

Dále způsobů jak otestovat, zda-li je proměnlivý koeficient konstantní či nikoliv bylo navrhnuto několik. Výběr vhodného testu pak závisí primárně na volbě metody odhadu a sekundárně na preferencích uživatele algoritmu.

Taktéž volba korekce mnohonásobného testování někomu nemusí vyhovovat. Pokud bychom nepovažovali dva použité testy za nezávislé, pak lze použít kupříkladu Benjamini-Hochberg-Yekutieli modifikovanou korekci.

5. Aplikace

V předchozí kapitole jsme shrnuli teoretické podklady modelů s proměnlivými koeficienty a v této části se podíváme na jejich praktickou aplikaci. Kombinace neparametrických odhadových metod a interpretovatelnosti zobecněného lineárního regresního modelu dělá z modelů s proměnlivými koeficienty všestranný nástroj. Jejich využití lze proto najít v širokém spektru oblastí, od financí a politiky po lékařství a ekologii.

(Hastie a Tibshirani, 1993) představili model s proměnlivými koeficienty pro zkoumání závislosti koncentrace oxidu dusnatého v motorech spalujících etanol. Parametry modelu autoři zvolili jako závislé na poměru ekvivalence vyjadřujícím míru směsi palivo-vzduch. Dále autoři zkoumali pravděpodobnost infarktu myokardu s proměnlivostí koeficientů dle systolického krevního tlaku a cholesterolu. Jako poslední příklad byla prezentována studie doby přežití pacientů s rakovinou plic, kde byl k již existujícímu modelu přidán proměnlivý parametr dle času.

(Cai a kol., 2000) se zabývali počtem návštěv nemocnice s oběhovými a respiračními potížemi vzhledem ke koncentraci znečišťujících látek ve vzduchu. Autoři použili model s časově proměnlivými koeficienty. V druhém příkladu analyzovali binární odezvu indikující jestli pacient přežil popáleniny, přičemž se zkoumala závislost koeficientů na věku pacienta.

(Fan a Zhang, 2008) se taktéž zaměřili na počet návštěv nemocnic pacientů s oběhovými a respiračními potížemi v Hong Kongu.

(Huang a kol., 2004) zkoumali procentuální podíl CD4 buněk u osob nakažených virem HIV a proměnlivost koeficientů v závislosti na době od nakažení.

(Marx, 2010) studovali možný výskyt kataklastické zlomové zóny (reprezentované odezvou obsahu kataklastické horniny) v závislosti na obsahu podzemní vody, grafitu, oxidu hlinitého, oxidu sodného a tepelné vodivosti. Data pocházela z měření až do hloubky 9,1 kilometrů pod povrchem. Jako modifikátor vlivu parametrů byla použita právě hloubka měření. Jejich výsledný model byl o 12%-21% lepší (podle koeficientu determinance) než ostatní klasické modely odhadnuté na těchto datech.

Většina uvedených příkladů se týkala lékařských dat a jejich modely se vyznačovaly parametry proměnlivými s časovou proměnnou. Proměnlivost parametrů s časem je jistě velice intuitivní, avšak i jiné proměnné mohou přinést zajímavé výsledky.

V první části této kapitoly prozkoumáme dostupný software obsahující funkce pro vytváření modelů s proměnlivými koeficienty. V části druhé budeme diskutovat empirické vlastnosti testů konstantnosti koeficientu zjištěných pomocí simulací. Ve třetí části pak představíme příklad demonstrující využití modelů s proměnlivými koeficienty.

5.1 Software

Většina komerčně využívaných statistických softwarů, jako *SAS*, *STATA*, *SPSS* či *Eviews* v sobě doposud nemá implementovány modely s proměnlivými koeficienty. Z těch nejběžněji užívaných je zatím podporuje pouze volně dostupný software *R* (R Core Team, 2013). R operuje na principu balíčků ("package"), ve kterých různí vývojáři implementují teoretické principy a algoritmy ve formě funkcí. V současné době se nabízí několik balíčků s modely s proměnlivými koeficienty. (Sperlich a Theler, 2015) porovnali několik *R* balíčků schopných odhadnout modely s proměnlivými koeficienty. Z nich autoři považují za nejkvalitnější balíček *mgcv* (Wood, 2007) zaměřený na neparametrickou regresi. Ten disponuje funkcí *gam*, která v sobě má možnost použití modelů s proměnlivými koeficienty a vykreslení jejich koeficientů. Dále je zde rovněž k dispozici funkce *bam* využívající stejné principy, avšak za použití efektivnějších algoritmů s rychlejším výpočtem. Právě v tomto programu jsme vypracovali následující příklady.

Ještě je třeba zmínit zásadní věc a to výpočetní složitost modelů s proměnlivými koeficienty v softwaru *R*. Použité výpočetní algoritmy v současné době nejsou zcela optimalizované a to společně se složitostí těchto modelů způsobuje vysoké nároky na výpočetní výkon použitého počítače. Při odhadu pouze s několika prediktory algoritmus doběhne rychle, avšak jakmile se počet regresorů vyšplhá do řádu desítek či dokonce stovek, tak nastávají problémy. Software *R* je známý tím, že využívá především paměť RAM. Při desítkách regresorů je pak nutné disponovat buď špičkovým počítačem nebo serverovou verzí *R*, která je však již placená. Pokud tomu tak není, pak se nelze dobrat výsledku, jelikož R skončí s chybovou hláškou oznamující, že nebylo možné alokovat další paměť. Právě tato náročnost brání použití modelů s proměnlivými koeficienty pro širší komerční využití. Je otázkou, zda-li se dříve podaří najít rychlejší algoritmy pro jejich odhad nebo výpočetní technika pokročí natolik dopředu, že vyžadovaný výkon nebude problémem.

5.2 Empirické vlastnosti

V práci jsme podrobněji popsali dva způsoby testování hypotézy o konstantnosti proměnlivého koeficientu. Pro odhad pomocí lokální regrese se jednalo o test s testovou statistikou a kritickou hladinou založenou na distribuci maximálního rozdílu mezi odhadnutým a skutečným funkcionálním koeficientem. Pro odhad na longitudinálních datech přes polynomiální splajny byl popsán test operující na principu bootstrappingu. Odhad pomocí polynomiálních splajnů lze však samozřejmě aplikovat i na standardní data s nezávislými pozorováními, k čemuž budeme používat výše zmíněný balíček *mgcv* v softwaru *R*. Ten však v sobě obsahuje odlišný test pro ověření nekonstantnosti koeficientů. Jakým způsobem je tento test navržen bude blíže popsáno v následujícím příkladě, případně celé odvození je k dispozici v (Wood, 2013). Vzhledem k tomu, že pro odhad pomocí splajnů budeme využívat tento balíček, rozhodli jsme se prozkoumat empirické vlastnosti tohoto testu a navíc ještě námi naprogramovaného testu z metody odhadu lokální regresí.

Jako nejvhodnější způsob pro porovnání vstupní hladiny významnosti testu α a skutečně pozorované hladiny α_R se jeví simulace. Princip této metody je následující. Nejprve si náhodně vygenerujeme data odpovídající předem zvolenému podkladovému modelu. Na těchto datech následně odhadneme model s proměnlivými koeficienty (pomocí polynomiálních splajnů a lokální regrese) a otestujeme vybranou hypotézu o konstantnosti koeficientu na zvolené hladině α. Výsledek testu (zamítnutí/nezamítnutí nulové hypotézy) si zapamatujeme a celý postup opakujeme. Nakonec se podíváme na výsledky testování ze všech těchto simulací, ze kterých spočteme skutečnou hladinu významnosti α_R. Čím větší počet simulací provedeme, tím přesnější dostaneme výsledky. Na druhou stranu jsou ale simulace výpočetně náročná záležitost a při dostatečně vysokém počtu už jsou rozdíly v přesnosti minimální.

