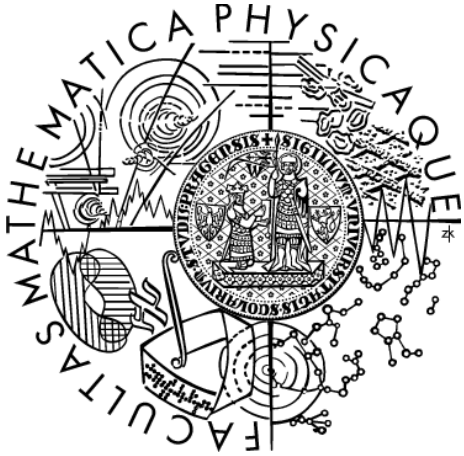
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Bronislava Hyklová

Odhady varianční funkce v neparametrických modelech

Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: *Prof. RNDr. Marie Hušková, DrSc.*
Studijní program: *Matematika*
Studijní obor: *Učitelství matematiky pro střední školy v kombinaci s odbornou matematikou*
Studijní plán: *Matematická statistika*

Děkuji Prof. RNDr. Marii Huškové, DrSc., za volbu podnětného tématu a za její cenné poznámky a připomínky k obsahu. Dále děkuji své rodině a příteli, jejichž podpora po dobu mého studia mi byla nedocenitelnou pomocí.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 5.12.2006 Bronislava Hyklová

Obsah

Název práce: Odhady varianční funkce v neparametrických modelech
Autor: Bronislava Hyklová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Prof. RNDr. Marie Hušková, DrSc.
e-mail vedoucího: Marie.Huskova@mff.cuni.cz

Abstrakt: Tato diplomová práce se zabývá odhadováním varianční funkce v neparametrickém regesním modelu. Je zaměřena především na lokálně polynomické odhady. Odvozena jsou přesná vyjádření podmíněného vychýlení a kovariance odhadů varianční funkce a uvedeny jsou také významné asymptotické aproximace těchto charakteristik. Dále je popsána metoda EBBS pro výběr šířky okénka a Detteho test homoskedasticity. Na závěr jsou prezentovány výsledky praktického zpracování dat z pražského Klementina.

Klíčová slova: homoskedasticita, lokálně polynomický odhad, neparametrická regrese, odhad varianční funkce, výběr šířky okénka

Title: Variance-function estimation in nonparametric models
Author: Bronislava Hyklová
Departement: Department of Probability and Mathematical Statistics
Supervisor: Prof. RNDr. Marie Hušková, DrSc.
Supervisor's e-mail address: Marie.Huskova@mff.cuni.cz

Abstract: The thesis studies variance function estimation in nonparametric regression model. It focuses on local polynomial estimators particularly. Exact expressions of conditional variance function estimator bias and covariance are derived and important asymptotical aproximations of these characteristics are also provided. Further the EBBS method for bandwidth selection and Dette's homoscedasticity test are described. Results of Prague Klementinum data processing are presented at the end of the thesis.

Keywords: Bandwidth selection, Homoscedasticity, Local polynomial estimator, Nonparametric regression, Variance function estimation

Úvod

Tato diplomová práce se zabývá odhadováním varianční funkce v neparametrickém regesním modelu. Je zaměřena především na lokálně polynomické odhady.

První kapitola nás uvádí do problému nelineární regrese. Je zde zmíněna souvislost s regresí lineární a také jsou zde popsány jádrové funkce, které se v nelineárních modelech využívají.

Druhá kapitola popisuje vlastní lokální polynomický odhad varianční funkce a jeho odlišnost od odhadů střední hodnoty stejnou metodou.

Ve třetí kapitole získáme přesná vyjádření podmíněného vychýlení a kovariance odhadů varianční funkce. Uvádíme také významné asymptotické aproximace, které je možné využít například pro výběr šířky okénka.

Čtvrtá kapitola je věnována výběru šířka okénka. Hlavní důraz je kladen na metodu EBBS (Empirical Bias Bandwidth Selection), která je také použita v praktickém zpracování dat z pražského Klementina.

Pátá kapitola se zabývá Detteho testem homoskedasticity. Je zde uvedena jeho základní idea a také testové statistiky. Bez důkazu jsou i uvedeny teoretické podklady pro tento test.

V poslední, šesté kapitole prezentujeme zpracování dat z pražského Klementina metodami z teoretické části diplomové práce.

V dodatku uvádíme některé důkazy a odvození vztahů, které se v práci objevují. V příloze je hlavní část programu použitého pro zpracování dat.

Kapitola 1

Od lineární regrese k nelineární

1.1 Lineární regresní model

Lineární regrese je jedna z nejklasičtějších a nejpoužívanějších statistických metod. Jedním z hlavních cílů je popsat vztah mezi nezávisle a závisle proměnnými. Tyto proměnné často označujeme jako plán experimentu a pozorování.

Data jsou považována za realizace modelu

$$\mathbf{Y} = \mathbf{X}\beta + \sigma\varepsilon \quad , \tag{1.1}$$

kde $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ je náhodný vektor, $\mathbf{X} = (X_{ij})$ je pevně zvolená matice typu $n \times k,\ k < n$, $\beta = (\beta_1, \ldots, \beta_k)$ je vektor neznámých parametrů a $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ je náhodný vektor chyb, který splňuje podmínky $\mathbb{E}(\varepsilon) = \mathbf{0}$, $\operatorname{var}\varepsilon = \mathbf{I},\ \sigma^2 > 0$ neznámý parametr.

Parametry $\beta_1, \ldots, \beta_k$ se odhadují *metodou nejmenších čtverců*, tj. z podmínky, že výraz

$$(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

má být minimální. Označíme tyto odhady $\mathbf{b} = (b_1, \ldots, b_k)^T$. Metodou nejmenších čtverců dostáváme

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Další poznatky a důkazy nalezneme v Anděl [1].

Hlavním účelem takové regresní analýzy je určit kvantitativně vliv plánu experimentu $\mathbf{X}$ na pozorování $\mathbf{Y}$, shrnout vztah mezi těmito dvěma veličinami, předpovědět střední hodnotu pozorování na daném $\mathbf{X}$ a extrapolovat výsledky mimo rozsah pozorovaných hodnot regresorů.

Uveďme jednoduchý příklad proložení dat $(X_i, Y_i),\ i = 1, \ldots, n$ přímkou. Nechť tedy platí závislost

$$Y_i = \beta_0 + \beta_1 X_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n. \tag{1.2}$$

Cílem je odhadnout parametry $\beta_0,\ \beta_1$ a regresní funkci $m(x)$:

$$m(x) = \mathbb{E}\,(\mathbf{Y}|X = x) = \beta_0 + \beta_1 x \quad .$$

Ne vždy lze ovšem data přímkou proložit. V takovém případě je přirozené zvážit stupeň polynomu, kterým aproximujeme regresní funkci. Tímto způsobem se dostáváme k polynomické regresi. Tento postup je široce užívaný, přesto skrývá několik nevýhod. Uveďme si některé z nich:

1. Polynomy mají spojité derivace na celém definičním oboru, což může zabránit uspokojivě vyrovnat data, např. data, v nichž se vyskytují skoky.

2. Prokládání polynomickou regresí je citlivé na výskyt odlehlých pozorování.

3. Není možné měnit stupeň polynomu v závislosti na průběhu dat.

Jedna z metod, která netrpí těmito nedostatky, se obecně nazývá lokální modelování. Základní myšlenka lokálního modelování se dá vyjádřit následovně. Zatímco v klasické regresní analýze zpřesňujeme odhad regresní funkce zvýšením počtu odhadovaných parametrů, v případě lokální regrese používáme k výpočtu hodnoty regresní funkce v bodě x pouze data z jeho

jistého okolí $(x-h, x+h)$, kde h je parametr, který se v češtině nazývá šířka okénka (z anglického bandwidth). Šířka okénka může být závislá na plánu experimentu $\mathbf{X}$, může být určena polohou bodu x_0, ve kterém sestrojujeme odhad regresní funkce, a může to být i pevně zvolená konstanta. Volba šířky okénka je jednou z nejdůležitějších voleb v lokálním modelování. O některých metodách se zmíníme v kapitole 4, další lze nalézt např. ve Fan a Gijbels [6].

1.2 Lokálně polynomická regrese

Myšlenku lokálního modelování jsme zmínili v minulé části 1.1. Speciálním případem tzv. lokálního modelování je lokálně polynomická regrese. Pro odhad regresní funkce v bodě x používáme pouze data z jistého okolí tohoto bodu a prokládáme je polynomem, jehož stupeň je předem zvolený.

Předpokládejme, že regresní funkci m můžeme lokálně aproximovat užitím Taylorova rozvoje

$$m(t) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(t-x)^j \equiv \sum_{j=0}^{p} \beta_j (t-x)^j \quad , \tag{1.3}$$

kde t je bod v okolí bodu x a funkce m má v tomto bodě derivaci řádu $(p+1)$. Data (X_i, Y_i), $i = 1, \ldots, n$ tímto polynomem lokálně proložíme a řešíme minimalizaci váženého součtu čtverců

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j(t)(X_i - t)^j \right)^2 \frac{1}{h} K\left(\frac{X_i - t}{h} \right) \quad , \tag{1.4}$$

kde $K(\cdot)$ je jádrová funkce a h je šířka okénka. Je-li $\widehat{\beta}_j(t)$, $j = 0, \ldots, p$ řešením uvedené metody váženého součtu čtverců (1.4), potom je z Taylorova rozvoje (1.3) vidět, že odhadem $m^{(\nu)}(t)$, $\nu = 0, \ldots, p$ ν-té derivace funkce m je

$$\widehat{m}^{(\nu)}(t) = \nu! \widehat{\beta}_\nu(t) \quad . \tag{1.5}$$

Vše závisí na vhodné volbě šířky okénka, na jádrové funkci a na stupni regresního polynomu.

Šířka okénka hraje v lokálním modelování jednu z nejdůležitějších rolí. Příliš malá hodnota má za následek zvětšení rozptylu odhadovaných parametrů, zatímco při zvětšování šířky okénka roste vychýlení odhadu regresní funkce. V krajních případech, pro $h \to 0$ dostáváme interpolaci pozorovaných dat a pro $h \to \infty$ dostáváme nejjednodušší model, a to lineární (1.2). Vychýlení odhadu i rozptyl parametrů můžeme kontrolovat dle střední čtvercové chyby

$$\mathrm{MSE}^{(\nu)}(x_0) = \mathbb{E}\left(\widehat{m}^{(\nu)}(x_0) - m^{(\nu)}(x_0)\right)^2.$$

Proto minimalizace MSE je často uváděna jako vhodná metoda volby okénka. Chceme-li získat optimální globální šířku okénka, minimalizujeme integrovanou střední čtvercovou chybu

$$\mathrm{MISE}^{(\nu)} = \int \mathrm{MSE}^{(\nu)}(t) w(t) dt,$$

kde $w(\cdot) \geq 0$ je váhová funkce (viz Šalom [18]).

Stupeň regresního polynomu je také důležitou volbou. Přestože vychýlení

$$\mathrm{BIAS}^{(\nu)}(\widehat{m}^{(\nu)}(x_0)) = \mathbb{E}\left[\widehat{m}^{(\nu)}(x_0)\right] - m^{(\nu)}(x_0)$$

je primárně ovlivněno šířkou okénka, řád polynomu má na něj také vliv. S rostoucím řádem vychýlení klesá, zvyšuje se však výpočetní složitost a také rozptyl odhadu. V praxi se nejčastěji volí řád polynomu $p = 1$, případně $p = 2$ nebo 3, tak, aby rozdíl $p - \nu$ byl spíše lichý než sudý.

Jádrová funkce, viz další část 1.3, se volí obvykle tak, aby byla na svém definičním oboru nezáporná. Doporučuje se použití Epanechnikova jádra (1.10), to jest

$$K(t) = \frac{3}{4}(1 - t^2)_+ ,$$

které minimalizuje asymptotickou střední čtvercovou chybu MSE odhadovaných parametrů. Požadujeme-li neomezenost nosiče, lze s úspěchem využít Gaussova jádra.

1.3 Jádrové funkce

Uveďme ještě příklady jádrových funkcí, které se k lokálním odhadům používají. Standardně se o jádrové funkci předpokládá, že platí

$$K(\cdot) \geq 0, \quad \int K(u)du = 1 \quad a \quad \int uK(u)du = 0\,.$$

Tyto předpoklady také splňuje každá pravděpodobnostní hustota, která je symetrická kolem nuly.

Poznámka Užívají se i jádra, pro které nemusí být splněna podmínka $K(\cdot) \geq 0$, nebo dokonce i další zmíněné podmínky. Také existují jádrové funkce, které jsou jednostranné.

Uveďme nejčastěji používaná jádra. Definujme $u_+ := \max(0; u)$.

1. **Trojúhelníkové jádro**

$$K(t) = (1 - |t|)_+ \quad t \in \mathbb{R} \tag{1.6}$$

2. **Gaussovo jádro**

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \quad t \in \mathbb{R} \tag{1.7}$$

3. **Symetrická beta rodina**

$$K(t) = \frac{1}{B(\frac{1}{2}; \gamma + 1)}(1 - t^2)_+^{\gamma} \quad t \in \mathbb{R}, \quad \gamma = 0, 1, \ldots \quad , \tag{1.8}$$

kde B je beta funkce definovaná vztahem

$$B(a; b) = \int_0^1 u^{a-1}(1-u)^{b-1}du \quad a > 0,\ b > 0.$$

Při volbě:

(a) $\gamma = 0$ dostáváme **rovnoměrné jádro** na intervalu $\langle -1; 1 \rangle$, tj.

$$K(t) = \frac{1}{2}\mathbb{I}\{|t| \leq 1\} \tag{1.9}$$

(b) $\gamma = 1$ dostáváme **Epanechnikovo jádro**

$$K(t) = \frac{3}{4}(1 - t^2)_+ \tag{1.10}$$

(c) $\gamma = 2$ dostáváme tzv. **biweight kernel**

$$K(t) = \frac{15}{16}\left[(1 - t^2)_+\right]^2 \tag{1.11}$$

(d) $\gamma = 3$ dostáváme tzv. **triweight kernel**

$$K(t) = \frac{35}{32}\left[(1 - t^2)_+\right]^3 \tag{1.12}$$

Další informace o jádrových funkcích lze nalézt ve Fan a Gijbels [6] nebo ve Wand a Jones [19].

1.4 Použité značení

V diplomové práci zachováme úmluvu, že násobení a dělení sloupcových vektorů se provádí po složkách. Abychom zabránili záměně mezi klasickým násobením čtvercových matic a násobením po prvcích, budeme druhou možnost značit $\mathbf{A} \odot \mathbf{B}$. Tento součin se někdy nazývá *Hadamardův součin* matic $\mathbf{A}$ a $\mathbf{B}$.

Dále pro zkrácení zápisu podmíněných středních hodnot budeme značit $\chi = (X_1, \ldots, X_n)$ a výraz $\mathrm{cov}\,(\mathbf{U}|\mathbf{W})$ bude znamenat podmíněnou varianční matici vektoru $\mathbf{U}$ při daném vektoru $\mathbf{W}$, kde $\mathbf{U}$ a $\mathbf{W}$ jsou náhodné vektory.

Značení $X \xrightarrow{d} \mathcal{L}$ znamená, že asymptotické (limitní) rozdělení náhodné veličiny X je $\mathcal{L}$.

Kapitola 2

Formulace lokálně polynomického odhadu varianční funkce

2.1 Obecný tvar odhadu varianční funkce

Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ jsou dvojice nezávislých veličin. Předpokládejme, že vyhovují heteroskedastickému neparametrickému regresnímu modelu

$$Y_i = m(X_i) + v(X_i)\varepsilon_i \quad i = 1, \ldots, n \ , \tag{2.1}$$

kde $\varepsilon_1, \ldots, \varepsilon_n$ jsou nezávislé náhodné veličiny s $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = 1$ a $\mathbb{E}(\varepsilon_i^4) < \infty$. Funkci m nazýváme *regresní funkcí* a funkci v (někdy v^2) nazýváme *varianční funkcí*. Můžeme také jako $\mathbf{m}$ a $\mathbf{v}^2$ označit sloupcové vektory, které obsahují hodnoty $m(X_i)$ a $v^2(X_i)$, $1 \leq i \leq n$. A nakonec také $\mathbf{Y}$ může být vektor $n \times 1$ obsahující hodnoty Y_i.

Předpokládejme, že $\widehat{\mathbf{m}} = [\widehat{m}(X_1), \ldots, \widehat{m}(X_n)]'$ je odhad regresní funkce lineárního modelu, kterým prokládáme data (X_i, Y_i). Můžeme tedy vyjádřit $\widehat{\mathbf{m}} = \mathbf{SY}$, kde $\mathbf{S}$ je matice typu $n \times n$, kterou označujeme jako vyhlazovací matice (z anglického *smoother matrix*). Matice $\mathbf{S}$ závisí pouze na $\mathbf{X}$, nikoliv

na $\mathbf{Y}$. Předpokládejme, že $\mathbf{S}$ zachovává vektor konstant v tom smyslu, že $\mathbf{S1} = \mathbf{1}$, kde $\mathbf{1}$ značí vektor jedniček.

