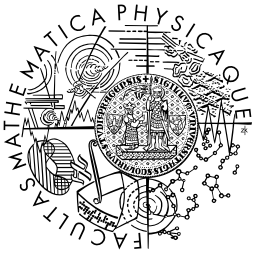


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Martin Měsíček

Neparametrické regresní odhady

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	doc. RNDr. Zdeněk Hlávka, Ph.D.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie

Praha 2017

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

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

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

Název práce: Neparametrické regresní odhady

Autor: Bc. Martin Měsíček

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Zdeněk Hlávka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá lokálně polynomickými odhady funkce podmíněného rozptylu v heteroskedastickém neparametrickém regresním modelu. Předpokládáme jistou hladkost regresní a rozptylové funkce, nikoliv však jejich příslušnost do nějaké parametrické rodiny. Základní idea je použít lokálně lineární regresi na kvadrát reziduí. Takový odhad má pak vysokou minimax eficienci a je adaptivní k neznámé regresní funkci. Nicméně při praktickém použití může nabývat záporných hodnot, což pro odhad rozptylu nedává smysl. Proto Xu a Phillips představili nový odhad rozptylu, který je asymptoticky ekvivalentní lokálně lineárnímu odhadu rozptylu pro vnitřní body a zároveň má zaručenu nezápornost. My jsme navíc srovnali asymptotiku obou odhadů pro hraniční body a prokázali podstatně lepší chování lokálně lineárního odhadu v těchto bodech. To nás motivovalo k představení modifikace lokálně lineárního odhadu, která zaručuje jeho nezápornost. Na závěr jsme srovnali všechny zmíněné odhady v simulační studii.

Klíčová slova: Podmíněný rozptyl; Lokálně lineární odhad; Eficientní odhad.

Title: Nonparametric regression estimators

Author: Bc. Martin Měsíček

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Zdeněk Hlávka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis is focused on local polynomial smoothers of the conditional variance function in a heteroscedastic nonparametric regression model. Both mean and variance functions are assumed to be smooth, but neither is assumed to be in a parametric family. The basic idea is to apply a local linear regression to squared residuals. This method, as we have shown, has high minimax efficiency and it is fully adaptive to the unknown conditional mean function. However, the local linear estimator may give negative values in finite samples which makes variance estimation impossible. Hence Xu and Phillips proposed a new variance estimator that is asymptotically equivalent to the local linear estimator for interior points but is guaranteed to be non-negative. We also established asymptotic results of both estimators for boundary points and proved better asymptotic behavior of the local linear estimator. That motivated us to propose a modification of the local linear estimator that guarantees non-negativity. Finally, simulations are conducted to evaluate the finite sample performances of the mentioned estimators.

Keywords: Conditional variance; Local linear estimator; Efficient estimator.

Rád bych na tomto místě poděkoval panu doc. RNDr. Zdeňkovi Hlávkovi, Ph.D. za důležité poznámky a rady při psaní tohoto textu.

Obsah

Úvod

V dnešní době máme k dispozici obrovské množství dat, která lze chápat jako realizace náhodných veličin. Abychom z těchto dat dokázali vytěžit maximum informací, je třeba zkoumat závislosti mezi jednotlivými veličinami. Nyní přichází na řadu regresní statistická analýza. Umožňuje nám v datech hledat různé trendy, periodicity nebo dokonce předpovídat budoucí realizace.

Budeme předpokládat, že námi vysvětlovaná náhodná veličina lze rozdělit na deterministickou část (reprezentovanou regresní funkcí) a náhodnou složku, u které předpokládáme nulovou střední hodnotu. Dále rozlišujeme dvě situace, buďto předpokládáme konstantní rozptyl náhodné složky (jedná se o tzv. homoskedastický model), a nebo ho modelujeme pomocí rozptylové funkce. Nás bude zajímat především druhý případ, který má významné uplatnění ve financích, neboť rozptylem můžeme charakterizovat riziko a je nezbytné jej umět správně odhadnout.

Regresní analýzu dělíme na parametrickou a neparametrickou. Parametrická regrese je častěji používaná, ale je nutné a priori znát charakter regresní funkce. Nejznámějším příkladem je lineární regrese, kdy vysvětlovanou proměnnou modelujeme jako lineární kombinaci předem známých funkcí a odhadujeme k nim příslušné koeficienty. V případě, že charakter regresní funkce dopředu neznáme, je vhodnější použít některou z metod neparametrické regrese. Nehrozí nesprávná specifikace modelu a z toho plynoucí vysoké vychýlení odhadu. V této práci se zaměříme především na metody jádrové neparametrické regrese pro odhad rozptylové funkce.

Diplomová práce je rozdělena do dvou kapitol. První kapitola se zabývá odhady regresní funkce, uvede nás do problematiky jádrových odhadů a především se zaměří na lokálně lineární odhad a jeho asymptotické vlastnosti. Položí teoretický základ pro druhou kapitolu. Mimo jiné také nastíní možnosti volby vyhlazovacího parametru.

Druhá kapitola se zabývá odhady rozptylové funkce. Jedná se o stěžejní kapitolu, neboť cílem celé práce je předložit čtenáři co možná nejlepší jádrový odhad rozptylové funkce. Prezentujeme v ní odhady založené na reziduích a srovnáváme jejich asymptotické vlastnosti. Kvůli absenci asymptotických výsledků v hraničních bodech si sami zformulujeme a dokážeme větu o podmíněném asymptotickém vychýlení a rozptylu v těchto bodech pro uvedené odhady. Na závěr srovnáme prezentované odhady na simulovaných datech.

Diplomová práce je napsána v systému LaTeX a na praktické výpočty používáme program R Core Team (2015). Všechny naprogramované funkce jsou součástí přiloženého CD.

1. Odhady regresní funkce

1.1 Formulace modelu

V celé této práci budeme uvažovat následující model

$$Y_t = m(X_t) + \sigma(X_t)\epsilon_t, \tag{1.1}$$

kde $\{(X_t, Y_t), t = 1, ..., n\}$ je dvoudimenzionální striktně stacionární náhodný proces a $\{\epsilon_t, t = 1, \ldots, n\}$ je náhodná složka splňující $\mathsf{E}[\epsilon_t|X_t] = 0$, $\mathrm{var}[\epsilon_t|X_t] = 1$. Dále $m(x) = \mathsf{E}[Y_t|X_t = x]$ je *regresní funkce* a $\sigma^2(x) = \mathrm{var}[Y_t|X_t = x]$ *rozptylová funkce*. Hustotu X označme f_X. Později budeme na tento model klást další předpoklady.

Model nazveme *parametrickým regresním modelem*, pokud a priori předpokládáme, že regresní funkce m náleží do parametrické rodiny $\{g(x, \theta) : \theta \in \Theta\}$, kde $g(\cdot, \cdot)$ je předem daná funkce a Θ je podmnožina $\mathbb{R}^k$, k fixní a nezávislé na n. Potom odhad funkce $m(\cdot)$ je ekvivalentní odhadu konečně rozměrného parametru θ. My však a priori žádnou informaci o $m(\cdot)$ nemáme, budeme se tedy zabývat *neparametrickým regresním modelem*. Model dále nazveme *homoskedastickým*, je-li rozptylová funkce $\sigma^2(\cdot)$ konstantní, v opačném případě jej nazveme *heteroskedastickým*.

Cílem této práce je srovnání různých odhadů $\sigma^2(\cdot)$, a to jak teoreticky, porovnáním asymptotického chování, tak pomocí simulačních studií. Jak by takový odhad mohl vypadat? Z rozkladu $\sigma^2(x) = \mathsf{E}[Y_t^2|X_t = x] - m^2(x)$ dostáváme přímočarý odhad

$$\hat{\sigma}_d^2(x) = \hat{v}(x) - \{\hat{m}(x)\}^2, \tag{1.2}$$

kde $\hat{m}(x)$ a $\hat{v}(x)$ jsou po řadě odhady $m(x)$ a $v(x) = \mathsf{E}(Y_t^2|X_t = x)$. Nicméně stále nevíme, jak bychom takové odhady mohli neparametricky získat. Proto se dále zaměříme na tzv. *lokálně polynomické odhady*, nejprve však definujeme základní pojmy.

1.2 Základní pojmy

Definice 1. *Odhad $\hat{m}_n(x)$ funkce $m(x)$ nazveme lineárním neparametrickým odhadem, jestliže může být zapsán ve tvaru*

$$\hat{m}_n(x) = \sum_{i=1}^{n} Y_i W_{ni}(x),$$

kde váhy $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ závisí pouze na n, i, x a na hodnotách $X_1, \ldots, X_n$.

Odhadovat regresní funkci lokálně nám umožní tzv. *jádrová funkce*.