Výchozí vzorec, podle kterého generujeme data, jsme zvolili velice jednoduše a to tak, že odezva je vždy nulová, tedy

$$Y_t = \eta_t,$$

kde náhodné chyby η_t jsou nezávislé a normálně rozdělené s nulovou střední hodnotou a rozptylem σ^2.

Data dle tohoto vzorce generujeme následovně – nejprve vygenerujeme bílý šum délky n, tedy vektor $\boldsymbol{\eta}$, jehož složky jsou navzájem nezávislé s normálním rozdělením s nulovou střední hodnotou a rozptylem σ^2. Dále pak generujeme vektor regresorů $\mathbf{X}$, u kterého ale chceme longitudinální strukturu. Vygenerujeme si tedy časovou řadu definovanou jako $X_t = X_{t-1} + \upsilon_t$, kde $\upsilon_t \sim N(1, 2)$ a $X_1 = 30$. Výsledný vektor odezev $\mathbf{Y}$ je pak roven vygenerovanému bílému šumu, jelikož koeficient modelu u regresoru X předpokládáme konstantně nulový.

Na takovýchto datech pak odhadneme model s proměnlivými koeficienty

$$Y_t = \beta(t)X_t + \varepsilon_t,$$

a to přes polynomiální splajny a pomocí lokální regrese. Na každém z nich pak budeme testovat příslušnou nulovou hypotézu

$$\begin{aligned} H_0 &: \beta(t) = 0 \\ H_1 &: \beta(t) \neq 0. \end{aligned} \tag{5.1}$$

Celý postup D-krát opakujeme a ve výsledku získáme skutečně pozorované hladiny významnosti testů α_R.

Z důvodu výpočetní náročnosti jsme se rozhodli zvolit počet simulací jako $D = 1000$. Domníváme se, že takovýto počet by při jednoduchosti modelu měl bohatě postačovat pro rozumnou přesnost. Dále budeme zkoušet různé možnosti konfigurace parametrů – délky generovaných dat $n = 10, 100, 500$, hladiny významnosti $\alpha = 0.01, 0.05, 0.1$, rozptyl bílého šumu $\sigma^2 = 0.05, 1, 2$ a u metody lokální regrese i šířky pásma = 3,7. V druhé fázi pak vyzkoušíme tytéž simulace, ale za neplatnosti nulové hypotézy, tj. s výchozím vzorcem pro podkladová data definovaným jako

$$Y_t = X_t + \varepsilon_t,$$

kde pak opět budeme odhadovat model s proměnlivými koeficienty ve tvaru

$$Y_t = \beta(t)X_t + \varepsilon_t.$$

Tam se však vzhledem k neplatnosti nulové hypotézy již budeme zabývat skutečnou silou testu $1 - \beta_R$. Pro všechny tyto kombinace provedeme simulace a změříme skutečné hladiny nebo síly testů a o výsledcích budeme diskutovat.

5.2.1 Odhad pomocí polynomiálních splajnů

První vlnu simulací jsme provedli pro model s proměnlivými koeficienty odhadnutý pomocí polynomiálních splajnů přes balíček *mgcv*. Níže uvádíme dvě tabulky, které obsahují různé kombinace výše vyjmenovaných vstupních parametrů naší simulace, respektive zvolených podkladových dat. V první tabulce testujeme

Tabulka 5.1: Simulace za platnosti nulové hypotézy, metoda polynomiálních splajnů

Tabulka výsledků simulací pro různé kombinace délky dat (n), hladiny významnosti (α) a rozptylu (σ^2). Pozorovaná α značí ze simulací pozorovanou α_R. Rozdíl je absolutní hodnota rozdílu α a α_R.

n	α	σ^2	pozorovaná α	rozdíl
10	0.01	0.05	0.036	0.026
		1.00	0.026	0.016
		2.00	0.022	0.012
	0.05	0.05	0.075	0.025
		1.00	0.105	0.055
		2.00	0.092	0.042
	0.10	0.05	0.182	0.082
		1.00	0.174	0.074
		2.00	0.208	0.108
100	0.01	0.05	0.017	0.007
		1.00	0.016	0.006
		2.00	0.011	0.001
	0.05	0.05	0.061	0.011
		1.00	0.075	0.025
		2.00	0.075	0.025
	0.10	0.05	0.119	0.019
		1.00	0.138	0.038
		2.00	0.112	0.012
500	0.01	0.05	0.012	0.002
		1.00	0.015	0.005
		2.00	0.009	0.001
	0.05	0.05	0.073	0.023
		1.00	0.059	0.009
		2.00	0.061	0.011
	0.10	0.05	0.125	0.025
		1.00	0.134	0.034
		2.00	0.106	0.006

nulovou hypotézu nevýznamnosti koeficientu β (tj. konstantní nulová hodnota) na datech, ve kterých je tento koeficient skutečně nulový.

Na výsledky simulací se budeme dívat vždy z pohledu vlivu jedné z našich proměnných. Vzhledem k počtu kombinací u každé jednotlivé proměnné je třeba tyto výsledky nějak zagregovat a to průměrným absolutním rozdílem mezi α a α_R (dále budeme zmiňovat jen jako rozdíl). Kupříkladu pokud se chceme zaměřit na vliv délky dat na tento rozdíl, pak pro každou délku dat n máme v tabulce 9 pozorování, ze kterých zprůměrujeme jejich rozdíly α a α_R a získáme tak jedinou hodnotu pro každou délku dat, které pak budeme porovnávat.

Nejprve se zaměřme na vliv různé délky generovaných dat n. Pro $n = 10$ nám průměrný rozdíl vyšel přibližně 0.049 a pro $n = 100$ pak 0.016, což je o 67% nižší. U délky $n = 500$ vychází 0.013, tedy o dalších 19% nižší oproti délce $n = 100$. Závěry tedy odpovídají našim očekáváním, a to že na větších datech statistická inference dává přesnější výsledky.

Dalším pohledem je vliv rozptylu σ^2 na testování hypotéz. Při $sd = 0.05$ nám průměrný rozdíl vyšel 0.024, pro $sd = 1$ je 0.029 a pro $sd = 2$ opět 0.024. Vliv rozptylu je zde tedy kupodivu minimální, jelikož pro všechny tři hladiny nám průměrný rozdíl vychází téměř shodně. Zdá se že vlivy ostatních parametrů vliv rozptylu převažují.

V poslední řade se podívejme na vliv různých hladin α. Volba tohoto parametru se ukazuje jako klíčová, jelikož rozdíly mezi jednotlivými hladinami jsou výrazné. Největší rozdíl nepřekvapivě vyšel pro nejvyšší hladinu $\alpha = 0{,}1$ a to 0.044. Pro hladinu 0,05 pak ten samý průměrný rozdíl vyšel 0.025, což je o 43% nižší. Rozdíl u hladiny 0,01 je pak ještě o dalších 66% nižší a to 0.008. S čím větší přesností tedy hodláme testovat, tím si jsou stanovená a skutečná hladina blíže.

Další část se věnuje těm samým simulacím, avšak za předpokladu neplatnosti testované nulové hypotézy. Hodnota koeficientu β v podkladových datech tedy neodpovídá nulové hypotéze $\beta(t) = 0$. Tím pádem ale už nelze měřit skutečné hladiny, ale díváme se na sílu testu $(1 - \beta)$. Hodnoty α a β jsou spolu provázané a při porovnávání teoreticky zadaných hodnot a skutečně naměřených se pak díváme na rozdíl $(1 - \alpha) - (1 - \beta)$.