Nechť $\mathbf{S}_1$ je vyhlazovací matice odpovídající počátečním vyhlazeným datům a položme $\mathbf{r} = \mathbf{Y} - \widehat{\mathbf{m}} = (\mathbf{I} - \mathbf{S}_1)\mathbf{Y}$ vektor reziduí. Víme, že $\mathbf{v}\varepsilon = \mathbf{Y} - \mathbf{m}$ (násobení a dělení vektorů se provádí po prvcích). Vidíme tedy, že $\mathbf{r}$ můžeme chápat jako odhad náhodného vektoru chyb $\mathbf{v}\varepsilon$, který jako takový není možné měřit. Protože $\mathbb{E}\,\varepsilon_i^2 = \operatorname{var}\varepsilon_i = 1$, je $\mathbb{E}\,(v(X_i)\varepsilon_i)^2 = v^2(X_i)$. Uvažujme tedy model

$$(v(X_i)r_i)^2 = v^2(X_i) + v^*(X_i)\varepsilon_i^* \quad i = 1, \ldots, n\ ,$$

kde ε^* je podobně jako v (2.1) vektor náhodných chyb s nulovou střední hodnotou a jednotkovým rozptylem.

Je tedy zřejmé, že přirozeným způsobem odhadu varianční funkce $\mathbf{v}^2 = [v^2(X_1), \ldots, v^2(X_n)]'$ je vyhlazení kvadratických reziduí $\mathbf{r}^2$, čímž získáme $\widehat{\mathbf{v}}^2 = \mathbf{S}_2\mathbf{r}^2$. Zde je $\mathbf{S}_2$ další vyhlazovací matice (z metody nejmenších vážených čtverců pro lokálně polynomický odhad varianční funkce) a $\mathbf{r}^2$ je druhá mocnina hodnot $\mathbf{r}$. Přirozeným požadavkem na náš odhad je jeho nestrannost v případě homoskedasticity chyb, tedy když $\mathbf{v}^2 = \sigma^2\mathbf{1},\ \sigma^2 > 0$, a nezávislost na vychýlení $\widehat{\mathbf{m}}$ způsobené počáteční maticí $\mathbf{S}_1$. V případě homoskedasticity platí

$$\mathbb{E}\,(\mathbf{S}_2\mathbf{r}^2|X_1, \ldots, X_n) = \mathbf{S}_2[\{\mathbb{E}\,(\mathbf{S}_1\mathbf{Y}|X_1, \ldots, X_n) - \mathbf{m}\}^2 + \sigma^2(1 + \Delta)]\ ,$$

kde $\Delta = \operatorname{diag}(\mathbf{S}_1\mathbf{S}_1' - 2\mathbf{S}_1)$ a $\operatorname{diag}(\mathbf{A})$ značí sloupcový vektor obsahující diagonální hodnoty čtvercové matice $\mathbf{A}$. Když je $\widehat{\mathbf{m}} = \mathbf{S}_1\mathbf{Y}$ podmíněně nestranné, platí

$$\mathbb{E}\,(\mathbf{S}_2\mathbf{r}^2|X_1, \ldots, X_n) = \sigma^2(\mathbf{1} + \mathbf{S}_2\Delta)\,.$$

A vhodný odhad varianční funkce je tedy

$$\widehat{\mathbf{v}}^2 = \frac{\mathbf{S}_2\mathbf{r}^2}{\mathbf{1} + \mathbf{S}_2\Delta}. \tag{2.2}$$

Jinou možností, jak opravit odhad $\widehat{\mathbf{v}}^2$ na vliv odhadu $\widehat{\mathbf{m}}$, je vydělení každého rezidua r_i opravným „studentizujícím" faktorem $\sqrt{1+\Delta_i}$, tj. využití vztahu

$$\widehat{\mathbf{v}}^2 = \mathbf{S}_2 \frac{\mathbf{r}^2}{\mathbf{1}+\Delta}. \tag{2.3}$$

Výhodou vztahu (2.3) je menší výpočetní náročnost. Na druhé straně však při zpracování dat metodou EBBS (viz část 4.1), kterou budeme používat, je většina času potřebná pro nalezení lokální šířky okénka, tudíž mírné zvýšení náročnosti samotného vyhlazení se téměř neprojeví, jak v simulaci ukázali Ruppert a kol [16].

2.2 Vztah k parametrickému modelování

Máme-li lineární model $Y_i = (\mathbf{X}\beta)_i + v(X_i)\varepsilon_i,\ \mathrm{var}\,(\varepsilon_i) = 1,\ i = 1,\ldots,n$, kde $\mathbf{X}$ je regulární matice typu $n \times p$ a β je matice koeficientů typu $p \times 1$, potom můžeme v odhadu varianční funkce nahradit vyhlazovací matici $\mathbf{S}_1$ maticí $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, která je symetrická a idempotentní — získáme tak

$$\widehat{m} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

Použitím této matice získáme ze vztahu (2.2)

$$\widehat{\mathbf{v}}^2 = \frac{\mathbf{S}_2\{(\mathbf{I}-\mathbf{R})\mathbf{Y}\}^2}{\mathbf{1}-\mathbf{S}_2\{\mathrm{diag}(\mathbf{R})\}}\,,$$

kde $\mathbf{S}_2$ je vyhlazovací matice definovaná v předchozí části 2.1.

Pro homoskedastický lineární regresní model se odhad zjednoduší na známý

$$\widehat{\sigma}^2 = \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{R})\mathbf{Y}}{n-p}\,.$$

2.3 Lokálně polynomický odhad

Abychom odhadli $m(x)$ v pevném bodě x, modelujeme data pomocí polynomu stupně p metodou vážených nejmenších čtverců, kde váhy přiřazené

(X_i, Y_i) klesají k nule s rostoucí vzdáleností X_i od x. Stejně tak, jak je obvyklé v literauře, i my použijeme váhy $\frac{1}{h}K\{(X_i - x)/h\}$, kde K je jádrová funkce a h je šířka okénka. Potom $\widehat{m}(x)$ je předpovídaná hodnota tohoto modelu v bodě x. Pro zjednodušení zápisu a výpočtů budeme pracovat s odchylkou od x, tedy s $(X_i - x)$. Pak $\widehat{m}(x)$ je absolutním členem tohoto modelu. Tedy s využitím maticového řešení metody vážených nejmenších čtverců (1.4) můžeme vyjádřit (i, j)-tý prvek vyhlazovací matice $\mathbf{S}_{p,h}$ jako

$$(\mathbf{S}_{p,h})_{ij} = \mathbf{e}_1'\{\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\}^{-1}\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{e}_j \ , \tag{2.4}$$

kde $\mathbf{e}_i$ je sloupcový vektor s jedničkou na i-té pozici a nulami jinde,

$$\mathbf{X}_p(x) = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix}$$

a

$$\mathbf{W}_h(x) = \mathrm{diag}_{1 \le i \le n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right) ,$$

kde $\mathrm{diag}_{\,1 \le i \le n} a_i$ značí diagonální matici typu $n \times n$ s prvky a_i na diagonále. Vynásobením vektorem $\mathbf{e}_1'$ zleva vybíráme z celého odhadu pouze hledaný průsečík. Odvození vztahu (2.4) uvedeme v dodatku A.1 Odhad střední hodnoty v bodě X_i pak bude dán vztahem

$$\widehat{m}(X_i) = \sum_{j=1}^{n} (\mathbf{S}_{p,h})_{ij}\, Y_j \, . \tag{2.5}$$

Odhad v libovolném bodě x dostaneme snadno tak, že ve vztahu (2.5) nahradíme bod X_i bodem x, tj

$$\widehat{m}(x) = \mathbf{e}_1'\{\mathbf{X}_p(x)'\mathbf{W}_h(x)\mathbf{X}_p(x)\}^{-1}\mathbf{X}_p(x)'\mathbf{W}_h(x)\,\mathbf{Y} \, . \tag{2.6}$$

Pro nejčastěji používané stupně $p = 1$ a $p = 2$ má (i, j)-tý prvek vyhlazovací matice tvar

$$(\mathbf{S}_{1,h})_{ij} = \frac{1}{\det_{(1,h)}} \times$$
$$\times \left(K_{j,i} \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^2 - K_{j,i}(X_j - X_i) \sum_{k=1}^{n} K_{k,i}(X_k - X_i) \right),$$

$$(\mathbf{S}_{2,h})_{ij} = \frac{1}{\det_{(2,h)}} \times$$
$$\times \left\{ K_{j,i} \left[\sum_{k=1}^{n} K_{k,i}(X_k - X_i)^2 \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^4 - \right.\right.$$
$$\left. - \left(\sum_{k=1}^{n} K_{k,i}(X_k - X_i)^3 \right)^2 \right] +$$
$$K_{j,i}(X_j - X_i) \left[\sum_{k=1}^{n} K_{k,i}(X_k - X_i)^2 \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^3 - \right.$$
$$\left. - \sum_{k=1}^{n} K_{k,i}(X_k - X_i) \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^4 \right] +$$
$$K_{j,i}(X_j - X_i)^2 \left[\sum_{k=1}^{n} K_{k,i}(X_k - X_i) \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^3 - \right.$$
$$\left.\left. - \left(\sum_{k=1}^{n} K_{k,i}(X_k - X_i)^2 \right)^2 \right] \right\},$$

kde

$$\det_{(1,h)} = \sum_{k=1}^{n} K_{k,i} \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^2 - \left(\sum_{k=1}^{n} K_{k,i}(X_k - X_i) \right)^2,$$

$$\det_{(2,h)} = \sum_{k=1}^{n} K_{k,i} \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^2 \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^4 +$$
$$+ 2 \sum_{k=1}^{n} K_{k,i}(X_k - X_i) \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^2 \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^3 -$$
$$- \left(\sum_{k=1}^{n} K_{k,i}(X_k - X_i)^2 \right)^3 - \sum_{k=1}^{n} K_{k,i} \left(\sum_{k=1}^{n} K_{k,i}(X_k - X_i)^3 \right)^2 -$$
$$- \left(\sum_{k=1}^{n} K_{k,i}(X_k - X_i) \right)^2 \sum_{k=1}^{n} K_{k,i}(X_k - X_i)^4$$

a $K_{j,i} = \frac{1}{h} K\left(\frac{X_j - X_i}{h}\right)$.

Použitím vztahů (2.2) a (2.4) definujeme lokálně polynomický odhad varianční funkce jako

$$\begin{aligned}\widehat{v}^2(x) &= \widehat{v}^2(x; p_1, p_2, h_1, h_2) = \\ &= \frac{\mathbf{e}_1'\{\mathbf{X}_{p_2}(x)'\mathbf{W}_{h_2}(x)\mathbf{X}_{p_2}(x)\}^{-1}\mathbf{X}_{p_2}(x)'\mathbf{W}_{h_2}(x)\mathbf{r}^2}{1 + \mathbf{e}_1'\{\mathbf{X}_{p_2}(x)'\mathbf{W}_{h_2}(x)\mathbf{X}_{p_2}(x)\}^{-1}\mathbf{X}_{p_2}(x)'\mathbf{W}_{h_2}(x)\Delta}\ ,\end{aligned}$$

kde $\mathbf{r} = (\mathbf{I} - \mathbf{S}_{p_1,h_1})\mathbf{Y}$ a $\Delta = \mathrm{diag}(\mathbf{S}_{p_1,h_1}\mathbf{S}'_{p_1,h_1} - 2\mathbf{S}_{p_1,h_1})$. Zde uvažujeme dvojí vyhlazení — nejprve odhadujeme střední hodnotu $m(x)$ lokálními polynomy stupně p_1 s šířkou okénka h_1 a poté odhadujeme varianční funkci $v^2(x)$ z kvadratických reziduí $r_i^2 = (Y_i - \widehat{m}(X_i))^2$ lokálními polynomy stupně p_2 s šířkou okénka h_2.

Definici tohoto odhadu si zapamatujeme lépe, když v definici (2.2) zaměníme $\mathbf{S}_1 = \mathbf{S}_{p_1,h_1}$ a $\mathbf{S}_2 = \mathbf{S}_{p_2,h_2}$.

Pro nejčastěji používaný lokálně lineární odhad varianční funkce dostáváme

$$\begin{aligned}&\widehat{v}^2(x) = \widehat{v}^2(x; p_1, 1, h_1, h_2) = \\ &= \frac{\displaystyle\sum_{j=1}^{n}\left(K_j \sum_{k=1}^{n} K_k(X_k - x)^2 - K_j(X_j - x)\sum_{k=1}^{n} K_k(X_k - x)\right) r_j^2}{\mathrm{det}_{(1,h)} + \displaystyle\sum_{j=1}^{n}\left(K_{j,i}\sum_{k=1}^{n} K_k(X_k - x)^2 - K_j(X_j - x)\sum_{k=1}^{n} K_k(X_k - x)\right)\Delta_j}\ ,\end{aligned}$$

kde $K_j = \frac{1}{h} K\left(\frac{X_j - x}{h}\right)$.

Kapitola 3

Teorie lokálně polynomického odhadu varianční funkce

V této kapitole nejprve získáme přesné algebraické maticové vyjádření podmíněné střední hodnoty a kovariance odhadu $\widehat{v}^2$ v obecné třídě odhadů varianční funkce. V lokálním modelování nám tyto výsledky dají významné asymptotické aproximace, které dále můžeme využít pro výběr šířky okénka nebo odhadnutí rozptylu odhadu.

Narozdíl od předchozích (a následujících) kapitol, v této kapitole předpokládáme poněkud obecnější variantu modelu (2.1) - X_i zde nemusí být pouze pevně zvolené hodnoty, ale mohou to být náhodné veličiny. Proto všechny charakteristiky odhadů, které v této kapitole odvozujeme, jsou podmíněny právě náhodnými veličinami X_i. V případě pevných, nenáhodných X_i se z podmíněných charakteristik samozřejmě stávají charakteristiky nepodmíněné.

Připomeňme, že násobení a dělení sloupcových vektorů se provádí po složkách a násobení matic po prvcích značíme $\mathbf{A} \odot \mathbf{B}$. Dále $\chi = (X_1, \ldots, X_n)$ a $\operatorname{cov}(\mathbf{U}|\mathbf{W})$ je podmíněná varianční matice vektoru $\mathbf{U}$ při daném vektoru $\mathbf{W}$, kde $\mathbf{U}$ a $\mathbf{W}$ jsou náhodné vektory.

3.1 Obecný odhad

Následující matice používáme pro zjednodušení reprezentace vychýlení a kovariance $\widehat{\mathbf{v}}^2$:

$$\mathbf{V}^2 = \mathrm{diag}(\mathbf{v}^2), \quad \mathbf{G} = \mathrm{diag}_{1\le i\le n}\{\mathbb{E}\,(\varepsilon_i^3)\}, \quad \mathbf{T} = \mathrm{diag}_{1\le i\le n}\{\mathbb{E}\,(\varepsilon_i^4)\}$$

VĚTA 3.1 *Nechť* $\mathbf{b}_1 = (\mathbf{S}_1 - \mathbf{I})\mathbf{m}$ *značí vektor vychýlení vyhlazení* $\mathbf{S}_1$. *Potom*

$$\mathbb{E}\,(\widehat{\mathbf{v}}^2 - \mathbf{v}^2|\chi) = \frac{(\mathbf{S}_1 - \mathbf{I})\mathbf{v}^2 + \mathbf{S}_2\{\mathbf{b}_1^2(\mathbf{S}_1\mathbf{V}^2\mathbf{S}_1' - 2\mathbf{S}_1\mathbf{V}^2\} - (\mathbf{S}_1\Delta)\mathbf{v}^2}{\mathbf{1} + \mathbf{S}_1\Delta} \tag{3.1}$$

a

$$\begin{aligned}\mathrm{cov}\,(\widehat{\mathbf{v}}^2|\chi) = \mathbf{S}_2[\{(\mathbf{S}_1 - \mathbf{I}) \odot (\mathbf{S}_1 - \mathbf{I})\}(\mathbf{T} - 3\mathbf{V}^4) \times \{(\mathbf{S}_1 - \mathbf{I}) \odot (\mathbf{S}_1 - \mathbf{I})\}' \\ + 2(\mathrm{diag}\,\mathbf{b}_1)(\mathbf{S}_1 - \mathbf{I}) \times \mathbf{G}\{(\mathbf{S}_1 - \mathbf{I}) \odot (\mathbf{S}_1 - \mathbf{I})\}' \\ + 2\{(\mathbf{S}_1 - \mathbf{I}) \odot (\mathbf{S}_1 - \mathbf{I})\} \times \mathbf{G}(\mathbf{S}_1 - \mathbf{I})'(\mathrm{diag}\,\mathbf{b}_1) \\ + 2\{(\mathbf{S}_1 - \mathbf{I})\mathbf{V}^2(\mathbf{S}_1 - \mathbf{I})'\} \odot \{(\mathbf{S}_1 - \mathbf{I})\mathbf{V}^2(\mathbf{S}_1 - \mathbf{I})'\} \\ + 4\{(\mathbf{S}_1 - \mathbf{I})\mathbf{V}^2(\mathbf{S}_1 - \mathbf{I})'\} \odot (\mathbf{b}_1\mathbf{b}_1')]\mathbf{S}_2'/\{(\mathbf{1} - \mathbf{S}_2\Delta)(\mathbf{1} - \mathbf{S}_2\Delta)'\}.\end{aligned} \tag{3.2}$$

Důkaz Je uveden v dodatku A.2.