Definice 2. *Nechť funkce $K : \mathbb{R} \to [0, \infty]$ je integrovatelná a splňuje $\int K(u)\,du = 1$, a $K(-u) = K(u)$ pro všechna u, pak ji nazýváme jádro nebo jádrová funkce.*

Příklady klasických jádrových funkcí:

(a) Trojúhelníkové jádro $K(x) = (1 - |x|)\,\mathrm{I}_{[|x|\leq 1]}$

(b) Rovnoměrné jádro $K(x) = \frac{1}{2}\,\mathrm{I}_{[|x|\leq 1]}$

(c) Epanechnikovo jádro $K(x) = \frac{3}{4}(1 - x^2)\,\mathrm{I}_{[|x|\leq 1]}$

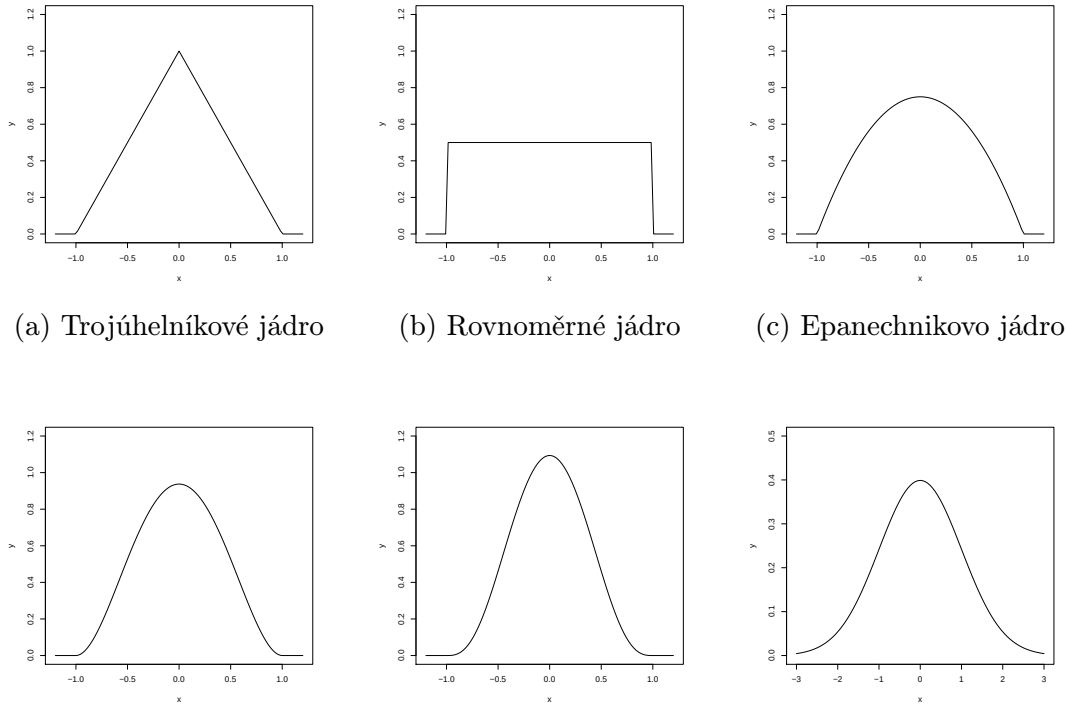
(d) Biweight jádro $K(x) = \frac{15}{16}(1 - x^2)^2\,\mathrm{I}_{[|x|\leq 1]}$

(e) Triweight jádro $K(x) = \frac{35}{32}(1 - x^2)^3\,\mathrm{I}_{[|x|\leq 1]}$

(f) Gaussovo jádro $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

Poznámka. Symbolem K_h rozumíme

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right), \ kde \quad h > 0.$$



(a) Trojúhelníkové jádro (b) Rovnoměrné jádro (c) Epanechnikovo jádro

(d) Biweight jádro (e) Triweight jádro (f) Gaussovo jádro

Obrázek 1.1: Příklady jádrových funkcí

1.3 Lokálně polynomická regrese

V této podkapitole přiblížíme metody lokálně polynomické regrese. Čerpáme z Fan a Gijbels (1996). Lokálně polynomická regrese má řadu výhod. Funguje dobře jak pro náhodný, tak i pro pevný design, jak pro rovnoměrně rozdělené regresory, tak i pro rozdělení s velkými shluky. Navíc lokálně polynomické odhady s lichým stupněm p, na rozdíl od jiných metod, netrpí tzv. *hraničními efekty*,

tedy řád vychýlení a rozptylu je stejný jak pro tzv. *vnitřní*, tak i pro tzv. *hraniční body*. To bylo empiricky doloženo už v článku Tibshirani a Hastie (1987) pro lokálně lineární odhad. Z výpočetního hlediska je výhodou také jejich jednoduchost. Řád operační složitosti algoritmů lokálně polynomického vyhlazování (jsou-li správně naprogramovány) je lineární, jak bylo ukázáno ve Fan a Marron (1994). A v neposlední řadě mají lokálně polynomiální odhady velmi dobrou eficienci. Jak ukážeme později, lokálně lineární odhad je dokonce v určitém smyslu asymptoticky eficientní mezi lineárními odhady.

Předpokládejme, že m je p-krát diferencovatelná na okolí bodu x, pak můžeme aproximovat $m(X_i)$ Taylorovým polynomem stupně p se středem v bodě x:

$$m(X_i) \doteq m(x) + m'(x)(X_i - x) + \frac{m''(x)}{2!}(X_i - x)^2 + \ldots + \frac{m^p(x)}{p!}(X_i - x)^p.$$

Označíme nejprve:

$$\beta_j(x) = \frac{m^{(j)}(x)}{j!}, \quad j = 0, \ldots, p, \quad \boldsymbol{\beta}(x) = (\beta_0(x), \ldots, \beta_p(x)).$$

Pomocí vážené metody nejmenších čtverců, kde funkci váhy plní nějaká jádrová funkce K, odhadujeme $\boldsymbol{\beta}(x)$ následovně:

$$\hat{\boldsymbol{\beta}}(x) = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} [Y_i - b_0 - b_1(X_i - x) - \ldots - b_p(X_i - x)^p]^2 \, K\left(\frac{X_i - x}{h_n}\right),$$

kde posloupnost h_n je tzv. *vyhlazovací parametr* a splňuje $0 < h_n \to 0$, $nh_n \to \infty$, pro $n \to \infty$. Odhadem regresní funkce $m(x)$ tedy je $\hat{m}_n(x) = \hat{\beta}_0(x)$. Může se nám hodit i maticový zápis $\hat{\boldsymbol{\beta}}(x)$, označme:

$$\boldsymbol{X}_p(x) = \begin{pmatrix} 1 & (X_1 - x) & \ldots & (X_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x) & \ldots & (X_n - x)^p \end{pmatrix},$$

$\boldsymbol{y} = (Y_1, \ldots, Y_n)^T$ a $\boldsymbol{K_{h_n}}(x) = \mathrm{diag}_{1\le i\le n}\{K_{h_n}(X_i - x)\}$, pak můžeme psát

$$\hat{\boldsymbol{\beta}}(x) = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p+1}} (\boldsymbol{y} - \boldsymbol{X}_p(x)\boldsymbol{b})^T \boldsymbol{K_{h_n}}(x)(\boldsymbol{y} - \boldsymbol{X}_p(x)\boldsymbol{b})$$

a pomocí teorie nejmenších čtverců dostáváme

$$\hat{m}_n(x) = \hat{\beta}_0(x) = \boldsymbol{e}_1^T(\boldsymbol{X}_p^T(x)\boldsymbol{K_{h_n}}(x)\boldsymbol{X}_p(x))^{-1}\boldsymbol{X}_p^T(x)\boldsymbol{K_{h_n}}(x)\boldsymbol{y},$$

kde $\boldsymbol{e}_1^T = (1, 0, 0, \ldots, 0)$. Dále označme:

$$\boldsymbol{m} = (m(X_1), \ldots, m(X_n))^T, \quad \hat{\boldsymbol{m}}_n = (\hat{m}_n(X_1), \ldots, \hat{m}_n(X_n))^T,$$

pak $\hat{\boldsymbol{m}}_n = \boldsymbol{S}_{p,h_n}\boldsymbol{y}$, kde $\boldsymbol{S}_{p,h_n}$ je tzv. *vyhlazovací matice* lokálně polynomiální regrese stupně p s vyhlazovacím parametrem h_n, tj.:

$$(\boldsymbol{S}_{p,h_n})_{ij} = \boldsymbol{e}_1^T(\boldsymbol{X}_p^T(X_i)\boldsymbol{K_{h_n}}(X_i)\boldsymbol{X}_p(X_i))^{-1}\boldsymbol{X}_p^T(X_i)\boldsymbol{K_{h_n}}(X_i)\boldsymbol{e}_j.$$

Lokálně polynomiální regrese s sebou přináší další volby, které musíme učinit. Především je to volba vyhlazovacího parametru, která hraje zásadní roli. Příliš

velký vyhlazovací parametr způsobí, že do odhadu zahrneme příliš vzdálená pozorování, v důsledku čehož se zvýší vychýlení odhadu. Takovou situaci nazýváme *přehlazením odhadu*. Podmíněným *vychýlením odhadu* myslíme:

$$\operatorname{bias}[\hat{m}_n(x)|\boldsymbol{X}] = \mathsf{E}\left[\hat{m}_n(x)\right) - m(x)|\boldsymbol{X}] = \mathsf{E}\left[\hat{m}_n(x))|\boldsymbol{X}\right] - m(x).$$

Naopak příliš malý vyhlazovací parametr způsobí, že do odhadu budou zahrnuta jen nejbližší pozorování, tedy vychýlení se sníží, ale na úkor rostoucího rozptylu. Takovou situaci nazýváme *podhlazení odhadu*. Vyhlazovacímu parametru se věnuje sekce 1.4.

Dále je třeba zvolit stupeň polynomické regrese. To nám částečně usnadní asymptotické výsledky z Fan a Gijbels (1996):

$$\operatorname{bias}[\hat{m}_n(x)|\boldsymbol{X}] = \begin{cases} O_P(h_n^{p+1}), & p = 2k+1, \\ O_P(h_n^{p+2}), & p = 2k, \end{cases}$$

$$\operatorname{var}[\hat{m}_n(x)|\boldsymbol{X}] = O_P\left(\frac{1}{nh_n}\right), \quad p = k, \quad k \in \mathbb{N}_0.$$

Tedy přechod z $p = 2k+1$ na $p+1$ snižuje řád podmíněného asymptotického vychýlení. Dále je patrné, že stupeň polynomické regrese nemá vliv na řád podmíněného asymptotického rozptylu, nicméně jeho konstantní člen se zvyšuje při přechodu z $p = 2k+1$ na $p+1$, jak je ukázáno ve Fan a Gijbels (1995). Naopak při přechodu z $p = 2k$ na $p+1$ se konstantní člen podmíněného asymptotického rozptylu nemění, nemění se ani řád podmíněného asymptotického vychýlení, nicméně v důsledku parametru navíc se naskýtá příležitost k redukci jeho konstantního členu (zejména v hraničních bodech). Tohle můžeme například nahlédnout srovnáním asymptotického vychýlení lokálně konstantního a lokálně lineárního odhadu. Navíc odhady se sudým stupněm polynomické regrese dosahují nižší eficience. Proto se doporučuje lichý stupeň. Protože však neexistuje žádné obecné srovnání lokálně polynomiálních modelů lichých stupňů, můžeme v literatuře (např. (Fan a Gijbels, 1996)) nalézt algoritmy na volbu stupňů polynomiální regrese. Připomeňme však, že komplexnost modelu regulujeme i volbou vyhlazovacího parametru, proto Fan a Gijbels doporučují použít nejnižší liché stupně, tedy $p = 1$ nebo výjimečně $p = 3$.