Vliv různých délek pozorovaných dat vyšel odlišně než v předchozí části. Rozdíl mezi teoretickou a skutečnou sílou testu nám vyšel stejně jako 0.053 pro délky $n = 100, 500$ a při délce 10 vychází 0.070, tedy o 31% vyšší. Síla testu za neplatnosti nulové hypotézy tedy na rozsahu dat zdá se příliš nezávisí. Uvědomme si ale, že nulová hypotéza předpokládala nulovost koeficientu, kdežto podkladový model ho definoval jako jednotkový. To je relativně velký rozdíl hodnot a test je dostatečně přesný aby nulovou hypotézu zamítal i při málo datech. Kdyby podkladový model počítal s koeficientem $\beta = 0.1$ nebo nějakými ještě menšími hodnotami, pak by rozdíly mezi silami testu byly pro různé délky dat nepochybně podstatnější.

Pohled na odlišnosti při použití různých rozptylů nám dává v podstatě stejné rozdíly pro všechny rozptyly. Pro $sd = 0.05, 1, 2$ vychází $0.058, 0.059, 0.060$. Pro vyšší rozptyly je průměrný rozdíl vyšší, ale jen velice nepatrně, a proto zde vliv rozptylu na sílu testu za neplatnosti nulové hypotézy není příliš podstatný.

Vliv stanovené hladiny α je v tomto případě obdobně důležitý jako v předchozí části. Nejvyšší hladina $\alpha = 0.10$ implikuje největší rozdíl 0.106. Její snížení na

Tabulka 5.2: Simulace za neplatnosti nulové hypotézy, metoda polynomiálních splajnů

Tabulka výsledků simulací pro různé kombinace délky dat (n), rozptylu (σ^2) a hladiny významnosti (α). Pozorovaná $1-\beta$ značí ze simulací pozorovanou $1-\alpha_R$. Rozdíl je absolutní hodnota rozdílu $(1-\alpha)-(1-\beta)$.

n	σ^2	α	pozorovaná $1-\beta$	rozdíl
10	0.05	0.01	0.984	0.026
		0.05	0.988	0.062
		0.10	0.988	0.112
	1.00	0.01	0.984	0.026
		0.05	0.984	0.066
		0.10	0.984	0.116
	2.00	0.01	0.976	0.034
		0.05	0.986	0.064
		0.10	0.976	0.124
100	0.05	0.01	1.000	0.010
		0.05	1.000	0.050
		0.10	1.000	0.100
	1.00	0.01	1.000	0.010
		0.05	1.000	0.050
		0.10	1.000	0.100
	2.00	0.01	1.000	0.010
		0.05	1.000	0.050
		0.10	1.000	0.100
500	0.05	0.01	1.000	0.010
		0.05	1.000	0.050
		0.10	1.000	0.100
	1.00	0.01	1.000	0.010
		0.05	1.000	0.050
		0.10	1.000	0.100
	2.00	0.01	1.000	0.010
		0.05	1.000	0.050
		0.10	1.000	0.100

hodnotu 0,05 pak vede ke snížení rozdílu o 48% na 0.055. Hladina $\alpha = 0.01$ pak rozdíl snižuje o dalších 70% a to na 0.016.

Vlivy vybraného rozptylu a stanovené hladiny nám sice dávají odlišné rozdíly, avšak v našem případě to na samotný výsledek testu nemá žádný vliv. Všimněme si, že pro větší délky pozorovaných dat ($n = 100, 500$) nám test zamítl nulovou hypotézu ve 100% případů a byl tedy ještě silnější než by teoreticky měl být. Pokud je tedy proměnlivý koeficient odlišný od nuly v řádu jednotek, pak je tento test velice silný v zamítání takové nulové hypotézy.

5.2.2 Odhad pomocí lokální regrese

Takovéto simulace nyní provedeme i pro odhad pomocí lokální regrese a pro ni navržený test založený na distribuci maximálního rozdílu mezi skutečným a odhadnutým proměnlivým koeficientem. V úvahu budeme navíc brát i další faktor, a to šířku pásma, která je pro odhad lokálním vyrovnáváním klíčová.

Opět se budeme dívat na vlivy jednotlivých proměnných, u kterých budeme operovat s průměrným rozdílem hladiny významnosti α a skutečně pozorované hladiny α_R.

V první části se podíváme na simulace modelu na datech, ve kterém nulová hypotéza platí. Právě na nich se ukázala důležitost volby šířky pásma. Při volbě šířky pásma 3 nám totiž simulované hladiny testu vycházejí velice nepřesně, téměř nezávisle na volbě ostatních faktorů. Zdá se tedy, že volba malého pásma má negativní dopad na statistickou inferenci a je vhodnější volit pásmo širší. Dále se budeme podrobně zabývat pouze výsledky pro volbu šířky pásma 7, kde již výsledky dosahují slušné přesnosti.

Nejprve se zaměřme na délku pozorovaných dat. Pro volbu $n = 10$ je výsledek vysoce nepřesný, jelikož nám v průměru v 63% případů simulací použitý test zamítal platnou nulovou hypotézu. U délky dat 100 však již pozorovaná α_R a testovaná α vycházejí podobně s rozdílem 0.037, tedy v řádu setin a pro volbu $n = 500$ se rozdíly pohybují v řádu tisícin s rozdílem 0.008, což je o 78% méně než

Tabulka 5.3: Simulace za platnosti nulové hypotézy, metoda lokální regrese

Tabulka výsledků simulací pro různé kombinace délky dat (n), rozptylu (σ^2), hladiny významnosti (α) a šířky pásma. Pozorovaná α značí ze simulací pozorovanou α_R. Rozdíl je absolutní hodnota rozdílu $\alpha - \alpha_R$. Rozděleno do dvou tabulek dle šířky pásma.

n	σ^2	α	š.pásma	pozorovaná α	rozdíl	n	σ^2	α	š.pásma	pozorovaná α	rozdíl
10	0.05	0.01	3	0.289	0.279	10	0.05	0.01	7	0.665	0.655
		0.05		0.261	0.211			0.05		0.712	0.662
		0.10		0.323	0.223			0.10		0.694	0.594
	1.00	0.01		0.296	0.286		1.00	0.01		0.657	0.647
		0.05		0.334	0.284			0.05		0.704	0.654
		0.10		0.337	0.237			0.10		0.723	0.623
	2.00	0.01		0.346	0.336		2.00	0.01		0.654	0.644
		0.05		0.325	0.275			0.05		0.721	0.671
		0.10		0.254	0.154			0.10		0.659	0.559
100	0.05	0.01	3	0.267	0.257	100	0.05	0.01	7	0.053	0.043
		0.05		0.210	0.160			0.05		0.091	0.041
		0.10		0.242	0.142			0.10		0.043	0.057
	1.00	0.01		0.237	0.227		1.00	0.01		0.063	0.053
		0.05		0.224	0.174			0.05		0.038	0.012
		0.10		0.268	0.168			0.10		0.078	0.022
	2.00	0.01		0.228	0.218		2.00	0.01		0.052	0.042
		0.05		0.185	0.135			0.05		0.043	0.007
		0.10		0.241	0.141			0.10		0.047	0.053
500	0.05	0.01	3	0.384	0.374	500	0.05	0.01	7	0.013	0.003
		0.05		0.373	0.323			0.05		0.054	0.004
		0.10		0.394	0.294			0.10		0.116	0.016
	1.00	0.01		0.418	0.408		1.00	0.01		0.017	0.007
		0.05		0.422	0.372			0.05		0.051	0.001
		0.10		0.368	0.268			0.10		0.118	0.018
	2.00	0.01		0.402	0.392		2.00	0.01		0.019	0.009
		0.05		0.336	0.286			0.05		0.058	0.008
		0.10		0.363	0.263			0.10		0.095	0.005

Tabulka 5.4: Simulace za neplatnosti nulové hypotézy, metoda lokální regrese

Tabulka výsledků simulací pro různé kombinace délky dat (n), rozptylu (σ^2), hladiny významnosti (α) a šířky pásma. Simulace značí ze simulací pozorovanou $1-\beta$. Rozdíl je absolutní hodnota rozdílu $(1-\alpha)-(1-\beta)$. Rozděleno do dvou tabulek dle šířky pásma.