□

DŮSLEDEK 3.2 *Jestliže mají* ε_i *normální rozdělení, potom*

$$\mathrm{cov}\,(\widehat{\mathbf{v}}^2|\chi) = \frac{2\mathbf{S}_2[\{(\mathbf{S}_1 - \mathbf{I})\mathbf{V}^2(\mathbf{S}_1 - \mathbf{I})'\} \odot \{(\mathbf{S}_1 - \mathbf{I})\mathbf{V}^2(\mathbf{S}_1 - \mathbf{I})' + 2\mathbf{b}_1\mathbf{b}_1'\}]\mathbf{S}_2'}{(\mathbf{1} - \mathbf{S}_2\Delta)(\mathbf{1} - \mathbf{S}_2\Delta)'}$$

Poznámka Podmíněná průměrná střední čtvercová chyba (MASE - *Mean Average Squared Error*) odhadu $\widehat{\mathbf{v}}^2$ je definována jako

$$\mathrm{MASE}(\widehat{\mathbf{v}}^2) = n^{-1}\mathbb{E}\left[\sum_{i=1}^{n}\{\widehat{v}^2(X_i) - v^2(X_i)\}^2|\chi\right].$$

Při značení $\mathrm{MASE}(\widehat{\mathbf{v}}^2) = n^{-1}\{||\mathbb{E}\,(\widehat{\mathbf{v}}^2|\chi) - \mathbf{v}^2||^2 + \mathrm{Tr}\,\mathrm{cov}\,(\widehat{\mathbf{v}}^2|\chi)\}$, kde $||\mathbf{x}||^2 = \mathbf{x}'\mathbf{x}$, můžeme předchozí výsledek využít pro nalezení exaktní formule pro $\mathrm{MASE}(\widehat{\mathbf{v}}^2)$ *libovolného* páru vyhlazovacích matic $\mathbf{S}_1$ a $\mathbf{S}_2$.

3.2 Asymptotické chování lokálně polynomického odhadu

Jak jsme poznamenali v úvodu této kapitoly, všechny výsledky podkapitoly (3.1) jsou podmíněny pomocí X_i, a tedy nezávisí na jejich rozdělení. Nepotřebovali jsme tedy žádné předpoklady o veličinách X_i. Nicméně pro získání asymptotických výsledků budeme již nějaké předpoklady o chování X_i pro $n \to \infty$ potřebovat. Nejjednodušším předpokladem, a jediným, který budeme používat v této podkapitole, je nezávislost a stejné rozdělení (iid) veličin X_i. Nechť f značí sdruženou hustotu veličin $X_1, \ldots, X_n$ a pro funkci η nechť platí $\eta(X_i) = \mathrm{var}\,(v^2(X_i)\varepsilon_i^2),\ i = 1, \ldots, n$. Potom definujme funkci

$$K_{(p)}(u) = \frac{|\mathbf{M}_p(u)|}{|\mathbf{N}_p|} K(u)\,,$$

kde $\mathbf{N}_p$ je matice typu $(p+1) \times (p+1)$, která má na pozici (i,j) hodnotu $\int u^{i+j-2} K(u)du$, a $\mathbf{M}_p(u)$ je stejná matice jako $\mathbf{N}_p$, kde první sloupec je nahrazen $(1, u, \ldots, u^p)$. $K_{(p)}$ je jádro p-tého řádu.

VĚTA 3.3 *Předpokládejme, že x je vnitřní bod nosiče f, m má spojité derivace až do řádu $p_1 + 2$ včetně, v^2 má spojité derivace do řádu $p_2 + 2$ včetně a f a η jsou diferencovatelné v okolí bodu x. Nechť dále $h_1, h_2 \to 0, \quad nh_1, nh_2 \to \infty$ a*

$$\{h_1^{2(p_1+1)} + (nh_1)^{-1}\} = o(h_2^{p_2+1}) \tag{3.3}$$

pro $n \to \infty$. Potom pro p_2 liché platí

$$\mathbb{E}\,\{\widehat{v}^2(x) - v^2(x)|\chi\} = \left\{\int u^{p_2+1} K_{(p_2)}(u)du\right\} \times \left\{\frac{(v^2)^{(p_2+1)}(x)}{(p_2+1)!}\right\} h_2^{p_2+1} + o_P(h_2^{p_2+1}).$$

a pro p_2 sudé

$$\mathbb{E}\left\{\widehat{v}^2(x) - v^2(x)|\chi\right\} = \left\{\int u^{p_2+2} K_{(p_2)}(u)du\right\} \times \left\{\frac{(v^2)^{(p_2+1)}(x) f'(x)}{f(x)(p_2+1)!} + \frac{(v^2)^{(p_2+2)}(x)}{(p_2+2)!}\right\} h_2^{p_2+2} + o_P(h_2^{p_2+2}).$$

V obou případech potom platí

$$\text{var}\left\{\widehat{v}^2(x)|\chi\right\} = \left\{\int K_{(p_2)}(u)^2 du\right\} \times \left\{\frac{n^{-1}h_2^{-1}\eta(x)}{f(x)}\right\} + o_P\{(nh_2)^{-1}\}.$$

Důkaz Je uveden v dodatku A.3.

□

Poznámka První výraz závisí pouze na šířce okénka h_2. To ukazuje, že počáteční šířka h_1 má na asymptotické chování $\widehat{v}^2(x)$ pouze druhotný vliv. Jestliže $p_1 = p_2$ a jestliže h_1 a h_2 jsou vybrány optimálně pro odhad funkcí m a v^2, potom $h_1^{2(p_1+1)}$ a $(nh_1)^{-1}$ budou mít pro $n \to \infty$ stejný řád a oba budou $o_P(h_2^{p_2+1})$, čímž budou splněny podmínky (3.3).

Poznámka Srovnání s Větou 4.1 v Ruppert a Wand [15] ukazuje, že předcházející výrazy pro vychýlení a rozptyl našeho lokálně polynomického odhadu varianční funkce jsou analogické vztahům pro lokálně polynomický odhad regresní funkce. Jediný rozdíl je, že asymptotické vychýlení závisí na derivaci v^2 místo m a asymptotický rozptyl $\widehat{v}^2(x)$ je úměrný rozptylu druhých mocnin ε_i místo Y_i.

Zde je důležitý důsledek: Asymptoticky se $\widehat{v}^2$ chová jako lokálně polynomické vyhlazení ε_i^2, tzn. v může být odhadnuta tak, jako kdyby m byla známá, takže neztrácíme asymptotickou efektivitu kvůli odhadu m. Důsledek má také důležitý význam pro výběr šířky okénka, protože ospravedlňuje použití standardních selektorů (které byly vyvořeny pro odhad regresní funkce) i pro vyhlazování kvadratických reziduí při odhadu varianční funkce. Nepotřebujeme tedy nové selektory šířky okénka pro varianční funkci, jak uvidíme v následující kapitole.

Kapitola 4

Výběr šířky okénka

Důležitým praktickým problémem je výběr šířky okénka. Můžeme použít buď tzv. lokální okénka, kde h_1 a h_2 jsou funkcí x, nebo globální okénka, která na x nezávisí. V našem případě předpokládáme, že okénka jsou lokální. V ideálním případě bychom obě okénka h_1 i h_2 získali minimalizací střední čtvercové chyby (MSE) veličiny $\widehat{v}^2$ v bodě x. Tento postup je ale v praxi složitý, protože vliv h_1 na MSE$(\widehat{v}^2)$ je druhořadý, a proto je těžké ho odhadnout.

Ruppert a kol.[16] s použitím věty 3.3 navrhli alternativní strategii, která vede k asymptoticky optimálnímu okénku. Nejprve použijeme výběrový algoritmus (selektor) pro lokální okénka k nalezení asymptoticky optimálního h_1 pro odhad $m(x)$. Můžeme například použít selektor podle Fan a Gijbels [7], my však budeme používat *metodu EBBS* (Empirical Bias Bandwidth Selection), neboli výběr šířky okénka na základě empirického vychýlení, o které se zmiňuje Ruppert [14]. Poté zpracujeme druhé mocniny reziduí, jako kdyby to byly přímo druhé mocniny ε, přičemž použijeme stejný selektor okénka jako pro odhad regresní funkce. Jestliže použijeme $p_1 \geq p_2$, potom bude splněna podmínka (3.3).

Obecnou radou při odhadu m a v^2, nikoliv jejich vyšších derivací, je použití $p_1 = 2$ (případně 3) a $p_2 = 1$. Ve většině případů nemá v^2 příliš

velkou křivost, takže postačuje lineární odhad. Na druhé straně bývá křivost m již dostatečně velká na to, aby vychýlení jejího lineárního odhadu výrazně ovlivnilo čtverce reziduí, čímž se zvyšuje vychýlení v^2. Toto vychýlení může dokonce ovládnout $\widehat{v}^2$ — pokud je v^2 malé, potom $\widehat{v}^2$ bude odhad spíše druhé mocniny vychýlení v oblastech, kde m má velkou křivost. V takových případech velmi pomáhá použití lokálního kvadratického modelu pro m.

4.1 Metoda EBBS

Jak už bylo zmíněno, teorie v minulé kapitole naznačovala, že kterákoliv šířka okénka vhodná pro odhad m nebo její derivaci je vhodná i pro vyhlazení čtvercových reziduí k odhadu v^2 a její derivace. V našem případě bude použita metoda EBBS. Tato metoda má následující výhody:

1. Okénko je lokální, tzn. může záviset na x. EBBS minimalizuje odhad MSE v x. Odhad MSE bere v úvahu i hraniční případy.

2. Je možné zahrnout i odhad derivací.

3. Může být použita jak v případě sudých, tak lichých stupňů polynomů, narozdíl od klasických selektorů, které jsou omezeny pouze na liché stupně.

4. Upřednostňuje přesná vyjádření před asymptotickými aproximacemi. Není zde žádný předpoklad, že X_i mají pravděpodobnostní hustotu — exaktní výrazy používají pouze jejich vzorek.

Předpokládejme, že si pro nějaké $\nu \geq 0$ přejeme odhadnout $m^{(\nu)}(x)$ pro všechna x v nějaké mřížce lokálními polynomy stupně p. Nechť $\mathrm{MSE}\,(h; x)$ je MSE pro $m^{(\nu)}(x)$ používající okénko h. Střední čtvercová chyba, vychýlení i rozptyl jsou podmíněny $X_1, \ldots, X_n$. K odhadu $\mathrm{MSE}\,(h; x)$ odhadujeme rozptyl a vychýlení samostatně, značíme je jako $\mathrm{var}\,(h; x)$ a $\mathrm{bias}\,(h; x)$. Pro

var $(h; x)$ existuje přesné vyjádření (Ruppert a Wand [15]), ve které nahradíme v^2 jejím odhadem $\widehat{v}^2(x)$. Pro základní a nejčastěji používaný případ $\nu = 0$, tedy odhad samotné střední hodnoty, dostáváme vyjádření

$$\begin{aligned}\widehat{\text{var}}\,(h; x) = \mathbf{e}_1'\{\mathbf{X}_p(x)'\mathbf{W}_h(x)\mathbf{X}_p(x)\}^{-1}\mathbf{X}_p(x)'\mathbf{W}_h(x)\times \\ \times\, \widehat{\mathbf{V}}^2 \times \mathbf{W}_h(x)\mathbf{X}_p(x)\{\mathbf{X}_p(x)'\mathbf{W}_h(x)\mathbf{X}_p(x)\}^{-1}\mathbf{e}_1\,, \qquad (4.1)\end{aligned}$$

kde $\widehat{\mathbf{V}}^2 = \text{diag}\,((\widehat{v}^2(X_1), \ldots, (\widehat{v}^2(X_n))$. Odvození tohoto vztahu je možné nalézt v důsledku A.4.

Z asymptotického vyjádření pro vychýlení (viz Ruppert a Wand [15])[1] můžeme získat vhodný model pro odhadnutí vychýlení

$$\mathbb{E}\,\widehat{m}^{(\nu)}(x; h) = b_0 + b_{p+1-\nu}h^{p+1-\nu} + \cdots + b_{p+t-\nu}h^{p+t-\nu} \qquad (4.2)$$

pro nějaké $t \geq 1$ (doporučuje se $t = 1$ pro $p - \nu$ liché a $t = 2$ pro $p - \nu$ sudé). Zde píšeme $\widehat{m}^{(\nu)}(x; h)$ místo $\widehat{m}^{(\nu)}(x)$, abychom zdůraznili závislost na šířce okénka. Opět pro $\nu = 0$ dostaneme

$$\mathbb{E}\,\widehat{m}(x; h) = b_0 + b_{p+1}h^{p+1}\,. \qquad (4.3)$$

Abychom odhadli bias $(h_0; x)$ pro nějaké h_0, spočteme $\widehat{m}^{(\nu)}(x; h)$ pro hodnoty $\{h_j\}_{j=1}^M$, $M \geq t + 1$ v okolí h_0. Potom vyhladíme model (4.2) metodou nejmenších čtverců na datech $\{(h_j, \widehat{m}^{(\nu)}(x; h_j))\}_{j=1}^M$ a použijeme $\widehat{b}_0, \ldots, \widehat{b}_{p+t-\nu}$ k odhadu $\widehat{\text{bias}}\,(h_0; x)$. Použitím

$$\widehat{\text{MSE}}\,(h_0; x) = \widehat{\text{var}}\,(h_0; x) + \widehat{\text{bias}}^{\,2}(h_0; x)$$

můžeme odhadnout MSE na pevných hodnotách h a x. Model (4.2) je jediné použití asymptotiky, přičemž koeficienty v modelu (4.2) odhadujeme raději přímo než vložením odhadů do vzorce pro asymptotické hodnoty těchto koeficientů.

Pro pevné x odhadneme MSE v mřížce pro hodnoty h, řekněme 12 hodnot mezi span $(x; 0, 1)$ a span $(x; 1)$. Zde span $(x; q)$ je nejmenší hodnota

[1] pro $\nu = 0$ je asymptotické vyjádření vychýlení uvedeno ve větě 3.3

h taková, že alespoň $100q\,\%$ z $X_1, \ldots, X_n$ je uvnitř okénka šířky h kolem bodu x ($q \in [0;1]$). Poté nechť $\tilde{h}(x)$ minimalizuje $\widehat{\mathrm{MSE}}\,(h;x)$ v již zmíněné mřížce hodnot h. Tedy $\widehat{h}(x)$ je lokální šířka okénka, která se snaží minimalizovat MSE pro každé x v mřížce. Poznamenejme, že je nutné hledat první lokální minimum, nikoliv globální. Výše zmíněná metoda odhadu vychýlení totiž silně podhodnocuje vychýlení v případě, kdy h je tak velké, že všechny vlastnosti m jsou úplně vyhlazeny.

Spočítáme $\widehat{m}^{(\nu)}(x;\widehat{h}(x))$ na stejné mřížce x, kterou jsme použili pro nalezení $\widehat{h}(x)$. Pro výpočet reziduí interpolujeme $\widehat{m}^{(\nu)}(x;\widehat{h}(x))$ z mřížky x na $X_1, \ldots, X_n$ pomocí kubických splajnů. Pokud n není dostačně malé, řekněme menší než 100, potom přímý výpočet $\widehat{m}^{(\nu)}(x;\widehat{h}(x))$ je dosti pomalý vzhledem k náročnosti výpočtu $\widehat{h}(x)$.

K odhadu hodnoty m i v^2 použitím EBBS se využívá tříkrokový algoritmus:

1. Odhad m použitím „malého" pevného spanu.
2. Odhad v^2 vyhlazením druhých mocnin reziduí. Předpokládáme, že chyby mají konstantní špičatost, takže rozptyl druhých mocnin chyb je úměrný druhé hocnině jejich průměru. Tento předpoklad umožňuje vyhnout se odhadování varianční funkce čtvercových chyb.
3. Odhad m použitím metody EBBS a odhadu varianční funkce.

Kroky 2 a 3 mohou být zopakovány, ačkoli to není obecně nutné. Zda *span* v kroku 1 je dostatečně malý lze ověřit tím, jestli je menší než span vybraný v kroku 3 pro $\widehat{m}$. Bližší detaily metody EBBS je možné nalézt v Ruppert [14].

4.2 Další metody

Existuje řada dalších možností, jak vybrat šířku okénka, např. zobecněné křížové ověřování a podobné odhady. Většina těchto metod je mnohem jednodušší a méně náročná na výpočty než metoda EBBS. Na druhé straně mnoho z těchto jednodušších výběrů šířky okénka vytváří globální vyhlazovací parametr spíše než lokálně adaptivní okénko a jsou téměř výhradně určeny pro odhadování pomocí polynomů lichého stupně. Mimoto většina selektorů má za cíl najít šířku okénka optimální pro odhad m nebo v^2, nikoliv pro jejich derivace. Z těcho důvodů se metoda EBBS jeví jako velmi zajímavá a i my jsme ji použili při praktickém zpracování dat (viz kapitola 6).

Podrobnější informace o různých způsobech volby šířky okénka lze nalézt např. ve Fan a Gijbels [7], Chiu [11] nebo Sheather [17].