Do třetice je třeba zvolit jádrovou funkci. Vliv této volby na kvalitu odhadu je zřejmě nejmenší, jak je ukázano v článku Bowman a Azzalini (1997). Nicméně autoři článku Fan a kol. (1995) dokázali, že Epanechnikova jádrová funkce minimalizuje podmíněnou asymptotickou střední čtvercovou chybu polynomiálního odhadu lichého stupně (za použití asymptotického lokálně optimálního vyhlazovacího parametru) mezi všemi nezápornými, symetrickými, Lipschitzovskými spojitými funkcemi. Epanechnikova jádrová funkce je tedy ve zmíněném smyslu optimální pro lokálně polynomickou regresi. Naopak nevýhodou může být, že už její třetí derivace je identicky rovná nule, proto se v některých aplikacích dává přednost např. Gaussově jádrové funkci, která má nekonečně mnoho (nenulových) derivací.

1.3.1 Nadaraya-Watsonův odhad

Již v roce 1964 navrhli Nadaraya a Watson, oba nezávisle na sobě v článcích Nadaraya (1964) a Watson (1964), lokálně konstantní odhad regresní funkce.

Tedy, jedná se o lokálně polynomiální regresi stupně $p = 0$, tj.:

$$\hat{\beta}_0(x) = \arg\min_{b_0 \in \mathbb{R}} \sum_{i=1}^{n} [Y_i - b_0]^2 K\left(\frac{X_i - x}{h_n}\right) = \sum_{i=1}^{n} Y_i W_{ni}^{NW}(x) = \hat{m}_n^{NW}(x), \quad (1.3)$$

kde

$$W_{ni}^{NW}(x) = \frac{K(\frac{X_i - x}{h_n})}{\sum_{j=1}^{n} K(\frac{X_j - x}{h_n})}, \quad (1.4)$$

jestliže

$$\sum_{j=1}^{n} K\left(\frac{X_j - x}{h_n}\right) > 0. \quad (1.5)$$

Všimněme si, že $\sum_{i=1}^{n} W_{ni}^{NW}(x) = 1$. Výhodou může být, že Nadaraya-Watsonův odhad (dále jen NW odhad) vždy leží v rozmezí minimální a maximální pozorované odezvy a je velmi snadný na implementaci.

1.3.2 Lokálně lineární odhad

Idea lokálně lineárního odhadu (dále jen LL odhadu) byla poprvé prezentována v článku Stone (1977) pro rovnoměrné jádro a studována v článku Cleveland (1979). Jedná se o lokálně polynomiální regresi stupně $p = 1$, tj.:

$$\hat{\beta}(x) = \arg\min_{(b_0,b_1)^T} \sum_{i=1}^{n} [Y_i - b_0 - b_1(X_i - x)]^2 K\left(\frac{X_i - x}{h_n}\right),$$

$$\hat{m}_n^{LL}(x) = \hat{\beta}_0(x) = \sum_{i=1}^{n} Y_i W_{ni}^{LL}(x), \quad (1.6)$$

kde

$$W_{ni}^{LL}(x) = \frac{\frac{1}{nh_n} K(\frac{X_i - x}{h_n})[S_{n,2}(x) - \frac{X_i - x}{h_n} S_{n,1}(x)]}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)}, \quad i = 1, \ldots, n, \quad (1.7)$$

kde

$$S_{n,l} = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right)\left(\frac{X_i - x}{h_n}\right)^l, \quad l = 0, 1, 2. \quad (1.8)$$

Opět si můžeme lehce ověřit, že $\sum_{i=1}^{n} W_{ni}^{LL}(x) = 1$. Váhy LL odhadu dále splňují:

$$\sum_{i=1}^{n} (X_i - x) W_{ni}^{LL}(x) = 0. \quad (1.9)$$

Vlastnost (1.9) je považována za klíčovou pro redukci vychýlení odhadu, neboť díky ní vychýlení LL odhadu nezávisí na první derivaci regresní funkce $m(\cdot)$. Můžeme to nahlédnout z vyjádření podmíněného vychýlení:

$$\text{bias}[\hat{m}_n^{LL}(x)|\boldsymbol{X}] = \mathsf{E}\,[\sum_{i=1}^{n} Y_i W_{ni}^{LL}(x)|\boldsymbol{X}] - m(x) = \sum_{i=1}^{n}(m(X_i) - m(x))W_{ni}^{LL}(x), \tag{1.10}$$

pokud navíc uděláme Taylorův rozvoj pro $m(X_i)$ v bodě x, první derivace regresní funkce $m(\cdot)$ nebude ve vychýlení lokálně lineárního odhadu obsažena.

Lokálně lineární odhad patří k nejoblíbenějším neparametrickým odhadům především kvůli jeho skvělým asymptotickým vlastnostem. Protože z něj vycházejí i metody pro odhad rozptylové funkce, kterými se budeme zabývat později, popíšeme v následujících podkapitolách jeho asymptotické vlastnosti podrobněji. K tomu budeme potřebovat, aby pro model (1.1) platily navíc následující podmínky:

(a) $(X_1, Y_1), \ldots, (X_n, Y_n)$ jsou i.i.d. vektory z populace (X, Y).

(b) Vektor (X, Y) má spojitou neznámou hustotu $f(\cdot, \cdot)$.

Asymptotika LL odhadu

Asymptotikou odhadu myslíme automaticky nepodmíněnou asymptotiku. Z Fan a Gijbels (1996) (§ 6.2.2) plyne, že LL odhad je (za určitých předpokladů) asymptoticky normální a platí:

$$\text{bias}[\hat{m}_n^{LL}(x_0)] = \frac{1}{2}h_n^2 \left(\int u^2 K(u)\,du\right) m''(x_0) + o(h_n^2),$$
$$\text{var}[\hat{m}_n^{LL}(x_0)] = \frac{\sigma^2(x_0)\int K^2(u)\,du}{f_X(x_0)nh_n} + o\left(\frac{1}{nh_n}\right).$$

Eficience LL odhadu

Většina neparametrických odhadů regresní funkce je váženým průměrem odezvy, tedy splňuje definici 1 lineárního odhadu. Patří sem NW odhad, LL odhad, ale také např. vyhlazovací spliny. Abychom dokázali odpovědět na otázku, jak si vede LL odhad vůči ostatním lineárním odhadům a také vůči libovolným odhadům regresní funkce, budeme studovat tzv. *minimax risk*. Ukážeme, že LL odhad má vysokou minimax eficienci, mezi lineárními odhady dokonce 100%.

Definice 3. *Je-li F neparametrická třída hustot obsahující neznámou hustotu f vektoru (X, Y), definujeme maximální riziko odhadu $\hat{m}_n(x_0)$ přes všechny hustoty z F jako*

$$r(\hat{m}_n, F) = \sup_{f\in F} \text{MSE}_f[\hat{m}_n(x_0)] = \sup_{f\in F} \mathsf{E}_f[\hat{m}_n(x_0) - m(x_0)]^2,$$

kde E_f značí střední hodnotu vzhledem k hustotě f.

Pro výpočet maximálního rizika odhadu $\hat{m}_n(x_0)$ využíváme známého vztahu pro střední čtvercovou chybu:

$$\text{MSE}[\hat{m}_n(x_0)] = \text{var}[\hat{m}_n(x_0)] + \left(\text{bias}[\hat{m}_n(x_0)]\right)^2.$$

Definice 4. *Minimax risk přes všechny hustoty z F definujeme jako*

$$R_n(F) = \inf_{\hat{m}_n} \sup_{f \in F} \mathrm{MSE}_f[\hat{m}_n(x_0)] = \inf_{\hat{m}_n} r(\hat{m}_n, F).$$

Definice 5. *Kladnou posloupnost $\{a_n\}_{n=1}^{\infty}$ konvergující k nule, nazveme optimální řád konvergence odhadů vzhledem k F, jestliže existují $C < \infty$ a $c > 0$ tak, že:*

$$\limsup_{n\to\infty} a_n^{-2} R_n(F) \leq C, \quad \liminf_{n\to\infty} a_n^{-2} R_n(F) \geq c.$$

Definice 6. *Odhad $\hat{m}_n^*$ nazveme odhadem s optimálním řádem konvergence vzhledem k F, jestliže platí*

$$r(\hat{m}_n^*, F) \leq C a_n^2,$$

kde $\{a_n\}_{n=1}^{\infty}$ je optimální řád konvergence vzhledem k F a $C < \infty$ je konstanta.

Věta 1. *Nechť platí podmínky (a) a (b) a nechť jádrová funkce $K(\cdot)$ je omezená, spojitá a splňuje:*

$$\int u^2 K(u)\, du \neq 0, \quad \int u^{2r} K(u)\, du < \infty, \quad r = 1, 2, \ldots$$

a nechť $h_n = dn^{-\beta}$, kde d, β jsou konstanty $d > 0, 0 < \beta < 1$. Pak

$$r(\hat{m}_n^{LL}, \mathcal{L}_2) = \sup_{f \in \mathcal{L}_2} \mathrm{MSE}_f[\hat{m}_n^{LL}(x_0)] \leq \tag{1.11}$$

$$\leq \frac{1}{4} h_n^4 C^2 \left[\int u^2 K(u)\, du\right]^2 + \frac{B \int K^2(u)\, du}{bnh_n} + o\left(h_n^4 + \frac{1}{nh_n}\right), \tag{1.12}$$

kde C, C^*, B, b, c, α jsou kladné konstanty a

$$\mathcal{L}_2 = \{f(\cdot,\cdot) : |m(x) - m(x_0) - m'(x_0)(x - x_0)| \leq C(x - x_0)^2/2, |m(x_0)| \leq C^*\}$$
$$\cap \{f(\cdot,\cdot) : \sigma^2(x) \leq B, f_X(x_0) \geq b, |f_X(x) - f_X(y)| \leq c|x - y|^{\alpha}\}.$$

Důkaz. Ve Fan (1993). □

Poznámky ke třídě hustot $\mathcal{L}_2$.