n	σ^2	α	š.pásma	simulace	rozdíl	n	σ^2	α	š.pásma	simulace	rozdíl
10	0.05	0.01	3	0.992	0.002	10	0.05	0.01	7	1.000	0.010
		0.05		0.998	0.048			0.05		0.981	0.031
		0.10		0.994	0.094			0.10		0.995	0.095
	1.00	0.01		0.984	0.006		1.00	0.01		0.997	0.007
		0.05		0.993	0.043			0.05		0.974	0.024
		0.10		0.984	0.084			0.10		0.989	0.089
	2.00	0.01		1.000	0.010		2.00	0.01		0.987	0.003
		0.05		0.995	0.045			0.05		1.000	0.050
		0.10		0.993	0.093			0.10		0.983	0.083
100	0.05	0.01	3	1.000	0.010	100	0.05	0.01	7	1.000	0.010
		0.05		1.000	0.050			0.05		1.000	0.050
		0.10		1.000	0.100			0.10		1.000	0.100
	1.00	0.01		1.000	0.010		1.00	0.01		1.000	0.010
		0.05		1.000	0.050			0.05		1.000	0.050
		0.10		1.000	0.100			0.10		1.000	0.100
	2.00	0.01		1.000	0.010		2.00	0.01		1.000	0.010
		0.05		1.000	0.050			0.05		1.000	0.050
		0.10		1.000	0.100			0.10		1.000	0.100
500	0.05	0.01	3	1.000	0.010	500	0.05	0.01	7	1.000	0.010
		0.05		1.000	0.050			0.05		1.000	0.050
		0.10		1.000	0.100			0.10		1.000	0.100
	1.00	0.01		1.000	0.010		1.00	0.01		1.000	0.010
		0.05		1.000	0.050			0.05		1.000	0.050
		0.10		1.000	0.100			0.10		1.000	0.100
	2.00	0.01		1.000	0.010		2.00	0.01		1.000	0.010
		0.05		1.000	0.050			0.05		1.000	0.050
		0.10		1.000	0.100			0.10		1.000	0.100

pro délku dat 100. Tím se potvrzuje zřejmý fakt, že čím větší počet pozorování, ze kterých odhadujeme model, tím přesnější statistická inference. Na rozdíl od odhadu pomocí polynomiálních splajnů je ale tento test nepoužitelný pro velmi malé rozsahy dat.

Dalším faktorem, na jehož vliv se podíváme, je rozptyl σ^2. Ten zde překvapivě nemá zásadnější vliv. Když zprůměrujeme rozdíly testované a simulované α pro jednotlivé rozptyly přes všechny ostatní kombinace faktorů, tak výsledné hodnoty vycházejí pro všechny simulované rozptyly téměř shodně, tj. $0.231, 0.226, 0.222$ pro $sd = 0.05, 1, 2$ respektive. Vliv ostatních faktorů tedy na výslednou inferenci převažuje vliv rozptylu.

Vliv volby testované hladiny α taktéž nemá zásadnější vliv na výsledný rozdíl. Průměrné rozdíly pro $\alpha = 0.01, 0.05, 0.1$ vyšly $0.234, 0.229, 0.216$, tedy opět téměř shodně.

Z těchto simulací za platnosti nulové hypotézy tedy vyplývá jednoznačný závěr – šířka pásma a délka pozorovaných dat mají zcela klíčový vliv na přesnost testu založeného na distribuci maximálního rozdílu skutečného a odhadnutého proměnlivého koeficientu. V porovnání s těmito dvěma faktory pak zbylé proměnné mají již jen minimální vliv. Pro odhad pomocí lokální regrese tedy doporučujeme rozumně velkou šířku pásma a velký počet pozorování.

V druhé části se budeme zabývat simulacemi na datech, ve kterých nulová

hypotéza neplatí. V tomto případě již nezkoumáme hladinu α, avšak sílu testu $(1 - \beta)$. V tomto případě se použitý test ukazuje jako vysoce přesný ve smyslu správného zamítání neplatné nulové hypotézy. Na výslednou simulovanou sílu testu zde má hlavní vliv délka dat. Pro velmi malý rozsah dat $n = 10$ v malém procentu simulací dojde k nezamítnutí nulové hypotézy, avšak pro delší rozsahy dat test nulovou hypotézu zamítá ve 100% případů.

Volba rozptylu zde stejně jako v předchozí části nemá na průměrný rozdíl zásadnější vliv, pro všechny zkoumané možnosti vychází téměř shodně.

Hladina významnosti se u testování nulové hypotézy na datech, ve kterých neplatí ukázala jako významná. Pro $\alpha = 0.01$ nám průměrný rozdíl vyšel 0.009, pro hladinu $\alpha = 0.05$ vychází 0.045 a nakonec pro $\alpha = 0.1$ jako 0.096. Uvědomme si ale, že tento rozdíl není narozdíl od předchozí části rozdílem hladin, ale rozdílem sil testu. S rostoucí hladinou α tedy očekáváme rostoucí sílu testu.

Dále je třeba poznamenat, že závěry našich zkoumání zde budou poněkud zkreslené vzhledem k faktu, že test pro dostatečně velké počty pozorování v datech nulovou hypotézu vždy zamítá. Na druhou stranu to, že je test silnější než by měl být, je pro nás pozitivní.

5.2.3 Porovnání testů

Simulacemi jsme otestovali skutečné hladiny a síly testů na konstantnost proměnlivého koeficientu u odhadů modelu s proměnlivými koeficienty přes polynomiální splajny a lokální regresi. Pojďme si nyní tyto 2 testy porovnat.

Při simulacích na podkladových datech, které nesplňují nulovou hypotézu vychází oba přístupe stejně kvalitně. Pro větší rozsahy dat oba testy vždy zamítají neplatnou nulovou hypotézu. Při rozsahu dat $n = 10$ došlo v malém počtu simulací k nezamítnutí nulové hypotézy, avšak síla testu stále dosahovala hodnot blízkých 1.

U simulací na datech s neplatnou nulovou hypotézou jsem zkusili i jinou volbu vzorce, podle kterého je generujeme, a to

$$Y_t = \frac{t}{n} X_t + \varepsilon_t,$$

kde n značí délku dat. Model s proměnlivými koeficienty odhadnutý na takovýchto datech by tedy měl vyjít s proměnlivým koeficientem $\beta(t)$. V předchozím textu ale výsledky těchto simulací nezmiňujeme a to z toho důvodu, že vyšly téměř shodně s výsledky simulací na datech s konstantně jednotkovým koeficientem $\beta(t)$. Při malém počtu pozorování ($n = 10$) došlo v malé části simulací k nezamítnutí nulové hypotézy, avšak pro delší data již byla nulová hypotéza zamítnuta vždy. Oba zkoumané testy jsou tedy při zamítání neplatné nulové hypotézy skutečně velice silné.

Výraznější rozdíly se nám ale objevují u simulací na podkladových datech s platnou nulovou hypotézou. Test pro odhad pomocí polynomiálních splajnů vychází konzistentně pro všechny délky dat, tj. pro všechny délky dat vychází simulovaná hladina α podobně jako teoretická hladina a průměrné rozdíly jsou samozřejmě vyšší pro kratší délky. Naproti tomu u testu pro odhad lokální regresí je míra přesnosti ve smyslu rozdílu mezi teoretickou a simulovanou hladinou dosti odlišná pro naše tři simulované délky dat. U délky $n = 10$ je tento rozdíl závratný a test se ukazuje jako nepoužitelný. U délky 100 dostáváme podobné rozdíly jako

u testu u metody polynomiálních splajnů, avšak u poslední zkoumané délky 500 se test pro lokální regresi ukazuje o řád přesnější (tj. rozdíly hladin jsou o řád nižší). Další rozdíl je u volby testované hladiny α. U testu pro odhad polynomiálními splajny má volba hladiny podstatný vliv na výsledné rozdíly, kdežto u metody lokální regrese tento faktor nemá tak zásadní vliv, jelikož jej zastiňují volba šířky pásma a délka dat.