Kapitola 5

Detteho test homoskedasticity

Mějme obecný neparametrický regresní model

$$Y_{i,n} = m(x_{i,n}) + v(x_{i,n})\varepsilon_{i,n} \quad i = 1, \ldots, n\,, \tag{5.1}$$

kde $0 \leq x_{1,n} < x_{2,n} < \cdots < x_{n,n} \leq 1$ jsou pevně zvolené body plánu experimentu (body volíme pomocí tzv. návrhové hustoty $f(x)$), m je neznámá regresní funkce, v značí neznámou varianční funkci a $\varepsilon_{i,n}$ jsou náhodné veličiny, které jsou zadány ve formě trojúhelníkového schématu, jsou po řádcích nezávislé (tj. $\forall n \in \mathbb{N}\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ jsou nezávislé) a jejichž střední hodnota je nulová a rozptyl $\mathbb{E}\left[\varepsilon_{i,n}^2\right] = 1$. Problém heteroskedasticity v modelu (5.1) je široce zkoumán, protože pokud požadavek homoskedasticity není splněn, může dojít k podstatné ztrátě efektivnosti použitých procedur pro homoskedastické modely.

Kontroly předpokladu homoskedasticity jsou obvykle založeny na vizuálním prozkoumání grafů reziduí po parametrickém nebo neparametrickém vyhlazování. Existují formální procedury pro zcela parametricky specifikovanou regresní funkci, pro lineární model s neparametrickou varianční funkcí nebo semiparametrický model, kde neparamerická regresní funkce a varianční funkce jsou svázány specifickou spojovací funkcí. V poslední době byly navrženy také testy pro homoskedasticitu v celkově neparametrickém regresním modelu (5.1) — např. Eubank a Thomas [5] nebo Dette

a Munk [4]. Nicméně obě procedury mají vážné praktické nevýhody. Test prvně jmenovaných autorů požaduje normální rozdělení chyb, zatímco druhý test umí pouze odhalit alternativy konvergující k nule řádu $n^{-1/4}$.

My se zaměříme na další možný test zkonstruovaný v Dette [2], který netrpí těmito nedostatky. Základní idea je velice jednoduchá a bude pečlivě vysvětlena v části 5.1. Ukáže se, že problém testování homoskedasticity v modelu (5.1) je ekvivalentní s problémem testování tzv. pseudoreziduí na konstantní střední hodnotu. Toto přináší velmi jednoduchou testovací statistiku, jejíž asymptotická normalita je ukázána v části 5.2. Získáme různé stupně konvergence pro hypotézu a alternativu homoskedasticity. Navíc nový postup umožňuje odhalit alternativy konvergující k nule řádu $(n\sqrt{h})^{-1}$ (h označuje šířku okénka použitou v konstrukci testové statistiky) a v tomoto smyslu je efektivnější než test představený v Dette a Munk [4].

5.1 Základní idea a statistický test

Jednoduchý odhad integrované varianční funkce je dán v Rice [13], který vzal v úvahu součet druhých mocnin lokálních reziduí $1/(n-1)\sum_{i=2}^{n} R_{i,n}^2$, kde

$$R_{i,n}^2 = \frac{1}{2}(Y_{i,n} - Y_{i-1,n})^2 \,. \tag{5.2}$$

Za předpokladu hladkosti regresní a variační funkce máme

$$\mathbb{E}\,[R_{i,n}^2] = \frac{1}{2}[m(x_{i,n}) - m(x_{i-1,n})]^2 + \frac{1}{2}\mathbb{E}\,[v(x_{i,n})\varepsilon_{i,n} - v(x_{i-1,n})\varepsilon_{i,n}]^2 \approx v^2(x_{i,n})\,. \tag{5.3}$$

Proto tedy problém testování hypotézy homoskedasticity

$$H_0 :\ v^2(x) = v^2 \quad \forall x \in [0;1] \tag{5.4}$$

je (přibližně) ekvivalentní problému testování hypotézy, že regresní funkce náhodných proměnných $R_{i,n}^2$ na $x_{i,n}$ je konstantní.

V posledních dvou desetiletích bylo významné úsilí věnováno problému testování předpokladů o parametrech regresní funkce v modelu (5.1), např.

Dette a Munk [3] nebo Zheng [20], a vzhledem k argumentům v minulém odstavci se zdá být přirozené použití některé z těchto procedur pro pseudorezidua $R^2_{(i,n)}$. Zřejmým problémem v této konstrukci je fakt, že všechny tyto postupy jsou odvozeny za předpokladu nezávislosti přírůstků, zatímco výpočet pseudoreziduí zjevně vytváří m-závislé náhodné veličiny (zde $m = 1$, v zobecnění uvedeném dále již bude $m \geq 1$).

V následující části nicméně ukážeme, že tato základní myšlenka může být úspěšně aplikována pro konstrukci testu pro heteroskedasticitu s určitými technickými obtížnostmi v asymptotické analýze. V této chvíli použijeme modifikovanou verzi statistického testu představeného Zhengem [20]

$$T_n = \frac{1}{(n-1)(n-2)h} \sum_{|i-j|\geq 2} K\left(\frac{x_{i,n} - x_{j,n}}{h}\right)(R^2_{i,n} - \overline{R^2_n})(R^2_{j,n} - \overline{R^2_n}), \quad (5.5)$$

kde K označuje jádro splňující běžný předpoklad hustoty odhadu (viz část 5.2), h je šířka okénka a

$$\overline{R^2_n} = \frac{1}{n-1} \sum_{i=2}^{n} R^2_{i,n} \quad (5.6)$$

je střední hodnota pseudoreziduí. Předtím než ukážeme, že zamítnutí nulové hypotézy homoskedasticity pro velké hodnoty T_n přináší konzistentní test, předložíme zobecnění statistiky T_n, která je založena na pseudoreziduích vyšších řádů a která je asymptoticky účinnější.

Pohledem na (5.3) se snadno ukáže, že za platnosti hypotézy homoskedasticity a konstantní regrese $v^2(x) \equiv c$ je statistika $1/(n-1)\sum_{i=2}^{n} R^2_{i,n}$ nestranným odhadem v^2. Za účelem snížení vychýlení při nekonstantních regresních funkcích je nutné zprůměrovat v (5.2) více výrazů. Stejně jako v Dette [2] budeme nazývat vektor $(d_k)_{i=0}^{r}$ *diferenční posloupností řádu* r, jestliže

$$\sum_{i=0}^{r} d_i = 0\,, \quad \sum_{i=0}^{r} d_i^2 = 1 \quad (5.7)$$

a budeme definovat pseudorezidua řádu r jako

$$R^2_{i,n,r} = \left(\sum_{j=0}^{r} d_j Y_{i-j,n}\right)^2 . \quad (5.8)$$

Poznamenejme, že případ $r = 1$ dává pseudorezidua tvaru (5.2) zavedené v Rice [13]. Další oblíbený příklad získáme z posloupnosti $(d_0, d_1, d_2) = (1, -2, 1)/\sqrt{6}$, který je rozebírán v Gasser a kol. [8]. V případě rovnoměrného plánu experimentu (tj. $f \equiv 1$) je odpovídající odhad $1/(n-r)\sum_{i=2}^{n} R_{i,n}^2$ ve skutečnosti nestranným odhadem také pro lineární regresi, tj. $m(x) = a_0 + a_1 x$. Je-li návrh nerovnoměrný, je nutné odhad příslušně modifikovat (viz Gasser a kol. [8]).

Na základě r-tého řádu pseudoreziduí (5.8) definujeme následující zobecnění statistiky T_n:

$$T_n^{(r)} = \frac{1}{(n-r)(n-r-1)h} \sum_{|i-j|\geq r+1} K\left(\frac{x_{i,n} - x_{j,n}}{h}\right)(R_{i,n,r}^2 - \overline{R_{n,r}^2})(R_{j,n,r}^2 - \overline{R_{n,r}^2}), \tag{5.9}$$

kde

$$\overline{R_{n,r}^2} = \frac{1}{n-r}\sum_{i=r+1}^{n} R_{i,n,r}^2 \tag{5.10}$$

je střední hodnota reziduí. Statistické vlastnosti T_n a $T_n^{(r)}$ budou odvozené v následující kapitole.

5.2 Teorie Detteho testu

Abychom mohli vyslovit hlavní tvrzení této kapitoly, potřebujeme několik podmínek regularity týkajících se modelu (5.1). Nejprve návrhové body $x_{i,n}$ mají tvar regulární posloupnosti, tj.

$$\frac{i-1}{n-1} = \int_0^{x_{i,n}} f(x)dx, \quad i = 1, \ldots, n, \quad n \in \mathbb{N}, \tag{5.11}$$

kde f je kladná hustota na intervalu $[0; 1]$, která je Lipschitzovsky spojitá řádu $\gamma \geq \frac{1}{2}$, $f \in \mathrm{Lip}_\gamma[0; 1]$, tj. $|f(x) - f(y)| \leq c|x - y|^\gamma \ \forall x, y \in [0; 1]$. Dále předpokládáme, že $\varepsilon_{i,n}$ jsou iid, mají konečné momenty čtvrtého řádu

$$m_4(x_{i,n}) = \mathbb{E}\left[\varepsilon_{i,n}^4\right] < \infty, \tag{5.12}$$

kde funkce m_4 je Lipschitzovsky spojitá řádu $\gamma \geq \frac{1}{2}$, tj. $m_4 \in \mathrm{Lip}_\gamma[0;1]$, a rovnoměrně omezené momenty až do řádu 8, tj.

$$\mathbb{E}\left[|\varepsilon_{i,n}|^k\right] \leq c_k < \infty, \quad k = 5, 6, 7, 8\,. \tag{5.13}$$

Podobně se i o varianční a regresní funkci předpokládá, že jsou Lipschitzovsky spojité

$$v^2,\, m \in \mathrm{Lip}_\gamma[0;1]; \quad \gamma \geq \frac{1}{2}\,. \tag{5.14}$$

Dále pro jednoduchost předpokládejme, že jádrová funkce K používaná ve statistikách (5.5) a (5.9) je definovaná na $[-1;1]$, že je Lipschitzovsky spojitá řádu $\gamma \geq \frac{1}{2}$ a splňuje následující momentové podmínky:

$$\int_{-1}^{1} K(x)\, x^r\, dx = \begin{cases} 1 & \text{pro} \quad r = 0\,, \\ 0 & \text{pro} \quad r = 1\,, \\ 0 \leq \kappa_r < \infty & \text{pro} \quad r = 2 \end{cases} \tag{5.15}$$

Nakonec o šířce okénka použité v (5.5) a (5.9) předpokládáme, že

$$h \to 0\,, \quad nh^2 \to \infty\,. \tag{5.16}$$

První výsledek dává asymptotické rozdělení statistiky T_n definované v (5.5) a ukazuje konzistenci testu, který zamítá hypotézu homoskedasticity pro velké hodnoty T_n.

VĚTA 5.1 *Předpokládejme, že jsou splněny podmínky (5.11)-(5.16) a* $n \to \infty$.

(i) Za platnosti hypotézy (5.4) o homoskedasticitě ($v^2(x) \equiv v^2$) *máme pro statistiku* T_n *definovanou v (5.5)*

$$n\sqrt{h}T_n \xrightarrow{\mathcal{D}} \mathcal{N}(0; \lambda_0^2)\,, \tag{5.17}$$

kde je tzv. asymptotický rozptyl dán předpisem

$$\lambda_0^2 = 2v^8 \int_{-1}^{1} K^2(x)dx \int_0^1 m_4^2(x) f^2(x)dx\,. \tag{5.18}$$

(ii) *Za platnosti alternativy heteroskedasticity máme*

$$\sqrt{n}\left\{T_n - \frac{1}{h}\int\int K\left(\frac{x-y}{h}\right)\Delta(x)\Delta(y)f(x)f(y)dxdy\right\} \xrightarrow{\mathcal{D}} \mathcal{N}(0;\lambda_1^2)\,, \quad (5.19)$$

kde je tzv. asymptotický rozptyl dán vztahem

$$\lambda_1^2 = 4\int_0^1 m_4(x)v^4(x)\{(\Delta f)(x) - \overline{\Delta f}\}^2 f(x)dx\,, \quad (5.20)$$

funkce Δ *označuje odchylku* $v^2(\cdot)$ *od její střední hodnoty, tj.*

$$\Delta(x) = v^2(x) - \int_0^1 v^2(t)f(t)dt \quad (5.21)$$

a $\overline{\Delta f}$ *odpovídající střední hodnota* $\Delta f = \Delta\cdot f$ *vzhledem k návrhové hustotě, tj.*

$$\overline{\Delta f} = \int_{-1}^1 \Delta(x)f^2(x)dx\,. \quad (5.22)$$

Důkaz Lze nalézt v Dette [2].

□

Poznámka Poznamenejme, že vychýlení v (5.19) lze rozepsat jako $\mathbb{E}[T_n] = M^2(h) + o(1),\ n\to\infty$, kde

$$\begin{aligned} M^2(h) &= \frac{1}{h}\int\int K\left(\frac{x-y}{h}\right)(\Delta f)(x)(\Delta f)(y)dxdy \\ &= \int_0^1 \Delta^2(x)f^2(x)dx + o(1) \end{aligned} \quad (5.23)$$

a Δ je definována v (5.21). Pravá strana rovnice je řádu $o(1)$ právě tehdy, když platí hypotéza o homoskedasticitě, a konzistentní test získáme zamítnutím hypotézy (5.4), jestliže

$$n\sqrt{h}T_n > u_{1-\alpha}\widehat{\lambda}_{0,n}\,, \quad (5.24)$$

kde $u_{1-\alpha}$ značí $(1-\alpha)$ kvantil normálního rozdělení a $\widehat{\lambda}_{0,n}$ je nějaký konzistentní odhad pro asymptotický rozptyl (5.18). Jednoduchý odhad je získán obdobně jako v Dette a Munk [4], tj.

$$\widehat{\lambda}_{0,n} = 2(\widehat{A}_{1,n} - 6\widehat{A}_{2,n} + 9(\overline{R_n^2})^4)\int_{-1}^1 K^2(x)dx\,, \quad (5.25)$$

kde $\overline{R_n^2}$ je definován v (5.6),

$$\widehat{A}_{1,n} = \frac{4}{n-3} \sum_{i=2}^{n-2} R_{i,n}^4 R_{i+2,n}^4 \,, \tag{5.26}$$

a

$$\widehat{A}_{2,n} = \frac{2}{n-5} \sum_{i=2}^{n-4} R_{i,n}^4 R_{i+2,n}^2 R_{i+4,n}^2 \,. \tag{5.27}$$

Poznámka Poznamenejme, že rychlosti konvergence jsou rozdílné za platnosti nulové hypotézy a alternativy heteroskedasticity. Navíc pečlivá analýza důkazu první části věty 5.1 ukazuje, že asymptotická normalita je stále platná při lokální alternativě ve formě $v_n^2(t) = v^2 + (n\sqrt{h})^{-1/2} s(t)$, kde $s \in \mathrm{Lip}_\gamma[0;1]\,(\gamma \geq \frac{1}{2})$ je funkce taková, že $v_n^2(t)$ je nezáporná. V tomto případě máme

$$n\sqrt{h} T_n \xrightarrow{\mathcal{D}} \mathcal{N}(\mu_s; \lambda_0^2) \,, \tag{5.28}$$

kde $\mu_s = \int_0^1 s^2(x) f^2(x) dx$. Vzhledem k tomu, že test navržený v Dette a Munk [4] má rychlost konvergence řádově $n^{1/2}$, je schopný detekovat lokální alternativy blížící se k nule pomaleji než řádu $n^{-1/4}$. V tomto směru je nový test asymptoticky efektivnější a při vhodné volbě šířky okénka téměř dosahuje řádu, které odpovídají testům v parametrických modelech (viz Dette [2]).

VĚTA 5.2 *Předpokládejme, že jsou splněny podmínky (5.11)-(5.16) a* $n \to \infty$.

(i) Za platnosti hypotézy (5.4) o homoskedasticitě ($v^2(x) \equiv v^2$) *máme pro statistiku* $T_n^{(r)}$ *definovanou v (5.9)*

$$n\sqrt{h} T_n^{(r)} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0; \lambda_0^2(r)\right) \,, \tag{5.29}$$

kde je tzv. asymptotický rozptyl dán jako

$$\lambda_0^2(r) = 2v^8 \int_{-1}^{1} K^2(x) dx \int_0^1 \{m_4(x) - 1 + 4\delta_r\}^2 f^2(x) dx \tag{5.30}$$

a konstanta δ_r *je definována jako*

$$\delta_r = \sum_{k=1}^{r} \left(\sum_{j=0}^{r-k} d_j d_{j+k} \right)^2 \quad (r \geq 1)\,. \tag{5.31}$$

(ii) Za platnosti alternativy heteroskedasticity máme

$$\sqrt{n} \left\{ T_n^{(r)} - \frac{1}{h} \int \int K\left(\frac{x-y}{h} \right) \Delta(x)\Delta(y) f(x) f(y) dx dy \right\} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0; \lambda_1^2(r)\right)\,, \tag{5.32}$$

kde je tzv. asymptotický rozptyl dán jako

$$\lambda_1^2(r) = 4 \int_0^1 \left(m_4(x) - 1 + 4\delta_r\right) v^4(x) \{(\Delta f)(x) - \overline{\Delta f}\}^2 f(x) dx \tag{5.33}$$

a Δ, $\overline{\Delta f}$ *jsou definovány dle vztahů (5.21) a (5.22).*

Důkaz Lze nalézt v Dette [2].