- První podmínka omezuje shora $m''(x_0)$, třetí podmínka $\sigma^2(x_0)$ a čtvrtá podmínka omezuje $f_X(x_0)$ pro změnu zdola, tedy třída hustot je formulovaná tak, aby nám umožnila shora odhadnout $r(\hat{m}_n^{LL}, \mathcal{L}_2)$. Lipschitzovská podmínka na $f_X(\cdot)$ je přidána pouze z technických důvodů.
- Naopak třída hustot $\mathcal{L}_2$ neomezuje $m'(x_0)$. Pokud vychýlení odhadu $\hat{m}_n(x_0)$ závisí na $m'(x_0)$, pak střední čtvercová chyba odhadu $\hat{m}_n(x_0)$ je neomezená vzhledem k $\mathcal{L}_2$, tedy $r(\hat{m}_n, \mathcal{L}_2) = \infty$.

Nyní ukážeme, že $\hat{m}_n^{LL}(x_0)$ je odhadem s optimálním řádem konvergence vzhledem k $\mathcal{L}_2$. Ve Stone (1980) je ukázáno, že optimální řád konvergence vzhledem k $\mathcal{L}_2$ je $n^{-2/5}$. Tedy, že

$$R_n(\mathcal{L}_2) = \inf_{\hat{m}_n} \sup_{f \in \mathcal{L}_2} \mathsf{E}_f[\hat{m}_n(x_0) - m(x_0)]^2 \asymp n^{-4/5}.$$

Podle definice 6 chceme ukázat, že

$$r(\hat{m}_n^{LL}, \mathcal{L}_2) = \sup_{f \in \mathcal{L}_2} \mathsf{E}_f[\hat{m}_n^{LL}(x_0) - m(x_0)]^2 \leq Ln^{-4/5},$$

kde $\{n^{-2/5}\}_{n=1}^{\infty}$ je optimální řád konvergence a $L < \infty$ je nějaká konstanta. Minimalizujeme-li (1.12), dostáváme optimální volbu vyhlazovacího parametru a jádrové funkce:

$$h_{n,0} = \left(\frac{15B}{bC^2n}\right)^{1/5}, \quad K_0(x) = \frac{3}{4}(1-x^2)\,\mathrm{I}_{[|x|\leq 1]}\,. \tag{1.13}$$

Dosazením do (1.12) dostáváme

$$r(\hat{m}_n^{LL}, \mathcal{L}_2) = \sup_{f \in \mathcal{L}_2} \mathrm{MSE}_f[\hat{m}_n(x_0)] \leq \frac{3}{4} 15^{-1/5} C^{2/5} \left(\frac{B}{bn}\right)^{4/5} (1 + o(1)) \leq Ln^{-4/5},$$

pro vhodné $L < \infty$ a dostatečně velké n. Tedy $\hat{m}_n^{LL}(x)$ je odhad s optimálním řádem konvergence.

Definice 7. *Eficienci odhadu $\hat{m}_n^*(x_0)$ vzhledem k $\mathcal{L}_2$ definujeme jako*

$$\mathrm{ef}[\hat{m}_n^*(x_0)] = \left(\frac{R_n(\mathcal{L}_2)}{\sup_{f \in \mathcal{L}_2} \mathsf{E}_f[\hat{m}_n^*(x_0) - m(x_0)]^2}\right)^{5/4}.$$

Pokud existuje limita $\mathrm{ef}[\hat{m}_n^*(x_0)]$, $n \to \infty$, *tak ji nazýváme asymptotická eficience odhadu $\hat{m}_n^*(x_0)$ vzhledem k $\mathcal{L}_2$.*

Eficience (resp. asymptotická eficience) značí, kolik procent dostupných dat odhad využívá. Jiným slovy, odhad $\hat{m}_n^*(x_0)$ na vzorku dat velikosti n funguje stejně dobře (ve smyslu maximálního rizika odhadu) jako 100% eficientní odhad na vzorku dat velikosti $n * ef[\hat{m}_n^*(x_0)]$.

Pokud vychýlení odhadu $\hat{m}_n(x_0)$ závisí na $m'(x_0)$ (např. Nadaraya-Watsonův odhad), pak

$$r(\hat{m}_n, \mathcal{L}_2) = \infty \Rightarrow ef[\hat{m}_n(x_0)] = 0, \tag{1.14}$$

tedy odhad $\hat{m}_n(x_0)$ má 0% eficienci vzhledem k $\mathcal{L}_2$.

Nyní formulujeme větu o asymptotické eficienci LL odhadu.

Věta 2. *Nechť platí podmínky (a) a (b), pak má LL odhad asymptotickou eficienci vzhledem k $\mathcal{L}_2$ nejméně* 76,0 %, *tj.:*

$$\frac{R_n(\mathcal{L}_2)}{\sup_{f \in \mathcal{L}_2} \mathsf{E}_f[\hat{m}_n^{LL}(x_0) - m(x_0)]^2} \geq 0{,}760^{4/5} + o(1).$$

Důkaz. Ve Fan (1993). □

LL odhad využívá (asymptoticky) nejméně 76,0 % dostupných dat. Nyní ukážeme, že mezi lineárními odhady (viz definice 1) je dokonce 100% asymptoticky eficientní.

Definice 8. *Minimax risk lineárních odhadů přes všechny hustoty z F definujeme jako*

$$R_n^L(F) = \inf_{\hat{m}_n linearni} \sup_{f \in F} \mathsf{E}_f[\hat{m}_n(x_0) - m(x_0)]^2.$$

Věta 3. *Nechť platí podmínky (a) a (b), pak je LL odhad 100% asymptoticky eficientní mezi lineárními odhady vzhledem k $\mathcal{L}_2$, tj.:*

$$\frac{R_n^L(\mathcal{L}_2)}{\sup_{f \in \mathcal{L}_2} \mathsf{E}_f[\hat{m}_n^{LL}(x_0) - m(x_0)]^2} \to 1, n \to \infty.$$

Důkaz. Ve Fan (1993). □

Z věty 3 mimo jiné plyne, že optimální řád konvergence vzhledem k $\mathcal{L}_2$ mezi lineárními odhady je opět $n^{-2/5}$, tedy že:

$$R_n^L(\mathcal{L}_2) \asymp n^{-4/5}.$$

Podmíněná asymptotika LL odhadu

Podmíníme-li (nejen) LL odhad náhodným vektorem $\boldsymbol{X}$, může být snazší dopočítat jeho asymptotické vychýlení a rozptyl. Nyní zformulujeme podmíněnou asymptotiku LL odhadu pro vnitřní body.

Věta 4. *Nechť platí podmínka (a), K má omezený nosič a nechť $f_X(x_0) > 0$ a $f_X(\cdot)$, $m''(\cdot)$ a $\sigma^2(\cdot)$ jsou spojité na okolí x_0. Pak platí:*

$$\text{bias}[\hat{m}_n^{LL}(x_0)|\boldsymbol{X}] = \frac{1}{2}h_n^2 \left(\int u^2 K(u)\, du \right) m''(x_0) + o_P(h_n^2),$$

$$var[\hat{m}_n^{LL}(x_0)|\boldsymbol{X}] = \frac{\sigma^2(x_0) \int K^2(u)\, du}{f_X(x_0) n h_n} + o_P\left(\frac{1}{nh_n}\right),$$

$$\text{MSE}[\hat{m}_n^{LL}(x_0)|\boldsymbol{X}] = \frac{1}{4}h_n^4 \left[m''(x_0) \int u^2 K(u)\, du) \right]^2 + \\ + \frac{\sigma^2(x_0) \int K^2(u)\, du}{f_X(x_0) n h_n} + o_P\left(h_n^4 + \frac{1}{nh_n}\right).$$

Důkaz. Ve Fan a Gijbels (1996). □

Podmíněná asymptotika LL odhadu v hraničních bodech

Má-li jádrová funkce K omezený nosič $[-s_0, s_0]$, využívá LL odhad $\hat{m}_n^{LL}(x_0)$ pouze pozorování z intervalu $[x_0 - s_0 h_n, x_0 + s_0 h_n]$. Navíc nosič hustoty $f_X(\cdot)$ nemusí být vždy neomezený. Právě naopak, v případě praktických aplikací má $f_X(\cdot)$ vždy omezený nosič, bez újmy na obecnosti předpokládejme $[0, 1]$. Pak ovšem LL

odhad pro body $x_0 \in H_n := [0, s_0 h_n) \cup (1 - s_0 h_n, 1]$ nemůže využít pozorování ze stejně velkého a symetrického okolí jako pro body $x_0 \in V_n := [s_0 h_n, 1 - s_0 h_n]$. Samozřejmě pro libovolný bod $x_0 \in (0, 1)$ můžeme najít dostatečně velké $n \in \mathbb{N}$ tak, že $[x_0 - s_0 h_n, x_0 + s_0 h_n] \subset [0, 1]$. V praxi ovšem nemáme neomezený počet pozorování a tzv. *hraniční oblast* H_n tvoří zhruba $(2 s_0 h_n) * 100\,\%$ datového rozsahu.

Mluvíme-li o *asymptotice v hraničních bodech* intervalu $[0, 1]$, myslíme tím vždy v posloupnosti bodů $x_n = c h_n$ (levé hraniční body) nebo $x_n = 1 - c h_n$ (pravé hraniční body) pro nějakou konstantu $0 \leq c < 1$. *Asymptotikou ve vnitřním bodě* intervalu $[0, 1]$ myslíme v libovolném pevném bodě $x_0 \in (0, 1)$. Naopak pro pevné $n \in \mathbb{N}$ nazveme x_0 *hraničním bodem* intervalu $[0, 1]$, jestliže $x_0 \in H_n$, a *vnitřním bodem* intervalu $[0, 1]$, jestliže $x_0 \in V_n$.

Nyní zformulujeme podmíněnou asymptotiku LL odhadu v hraničních bodech.