Ve výsledku se jako vhodnější způsob jeví test navržený pro odhad pomocí polynomiálních splajnů, který se při simulacích chová konzistentně a jehož porovnání testované a pozorované hladiny α při pohledu na vlivy různých faktorů dává intuitivně smysl. Test přes distribuci maximálního rozdílu pro odhad lokální regresí nám sice při velkém počty dat a vhodně zvolené šířce pásma dává přesnější výsledky, avšak v opačném případě dostaneme výsledky velmi nepřesné, což je v případě tvorby modelů a používání výsledků statistického testování pro určení konečného tvaru modelu dosti nežádoucí.

5.3 Příklad

Cílem této práce je srozumitelně a kompaktně shrnout dosavadní teoretické poznatky týkající se modelů s proměnlivými koeficienty a jejich statistické inference. Jedním z možných použití těchto modelů je model, kde se koeficienty liší s časem. Ten je vhodný zejména pro dlouhodobá data, kdy se s velkou pravděpodobností vlivy jednotlivých regresorů s postupem času mění.

Zvolili jsme si proto relativně jednoduchý příklad, na kterém budeme demonstrovat sílu modelů s proměnlivými koeficienty a jejich inferenci. Podíváme se hlavně na přesnost výsledných odhadů z hlediska vystihnutí dat a z hlediska přesnosti odhadu dle konfidenčního pásma. Dále otestujeme hypotézu, zda-li výsledný koeficient není konstantně nulový. Jako data jsme použili čtvrtletní makroekonomická pozorování z USA v letech 1960-2000 (Greene, 2003). Naším cílem je modelovat závislost objemu spotřeby domácností ($CONS$) na hrubém domácím produktu (GDP) a čase (t) pomocí metody proměnlivých koeficientů a následně porovnat předpovědní schopnosti tohoto modelu s obdobným modelem lineární regrese.

Výsledné modely jsou ve tvaru:

$$CONS_t = \beta_0 + \beta_1 GDP + \beta_2 t + \beta_3 GDP * t + \varepsilon_t$$

$$CONS_t = \beta_1(t) GDP + \varepsilon_t$$

V kapitole Modely s proměnlivými koeficienty bylo popsáno několik způsobů odhadu modelu. Námi zvolený model odhadneme všemi těmito způsoby – pomocí polynomiálního splajnu, penalizovaného splajnu, vyhlazovacího splajnu a pomocí lokální regrese, které mezi sebou vzájemně porovnáme.

Nejprve však řekněme několik slov k volbě modelu aplikaci jednotlivých metod odhadu. V datech se nachází více proměnných než pouze hrubý domácí produkt GDP, avšak tato proměnná je z nich vůči zadané odezvě jediná relevantní. Všimněme si také, že v našem modelu s proměnlivými koeficienty není použit intercept. Je tomu tak ze dvou důvodů. Za prvé proměnlivý intercept nám nepřijde

jako zcela vhodný. Z matematického hlediska sice zlepšuje odhad, ale z praktického pohledu postrádá smysluplnou interpretaci. Přijde nám proto smysluplnější intercept buď nepoužít, nebo ho definovat jako konstantní, aby se zkoumaná proměnlivost přenesla do koeficientů k příslušným interpretovatelným proměnným. Druhým důvodem je odhad pomocí lokální regrese. Narozdíl od odhadů pomocí splajnů pro něj prozatím není k dispozici ucelený balíček v softwaru R. Byli jsme proto nuceni si tuto metodu sami naprogramovat. Při konstantním interceptu by se však nejednalo o model s proměnlivými koeficienty, ale o model se semi-proměnlivými koeficienty, kterými jsme se v této práci více nezabývali. Rozhodli jsme se proto intercept zcela vynechat u všech testovaných metod pro jejich porovnání.

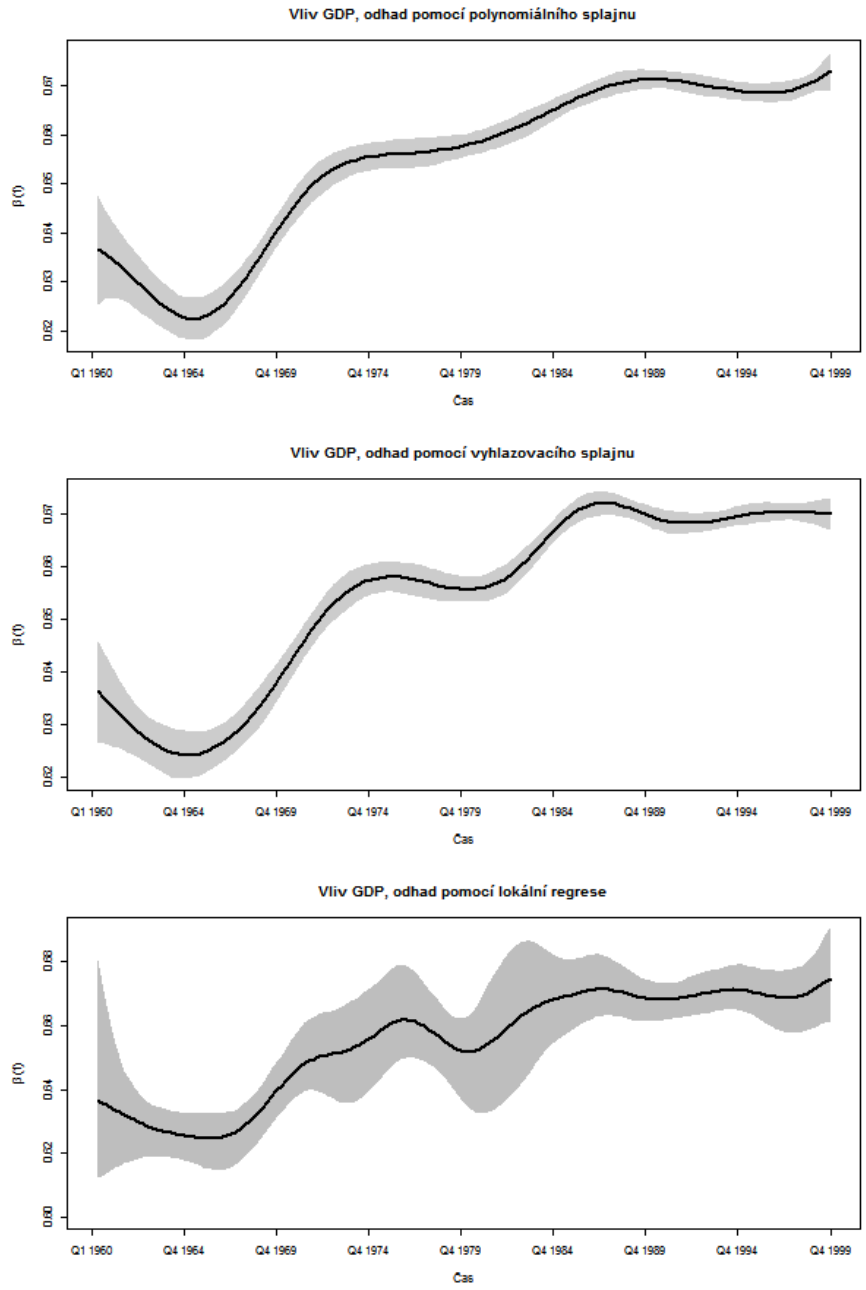
Na obrázku 5.1 jsme vykreslili výsledné odhady časově proměnlivého koeficientu β_1 pro odhady pomocí polynomiálního splajnu, vyhlazovacího splajnu a lokální regrese. Odhad pomocí penalizovaného splajnu byl téměř shodný s odhadem pomocí polynomiálního splajnu a proto jsme ho do grafu nezahrnuli.