□

Poznámka Poznamenejme, že situace diskutovaná ve větě 5.1 je získána pro volbu $r = 1$, $d_0 = -d_1 = 1/\sqrt{2}$, pro které $\delta_1 = \delta_0^4 = 1/4$. Jiné schéma vah navrhli Gasser a kol. [8], kteří použili pro rovnoměrný návrh

$$(d_0, d_1, d_2) = \frac{1}{\sqrt{6}}(1, -2, 1) \tag{5.34}$$

v definici (5.8) a argumentovali tím, že posloupnost (5.34) dává menší vychýlení, je-li $\overline{R_{n,r}^2}$ použito jako odhad rozptylu v neparametrické homoskedastické regresi. Upozorněme, že tuto posloupnost je nutné modifikovat v případě nerovnoměrného návrhu, abychom získali rozumně malé vychýlení.

S volbou (5.34) dostáváme ve Větě 5.2 ($r = 2$) $\delta_2 = \frac{17}{36}$, tedy

$$\lambda_0^2(2) = 2v^8 \int_{-1}^{1} K^2(x) dx \int_0^1 \{m_4(x) + \frac{8}{9}\}^2 f^2(x) dx$$

$$\lambda_1^2(2) = 4 \int_0^1 \left(m_4(x) + \frac{8}{9} \right) v^4(x) \{(\Delta f)(x) - \overline{\Delta f}\}^2 f(x) dx\,.$$

5.3 Praktická implementace

Ukázalo se, že testování parametrů regresní funkce na základě asymptotických vztahů s úrovní aproximace jako ve větách 5.1 a 5.2 je nevhodné pro reálné rozsahy vzorků dat. Řešením tohoto problému je použití bootstrapové procedury. My využijeme bootstrap založený na reziduích (viz Härdle a Bowman [10]).

Označme $\widehat{m}$ lokální polynomický odhad regresní funkce m se šířkou okénka h_m a definujme neparametrická rezidua jako

$$r_{i,n} = Y_{i,n} - \widehat{m}(x_{i,n})\,, \quad i = 1, \ldots, n\,. \tag{5.35}$$

Bootstrapový výběr pak definujme jako

$$Y^*_{i,n} = \widehat{m}(x_{i,n}) + r^*_i\,, \quad i = 1, \ldots, n \tag{5.36}$$

kde r^*_i je iid výběr z empirické distribuční funkce (centrovaných) $r_{i,n}$. Nulovou hypotézu zamítáme, jestliže je T_n větší než odpovídající kvantil bootstrapového rozdělení T_n.

Poznamenejme, že použití bootstrapové procedury vyžaduje volbu dvou jádrových funkcí a šířek okénka — pro definici (5.5) a pro výpočet reziduí v (5.35) z lokálního polynomického odhadu regresní funkce. Obvykle se v obou případech používá Epanechnikovo jádro (1.10). Jako šířky okénka navrhuje Dette [2] použít

$$h = h_m = \left(\frac{1}{n} \int_0^1 v^2(t) f(t) dt \right)^{\frac{1}{5}}\,.$$

Kapitola 6

Zpracování dat z pražského Klementina

V této kapitole se zabýváme praktickým zpracováním dat pomocí metod zmíněných v předchozích kapitolách. Daty jsou průměrné měsíční teploty z pražského Klementina měřené v letech 1771–2000. Základem bude model (2.1). Pro něj odhadneme regresní a varianční funkci metodou EBBS (v podkapitole 6.1) a otestujeme homoskedasticitu modelu pomocí Detteho testu (v podkapitole 6.2). Na závěr v podkapitole 6.3 porovnáme odhady varianční funkce s různými parametry metody EBBS a odhady získané s pevnou globální šířkou okénka (resp. s pevným globálním spanem).

Podobně jako v Šalom [18] nebo Goderniaux [9] budeme pracovat pouze s průměrnými ročními teplotami z let 1775–1989, neboť ta byla již několikrát zpracovávána.

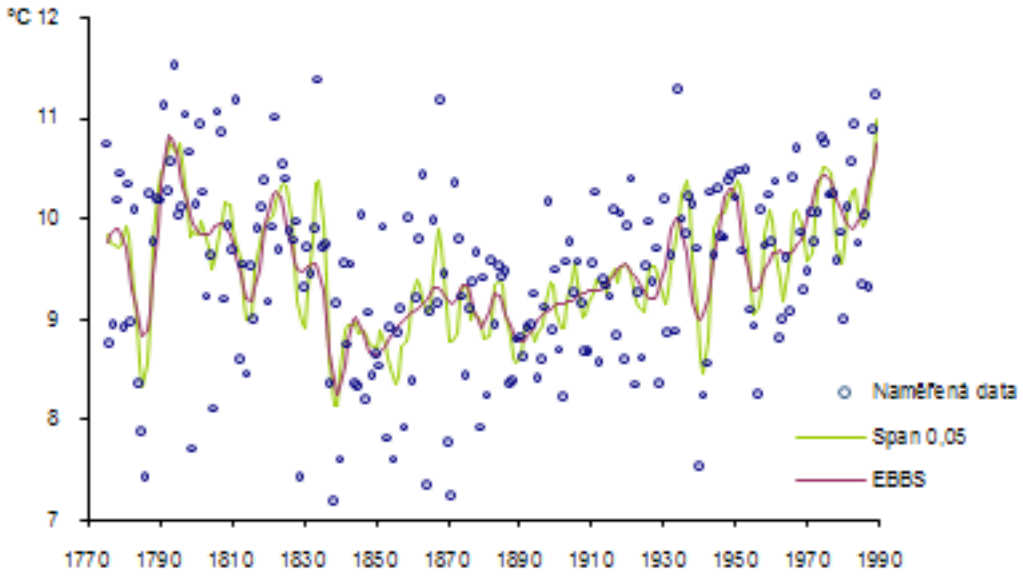
6.1 Odhad regresní a varianční funkce

Uvažujeme model (2.1), v našem případě konkrétně

$$Y_i = m(X_i) + v(X_i)\varepsilon_i \quad i = 1, \ldots, 215 \ . \tag{6.1}$$

Regresory X_i jsou pro jednoduchost přímo indexy jednotlivých pozorování, tedy $X_i = i,\ i = 1, \ldots, 215$. Chyby ε_i jsou nezávislé, stejně rozdělené s nulovou střední hodnotou a jednotkovým rozptylem.

Regresní funkci chceme odhadovat lokálními kvadratickými polynomy, varianční funkci pak pomocí lokálních lineárních polynomů. Pro výběr šířky okénka použijeme v obou případech metodu EBBS (viz kapitola 4.1). Podle této metody nejprve odhadneme regresní funkci s pevným „malým" spanem, v našem případě volíme span $= 0{,}05$. To znamená, že pro odhad v bodě X_i využíváme 5 % dat z okolí bodu X_i. Pro zpracování dat v této i následujících podkapitolách používáme Epanechnikovo jádro (1.10). Tím získáme počáteční odhad regresní funkce $\widehat{\mathbf{m}_0}$ (viz zelená křivka na obr. 6.1, který zobrazuje oba odhady regresní funkce a původní data).



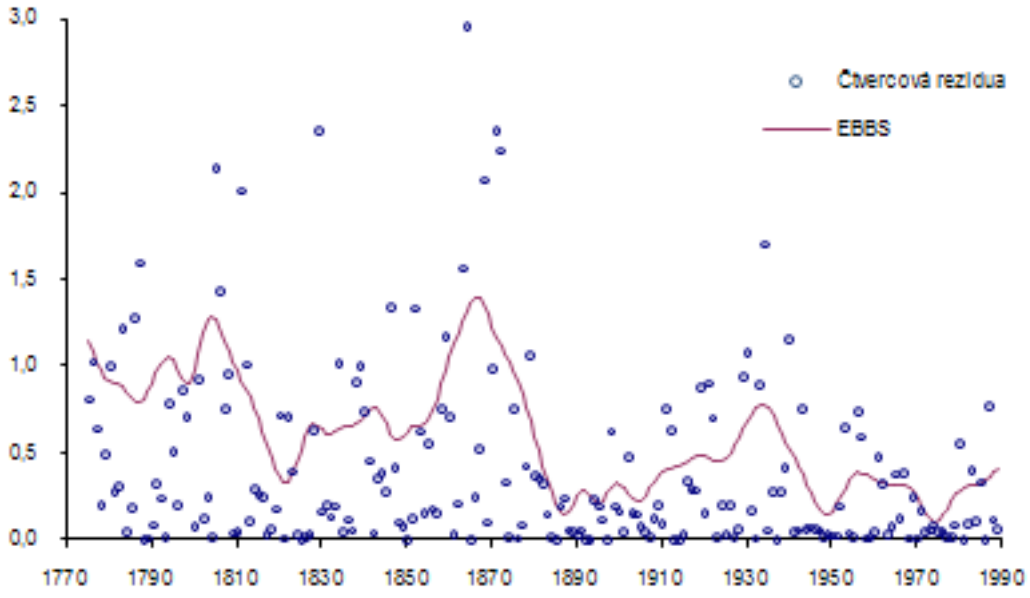
Obrázek 6.1: Odhad m metodou EBBS

Druhým krokem je odhad varianční funkce metodou EBBS. Jako data, ze kterých ji budeme odhadovat, použijeme kvadratická rezidua získaná z počátečního odhadu, tj. $R_{mi}^2 = (Y_i - \widehat{m}_{0i})^2$. Ta modelujeme vztahem

$$R_{mi}^2 = v^2(X_i) + \varsigma\varepsilon_i' \quad i = 1, \ldots, 215\ , \tag{6.2}$$

kde ε'_i jsou opět nezávislé, stejně rozdělené náhodné veličiny s nulovou střední hodnotou a jednotkovým rozptylem a $\varsigma^2 < \infty$ neznámý parametr[1].

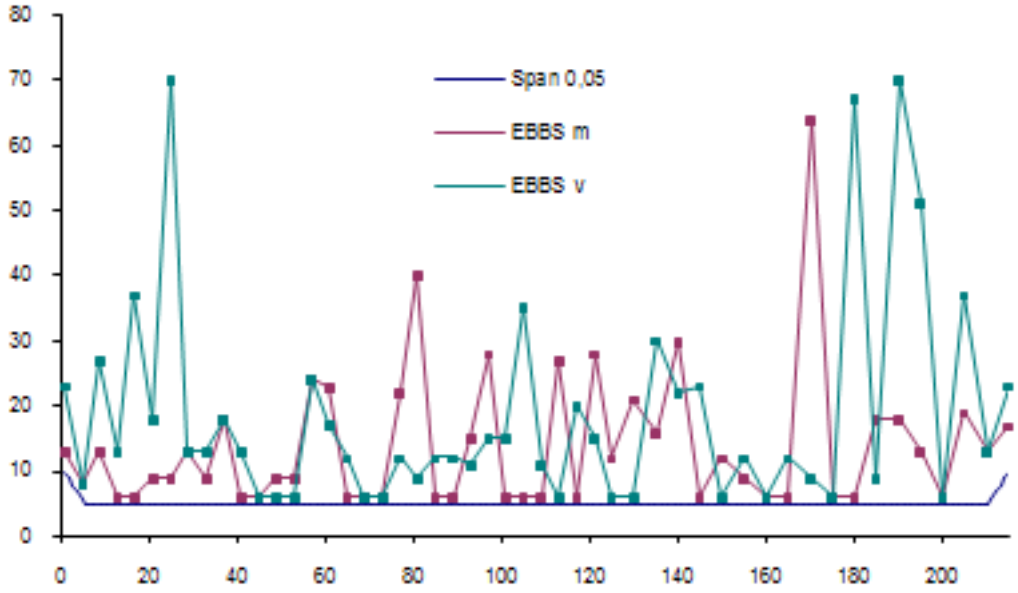
Celou metodu EBBS aplikujeme na mřížce padesáti hodnot X_{i_k} rovnoměrně vybraných z X_i. Pro každý bod generujeme mřížku dvanácti hodnot šířek okénka od span $= 0{,}05$ do span $= 1$. Pro každou z nich odhadneme velikost střední kvadratické chyby MSE, přičemž vychýlení odhadu modelujeme s parametry $t = 2$ a $M = 4$ (význam parametrů viz kapitola 4.1) a pro odhad rozptylu odhadu použijeme výběrový rozptyl kvadratických reziduí. Nejlepší šířku v daném bodě volíme jako argument prvního lokálního minima MSE. Takto získané šířky okénka jsou zobrazeny na obr. 6.3 (zelená křivka). S takto získanými šířkami okénka pak odhadujeme varianční funkci kvadratickými polynomy s korekcí dle (2.2). Získaný odhad poté interpolujeme z mřížky na všechna X_i kubickými splajny. Výsledný odhad $\widehat{\mathbf{v}}$ je zobrazen společně s kvadratickými rezidui na obr. 6.2.



Obrázek 6.2: Odhad v^2 metodou EBBS

[1]Předpokládáme konstantní špičatost veličin ε_i z (6.1), tedy ε'_i z (6.2) mají konstantní rozptyl.

V posledním kroku odhadujeme znovu regresní funkci, tentokrát již s volbou šířky okénka metodou EBBS. Opět tuto metodu aplikujeme na stejných padesáti hodnotách X_{i_k} s dvanácti možnými hodnotami šířek od span $= 0{,}05$ do span $= 1$ a parametry pro odhad MSE $t = 2$ a $M = 4$. Pro odhad rozptylu odhadu využijeme $\widehat{\mathbf{v}}$ získaný v předchozím kroku. Na závěr odhad regresní funkce opět interpolujeme z mřížky na celá data pomocí kubických splajnů. Výsledný odhad $\widehat{\mathbf{m}}$ je zobrazen červenou křivkou na již zmíněném obr. 6.1. Na posledním grafu této podkapitoly, obr. 6.3, jsou červenou křivkou znázorněny optimální šířky okénka pro odhad regresní funkce.



Obrázek 6.3: Šířky okénka

6.2 Detteho test

Detteho test provádíme na základě kapitoly 5. Použijeme bootstrapovou metodu s 500 opakováními. Hodnota testové statistiky T_n odpovídá přibližně 70% kvantilu bootstrapového rozdělení, tedy homoskedasticitu v modelu (6.1) nemůžeme zamítnout.

6.3 Vliv parametrů na odhad varianční funkce

V této podkapitole budeme odhadovat varianční funkci lokálními polynomy. Použijeme různé parametry a porovnáme takto získané výsledky. Opět vycházíme z modelů (6.1) a (6.2). Jako odhad regresní funkce, ze kterého získáváme kvadratická rezidua, použijeme odhad $\widehat{\mathbf{m}}_0$ z přechozích podkapitol.

Nejprve odhadujeme $\mathbf{v}$ lokálními lineárními polynomy s různou šířkou okénka. Volíme pevný globální span od 0,05 do 1,00, viz tab.6.1.

0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,60	0,70	0,80	0,90	1,00

Tabulka 6.1: Volby span

Jádro používáme opět Epanechnikovo a odhady počítáme se oběma typy korekce podle vztahů (2.2) a (2.3) i bez použití korekčního faktoru. Odhadujeme tedy podle vztahů (pro význam jednotlivých symbolů viz kapitola 2):

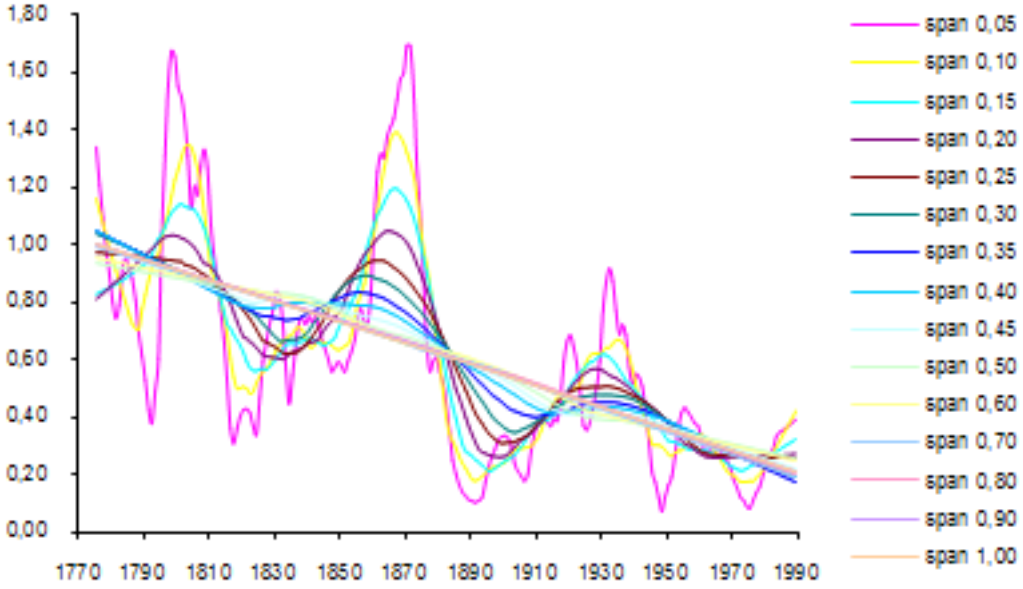
$$\widehat{\mathbf{v}}^2 = \frac{\mathbf{S}_2\mathbf{r}^2}{\mathbf{1} + \mathbf{S}_2\Delta} \tag{6.3a}$$

$$\widehat{\mathbf{v}}^2 = \mathbf{S}_2\frac{\mathbf{r}^2}{\mathbf{1} + \Delta} \tag{6.3b}$$

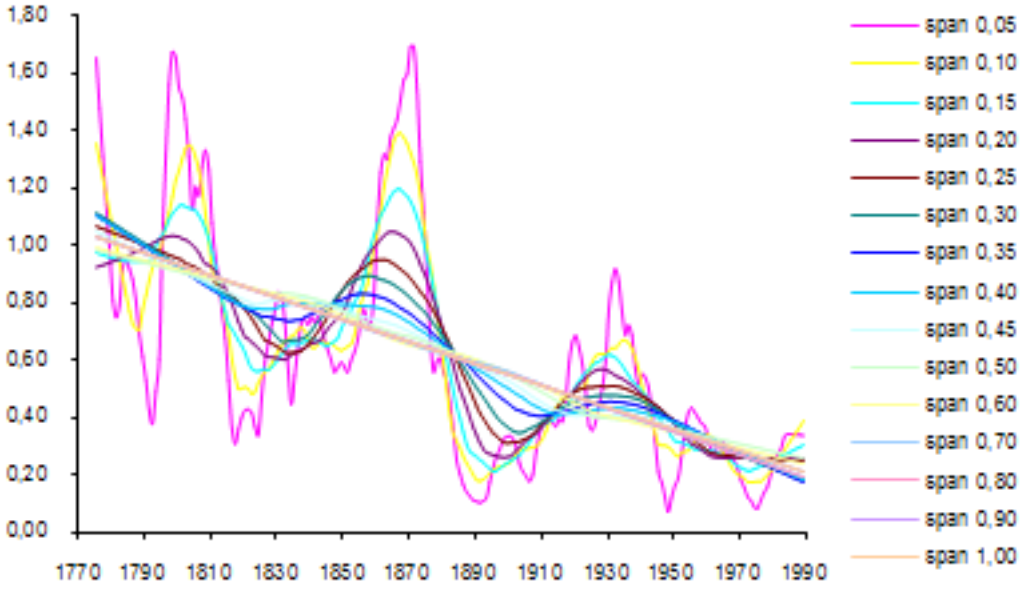
$$\widehat{\mathbf{v}}^2 = \mathbf{S}_2\mathbf{r}^2 \tag{6.3c}$$

Poznamenejme, že v grafech značíme odhad (6.3a) jako „normal", odhad (6.3b) jako „student" a odhad bez korekce (6.3c) jako „none". Výsledky jsou v grafech na obr. 6.4, obr. 6.5 a obr. 6.6.