Věta 5. *Nechť platí podmínka (a), $f_X(\cdot)$ má nosič $[0, 1]$ a $K(\cdot)$ má omezený nosič $[-s_0, s_0]$ a c je konstanta splňující $c < s_0$.*

(a) Dále nechť $f_X(\cdot)$, $m''(\cdot)$ a $\sigma^2(\cdot)$ jsou zprava spojité v bodě 0, pak platí:

$$\mathrm{bias}[\hat{m}_n^{LL}(c h_n) | \boldsymbol{X}] = \frac{h_n^2}{2} m''(0+) b_L^K(c) + o_P(h_n^2),$$

$$var[\hat{m}_n^{LL}(c h_n) | \boldsymbol{X}] = v_L^K(c) \frac{\sigma^2(0+)}{f_X(0+) n h_n} + o_P\left(\frac{1}{n h_n}\right),$$

$$b_L^K(c) = \tfrac{\mu_{2,c}^2 - \mu_{1,c}\mu_{3,c}}{\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2},\ v_L^K(c) = \tfrac{\int_{-c}^{\infty} (\mu_{2,c} - u\mu_{1,c})^2 K^2(u)\, du}{(\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2)^2} \text{ a } \mu_{j,c} = \textstyle\int_{-c}^{\infty} u^j K(u)\, du.$$

(b) Dále nechť $f_X(\cdot)$ a $m''(\cdot)$ a $\sigma^2(\cdot)$ jsou zleva spojité v bodě 1, pak platí:

$$\mathrm{bias}[\hat{m}_n^{LL}(1 - c h_n) | \boldsymbol{X}] = \frac{h_n^2}{2} m''(1-) b_P^K(c) + o_P(h_n^2),$$

$$var[\hat{m}_n^{LL}(1 - c h_n) | \boldsymbol{X}] = v_P^K(c) \frac{\sigma^2(1-)}{f_X(1-) n h_n} + o_P\left(\frac{1}{n h_n}\right),$$

$$b_P^K(c) = \tfrac{\nu_{2,c}^2 - \nu_{1,c}\nu_{3,c}}{\nu_{2,c}\nu_{0,c} - \nu_{1,c}^2},\ v_P^K(c) = \tfrac{\int_{-\infty}^{c} (\nu_{2,c} - u\nu_{1,c})^2 K^2(u)\, du}{(\nu_{2,c}\nu_{0,c} - \nu_{1,c}^2)^2} \text{ a } \nu_{j,c} = \textstyle\int_{-\infty}^{c} u^j K(u)\, du.$$

Důkaz. Ve Fan a Gijbels (1992). □

Poznámky k větě 5:

- Věta platí i pokud má $K(\cdot)$ neomezený nosič, stačí předpokládat navíc, že $\limsup_{u \to -\infty} |K(u) u^5| < \infty$.
- Intuitivně bychom očekávali, že asymptotické chování LL odhadu nebude záležet na tom, jestli se jedná o levý nebo pravý hraniční bod, a skutečně ze symetrie K plyne $b_L^K(c) = b_P^K(c)$ a $v_L^K(c) = v_P^K(c)$ pro všechna c.
- Řád podmíněné asymptotické střední čtvercové chyby lokálně lineárního odhadu v hraničních bodech se rovná řádu podmíněné asymptotické střední čtvercové chyby vnitřních bodů (viz věta 4), a proto LL odhad netrpí hraničními efekty (na rozdíl od např. Nadaraya-Watsonova odhadu).

Podmíněná eficience LL odhadu

Jedná se o analogii (nepodmíněné) eficience. Prezentujeme výsledek o podmíněné eficienci LL odhadu v krajních bodech nosiče z článku Cheng a kol. (1993).

Definice 9. *Podmíněnou eficienci odhadu $\hat{m}_n^*(x_0)$ mezi lineárními odhady vzhledem k třídě regresních funkcí $\mathcal{M}$ definujeme jako*

$$\mathrm{ef}[\hat{m}_n^*(x_0)|\boldsymbol{X}] = \left(\frac{\inf_{\hat{m}_n linearni} \sup_{m\in\mathcal{M}} \boldsymbol{E}[(\hat{m}_n(x_0) - m(x_0))^2|\boldsymbol{X}]}{\sup_{m\in\mathcal{M}} \boldsymbol{E}[(\hat{m}_n^*(x_0) - m(x_0))^2|\boldsymbol{X}]} \right)^{5/4}.$$

Pokud navíc existuje limita $\mathrm{ef}[\hat{m}_n^*(x_0)]$ *v pravděpodobnosti pro $n \to \infty$, tak ji nazýváme asymptotická podmíněná eficience odhadu $\hat{m}_n^*(x_0)$ mezi lineárními odhady vzhledem k $\mathcal{M}$. Interpretace je analogická jako v definici 7.*

Cheng, Fan a Marron ukázali v článku Cheng a kol. (1993), že (za předpokladu jednostranné spojitosti $f_X(\cdot)$ a $\sigma^2(\cdot)$ v krajních bodech nosiče) LL má nejméně 77,77% asymptotickou podmíněnou eficienci mezi lineárními odhady v krajních bodech nosiče $x_0 = 0$ a $x_0 = 1$ za použití příslušné optimální jádrové funkce a příslušného optimálního vyhlazovacího parametru a vzhledem k třídě regresních funkcí $\mathcal{M}_0$:

$$\mathcal{M}_0 = \{m(\cdot) : |m(x) - m(0+) - m'(0+)x| \leq \frac{Cx^2}{2}, x > 0\}.$$

Ve Fan a Gijbels (1996) je vylepšení tohoto výsledku. Autoři dokázali, že podmíněná asymptotická eficience LL odhadu v krajních bodech nosiče je dokonce 94,64 %.

Další vlastnosti lokálně lineárního odhadu můžeme najít v článcích:
Fan a Gijbels (1992) a Ruppert a Wand (1994).

1.3.3 Převážený Nadaraya-Watsonův odhad

Další možností, navrženou v článku Cai (2001), je zkombinovat dobré vlastnosti NW a LL odhadu. Jedná se o převážený NW odhad tak, aby splňoval klíčovou podmínku pro redukci asymptotického vychýlení. Nechť $\{w_{ni}(x)\}_{i=1}^n$ značí váhové funkce pozorování $X_1, \ldots, X_n$ v pevném bodě x, splňující:

$$w_{ni}(x) \geq 0, \quad \sum_{i=1}^n w_{ni}(x) = 1, \tag{1.15}$$

$$\sum_{i=1}^n w_{ni}(x)(X_i - x)K\left(\frac{X_i - x}{h_n}\right) = 0. \tag{1.16}$$

Právě podmínka (1.16) je zobecněním podmínky (1.9). Ovšem takto definované váhy nejsou jednoznačně určeny, proto je hledáme jako řešení optimalizační úlohy:

$$\{\hat{w}_{ni}(x)\}_{i=1}^n = \mathrm{argmax}_{\{w_{n1}(x),\ldots,w_{nn}(x)\}} \Pi_{i=1}^n w_{ni}(x),$$

vzhledem k podmínkám (1.15) a (1.16). Převážený Nadaraya-Watsonův odhad pak definujeme jako

$$\hat{m}_{PNW,n}(x) = \frac{\sum_{i=1}^{n} \hat{w}_{ni}(x) K(\frac{X_i - x}{h_n}) Y_i}{\sum_{i=1}^{n} \hat{w}_{ni}(x) K(\frac{X_i - x}{h_n})}.$$

Protože váhy pro každý pevný bod x splňují podmínku (1.15), můžeme na ně nahlížet jako na pravděpodobnosti a na $\Pi_{i=1}^{n} w_{ni}(x)$ jako na empirickou věrohodnostní funkci. Pro jejich hledání tedy používáme metodu maximální věrohodnosti, maximalizujeme logaritmickou věrohodnostní funkci vzhledem k podmínkám (1.15) a (1.16). V článku Cai (2001) autor odvodil asymptotické rozdělení odhadu pro vnitřní i pro hraniční body. Ze stejného článku plyne, že převážený Nadaraya-Watsonův odhad je (pro vnitřní body) asymptoticky ekvivalentní LL odhadu.

1.4 Volba vyhlazovacího parametru při odhadu regresní funkce

Volba vyhlazovacího parametru je zásadní pro kvalitu odhadu, proto bylo této problematice v posledních desítkách let věnováno mnoho úsilí a bylo představeno nespočet metod. V této podkapitole stručně popíšeme metody, které slouží k volbě vyhlazovacího parametru při odhadu regresní funkce a které mohou být užitečné i pro odhady prezentované v kapitole 2.

1.4.1 Optimální vyhlazovací parametry

Teoreticky by bylo nejlepší volit tzv. *optimální vyhlazovací parametry* (např. (Fan a Gijbels, 1996)).