Vidíme, že do roku 1965 se vliv GDP snižoval, ale dále od tohoto roku se již, až na menší pokles kolem roku 1980, zvyšoval nebo stagnoval v 90. letech. Z těchto odhadů prvotně soudíme, že se zde proměnlivost s časem jistě nachází.

Taktéž se zde ukazují rozdíly mezi jednotlivými typy odhadů. Vyhlazovací splajn na rozdíl od polynomiálního není tak hladký. To je způsobeno tím, že při jeho odhadu se jako uzly volí všechny pozorované časy t, kdežto u polynomiálního splajnu je vybrán pouze určitý počet ekvidistantních uzlů. Avšak i přes drobné rozdíly se dá říci, že všechny odhady pomocí splajnů jsou si relativně dosti podobné. Odhad pomocí lokální regrese je ale už ve svém principu fundamentálně odlišný, což je zřetelně vidět u porovnání se splajnovými odhady v grafu 5.1. 95 konfidenční pásmo u lokální regrese je taktéž mnohem širší než u odhadu pomocí splajnů.

U metody odhadu pomocí lokální regrese hraje velkou roli výběr šířky pásma h. Při menší šířce pásma sice dostáváme přesnější odhadnutou křivku, která však není příliš hladká a její konfidenční pásmo je širší než při odhadu s větší šířkou pásma. (Fan a Zhang, 2008) doporučují volit šířku pásma minimalizující střední čtvercovou chybu MSE. Nejprve jsme ručně vyzkoušeli několik odhadů pro různé šířky pásma. Jako dobrá volba nám přišla šířka pásma 10 čtvrtletí, jejíž odhad jsme vykreslili na obrázku 5.1. Nejmenší střední čtvercová chyba MSE nám ale vyšla pro šířku pásma 3 čtvrtletí. Podívejme se na rozdíl mezi těmito volbami na obrázku 5.2. Šířka pásma 3 sice vykresluje datům přesněji odpovídající křivku, avšak za cenu znatelného rozšíření konfidenčního pásma. Jak volit šířku pásma se tedy při aplikaci odhadu modelu s proměnlivými koeficienty pomocí lokální regrese jeví jako hlavní otázka. Volba nejpřesnějšího odhadu při minimalizaci střední čtvercové chyby MSE se za cenu stability výsledného odhadu nemusí jevit jako zcela optimální. Při časové proměnlivosti v dlouhodobém horizontu jsou prudké změny chování regresorů spíše nežádoucí. Významné události (v případě tohoto příkladu globálního charakteru, jinak události s velkým dopadem na zkoumaný jev) mohou způsobit výraznou změnu, avšak obecně bychom čekali změnu pozvolnou. Proto jsme se nakonec rozhodli použít šířku pásma 10.

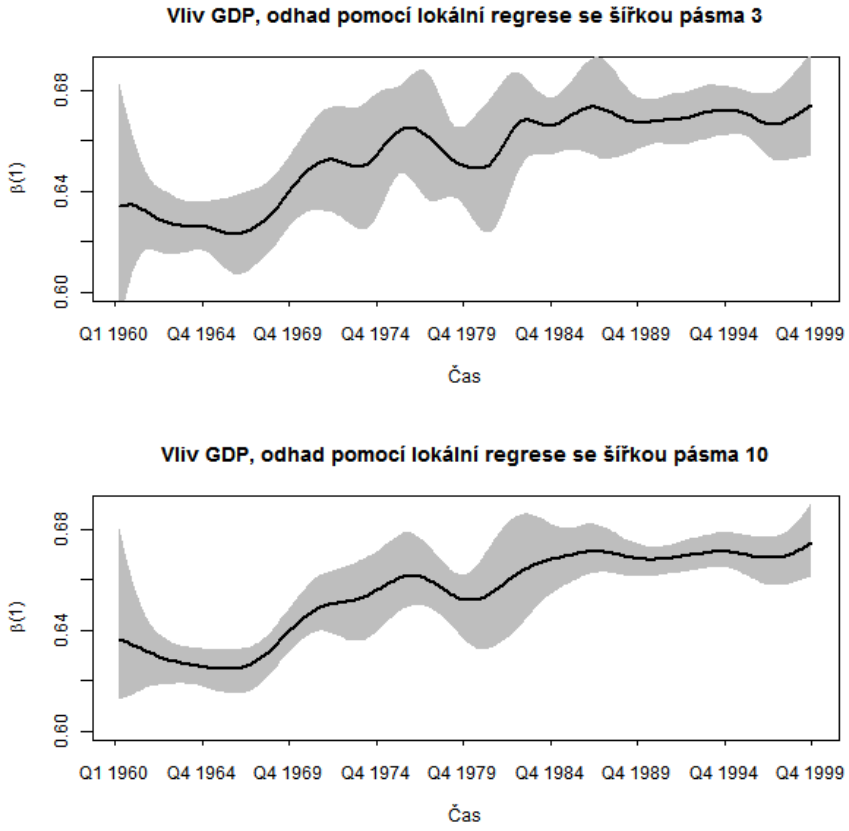
Pojďme se podívat na přesnost jednotlivých odhadů pro porovnání s modelem klasické lineární regrese. Jako ukazatel přesnosti si vezměme reziduální součet čtverců.



Obrázek 5.1: Vliv regresoru hrubého domácího produktu GDP na odezvu spotřeby domácností CONS a jeho 95% konfidenční pásmo dle různých typů odhadů modelu

Tabulka 5.5 nám ukazuje několik poznatků. Modely s proměnlivými koeficienty dosahují lepších výsledků než klasická lineární regrese i na takto jednoduchém příkladu a při příkladech složitějších očekáváme obdobný či ještě větší rozdíl. Vyhlazovací splajny díky volbě většího počtu kořenů dosahují lepšího podchycení dat než splajny polynomiální či penalizované. Lokální regrese má vůči datům nejlepší odhad, ale za cenu jeho velké nestability. Ta se pak projeví ve veškeré inferenci a může velice nepříjemně zkreslovat výsledky. Splajnové odhady jsou naopak velice přesné soudě dle jejich úzkých konfidenčních pásem. Každá z těchto metod tedy má své pro a proti a je na čtenáři, kterou považuje za nejlepší.

V tabulce 5.6 porovnáváme predikce jednotlivých metod. Z nich je v tomto



Obrázek 5.2: Vliv volby šířky pásma při odhadu pomocí lokální regrese na výsledný odhad proměnlivého koeficientu β_1 pro regresor hrubého domácího produktu GDP a jeho 95% konfidenční pásmo

ohledu zdaleka nejpřesnější odhad pomocí lokální regrese. Odhad pomocí vyhlazovacího splajnu byl podle reziduálního součtu čtverců o něco lepší než odhady pomocí polynomiálního a penalizovaného splajnu, ale při predikci dává horší výsledky. Lineární model dává nejhorší výsledky.

Při predikci v tomto případě si ale musíme uvědomit použitou metodologii. Výsledné splajny byly odhadnuty na časovém intervalu mezi roky 1960 a 2000. Všechny splajnové metody ke své konstrukci používají B-splajnovou bázi. Ta však neexistuje mimo své vnější uzly, což jsou právě roky 1960 a 2000. Mimo tyto meze jsou její hodnoty lineárně interpolovány pomocí krajních hodnot a derivací v těchto bodech. Stejně tak u odhadnuté křivky pomocí lokální regrese jsme

Metoda odhadu	**Reziduální součet čtverců**
Lineární regrese	167 475.4
Polynomiální splajn	134 947.5
Penalizovaný splajn	134 870.8
Vyhlazovací splajn	122 075.7
Lokální regrese, $h = 10$	83 579.5
Lokální regrese, $h = 3$	51 742.6

Tabulka 5.5: Porovnání přesnosti odhadu jednotlivých metod dle reziduálního součtu čtverců

pro odhad použili tutéž lineární interpolaci. Je proto třeba dávat si pozor na směrnici odhadnuté křivky proměnlivého koeficientu v krajních bodech, jelikož má na predikce mimo odhadnutý interval klíčový vliv.