Výsledky ukazují, že čím větší je šířka okénka (resp. span), tím více se „vyhladí" veškeré informace z kvadratických reziduí. Pro vysoké spany se tak blížíme ke klasické lineární regresi. Tento trend je patrný už pro spany větší než 0,50. Odhady pro span větší než 0,70 téměř splývají a tvoří přímku.

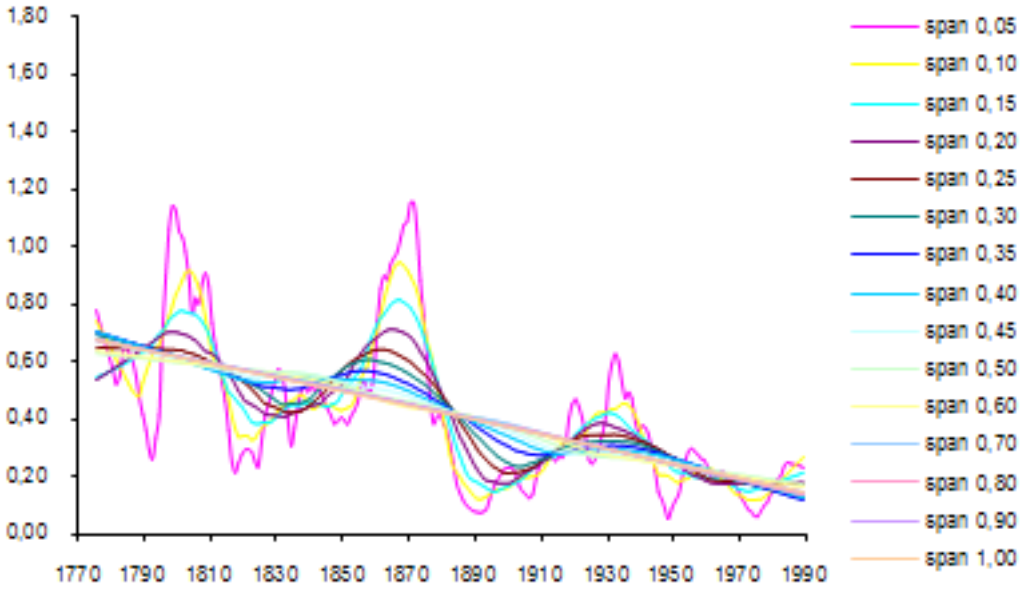


Obrázek 6.4: Odhad v^2 — pevný span, korekce "normal"

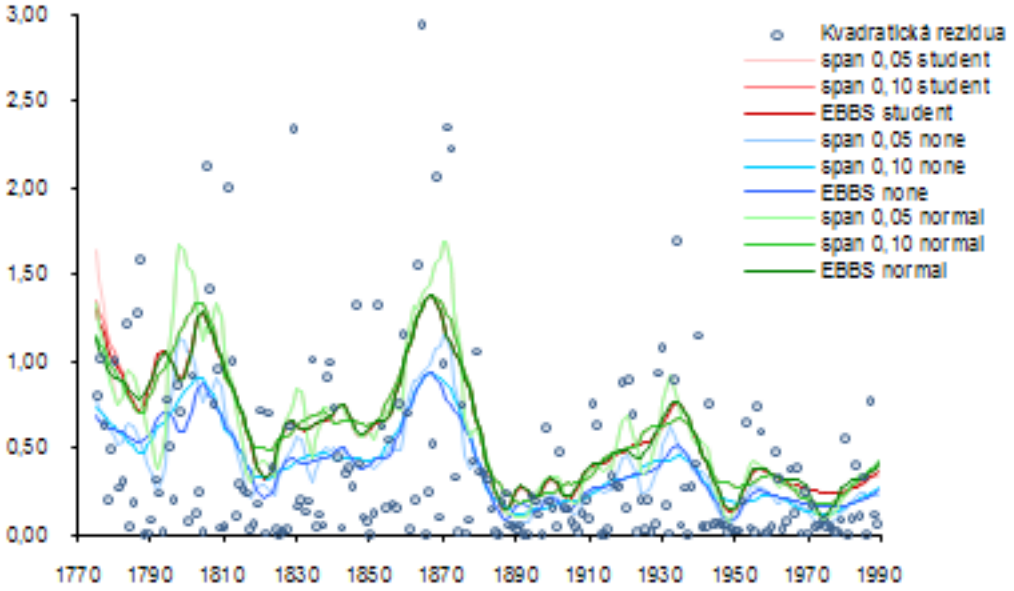


Obrázek 6.5: Odhad v^2 — pevný span, korekce "student"

Zajímavým zjištěním je také to, že se příliš neprojevil vliv jednotlivých korekcí. Proto dalším krokem je porovnání vlivu korekcí na odhad. Použijeme jak pevné šířky okénka odpovídající span $= 0{,}05$ a span $= 0{,}10$, tak výběr šířky metodou EBBS (viz obr. 6.7).



Obrázek 6.6: Odhad v^2 — pevný span, bez korekce



Obrázek 6.7: Odhad v^2 — porovnání korekcí

Vidíme, že rozdíl mezi korekcemi „normal" (6.3a) a „student" (6.3b) je minimální, grafy odhadů téměř splývají. Pokud však žádnou korekci nepoužijeme (vztah (6.3c)), je odhad varianční funkce poněkud menší a zdá se, že varianční funkci podhodnocuje.

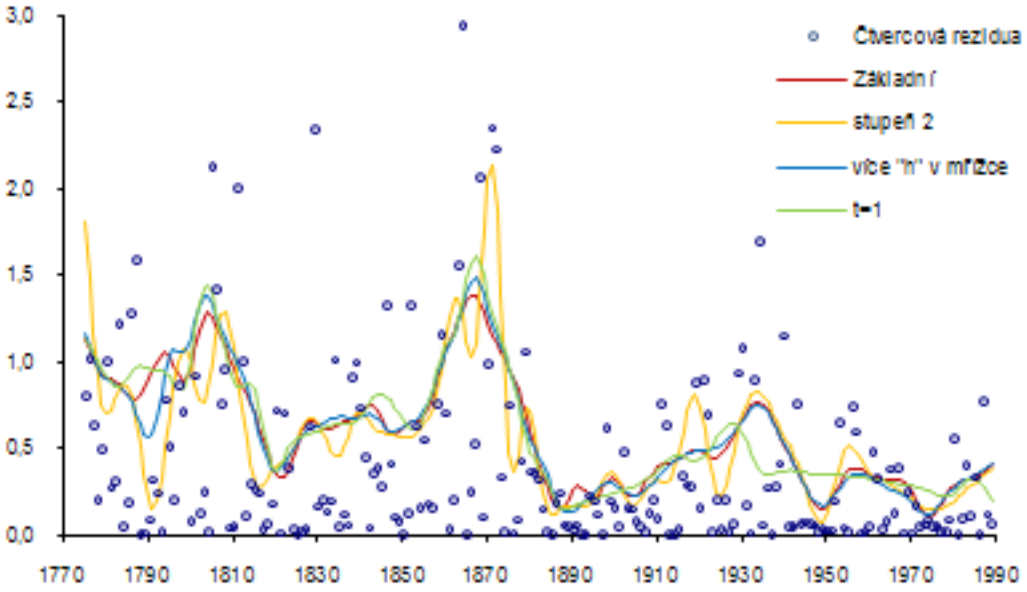
Posledním krokem je porovnání odhadů varianční funkce získaných metodou EBBS s různými parametry. Základním odhadem (*Odhad 1.* — červená barva v grafech na obr. 6.8, obr. 6.9 a obr. 6.10) zůstává $\widehat{\mathbf{v}}$ z přechozí části 6.1, tedy počet bodů X_i je 50, šířky okénka volíme z mřížky 12 hodnot od span $= 0{,}05$ do span $= 1{,}00$, odhad je lokálně lineární s korekcí „normal" (6.3a) a pro odhad MSE jsou použity parametry $t = 2$ a $M = 4$.

Variantami jsou odhady s následujícími změnami parametrů od základního odhadu:

Odhad 2. Odhad je kvadratický, tj. $p = 2$ (žlutá barva).

Odhad 3. Pro odhad MSE používáme $M = 6$ bodů z okolí h_0 a volíme jemnější mřížku šířek okénka s 20 hodnotami (modrá barva).

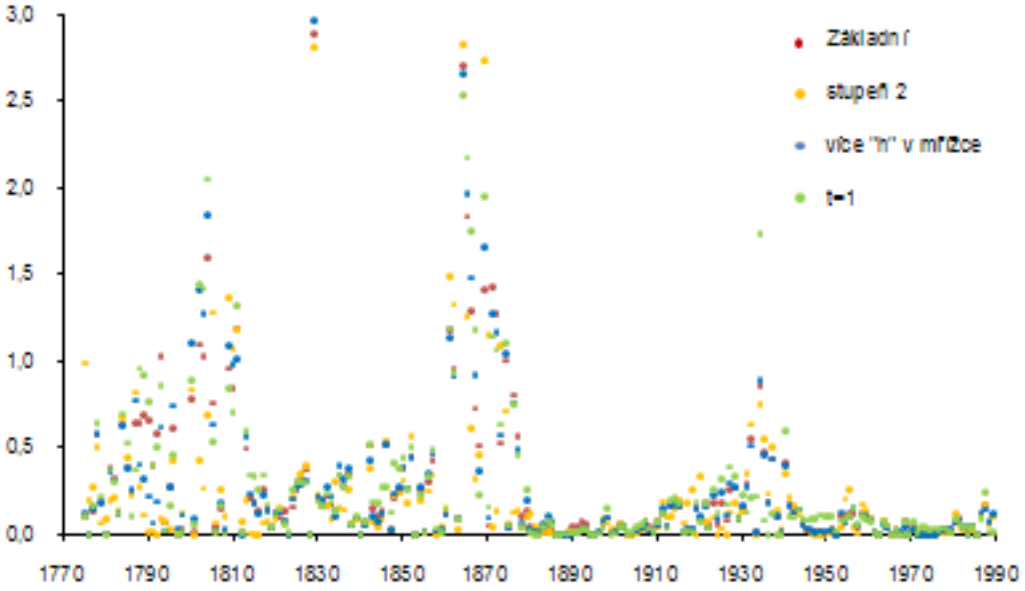
Odhad 4. Pro odhad MSE volíme parametr $t = 1$ (zelená barva).



Obrázek 6.8: Odhad v^2 — různé parametry EBBS

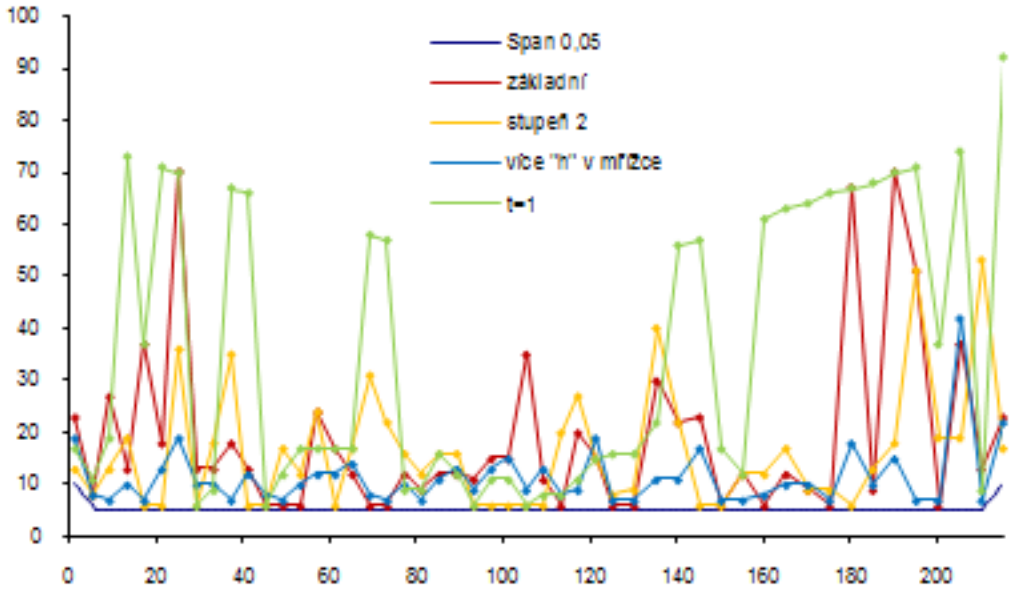
Na obr. 6.8 vidíme získané odhady varianční funkce. Pouze změna stupně polynomu použitého k odhadu varianční funkce má větší vliv na výsledný odhad. Vliv stupně polynomu není nijak překvapivý - vyšší stupeň polynomu umožní lépe zachytit průběh dat. Ovšem na grafu na obr. 6.9, kde

jsou zobrazeny kvadratické odchylky $(R_i^2 - \widehat{v}_i)^2$, vidíme, že změna stupně polynomu neměla na velikost odchylek příliš velký vliv. Zvýšení stupně tedy přes „lepší vyhlazení" nepřinesla významné zlepšení odhadu. Změna parametrů vlastní metody EBBS na odhad varianční funkce ani na odchylky vliv v podstatě nemá.



Obrázek 6.9: Rezidua $(r^2 - v^2)^2$ — různé parametry EBBS

Zajímavým výsledkem jsou velké hodnoty odchylek kolem roků 1800, 1830, 1870 a 1940. Při porovnání s výsledky týkajícími se skoků regresní funkce stejných dat, které jsou uvedeny v Šalom [18], si můžeme všimnout, že tyto roky přibližně odpovídají nalezeným skokům regresní funkce (ta byla odhadována se šířkou okénka 4–8 a možné skoky byly detekovány v letech 1786–1787, 1836–1837, 1871–1872 a 1939–1940). Proto se domníváme, že zmíněné vysoké odchylky odhadu varianční funkce jsou způsobeny spíše skoky v regresní funkci (a tedy velkým vychýlením odhadu regresní funkce v daných letech) než vysokou hodnotou skutečné varianční funkce.



Obrázek 6.10: Šířky okénka — různé parametry EBBS

Na posledním grafu (obr. 6.10) si můžeme prohlédnout šířky okénka pro jednotlivé odhady tak, jak byly zvoleny metodou EBBS. Pro větší přehlednost byly šířky v bodech mřížky pospojovány úsečkami a pro porovnání jsou zobrazeny i šířky odpovídající pevnému span $= 0{,}05$.