1. Chceme-li volit *optimální lokální vyhlazovací parametr* pro odhad $\hat{m}_n(x_0)$, tedy v závislosti na poloze bodu x_0, minimalizujeme podmíněnou střední čtvercovou chybu (MSE):

   $$(\text{bias}[\hat{m}_n(x_0)|\boldsymbol{X}])^2 + \text{var}[\hat{m}_n(x_0)|\boldsymbol{X}].$$

   Tato ideální volba lokálního vyhlazovacího parametru může být aproximována tzv. *asymptoticky optimálním lokálním vyhlazovacím parametrem*:

   $$h_n^{opt}(x_0) = \arg\min_{h_n>0} \text{AMSE}[\hat{m}_n(x_0)|\boldsymbol{X}],$$

   kde $\text{AMSE}[\hat{m}_n(x_0)|\boldsymbol{X}]$ značí *asymptotickou podmíněnou střední čtvercovou chybu*. Přesností této aproximace se zabývá článek Fan a kol. (1996). Snadno ověříme, že speciálně pro lokálně lineární odhad je asymptoticky optimální lokální vyhlazovací parametr:

   $$h_{LL,n}^{opt}(x_0) = \left[\frac{\sigma^2(x_0) \int K^2(u)\, du}{n f_X(x_0)(m''(x_0)[\int u^2 K(u)\, du]^2} \right]^{1/5}.$$

2. Chceme-li naopak volit *optimální globální vyhlazovací parametr*, tedy nezávisle na poloze bodu x_0, minimalizujeme *podmíněnou váženou střední integrovanou čtvercovou chybu* (MISE):

$$\int ((\text{bias}[\hat{m}_n(x)|\boldsymbol{X}])^2 + \text{var}[\hat{m}_n(x)|\boldsymbol{X}])f_X(x)w(x)\,dx,$$

kde $w(\cdot) \geq 0$ je vhodně zvolená váhová funkce, např. 1 na intervalu, který nás zajímá, a jinak 0. Analogicky jako v předchozím případě aproximujeme MISE pomocí *podmíněné asymptotické vážené střední integrované čtvercové chyby* (AMISE) a získáme tzv. *asymptoticky optimální globální vyhlazovací parametr*:

$$h_n^{opt} = \arg\min_{h_n>0} \text{AMISE}[\hat{m}_n|\boldsymbol{X}],$$

kde

$$\text{AMISE}[\hat{m}_n|\boldsymbol{X}] = \int \text{AMSE}[\hat{m}_n(x)|\boldsymbol{X}]f_X(x)w(x)\,dx.$$

Snadno ověříme, že speciálně pro lokálně lineární odhad je globálně optimální vyhlazovací parametr:

$$h_{LL,n}^{opt} = \left[\frac{\int \sigma^2(x)w(x)\,dx \int K^2(u)\,du}{n[\int u^2K(u)\,du]^2 \int [m''(x)]^2 w(x)f_X(x)\,dx}\right]^{1/5}.$$

Asymptoticky optimální vyhlazovací parametry není složité získat, ovšem jedná se o teoretické volby, které závisí na neznámých funkcí jako je $f_X(\cdot)$, $\sigma^2(\cdot)$ nebo $m''(\cdot)$. V simulačních příkladech, kdy zmíněné funkce a priori známe, je můžeme použít pro kontrolu kvality metod volby vyhlazovacího parametru. Metody praktického výpočtu vyhlazovacího parametru se většinou dělí do dvou kategorií: klasické a plug-in metody.

1.4.2 Klasické metody

Čerpáme z článku Hurvich a kol. (1998) a z Koláček (2004). Klasické metody jsou založeny na minimalizaci buďto tzv. Kullback-Leiblerovy informace (např. vylepšené Akeikeho informačního kritérium), a nebo střední průměrné čtvercové chyby (zkráceně MASE, např. penalizační metody):

$$\text{MASE}(h_n) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left[(\hat{m}_n(X_i) - m(X_i))^2\,|\boldsymbol{X}\right].$$

V druhém případě jsou metody založené na myšlence odhadnout $\text{MASE}(h_n)$ pomocí reziduálního středního čtverce:

$$\text{RMS}(h_n) = \frac{1}{n}\sum_{i=1}^{n} (\hat{m}_n(X_i) - Y_i)^2.$$

Ovšem funkce $\text{RMS}(\cdot)$ je rostoucí, tedy kdybychom minimalizovali ji samotnou, metoda by vedla k volbě příliš malých vyhlazovacích parametrů. Navíc její podmíněné vychýlení obsahuje členy závisející na h_n. Následující metody vyřeší oba zmíněné problémy.

Penalizační metody

Idea penalizačních metod je vhodná úprava funkce RMS$(\cdot)$ tak, aby došlo k asymptotickému zanedbání členů vychýlení, které závisí na h_n. K tomu nám pomůže tzv. *penalizační funkce*.

Definice 10. *Libovolnou funkci $\psi(u)$, jejíž Taylorův rozvoj 1. řádu se středem v nule je tvaru*

$$\psi(u) = 1 + 2u + O(u^2),$$

nazýváme penalizační funkcí.

Příklady penalizačních funkcí:

1. Zobecněná metoda křížového ověřování (Craven, Wahba 1979)

$$\psi(u) = \frac{1}{(1-u)^2}$$

2. Akaikeho informační kritérium (Akaike 1970)

$$\psi(u) = e^{2u}$$

3. Riceho metoda (Rice 1984)

$$\psi(u) = \frac{1}{2-u}$$

Pro odhad $\hat{m}_n(\cdot)$ se vyhlazovací parametr volí jako:

$$h_n^{pen} = \arg\min_{h_n>0} P(h_n) = \arg\min_{h_n>0} \frac{1}{n}\sum_{i=1}^{n} \left(\hat{m}_n(X_i) - Y_i\right)^2 \psi(W_{ni}(X_i)),$$

kde $\psi(\cdot)$ je nějaká penalizační funkce a $W_{ni}(X_i)$ jsou váhy odhadu $\hat{m}_n(\cdot)$. Hodnota $\psi(W_{ni}(X_i))$ roste s komplexností (podhlazením) odhadu $\hat{m}_n(\cdot)$. Tím se penalizuje příliš malá volba h_n. Nyní již asymptotické podmíněné vychýlení minimalizované funkce $P(h_n)$ neobsahuje členy závislé na h_n, jak ukáže následující věta speciálně pro LL odhad.

Věta 6. *Podmíněné asymptotické vychýlení funkce $P(h_n)$ pro LL odhad nezávisí na h_n.*

Důkaz. Větu dokážeme později.

□

Klasické metody volby vyhlazovacího parametru trpí jistými nedostatky, především při použití na lokálně polynomické odhady. Jednak tato metoda vede k velmi variabilním volbám vyhlazovacího parametru a často také k podhlazení odhadu (Hurvich a kol., 1998). Tyto nedostatky vedly k formulaci plug-in metod.

1.4.3 Plug-in metody

Tyto metody využívají znalosti asymptoticky optimálních vyhlazovacích parametrů s použitím odhadů neznámých funkcí. Konkrétně pro výpočet $\hat{h}_{LL,n}^{opt}$ bychom museli nejprve odhadnout $\int \sigma^2(x)w(x)\,dx$ a $\int f_X(x)m''(x)w(x)\,dx$. Jedná se tedy o dvoustupňový odhad, nejprve totiž musíme zvolit počáteční vyhlazovací parametr pomocí jiné metody a odhadnout neznáme hodnoty.

Plug-in metody vedou obecně k méně variabilním volbám vyhlazovacího parametru než klasické metody a také nevedou k podhlazení odhadu (Hurvich a kol., 1998). Na druhou stranu jsou tyto metody definované pouze pro odhady, u nichž asymptotický optimální vyhlazovací parametr má jednoduchý tvar. Ačkoliv jsou plug-in metody podstatně lepší v odhadování h_n^{aopt}, tato výhoda se nepřenáší do odhadu $\hat{h}_n^{opt}$, kde

$$\hat{h}_n^{opt} = \arg\min_{h_n>0} \mathrm{ISE} = \arg\min_{h_n>0} \int (\hat{m}_n(x) - m(x))^2 f_X(x)w(x)\,dx$$

nebo

$$\hat{h}_n^{opt} = \arg\min_{h_n>0} \mathrm{ASE} = \arg\min_{h_n>0} \frac{1}{n}\sum_{i=1}^{n}(m(X_i) - \hat{m}_n(X_i))^2 w(X_i).$$

Navíc podle autorů článku Hurvich a kol. (1998) je odhad $\hat{h}_n^{opt}$ smysluplnější. Diskuze, zda odhadovat $\hat{h}_n^{opt}$ nebo h_n^{opt}, je např. v Hall a Marron (1991).

1.4.4 Vylepšené Akaikeho informační kritérium

Patří mezi klasické metody minimalizující Kullback-Leiblerovu informaci.
V článku Hurvich a kol. (1998) byla pro homoskedastický model představena metoda založená na vylepšené verzi Akeikeho informačního kritéria (zkráceně AIC_C):

$$h_n^{\mathrm{AIC}_c} = \arg\min_{h_n>0} \mathrm{AIC}_C(h_n),$$

kde

$$\mathrm{AIC}_C(h_n) = \log(\hat{\sigma}^2) + 1 + \frac{2(\mathrm{tr}(\boldsymbol{S}_{h_n}) + 1)}{n - \mathrm{tr}(\boldsymbol{S}_{h_n}) - 2},$$

kde $\mathrm{tr}(\boldsymbol{S}_{h_n})$ značí stopu vyhlazovací matice $\boldsymbol{S}_{h_n}$, ta může být interpretována jako efektivní počet použitých parametrů (viz Hastie a Tibshirani (1990)). Na rozdíl od plug-in metod může být tato metoda použita k volbě vyhlazovacího parametru pro libovolný lineární odhad regresní funkce $m(\cdot)$. Navíc již nevede k variabilním volbám vyhlazovacího parametru ani k podhlazení odhadu (jako jiné klasické metody). Autoři pomocí Monte Carlo simulací dokonce ukázali, že AIC_C metoda má srovnatelné výsledky s plug-in metodami, pokud nějaká z nich existuje a funguje dobře, a zároveň podává dobré výsledky, pokud plug-in metody selhávají nebo jsou nedostupné.

1.4.5 Metoda křížového ověřování

Jednou z nejpopulárnějších metod volby vyhlazovacího parametru je tzv. *metoda křížového ověřování*. Idea metody (např. v (Omelka, 2015)) je odhadnout:

$$\mathrm{MASE}(h_n) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left[(\hat{m}_n(X_i) - m(X_i))^2 \,|\boldsymbol{X}\right]$$

pomocí:

$$\mathrm{CV}(h_n) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_n^{(-i)}(X_i))^2 w(X_i),$$

kde $\hat{m}_n^{(-i)}(X_i)$ je odhad $m(X_i)$ spočtený na podvýběru $\{(X_j, Y_j)\}_{j=1, j\neq i}^n$ a $w(\cdot)$ je námi zvolená váhová funkce. Funkce $\mathrm{CV}(\cdot)$ se nazývá *funkce křížového ověřování* a vyhodnocuje schopnost odhadu predikovat $\{Y_i\}_{i=1}^n$ s využitím podvýběrů $\{(X_j,Y_j)\}_{j=1,j\neq i}^n$, $i = 1, \ldots, n$. Je-li $\hat{m}_n(\cdot)$ lineární odhad, pak

$$\hat{m}_n^{(-i)}(X_i) = \sum_{j=1, j\neq i}^{n} Y_j W_{nj}^{(-i)}(X_i),$$

kde $W_{nj}^{(-i)}(X_i)$ jsou váhy v bodě X_i spočtené na podvýběru $\{(X_j,Y_j)\}_{j=1,j\neq i}^n$.