Z hlediska použití modelu s proměnlivými koeficienty pro predikce se tedy jeví jako prozíravější volit takový modifikátor vlivu, u kterého máme k dispozici data pokrývající jeho definiční obor. Dobrým příkladem je třeba věk. Při využití v bankovním sektoru vytvoříme model s věkem klienta jako modifikátorem vlivu. Ten odhadneme na klientských datech, ve kterých se pravděpodobně nachází většina hodnot, jichž věk může nabývat. Při predikci pro nové klienty pak příslušné koeficienty budou spadat do odhadnutých křivek bez nutnosti jakékoliv interpolace.

Na závěr budeme testovat relevantnost regresoru GDP. Jelikož se jedná o křivku, tak jediný případ, kdy by tento koeficient nebyl významný je takový, kdy by byl konstantně nulový. Otestujme tedy hypotézu

$$\begin{aligned} H_0 &: \beta_1(t) = 0 \\ H_1 &: \beta_1(t) \neq 0. \end{aligned} \tag{5.2}$$

Odhad přes polynomiální splajny je součástí balíčku *mgcv* a má v sobě zakomponovaný i takovýto test, jehož testová statistika je definována jako

$$T_r = \hat{\boldsymbol{f}}_j^T \boldsymbol{V}_{f_j}^{r-} \hat{\boldsymbol{f}}_j,$$

kde

$\hat{\boldsymbol{f}}_j = (\hat{\beta}_j(t_1),...,\hat{\beta}_j(t_n))$,
$\boldsymbol{V}_{f_j} = \mathcal{B}_j \boldsymbol{V}_\gamma \mathcal{B}_j^T$,
$\mathcal{B}_j =$ matice vyhodnocené B-splajnové báze pro koeficient β_j,
$\boldsymbol{V}_\gamma =$ kovarianční matice koeficientů γ_j,
$\boldsymbol{V}_{f_j}^{r-}$ je pseudoinverze řádu r.

Testová statistika T_r pak má χ_r^2 rozdělení. Odvození tohoto testu a jeho testové statistiky je podrobněji popsáno v (Wood, 2013). P-hodnota testu vyšla pro odhad polynomiálním splajnem menší než 0,001 a nulovou hypotézu nevýznamnosti regresoru GDP proto zamítáme.

Pro odhad pomocí lokální regrese je takovýto test konstruován přes distribuci maximálního rozdílu mezi odhadnutou a skutečnou funkcí, kde nulovou hypotézu

Čas	**Data**	**Lin. model**	**Pol. splajn**	**Penal. splajn**	**Vyhl. splajn**	**Lokální regrese (10)**
Q1 2000	6171,1	6081,34	6131.87	6133.57	6099.24	6149.59
Q2 2000	6226,3	6160,82	6224.12	6226.26	6183.79	6244.46
Q3 2000	6292,1	6185,46	6251.61	6254.17	6203.87	6274.39
Q4 2000	6341,1	6217,59	6288.13	6290.99	6232.72	6313.27

Tabulka 5.6: Porovnání skutečných pozorování a predikce dle lineárního modelu a modelů s proměnlivými koeficienty odhadnutými pomocí polynomiálního, penalizovaného a vyhlazovacího splajnu a pomocí lokální regrese se šířkou pásma 10

z (5.2) zamítáme při testové statistice překračující kritickou hladinu. Testová statistika nám vyšla 14.571 a kritická hladina 3.663, takže nulovou hypotézu nevýznamnosti regresoru GDP zamítáme i při odhadu lokální regresí.

Takovéto testování pro odhad lokální regresí ale bylo navrženo obecně pro jakoukoliv konstantu. Můžeme proto zkoumat i to, jestli odhadnutý proměnlivý koeficient není konstantní. Prvně je třeba zvolit nějakou hodnotu dané konstanty, oproti které chceme koeficient testovat. Vezměme si tedy průměr odhadnutých koeficientů $\beta_1(t)$ přes všechny pozorované časy $t = 1,...,160$, který je roven hodnotě $c = 0{,}6546$. Budeme tedy testovat hypotézu

$$\begin{aligned} H_0 &: \beta_1(t) = 0{,}6546 \\ H_1 &: \beta_1(t) \neq 0{,}6546, \end{aligned} \tag{5.3}$$

kterou opět zamítáme při testové statistice překračující kritickou hladinu. Testová statistika nám vyšla 5.278 a kritická hladina 3.663, a tím pádem zamítáme nulovou hypotézu, že proměnlivý koeficient je konstantní s hodnotou 0,6546.

Závěr

Modely s proměnlivými koeficienty přicházejí s převratnou myšlenkou změny koeficientů v závislosti na nějakém modifikátoru vlivu. Převratný není ani tak samotný odhad křivek, se kterým jsme se mohli setkat v neparametrické regresi, avšak začlenění takového odhadu do struktury klasické lineární regrese. Snadná interpretace modelu lineární regrese je tím pádem zachována. Navíc dostáváme i dodatečnou informaci o změně koeficientů, nemluvě o flexibilnějším neparametrickém odhadu. V první kapitole jsme definovali základní tvar modelů s proměnlivými koeficienty se standardními daty s nezávislými pozorováními a k němu i několik speciálních případů. Taktéž jsme definovali model s proměnlivými koeficienty s daty longitudinálními, kde pro jednotlivé subjekty je k dispozici vícero pozorování.

V druhé kapitole jsme stručně popsali několik v literatuře zpracovaných metod odhadu modelů s proměnlivými koeficienty. Důležité je uvědomit si charakter tématu modelů s proměnlivými koeficienty. Podstata neparametrické regrese nám dává možnost konstruovat odhady téměř jakýmkoliv způsobem. Kvůli kletbě dimenzionality se omezujeme pouze na hladké křivky, ale i tak zde stále existuje nespočet možných postupů. Uvedli jsme proto pouze takové metody, které již byly dostatečně vypracovány v akademických publikacích. Mějme ale na pozoru, že tyto metody nejsou konzistentní. Každý z autorů se vydal svým vlastním směrem a jediné, co tyto metody spojuje, je proměnlivost odhadnutých koeficientů. Tuto kapitolu dělíme dle dvou kritérií – dle metody odhadu (splajny/lokální regrese) a dle použitých dat (standardní/longitudinální). (Hastie a Tibshirani, 1993) navrhli první metodu odhadu pomocí vyhlazovacích splajnů na standardních datech s nezávislými pozorováními a s vlastním modifikátorem vlivu pro každý z regresorů. (Marx, 2010) použili stejná data i vlastní modifikátory vlivu, ale namísto vyhlazovacího splajnu se rozhodli pro splajn penalizovaný. (Fan a Zhang, 2000) považují modely s proměnlivými koeficienty ve své podstatě za lokální, a proto odhad konstruují pomocí lokální regrese, avšak pouze s jediným modifikátorem vlivu stejným pro všechny regresory. (Huang a kol., 2004) pro svůj odhad použili komplikovanější longitudinální data namísto dat s nezávislými pozorováními. Kvůli charakteru těchto dat zde mají taktéž jediný modifikátor vlivu, a to čas daného pozorování. Jako odhadovanou křivku si zvolili polynomiální splajn, často označovaný jako regresní splajn. Každou z těchto metod je tedy nutno považovat za jeden z mnoha směrů, jimiž se modely s proměnlivými koeficienty mohou vydat.