Dodatek A

Důkazy a odvození

A.1 Odvození vztahu (2.4)

Při odvozování řešení metody nejmenších vážených čtverců v bodě X_i, tedy takového vektoru $\widehat{\beta(X_i)} = (\widehat{\beta}_0(X_i), \ldots, \widehat{\beta}_p(X_i))$, který minimalizuje výraz

$$\sum_{j=1}^{n} \left(Y_j - \sum_{k=0}^{p} \beta_k(X_i)(X_j - X_i)^k \right)^2 \frac{1}{h} K\left(\frac{X_j - X_i}{h} \right) \quad , \tag{A.1}$$

postupujeme následujícím způsobem. Nejprve pro zjednodušení značení píšme $K_j = \frac{1}{h} K\left(\frac{X_j - X_i}{h}\right)$, $x_j = (X_j - X_i)$ a $\beta_k = \beta_k(X_i)$. Chceme tedy minimalizovat

$$\sum_{j=1}^{n} K_j \left(Y_j - \sum_{k=0}^{p} \beta_k x_j^k \right)^2 \quad . \tag{A.2}$$

Spočteme derivace vztahu (A.2) podle jednotlivých koeficientů β_m a položíme je rovné nule,

$$-2 \sum_{j=1}^{n} x_j^m K_j \left(Y_j - \sum_{k=0}^{p} \beta_k x_j^k \right) \stackrel{!}{=} 0\,, \quad m = 0, \ldots, p\,, \tag{A.3}$$

a po snadné úpravě získáme

$$\sum_{j=1}^{n} x_j^m K_j Y_j = \sum_{k=0}^{p} \beta_k \sum_{j=1}^{n} K_j x_j^{k+m}\,, \quad m = 0, \ldots, p\,. \tag{A.4}$$

Zaveďme nyní značení jako v kapitole 2, tedy

$$\mathbf{X}_p(X_i) = \begin{bmatrix} 1 & X_1 - X_i & \cdots & (X_1 - X_i)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - X_i & \cdots & (X_n - X_i)^p \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p \end{bmatrix}$$

a

$$\mathbf{W}_h(X_i) = \mathrm{diag}_{1\le j\le n} \frac{1}{h} K\left(\frac{X_j - X_i}{h}\right) = \mathrm{diag}_{1\le j\le n} K_j \ ,$$

kde $\mathrm{diag}_{\,1\le i\le n} a_i$ značí diagonální matici typu $n \times n$ s prvky a_i na diagonále. Potom vztah (A.4) můžeme přepsat s využitím násobení matic na

$$(x_1^m, \ldots, x_n^m)\mathbf{W}_h(X_i)\mathbf{Y} = (x_1^m, \ldots, x_n^m)\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\beta\,, \quad m = 0, \ldots, p$$

a celou soustavu pak najednou jako

$$\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{Y} = \mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\beta\,. \tag{A.5}$$

Bude-li matice $\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)$ regulární, pak řešením maticové rovnice (A.5) je

$$\widehat{\beta} = \{\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\}^{-1}\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{Y}$$

a hledaný odhad regresní funkce

$$\widehat{m}(X_i) = \widehat{\beta}_0 = \mathbf{e}_1'\{\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\}^{-1}\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{Y}$$

.

Proto (i, j)-tý prvek vyhlazovací matice $\mathbf{S}_{p,h}$ je

$$(\mathbf{S}_{p,h})_{ij} = \mathbf{e}_1'\{\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\}^{-1}\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{e}_j \ , \tag{A.6}$$

tedy vztah (2.4).

A.2 Důkaz Věty 3.1

Připomeňme na úvod, že zachováváme značení z předchozích kapitol, tj. násobení vektorů je vždy po složkách, násobení matic je klasické a násobení matic po složkách (Hadamardův součin) značíme symbolem $\odot$.

Nejdříve poznamenejme, že

$$\widehat{\mathbf{v}}^2 = \frac{\mathbf{S}_2 \mathrm{diag}\{(\mathbf{S}_1 - \mathbf{I})\mathbf{Y}\mathbf{Y}'(\mathbf{S}_1 - \mathbf{I})'\}}{\mathbf{1} + \mathbf{S}_2\Delta}.$$

Pro vychýlení platí

$$\begin{aligned}\mathbb{E}\,(\widehat{\mathbf{v}}^2|\chi) &= \frac{\mathbf{S}_2 \mathrm{diag}\{(\mathbf{S}_1 - \mathbf{I})(\mathbf{m}\mathbf{m}' + \mathbf{V}^2)(\mathbf{S}_1 - \mathbf{I})'\}}{\mathbf{1} + \mathbf{S}_2\Delta} \\ &= \frac{\mathbf{S}_2\{\mathrm{diag}(\mathbf{b}_1\mathbf{b}_1') + \mathbf{v}^2 + \mathrm{diag}(\mathbf{S}_1\mathbf{V}^2\mathbf{S}_1' - 2\mathbf{S}_1\mathbf{V}^2)\}}{\mathbf{1} + \mathbf{S}_2\Delta}\end{aligned}$$

Přímým výpočtem pak dostáváme první rovnost věty.

Vztah pro $\mathrm{cov}\,(\widehat{\mathbf{v}}^2|\chi)$ závisí hlavně na následujícím lemma.

LEMMA A.1 *Nechť $\mathbf{Y}$ je náhodný vektor, který má všechny složky nezávislé. Definujeme $\mathbf{m} = \mathbb{E}\,(\mathbf{Y})$, $\mathbf{V}^2 = \mathrm{diag}[\mathbb{E}\,\{(\mathbf{Y} - \mathbf{m})^2\}]$, $\mathbf{G} = \mathrm{diag}[\mathbb{E}\,\{(\mathbf{Y} - \mathbf{m})^3\}]$ a $\mathbf{T} = \mathrm{diag}[\mathbb{E}\,\{(\mathbf{Y} - \mathbf{m})^4\}]$. Potom pro libovolnou čtvercovou matici konstant $\mathbf{A}$, která má stejný počet řádků jako $\mathbf{Y}$, platí*

$$\begin{aligned}\mathrm{cov}\,\{(\mathbf{A}\mathbf{Y})^2\} &= (\mathbf{A}\odot\mathbf{A})(\mathbf{T} - 3\mathbf{V}^4)(\mathbf{A}\odot\mathbf{A})' \\ &\quad + 2\{\mathrm{diag}(\mathbf{A}\mathbf{m})\mathbf{A}\mathbf{G}(\mathbf{A}\odot\mathbf{A})' + (\mathbf{A}\odot\mathbf{A})\mathbf{G}\mathbf{A}'\mathrm{diag}(\mathbf{A}\mathbf{m})\} \\ &\quad + 2(\mathbf{A}\mathbf{V}^2\mathbf{A}')\odot(\mathbf{A}\mathbf{V}^2\mathbf{A}') + 4(\mathbf{A}\mathbf{V}^2\mathbf{A}')\odot\{(\mathbf{A}\mathbf{m})(\mathbf{A}\mathbf{m})'\}\,.\end{aligned}$$

Důkaz Budeme používat tenzorové značení a výsledky McCullagha [12]. Nechť a_{ij} označuje (i,j)-tou pozici matice $\mathbf{A}$. Zobecněný kumulant souboru náhodných veličin $(Y_1,\ldots,Y_n)$ je běžný kumulant náhodných veličin vytvořený pomocí součinů z tohoto souboru. Zobecněný kumulant budeme značit pomocí dělených horních indexů, např. $\kappa^i = \mathrm{cum}(Y_i) = \mathbb{E}\,(Y_i)$, $\kappa^{i,j} = \mathrm{cum}(Y_i, Y_j) = \mathrm{cov}\,(Y_i, Y_j)$ a $\kappa^{i,j,kl} = \mathrm{cum}(Y_i, Y_j, Y_k Y_l)$.

Existuje mnoho různých vztahů spojujících zobecněné kumulanty s běžnými kumulanty a momenty, viz McCullagh [12].

Snadno se ukáže, že (m,n)-tý prvek matice $\operatorname{cov}\{(\mathbf{AY})^2\}$ je

$$\operatorname{cov}\{(\mathbf{AY})^2\}_{mn} = \sum_i \sum_j \sum_k \sum_l a_{mi} a_{mj} a_{nk} a_{nl} \kappa^{ij,kl} .$$

Jednou ze základních identit pro zobecněné kumulanty je

$$\begin{aligned}\kappa^{ij,kl} = {} & \kappa^{i,j,k,l} + \kappa^i \kappa^{j,k,l} + \kappa^j \kappa^{i,k,l} + \kappa^k \kappa^{i,j,l} + \kappa^l \kappa^{i,j,k} + \kappa^{i,k} \kappa^{j,l} \\ & + \kappa^{i,l} \kappa^{j,k} + \kappa^i \kappa^k \kappa^{j,l} + \kappa^i \kappa^l \kappa^{j,k} + \kappa^j \kappa^k \kappa^{i,l} + \kappa^j \kappa^l \kappa^{i,k} ,\end{aligned}$$

jejíž odvození lze nalézt na str. 58 v McCullagh [12].

Toto implikuje, že díky vzájemné nezávislosti Y_i platí

$$\begin{aligned}\operatorname{cov}\{(\mathbf{AY})^2\}_{mn} = {} & \sum_i a_{mi}^2 a_{ni}^2 \kappa^{i,i,i,i} \\ & + 2 \sum_i \sum_j (a_{mi} a_{mj} a_{nj}^2 \kappa^i \kappa^{j,j,j} + a_{mi}^2 a_{ni} a_{nj} \kappa^j \kappa^{i,i,i}) \\ & + 2 \sum_i \sum_j a_{mi} a_{mj} a_{ni} a_{nj} \kappa^{i,i} \kappa^{j,j} \\ & + 4 \sum_i \sum_j \sum_k a_{mi} a_{mk} a_{nj} a_{nk} \kappa^i \kappa^j \kappa^{k,k}\end{aligned}$$

Je snadné ověřit, že odtud již vyplývá uvedený vztah lemma.

□

Následující lemma ukazuje, jak jsou kovarianční matice ovlivněny násobením po prvcích.

LEMMA A.2 *Nechť* $\mathbf{a}$ *je vektor konstant stejné délky jako* $\mathbf{Y}$. *Potom* $\operatorname{cov}(\mathbf{aY}) = (\mathbf{aa}') \odot \operatorname{cov}(\mathbf{Y})$.

Důkaz Pišme $\mathbf{a} = (a_1, \ldots, a_n)'$ a $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. Bez újmy na obecnosti

předpokládejme, že $\mathbb{E}\, Y_i = 0$ pro všechna i. Potom

$$\operatorname{cov}(\mathbf{aY}) = \operatorname{cov}\begin{pmatrix} a_1 Y_1 \\ \vdots \\ a_n Y_n \end{pmatrix} = \begin{pmatrix} \mathbb{E}\, a_1^2 Y_1^2 & \cdots & \mathbb{E}\, a_1 a_n Y_1 Y_n \\ \vdots & \ddots & \vdots \\ \mathbb{E}\, a_n a_1 Y_n Y_1 & \cdots & \mathbb{E}\, a_n^2 Y_n^2 \end{pmatrix} =$$

$$= \begin{pmatrix} a_1^2 & \cdots & a_1 a_n \\ \vdots & \ddots & \vdots \\ a_n a_1 & \cdots & a_n^2 \end{pmatrix} \odot \begin{pmatrix} \mathbb{E}\, Y_1^2 & \cdots & \mathbb{E}\, Y_1 Y_n \\ \vdots & \ddots & \vdots \\ \mathbb{E}\, Y_n Y_1 & \cdots & \mathbb{E}\, Y_n^2 \end{pmatrix} =$$

$$= (\mathbf{aa}') \odot \operatorname{cov}(\mathbf{Y}).$$

□

Vztah pro $\operatorname{cov}(\widehat{\mathbf{v}}^2|\chi)$ z věty 3.1 pak ihned vyplývá z Lemma A.1, Lemma A.2 a ze známé rovnosti $\operatorname{cov}(\mathbf{AY}) = \mathbf{A} \operatorname{cov}(\mathbf{Y})\, \mathbf{A}'$.

A.3 Důkaz Věty 3.3

Pro r-krát diferencovatelnou funkci g pišme $\mathbf{g}^{(r)} = [g^{(r)}(X_1), \ldots, g^{(r)}(X_n)]'$. Také budeme používat úmluvu, že pro n-dimenzionální náhodné vektory $\mathbf{U}_n$ a $\mathbf{W}_n$ a pro posloupnost náhodných proměnných c_n výraz $\mathbf{U}_n = \mathbf{W}_n + o_p(c_n)$ znamená, že pro každé pevné i platí $|U_n(i) - W_n(i)| = o_p(c_n)$, $n \to \infty$, a podobně pro $O_p(\cdot)$. Základním kamenem pro přechod od Věty 3.1 k Větě 3.3 je následující lemma.

LEMMA A.3 *Předpokládejme, že funkce g má spojité derivace až do řádu $p+2$, že f je diferencovatelná a že $X_1, \ldots, X_n$ jsou vnitřní body nosiče f. Dále předpokládejme, že $h = h_n \to 0$ a $nh \to \infty$ pro $n \to \infty$. Potom*

1. $$\mathbf{S}_{p,h}\,\mathbf{g} = \begin{cases} \mathbf{g} + h^{p+1}\{\int u^{p+1}K_{(p)}(u)du \\ \qquad \times \frac{\mathbf{g}^{(p+1)}}{(p+1)!} + o_P(h^{p+1})\} & p \text{ liché} \\ \mathbf{g} + h^{p+2}\{\int u^{p+2}K_{(p)}(u)du \\ \qquad \times \left\{\frac{\mathbf{g}^{(p+1)}\mathbf{f}'}{\mathbf{f}(p+1)!} + \frac{\mathbf{g}^{(p+2)}}{(p+2)!}\right\} + o_P(h^{p+2})\} & p \text{ sudé} \end{cases}$$

2. $$\operatorname{diag}\{\mathbf{S}_{p,h}\,(\operatorname{diag}\mathbf{g})\,\mathbf{S}'_{p,h}\} = (nh)^{-1}\left\{\int K_{(p)}(u)^2du\right\}\left(\frac{\mathbf{g}}{\mathbf{f}}\right) + o_P\{(nh)^{-1}\}$$

3. $$\operatorname{diag}(\mathbf{S}_{p,h}) = O_P\{(nh)^{-1}\}$$

4. $$\mathbf{S}_{p,h}\,(\operatorname{diag}\mathbf{g})\,\mathbf{S}'_{p,h} = O_P\{(nh)^{-1}\}$$

Důkaz Body (1) a (2) jsou přímým důsledkem věty 4.1 z Ruppert a Wand [15]. Podobnými argumenty jako v uvedeném článku lze odvodit body (3) a (4).

□

Větu 3.3 lze odvodit z Věty 3.1 opakovaným použitím lemmatu A.3. To ukazuje, že dominujícím výrazem podmíněného vychýlení (3.1) je $(\mathbf{S}_2 - \mathbf{I})\mathbf{v}^2$. Protože poloha X_i je libovolná, ihned získáváme hledaný výsledek.

Vztah pro podmíněnou kovarianci vyžaduje poněkud více algebraických úprav, ale jinak je jeho odvození stejně přímé. Rozvinutím jmenovatele v (3.2) lze nahlédnout, že vedoucím členem je $S_2\{(\mathbf{T} - 3\mathbf{V}^4) + 2\mathbf{V}^4\}\mathbf{S}'_2 = \mathbf{S}_2 \operatorname{diag}(\eta)\,\mathbf{S}'_2$, kde $\eta = [\eta(X_1), \ldots, \eta(X_n)]$. Opět použitím lemmatu A.3 obdržíme výsledek věty.

A.4 Odvození vztahu (4.1)

V této části odvodíme vztah pro podmíněný rozptyl lokálně polynomického odhadu střední hodnoty $\widehat{m}(x;h) = \mathbf{S}_{p,h}\mathbf{Y}$, kde $\mathbf{S}_{p,h}$ je vyhlazovací matice použitá k odhadu. Pro zjednodušení zápisu budeme psát $\chi = (X_1, \ldots, X_n)'$ a $\widehat{\mathbf{m}} = (\widehat{m}(X_1;h), \ldots, \widehat{m}(X_n;h))$.