Vyhlazovací parametr volíme:

$$h_n^{\mathrm{CV}} = \arg\min_{h_n>0} \mathrm{CV}(h_n).$$

Náhled, proč tato metoda funguje, nám poskytne následující heuristika:

$$\begin{aligned}
\mathrm{CV}(h_n) =& \frac{1}{n}\sum_{i=1}^{n}(Y_i - m(X_i) + m(X_i) - \hat{m}_n^{(-i)}(X_i))^2 w(X_i) \\
=& \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\sigma^2(X_i)w(X_i) + \frac{2}{n}\sum_{i=1}^{n}\epsilon_i\sigma(X_i)[m(X_i) - \hat{m}_n^{(-i)}(X_i)]w(X_i) \\
&+ \frac{1}{n}\sum_{i=1}^{n}[m(X_i) - \hat{m}_n^{(-i)}(X_i)]^2 w(X_i),
\end{aligned}$$

první sčítanec nezávisí na h_n, druhý má nulovou podmíněnou střední hodnotu (platí-li podmínka (a)) a třetí odhaduje $\mathrm{MASE}(h_n)$.

Metoda křížového ověřování patří mezi penalizační metody, což je doloženo v článku Härdle a kol. (1988). V článku Härdle (1990) je navíc ukázano, že tato metoda může být použita i při odhadu derivací regresní funkce $m(\cdot)$.

2. Odhady rozptylové funkce

Problém odhadu rozptylové funkce v modelu (1.1) můžeme převést na problém odhadu regresní funkce v odvozeném modelu

$$r_i^2 = \sigma^2(X_i)\epsilon_i^2 = \sigma^2(X_i) + v(X_i)\widetilde{\epsilon}_i, \tag{2.1}$$

kde $v(X_i) = \sigma^2(X_i)\lambda(X_i)$, $\lambda^2(X_i) = \mathsf{E}\left[(\epsilon_i^2 - 1)^2 | X_i\right]$ a

$$\widetilde{\epsilon}_i = \frac{\sigma^2(X_i)\epsilon_i^2 - \sigma^2(X_i)}{\sigma^2(X_i)\lambda(X_i)}, \quad i = 1, \ldots, n.$$

Snadno můžeme ověřit, že analogicky jako u modelu (1.1) platí předpoklady:

- $(X_1, r_1^2), \ldots, (X_n, r_n^2)$ tvoří dvoudimenzionální striktně stacionární náhodný proces (pokud platí navíc podmínka (a), pak jsou navíc i.i.d.),
- $\mathsf{E}\left[r_i^2 | X_i\right] = \mathsf{E}\left[\sigma^2(X_i)\epsilon_i^2 | X_i\right] = \sigma^2(X_i)$,
- $\mathrm{var}[r_i^2 | X_i] = \sigma^4(X_i)\lambda^2(X_i) = v^2(X_i)$,
- $\mathsf{E}\left[\widetilde{\epsilon}_i | X_i\right] = 0$ a $\mathrm{var}[\widetilde{\epsilon}_i | X_i] = 1$.

Tento model je použitelný pro tzv. *oracle odhady rozptylové funkce*, tedy odhady, u nichž a priori známe regresní funkci, a tedy i náhodné chyby $r_i = Y_i - m(X_i)$. Protože model (2.1) splňuje výše zmíněné předpoklady, platí pro oracle odhady teorie z podkapitoly 1.3. Naopak v praktických příkladech neznáme předem regresní funkci, a tedy ani náhodné chyby, proto se místo náhodných chyb uvažují rezidua $\hat{r}_i = Y_i - \hat{m}_n(X_i)$. Nyní však již výše ověřené předpoklady neplatí, tedy ani teorie z podkapitoly 1.3. Takovéto odhady nazýváme *odhady rozptylové funkce založené na reziduích*. Jeden takový představili autoři článku Fan a Yao (1998).

2.1 Fan-Yaoův odhad

Nechť $\hat{m}_n^{LL}(x)$ je LL odhad regresní funkce s jádrovou funkcí $K(\cdot)$ a vyhlazovacím parametrem $h_{n,1}$ a $\hat{r}_i^2 = [Y_i - \hat{m}_n^{LL}(X_i)]^2$ rezidua, pak Fan-Yaoův odhad (dále jen FY odhad) definujeme jako $\hat{\sigma}_{FY,n}^2(x) = \hat{\alpha}(x)$, kde

$$(\hat{\alpha}, \hat{\beta})(x) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} [\hat{r}_i^2 - \alpha - \beta(X_i - x)]^2 W\left(\frac{X_i - x}{h_{n,2}}\right),$$

kde $W(\cdot)$ je jádrová funkce a $h_{n,2}$ je druhý vyhlazovací parametr. Ačkoliv FY odhad rozptylové funkce $\sigma^2(x)$ můžeme vyjádřit

$$\hat{\sigma}_{FY,n}^2(x) = \sum_{i=1}^{n} \hat{r}_i^2 W_{ni}^{LL}(x),$$

kde $W_{ni}^{LL}(x)$ jsou váhy LL odhadu (viz (1.7)), jeho asymptotika neplyne přímo z asymptotiky LL odhadu regresní funkce, neboť rezidua nesplňují $\mathsf{E}\left[\hat{r}_i^2 | X_i\right] = \sigma^2(X_i)$. Autoři článku tedy asymptotiku odhadu odvodili sami.

1. Podmínky:

   (a) Pro daný bod x je $f_X(x) > 0$, $\sigma^2(x) > 0$ a funkce $\mathsf{E}\,[Y^k|X = z]$ je spojitá v x pro $k = 3, 4$. Dále $\ddot{m}(z)$ a $\ddot{\sigma}^2(z)$ jsou stejnoměrně spojité na otevřené množině obsahující x.

   (b) $\mathsf{E}\,|Y|^{4(1+\delta)} < \infty$, kde $\delta \in [0, 1)$ je konstanta.

   (c) Jádrové funkce $W(\cdot)$ a $K(\cdot)$ mají omezený nosič a spolu s $f_X(\cdot)$ splňují Lipschitzovskou podmínku.

   (d) Proces $\{(X_i, Y_i)\}$ je striktně stacionární a absolutně regulární (viz Davidson (1994), str. 209-211) s mixing koeficientem $\beta(j)$ splňujícím $\sum_{i=1}^{n} j^2\beta^{\delta/(1+\delta)}(j) < \infty$, přičemž δ je stejná jako v podmínce 1b.

   (e) Pro $n \to \infty$, $\lim_{n\to\infty} h_{n,i} = 0$ a $\liminf_{n\to\infty} nh_{n,i}^4 > 0$ pro $i = 1, 2$.

Poznámky k podmínkám:

- Předpoklad omezenosti nosičů $W(\cdot)$ a $K(\cdot)$ není nutný a je uveden pro jednoduchost důkazů. Věta je platná speciálně i pro Gaussovu jádrovou funkci.

- Jsou-li $\{(X_i, Y_i)\}$ nezávislé, pak podmínka 1d platí pro $\delta = 0$ a podmínka 1b se reguluje na $\mathsf{E}\,[Y^4] < \infty$.

Věta 7. *Nechť platí podmínky (1). Pak*

$$(nh_{n,2})^{\frac{1}{2}}[\hat{\sigma}^2_{FY,n}(x) - \sigma^2(x) - \theta_n]$$

je asymptoticky normální se střední hodnotou 0 *a rozptylem*

$$f_X^{-1}(x)\sigma^4(x)\lambda^2(x)\int W^2(t)\,dt,$$

kde

$$\lambda^2(x) = \mathsf{E}[(\epsilon_i^2 - 1)^2|X_i = x], \quad \epsilon_i = \frac{Y_i - m(X_i)}{\sigma(X_i)}, \quad \sigma_W^2 = \int t^2 W(t)\,dt,$$

$$\theta_n = \frac{h_{n,2}^2}{2}\sigma_W^2\ddot{\sigma}^2(x) + o(h_{n,1}^2 + h_{n,2}^2). \tag{2.2}$$

Tedy střední čtvercová chyba je

$$\mathrm{MSE}[\hat{\sigma}^2_{FY,n}(x)] = \theta_n^2 + \frac{1}{nh_{n,2}}f_X^{-1}(x)\sigma^4(x)\lambda^2(x)\int W^2(t)\,dt + o\left(\frac{1}{nh_{n,2}}\right). \tag{2.3}$$

Důkaz. Ve Fan a Yao (1998).

□

Eficience Fan-Yaova odhadu

Řekneme, že odhad rozptylové funkce $\hat{\sigma}_n^2(\cdot)$ je *adaptivní k neznámé regresní funkci*, jestliže je asymptoticky ekvivalentní příslušnému oracle odhadu. Oracle odhad příslušný FY odhadu (dále také FYO odhad) je $\hat{\sigma}^2_{FYO,n}(x) = \hat{\alpha}(x)$, kde

$$(\hat{\alpha}, \hat{\beta})(x) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} [r_i^2 - \alpha - \beta(X_i - x)]^2 W\left(\frac{X_i - x}{h_{n,2}}\right),$$

$$\hat{\alpha}(x) = \sum_{i=1}^{n} r_i^2 W_{ni}^{LL}(x).$$

Přímo z asymptotiky LL odhadu (§ 6.2.2 ve Fan a Gijbels (1996)) plyne, že FYO odhad je asymptoticky normální a platí:

$$\text{bias}[\hat{\sigma}^2_{FYO,n}(x)] = \frac{h_{n,2}^2}{2} \sigma_W^2 \ddot{\sigma}^2(x) + o(h_{n,2}^2),$$

$$\text{var}[\hat{\sigma}^2_{FYO,n}(x)] = \frac{1}{nh_{n,2}} f_X^{-1}(x) \sigma^4(x) \lambda^2(x) \int W^2(t)\, dt + o\left(\frac{1}{nh_{n,2}}\right).$$

Tedy za předpokladu, že $h_{n,1}$ konverguje k 0 alespoň tak rychle jako $h_{n,2}$, je FYO odhad asymptoticky ekvivalentní FY odhadu. Z (1.13) plyne, že $h_{n,1}$ minimalizující $\text{AMSE}[\hat{m}_n^{LL}(x)]$ je řádu $n^{-1/5}$. Dosadíme-li jej do (2.3), pak také $h_{n,2}$ minimalizující $\text{AMSE}[\hat{\sigma}^2_{FY,n}(x)]$ je řádu $n^{-1/5}$. Tedy podmínka je splněna a FY odhad je adaptivní k neznámé regresní funkci.