V práci jsme se zabývali pouze základním tvarem modelu s proměnlivými koeficienty na standardních nebo longitudinálních datech. Téma modelů s proměnlivými koeficienty ale obsahuje mnohem více směrů. Základní tvar je možné zobecnit na množinu exponenciálních rozdělení. Můžeme předpokládat apriorní informaci u Bayessovského tvaru, použít tyto modely pro analýzu přežití, odhadovat spolu s křivkami koeficientů i jejich příslušné modifikátory vlivu nebo používat vícerozměrné modifikátory vlivu, a takto bychom mohli vyjmenovat ještě mnoho různých speciálních případů. Každý z nich ale vyžaduje vlastní metodu odhadu a lze ho tím pádem považovat za jeden z mnoha směrů tématu modelů s proměnlivými koeficienty. Očekávat od nějaké práce shrnutí celého tohoto té-

matu je v dnešní době vzhledem k obrovskému množství akademických publikací o modelech s proměnlivými koeficienty nereálné, ne-li přímo nemožné. Proto jsme se rozhodli zaměřit se pouze na základní tvar těchto modelů se skalárními modifikátory vlivu pro standardní a longitudinální data, a tyto poznatky přehledně shrnout.

Třetí kapitolu považujeme za stěžejní část této práce, která se zaměřila na statistickou inferenci v modelech s proměnlivými koeficienty. Popsána byla ale pouze inference pro odhady pomocí lokální regrese a polynomiálních splajnů. Dotyční autoři odhadů pomocí vyhlazovacích a penalizovaných splajnů se bohužel podrobnější inferencí pro své odhady nezabývali. Kapitolu jsme tedy rozdělili na dvě části dle použitých dat, jelikož lokální odhad pracuje se standardními daty s nezávislými pozorováními, zatímco odhad pomocí polynomiálních splajnů byl navržen pro data longitudinální. K oběma těmto metodám bylo odvozeno vychýlení a rozptyl a dokázána asymptotická normalita za platnosti určitých technických podmínek. Pomocí vychýlení a rozptylu je pak možno zkonstruovat konfidenční pásmo na nějaké hladině $(1-\alpha), \alpha \in (0,1)$. Bylo by možné zkonstruovat i konfidenční intervaly, avšak vzhledem ke křivkové povaze odhadovaných koeficientů zde použití konfidenčního pásma dává větší smysl. Taktéž jsme popsali testování hypotéz, které se převážně zabývá otázkou, zda-li je výsledný koeficient skutečně proměnlivý nebo zda-li je konstantní.

Ve čtvrté kapitole byla autory navržena procedura pro výběr proměnných. Ta funguje obdobně jako klasická "forward selekce" v lineární regresi. V každém kroku je do modelu přidán jeden regresor a procedura vyhodnotí, zda-li ho v modelu ponecháme nebo naopak vyloučíme. Oproti lineární regresi nám ale modely s proměnlivými koeficienty přidávají další dimenzi tohoto problému, a to proměnlivost daného koeficientu. Proto musíme vyhodnotit dvě otázky. Je koeficient proměnlivý? A pokud ne, není jeho konstantní odhad nulový? V každém kroku otestujeme tyto dvě hypotézy. Vzhledem k tomu, že se již jedná o problém vícenásobného testování, tak je na tyto testy aplikována patřičná korekce. Sledovaný regresor dle této procedury dosahuje tří stavů. Může se jednat o proměnlivý regresor, což znamená, že do modelu vstupuje s proměnlivým koeficientem, nebo se může jednat o konstantní regresor, čili regresor s konstantním koeficientem. Jako poslední možnost definujeme irelevantní regresor, který v modelu nepoužijeme. Vzhledem k vysoké výpočetní náročnosti modelů s proměnlivými koeficienty bude tato procedura ještě mnohem náročnější, a proto autoři diskutovali i několik algoritmických vylepšení a kompromisů pro omezení této náročnosti.

V kapitole Aplikace jsme zmínili několik aplikací modelů s proměnlivými koeficienty na praktické příklady. Podívali jsme se i na statistické softwary, které obsahují již připravené funkce pro odhad modelů s proměnlivými koeficienty. Ty lze zatím najít pouze v programu R (R Core Team, 2013), kde je k dispozici několik balíčků. Modely s proměnlivými koeficienty lze samozřejmě odhadnout téměř v jakémkoliv vhodném softwaru, ale uživatel si musí sám naprogramovat dané odhadové procedury. Takové procedury, například pro odhad zobecněných modelů s proměnlivými koeficienty, nejsou vůbec triviální. Pomocí simulací jsme prověřili empirické vlastnosti testů konstantnosti proměnlivého koeficientu pro odhady pomocí polynomiálních splajnů a lokální regrese. Test pro metodu odhadu přes polynomiální splajny nám přišel přišel přesnější a doporučujeme proto použití balíčku *mgcv* pro odhad modelů s proměnlivými koeficienty. Dále jsme

vypracovali relativně jednoduchý příklad pro demonstraci různých předností modelů s proměnlivými koeficienty, kde jsme porovnali všechny čtyři popsané metody odhadu těchto modelů mezi sebou a zároveň i oproti klasické lineární regresi.

Do budoucna očekáváme rozvoj těchto modelů nejen směrem do hloubky, tj. další poznatky pro již navržené metody, ale hlavně do šířky s novými metodami a použitími. Modely s proměnlivými koeficienty se v současné době těší velké přízni akademiků. Pokud tento trend bude pokračovat, můžeme se pak možná těšit na novou éru matematického modelování.

Modely s proměnlivými koeficienty tedy jednoznačně považujeme za velký krok kupředu v oblasti matematického modelování. Bezpochyby mají potenciál stát se nástupcem lineární regrese, avšak než se tak stane, je třeba nejprve vyřešit několik překážek. Odhad modelu pro data s více regresory či velkým množstvím pozorování je výpočetně náročný. Nemyslíme si, že odhadové procedury je možno nějak radikálněji optimalizovat. Problém neparametrického odhadu je prostě sama o sobě složitá záležitost, a proto lze nejspíš tuto překážku odstranit až s rozvojem výpočetní techniky. Druhým problémem zůstává nedostatečná informovanost o těchto modelech. Je třeba toto téma dostat mezi širší odbornou veřejnost, ať již začleněním modelů s proměnlivými koeficienty do výuky na matematických vysokých školách či přidáním funkcí pro jejich odhad do používaných statistických softwarů. Až se tyto problémy vyřeší, tak modely s proměnlivými koeficienty zajisté čeká dynamická budoucnost i v praktické aplikaci.

Seznam použité literatury

Cai, Z., Fan, J. a Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, **95**(451), 888–902.

Fan, J. a Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**(5), 1491–1518.

Fan, J. a Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27** (4), 715–731.

Fan, J. a Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, **1**, 179–195.

Fan, J., Zhang, C. M. a Zhang, J. (2001). Generalized likelihood ration statistic and wilks phenomenon. *The Annals of Statistics*, **29**(1), 153–193.

Fan, J., Yao, Q. a Cai, Z. (2003). Adaptive varying co-efficient linear models. *Journal of the Royal Statistical Society. Series B (statistical methodology)*, **65** (1), 57–80.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall, 5th edition.

Hastie, T. a Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, **55**(4), 757–796.

Huang, J. Z., Wu, C. O. a Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**(1), 111–128.

Huang, J. Z., Wu, C. O. a Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, **14**, 763–788.

Marx, B. D. (2010). P-spline varying coefficient models for complex data. *Statistical Modelling and Regression Structures*, pages 19–43.

Nelder, J. A. a Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, **135**(3), 370–384.

R Core Team (2013). R: A language and environment for statistical computing. URL `http://www.R-project.org/`. [cit. 27.12.2017].

Sperlich, S. a Theler, R. (2015). Modeling heterogeneity: a praise for varying-coefficient models in causal analysis. *Computational Statistics*, **30**(3), 693–718.

Wood, S. N. (2007). mgcv: Mixed gam computation vehicle with gcv/aic/reml smoothness estimation. r package version 1.8-22. URL `https://cran.r-project.org/web/packages/mgcv/mgcv.pdf`. [cit. 27.12.2017].

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, **100**(1), 221–228.