Z definice odhadu a vlastností kovarianční matice a podmíněné střední hodnoty dostáváme

$$\text{var}\,\{\widehat{\mathbf{m}}|\chi\} = \text{var}\,\{\mathbf{S}_{p,h}(\chi)\mathbf{Y}|\chi\} = \mathbf{S}_{p,h}\text{var}\,\{\mathbf{Y}|\chi\}\mathbf{S}'_{p,h}$$

a vzhledem k modelu (2.1) potom

$$\begin{aligned}\text{var}\,\{\widehat{\mathbf{m}}|\chi\} &= \mathbf{S}_{p,h}\text{var}\,\{\mathbf{m}+\mathbf{v}\varepsilon|\chi\}\mathbf{S}'_{p,h} = \\ &= \mathbf{S}_{p,h}\cdot\text{diag}\,\mathbf{v}\cdot\text{var}\,\{\varepsilon|\chi\}\cdot\text{diag}\,\mathbf{v}\cdot\mathbf{S}'_{p,h} = \\ &= \mathbf{S}_{p,h}\cdot\text{diag}\,\mathbf{v}\cdot\mathbf{1}\cdot\text{diag}\,\mathbf{v}\cdot\mathbf{S}'_{p,h} = \mathbf{S}_{p,h}\cdot\mathbf{V}^2\cdot\mathbf{S}'_{p,h}\,,\end{aligned}$$

kde jsme označili $\mathbf{V}^2 = \text{diag}_{\,1\le i\le n}(v^2(X_i))$.
Vyjádříme-li nyní rozptyl pouze v bodě X_i, potom s využitím vztahu pro vyhlazovací matici (2.4) získáme

$$\begin{aligned}\text{var}\,\{\widehat{m}(X_i)\} = \mathbf{e}'_1\{\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\}^{-1}\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\times \\ \times\,\mathbf{V}^2\times\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\{\mathbf{X}_p(X_i)'\mathbf{W}_h(X_i)\mathbf{X}_p(X_i)\}^{-1}\mathbf{e}_1\,.\end{aligned}$$

a nyní již stačí místo X_i dosadit libovolný bod x a místo matice $\mathbf{V}$ použít odhadnutou matici $\widehat{\mathbf{V}}^2 = \text{diag}_{\,1\le i\le n}(\widehat{v}^2(X_i))$. Tedy

$$\begin{aligned}\widehat{\text{var}}\,(h;x) = \mathbf{e}'_1\{\mathbf{X}_p(x)'\mathbf{W}_h(x)\mathbf{X}_p(x)\}^{-1}\mathbf{X}_p(x)'\mathbf{W}_h(x)\times \\ \times\,\widehat{\mathbf{V}}^2\times\mathbf{W}_h(x)\mathbf{X}_p(x)\{\mathbf{X}_p(x)'\mathbf{W}_h(x)\mathbf{X}_p(x)\}^{-1}\mathbf{e}_1\,.\end{aligned}$$

Literatura

[1] Anděl, J. (2002): Základy matematické statistiky. preprint, Praha

[2] Dette, H., Munk, A. (2002): A Consistent Test for Heteroscedasticity in Nonparametric Regression Based on the Kernel Method. *Journal of Statistical Planning and Inference* 103, str. 311–329

[3] Dette, H., Munk, A. (1998): Validation of Linear Regression Models. *The Annals of Statistics* 26, str. 778–800

[4] Dette, H., Munk, A. (1998): Testing Heteroscedasticity in Nonparametric Regresion. *Journal of the Royal Statistical Society* Series B 60, str. 693–708

[5] Eubank, R.L., Thomas, W. (1993): Detecting Heteroscedasticity in Nonparametric Regresion. *Journal of the Royal Statistical Society* Series B 55 (1), str. 145–155

[6] Fan, J., Gijbels, I. (1996): Local Polynomial Modelling and Its Applications. Chapman & Hall, London

[7] Fan, J., Gijbels, I. (1995): Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation. *Journal of the Royal Statistical Society* 57, str. 371–394

[8] Gasser, T., Sroka, L., Jennen-Steinmetz, G. (1986): Residual Variance and Residual Pattern in Nonlinear Regression. *Biometrika* 77, str. 625–633

[9] Goderniaux, A.C. (2002): Automatic Detection of Change-points in Nonparametric Regression. disertační práce, Université catholique de Louvain, Belgie

[10] Härdle, W., Bowman, A.W. (1988): Bootstraping in Nonparamtric Regression: Local Adaptive Smoothing and Confidence Bands. *Journal of the American Statistical Association* 83, str. 102–110

[11] Chiu, S.-T. (1992): Bandwidth Selection for Kernel Density Estimation. *The Annals of Statistics* 19, str. 1883–1905

[12] McCullagh, P. (1987): Tensor Methods in Statistics. Chapman & Hall, London

[13] Rice, J. (1984): Bandwidth Choice for Nonparamtric Regression. *The Annals of Statistics* 12, str. 1215–1230

[14] Ruppert, D. (1997): Empirical-bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation. *Journal of the American Statistical Association* 92, str. 1049–1062

[15] Ruppert, D., Wand, M.P. (1994): Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics* 22, str. 1346–1370

[16] Ruppert, D., Wand, M.P., Holst, U., Hoesjer, O. (1997): Local Polynomial Variance-function Estimation. *Technometrics* 39, str. 262–273

[17] Sheather, S.J. (1992): The Performance of Six popular Bandwidth Selection Methods on Some Real Data Sets. *Computational Statistics* 7, str. 225–250

[18] Šalom, L. (2004): Spojitá versus nespojitá nepramatrická regrese. diplomová práce, MFF UK Praha

[19] Wand, M.P., Jones, M.C. (1995): Kernel Smoothing. Chapman & Hall, London

[20] Zheng, J.X. (1996): A Consistent Test of a Functional Form via Nonparametric Estimation Techniques.. *Journal of Econometrics* 75, str. 263–289

Příloha I

Zdrojový kód programu

V této příloze uvádíme hlavní části zdrojového kódu programu, který byl použit pro zpracování dat z pražského Klementina (viz kapitola 6). Jednotlivé funkce jsme rozdělili do čtyř částí — funkce pro získání lokálně polynomického odhadu (podkapitola I.1), funkce pro výběr šířky okénka metodou EBBS (podkapitola I.2), funkce pro Detteho test (podkapitola I.3) a nakonec pomocné funkce (podkapitola I.4).

I.1 Lokálně polynomický odhad

```
LPEmPrim <- function(Xdata, Ydata, q, LPEpar) {
   # Lokálně polynomický odhad střední hodnoty Ydata na základě Xdata,
   #    stupeň LPEpar$p, pevný span q a jádro LPEpar$Kern
   # Vrací list(data, odhad, sirka, matice)

   h <- Span(Xdata, q, Xdata);
   Sm <- SM(Xdata, h, LPEpar);
   m <- LPE(Ydata, Sm, LPEpar);

   list(data=Ydata, odhad=m, sirka=h, matice=Sm)
}
```

```
LPE <- function(Ydata, S, LPEpar) {
   # spočítá lokálně polynomický odhad z dat Ydata
   #    pomocí vyhlazovací matice Sm
   #    s parametry LPEpar$korekce
   #        = "mean" - střední hodnota
   #          "normal" - korekce (1+S*delta)
   #          "student" - korekce (1+delta)
   #          "none" - bez korekce
   #    a korekční maticí LPEpar$Sk
   # je-li S sloupec vyhl.matice příslušný bodu Xi, potom
   #    výsledkem je odhad v bodě Xi

   if (LPEpar$korekce == "mean" | LPEpar$korekce == "none") {
       odhad <- crossprod(S, Ydata)
   } else {
       delta <- diag(crossprod(LPEpar$Sk)-2*LPEpar$Sk);
       if (LPEpar$korekce == "student") {
           odhad <- crossprod(S, Ydata/(1+delta))
       } else {
           odhad <- crossprod(S, Ydata)/(1+crossprod(S, delta))
       }
   };

   odhad
}


SM <- function(Xdata, h, LPEpar) {
   # Spočítá celou vyhlazovací matici pro data Xdata pro
   #    lokálně polynomický odhad stupně LPEpar$p s jádrem
   #    LPEpar$Kern a lokální šířkou okénka h
   # Vrací TRANSPONOVANOU MATICI - vhodné pro crossprod(Sm, Y)

   n <- length(Xdata);
   Sph <- array(0, c(n, n));
   for (i in 1:n) { Sph[,i] <- SMi(Xdata[i], Xdata, h[i], LPEpar) };

   Sph
}


SM2 <- function(Xgrid, Xdata, hgrid, LPEpar) {
   # Spočítá celou vyhlazovací matici v bodech Xgrid
   #    na základě Xdata pro lokálně polynomický odhad stupně
   #    LPEpar$p s jádrem LPEpar$Kern a lokální šířkou okénka h
   # Vrací TRANSPONOVANOU MATICI - vhodné pro crossprod(Sm, Y)

   Sph <- array(0, c(length(Xdata), length(Xgrid)));
   for (i in 1:length(Xgrid)) {
        Sph[,i] <- SMi(Xgrid[i], Xdata, hgrid[i], LPEpar)
   };

   Sph
}
```

```
SMi <- function(Xi, Xdata, hi, LPEpar) {
   # Spočítá řádek vyhlazovací matice v bode Xi na základě dat Xdata
   #    pro lokálně polynomický odhad stupně LPEpar$p
   #    s jádrem LPEpar$Kern a šířkou okénka hi
   # Vrací SLOUPEC - vhodné pro crossprod(Smi, Y)

   n <- length(Xdata);

   XminXi <- Xdata-Xi;
   Xpi <- array(1, c(n, LPEpar$p+1));
   for (j in 1:LPEpar$p) { Xpi[,j+1] <- XminXi^j };

   Whi <- diag(LPEpar$Kern(XminXi/hi)/hi);

   WhiXpi <- crossprod(Whi, Xpi);

   t(crossprod(ei(1,LPEpar$p+1), solve(crossprod(Xpi, WhiXpi)))
     %*% t(WhiXpi))
}
```

I.2 Metoda EBBS

```
EBBS <- function(Xdata, Ydata, Yvar,
                 GridXpar, GridHpar, LPEpar, MSEpar, Hpar) {
   # LPO dat Ydata s rozptylem Yvar na základě Xdata
   #    s automatickou volbou lokální šířky okénka metodou a parametry
   #      GridXpar(pocet) - počet bodů v mřížce X
   #      GridH1par(pocet, dolni, horni) - počet bodů v mřížce šířek
   #                                       a její rozsah
   #      LPEpar(p, Kern, korekce, Sk) - stupeň polynomu, jádro,
   #                                     korekce a korekční matice
   #                                     - viz funkce LPE
   #      MSEpar(t,J1,J2) - stupeň polynomu a počet generovaných bodů
   #                        před a za "h" pro odhad MSE
   #      Hpar(Kern, Bandspan) - jádro pro vyhlazení šířek
   #  Vrací list(data, odhad, sirka, matice)

   Xgrid <- GridX(Xdata, GridXpar);
   Hgrid <- numeric();
   MSE <- numeric();

   for (i in 1:length(Xgrid)) {
        H1grid <- GridH(Xgrid[i], Xdata, GridHpar);
        Hgrid <- c(Hgrid, H1grid);

        hatYgridH1 <- numeric();
        var <- numeric();
        bias <- numeric();

        for (j in 1:length(H1grid)) {
             Sij <- SMi(Xgrid[i], Xdata, H1grid[j], LPEpar);
             hatYgridH1 <- c(hatYgridH1, LPE(Ydata, Sij, LPEpar));
             var <- c(var, MSEvar(Yvar, Sij, LPEpar));
        }
```

```
        J1star <- max(0, MSEpar$J1)+1;
        J2star <- length(H1grid)-MSEpar$J2
        for (j in J1star:J2star) {
             bias <- c(bias, MSEbias(j, H1grid, hatYgridH1, MSEpar))
        }
        var <- var[J1star:J2star];
        MSE <- c(MSE, var + bias^2)
    }
    attr(MSE, "dim") <- c(J2star-J1star+1, length(Xgrid));
    attr(Hgrid, "dim") <- c(GridH1par$pocet, length(Xgrid));
    Hgrid <- Hgrid[J1star:J2star,];

    hatHgrid <- numeric();
    for (i in 1:length(Xgrid)) {
        hatHgrid <- c(hatHgrid, Hgrid[PrvniLokMin(MSE[,i]), i])
    }

    Sgrid <- SM2(Xgrid, Xdata, hatHgrid, LPEpar);
    hatYgrid <- LPE(Ydata, Sgrid, LPEpar);

    hatY <- Spline(hatYgrid, Xgrid, Xdata);
    hatH <- rep(0, length(Xdata));
    hatH[Indexy(Xgrid, Xdata)] <- hatHgrid;

    list(data=Ydata, odhad=hatY, grid=Xgrid, sirka=hatH, matice=Sgrid)
}



MSEvar <- function(var, Si, LPEpar) {
    # spočítá rozptyl odhadu dat v bodě Xi, pro který byla spočtena
    #    matice Si s použitím rozptylu dat var

    if (LPEpar$korekce == "mean" | LPEpar$korekce == "none") {
        odhad <- crossprod(Si, crossprod(diag(var), Si))
    } else {
        delta <- diag(crossprod(LPEpar$Sk)-2*LPEpar$Sk);
        if (LPEpar$korekce == "student") {
            odhad <- crossprod(Si, crossprod(diag(var), Si))
                     /(1+crossprod(Si, delta))^2
        } else {
            odhad <- crossprod(Si, crossprod(diag(var/(1+delta)), Si))
        }
    };

    odhad
}
```

```
MSEbias <- function(j, Hgrid, hatYgridH, MSEpar) {
   # odhad bias nejmenšími čtverci na Hgrid, harYgridH pro šířku h0
   #    s parametry MSEpar(t, J1, J2)

   Hdata <- Hgrid[(j-MSEpar$J1):(j+MSEpar$J2)];
   Ydata <- hatYgridH[(j-MSEpar$J1):(j+MSEpar$J2)];

   Ht <- array(1, c(MSEpar$J1+MSEpar$J2+1, MSEpar$t));
   for (k in 1:MSEpar$t) {
        Ht[, k] <- Hdata^k
   };

   b <- coef(lm(Ydata ~ Ht));

   crossprod(b[-1], Hgrid[j]^(1:MSEpar$t))
}
```

I.3 Detteho test

```
Dette <- function(Xdata, m, Dettepar) {
   # Provede Detteho test homoskedasticity na datech Xdata, m$data
   #    s využitím odhadu střední hodnoty m$odhad
   # Parametry Dettepar (Kern, sirka, Bootstrap) - jádro, šířka okénka
   #    a počet opakování bootstrapu
   # Vrací p-value odpovídající statistice z dat
   #    v bootstrapovém souboru statistik

   Tn <- DetteStat(Xdata, m$data, Dettepar);
   Tn;

   BootTn <- numeric();
   R <- m$data - m$odhad;

   for (i in 1:Dettepar$Bootstrap) {
        Rsample <- sample(R, replace=TRUE);
        Ysample <- m$odhad + Rsample;
        BootTn <- c(BootTn, DetteStat(Xdata, Ysample, Dettepar))
   }

   BootTn <- sort(abs(BootTn));
   BootTn;
   Pvalue(abs(Tn), BootTn)
}
```

```
DetteStat <- function(Xdata, Ydata, Dettepar) {
   # Spočítá statistiku pro Detteho test homoskedasticity z dat
   #    Xdata,Ydata s parametry Dettepar

   n <- length(Ydata);
   Xin <- (Xdata-1)/(n-1);

   R2in <- (Ydata[2:n]-Ydata[1:(n-1)])^2/2;
   R2n <- mean(R2in);

   Stat <- 0;
   for (i in 1:(n-3)) {
        for (j in (i+2):(n-1)) {
             Stat <- Stat +
                      Dettepar$Kern((Xin[i]-Xin[j])/Dettepar$sirka)
                      *(R2in[i]-R2n)*(R2in[j]-R2n)
        }
   }
   Stat <- 2*Stat/((n-1)*(n-2)*Dettepar$sirka);

   Stat
}



DetteH <- function(v) {
   # určí optimální šířku okénka pro Detteho test

   mean(v$odhad)^(1/5)
}
```

I.4 Pomocné funkce

```
ei <- function(i, n) {
   # Vytvoří vektor délky n s jedničkou na i-tém místě

   ei <- rep(0, times=n);
   ei[i] <- 1;
   ei
}



Indexy <- function(X, Xdata) {
   # Vyhledá pozice X v Xdata

   ind <- numeric();
   for (i in 1:length(X)) {
        ind <- c(ind, which(X[i]==Xdata))
   }

   ind
}
```

```
Span <- function(X, q, Xdata) {
   # Najde lokální šířky v bodech X odpovídající q*100% dat z Xdata

   n <-length(Xdata);
   pocet <- ceiling(n*q);
   indexy <- Indexy(X, Xdata);

   h <- rep(floor(pocet/2), times=n);
   h[indexy < pocet/2] <- pocet - indexy[indexy < pocet/2];
   h[indexy > n-pocet/2+1] <- indexy[indexy > n-pocet/2+1] - (n-pocet+1);

   h[indexy]
}



GridX <- function(Xdata, GridXpar) {
   # vrátí mřížku dat z Xdata o počtu GridXpar$pocet bodů

   n <- length(Xdata);
   dist <- floor((n-1)/(GridXpar$pocet-1));
   last <- n - 1 - (GridXpar$pocet-1)*dist;
   first <- GridXpar$pocet - last - 1;

   if (last == 0) {
       indexy <- c(1, 1+dist*1:first);
   } else if (first == 0) {
       indexy <- c(1, 1+(dist+1)*1:last);
   } else {
       indexy <- c(1, 1+dist*1:first, 1+dist*first + (dist+1)*1:last);
   }

   Xdata[indexy]
}



GridH <- function(Xi, Xdata, GridHpar) {
   # vrátí mřížku hodnot šířek okénka pro bod Xi na základě
   #    dat Xdata s počtem bodů GridHpar$pocet
   #    od span GridHpar$dolni do GridHpar$horni
   # body jsou rovnoměrně logaritmicky rozdělené

   dolnilog <- log(Span(Xi, GridHpar$dolni, Xdata));
   hornilog <- log(Span(Xi, GridHpar$horni, Xdata));

   floor(exp(dolnilog + (hornilog-dolnilog)
                        /(GridHpar$pocet-1)*(1:GridHpar$pocet-1)))
}
```

```
Spline <- function(Ygrid, Xgrid, Xdata) {
   # Proloží Xgrid,Hgrid kubickým splajnem a vrátí
   #    vektor odhadnutých dat pro SplineH.Xdata

   predict(interpSpline(Xgrid, Ygrid), Xdata)$y
}



PrvniLokMin <- function(Ydata) {
   # najde první lokální minimum v PrvniLokMin.Ydata, vrací jeho pozici

   last <- Ydata[1];

   for (i in 1:(length(Ydata)-1)) {
        if (Ydata[i+1] >= last) { break }
        else { last <- Ydata[i] }
   }

   i;
}



Pvalue <- function(Xi, Xdata) {
   # Spočítá 1-hodnotu empirické distribuční funkce dat Xdata v bodě Xi

   n <- length(Xdata);

   if (Xdata[n] < Xi) { alpha <- 0 }
   else { for (i in 1:n) { if (Xdata[i] > Xi) {break} };
          alpha <- 1-(i-1)/n }

   alpha
}
```