Z vět 2 a 3 navíc víme, že odhad $\hat{\sigma}^2_{FYO,n}$ je téměř asymptoticky eficientní vzhledem k $\mathcal{L}_2$ a je dokonce (100%) asymptoticky eficientní mezi lineárními odhady vzhledem k $\mathcal{L}_2$, proto také FY odhad je asymptoticky eficientní ve stejném smyslu.

Z (2.2) si můžeme všimnout, že chyba odhadu $m(x)$ přispívá k vychýlení $\hat{\sigma}^2_{FY,n}(x)$ pouze členem $o(h_{n,1}^2)$. Díky tomu můžeme dosáhnout adaptivního a eficientního odhadu za použití optimálních vyhlazovacích parametrů, aniž bychom museli podhladit odhad $\hat{m}_n^{LL}(x)$.

2.2 Přímočarý odhad

Vraťme se na chvíli k naivnímu přímočarému odhadu, jak jsme jej definovali v (1.2). Uvědomme si, že nebylo nic řečeno o tom, jakým způsobem odhadnout funkce $m(x)$ a $v(x)$. Za tímto účelem nejprve označíme odhady:

$$\hat{m}_n^{NW}(x) = \sum_{i=1}^{n} Y_i W_{n,i}^{NW}(x), \quad \hat{m}_n^{LL}(x) = \sum_{i=1}^{n} Y_i W_{n,i}^{LL}(x),$$
$$\hat{v}_n^{NW}(x) = \sum_{i=1}^{n} Y_i^2 W_{n,i}^{NW}(x), \quad \hat{v}_n^{LL}(x) = \sum_{i=1}^{n} Y_i^2 W_{n,i}^{LL}(x),$$

kde $W_{n,i}^{NW}(x)$ a $W_{n,i}^{LL}(x)$ jsou po řadě váhy NW a LL odhadu (viz (1.4) a (1.7)) při použití jádrové funkce $W(\cdot)$ a vyhlazovacího parametru $h_{n,2}$. Nyní definujeme dva přímočaré odhady:

$$\hat{\sigma}^2_{d:NW+NW,n}(x) = \hat{v}_n^{NW}(x) - [\hat{m}_n^{NW}(x)]^2, \quad \hat{\sigma}^2_{d:LL+LL,n}(x) = \hat{v}_n^{LL}(x) - [\hat{m}_n^{LL}(x)]^2.$$

Pomocí následující úvahy ukážeme, že vychýlení přímočarých odhadů je vyšší než vychýlení odhadů založených na reziduích. Uvažujme odhad založený na reziduích

$$\hat{\sigma}^2_{NW+NW,n}(x) = \sum_{i=1}^{n} [Y_i - \hat{m}_n^{NW}(X_i)]^2 \frac{W(\frac{X_i - x}{h_{n,2}})}{\sum_{j=1}^{n} W(\frac{X_j - x}{h_{n,2}})},$$

zaměníme-li všechny $\hat{m}_n^{NW}(X_i)$ za $\hat{m}_n^{NW}(x)$, dostáváme

$$\sum_{i=1}^{n} Y_i^2 \frac{W(\frac{X_i - x}{h_{n,2}})}{\sum_{j=1}^{n} W(\frac{X_j - x}{h_{n,2}})} - [\hat{m}_n^{NW}(x)]^2 = \hat{v}_n^{NW}(x) - [\hat{m}_n^{NW}(x)]^2 = \hat{\sigma}^2_{d:NW+NW,n}(x).$$

Zřejmě, $(Y_i - \hat{m}_n^{NW}(x))^2$ je více vychýlené od $\mathsf{E}\,[Y - m(X)]^2$ než $(Y_i - \hat{m}_n^{NW}(X_i))^2$, proto příspěvek vychýlení $\hat{m}_n(\cdot)$ k vychýlení přímočarého odhadu je vyšší než k vychýlení odhadu založeného na reziduích. Exaktněji to bylo ukázáno v Härdle a Tsybakov (1997) pro odhad $\hat{\sigma}_{d:LL+LL,n}(\cdot)$. Při použití stejné jádrové funkce $W(\cdot)$ a stejného vyhlazovacího parametru $h_{n,2}$ pro odhad m i v, je odhad asymptoticky normální a hlavní členy u asymptotického vychýlení a rozptylu jsou:

$$\text{bias}[\hat{\sigma}_{d:LL+LL,n}(x)] : \frac{h_{n,2}^2}{2} \sigma_W^2 [\ddot{\sigma}^2(x) + 2(\dot{m}(x))^2],$$

$$\text{var}[\hat{\sigma}_{d:LL+LL,n}(x)] : \frac{1}{n h_{n,2}} f_X^{-1}(x) \sigma^4(x) \lambda^2(x) \int W^2(t)\, dt.$$

Srovnáme-li výsledek s větou 7, zjistíme, že zatímco asymptotický rozptyl přímočarého odhadu je stejný jako u FY odhadu, asymptotické vychýlení obsahuje navíc člen $h_{n,2}^2 \sigma_W^2 (\dot{m}(x))^2$. Tento člen navíc může výrazně zhoršit kvalitu odhadu, zejména v případech, kdy první derivace $m(\cdot)$ je velká (např. je-li $m(\cdot)$ lineární funkce s velkou směrnicí).

Existence jednoho členu navíc ve vychýlení přímočarého odhadu můžeme vysvětlit následující heuristikou:

$$\begin{aligned} \hat{\sigma}^2_{d:LL+LL,n}(x) - \sigma^2(x) &= [\hat{v}_n^{LL}(x) - (\hat{m}_n^{LL}(x))^2] - [v(x) - m^2(x)] \\ &= [\hat{v}_n^{LL}(x) - v(x)] - 2m(x)[\hat{m}_n^{LL}(x) - m(x)] - [\hat{m}_n^{LL}(x) - m(x)]^2, \end{aligned} \tag{2.4}$$

pro $\{(X_i,\, Y_i^2), i = 1, \ldots, n\}$ můžeme (podobně jako pro $\{(X_i,\, r_i^2), i = 1, \ldots, n\}$) odvodit analogický model jako (1.1) a z asymptotiky LL odhadu pak dostáváme

$$\text{bias}[\hat{v}_{LL}(x)] = \frac{h_{n,2}^2}{2} \sigma_W^2 \ddot{v}(x) + o(h_{n,2}^2).$$

Odsud a z identity $v(x) = E[(Y - m(X))^2 + 2Ym(X) - m^2(X) | X = x]$ dostáváme

$$\begin{aligned} \mathsf{E}\,[\hat{v}_{LL}(x) - v(x)] &= \frac{h_{n,2}^2}{2} \sigma_W^2 \ddot{v}(x) + o(h_{n,2}^2) \\ &= \frac{h_{n,2}^2}{2} \sigma_W^2 \frac{d^2}{dx^2} [\sigma^2(x) + m^2(x)] + o(h_{n,2}^2) \\ &= \frac{h_{n,2}^2}{2} \sigma_W^2 [\ddot{\sigma}^2(x) + 2(\dot{m}(x))^2] + 2m(x) \frac{h_{n,2}^2}{2} \sigma_W^2 \ddot{m}(x) + o(h_{n,2}^2). \end{aligned} \tag{2.5}$$

Poslední výraz v (2.5) ruší střední hodnotu druhého výrazu v (2.4). Navíc z podmínky (1e) plyne, že $\mathrm{MSE}[\hat{m}_n^{LL}(x)] = O(h_{n,2}^4 + \frac{1}{nh_{n,2}}) = o(h_{n,2}^2)$. Proto po dosazení (2.5) do střední hodnoty (2.4) zůstává ve vychýlení $\hat{\sigma}^2_{d:LL+LL,n}(x)$ člen $h_{n,2}^2\sigma_W^2(\dot{m}(x))^2$.

Přímočarým odhadem se zabýval také článek Yao a Tong (1994).

2.3 Xu-Phillipsův odhad

V 2.1 jsme ukázali, že FY odhad je eficientní a adaptivní odhad. Mohlo by se tedy zdát, že hledání nejlepšího odhadu rozptylové funkce je u konce. Nicméně jedna nevýhoda zde existuje, a sice, že FY odhad může nabývat záporných hodnot pro konečné výběry! Takovou situaci ilustrujeme na empirickém příkladu v 2.7. Záporný odhad rozptylu samozřejmě nedává smysl. Z toho důvodu se můžeme v literatuře setkat s doporučením odhadnout rozptylovou funkci opět na základě reziduí (jako u FY odhadu), ale tentokrát místo vah LL odhadu použít váhy NW odhadu. Jejich nezápornost zaručí nezápornost celého odhadu.

Další možností je Xu-Phillipsův odhad (dále jen XP odhad) prezentovaný v článku Xu a Phillips (2012). Jedná se o převážený lokálně konstantní odhad založený na reziduích. Tento odhad, jak ukážeme později, je téměř asymptoticky ekvivalentní FY odhadu a navíc je vždy nezáporný. Pokud má navíc vysvětlující proměnná neomezený nosič, jsou oba odhady zcela asymptoticky ekvivalentní. XP odhad je navíc adaptivní, tedy je stejně eficientní, jako bychom regresní funkci a priori znali.

Podmíněná střední hodnota se odhadne opět pomocí lokálně lineárního odhadu $\hat{m}_n^{LL}(x)$ s jádrovou funkcí $K(\cdot)$ a vyhlazovacím parametrem $h_{n,1}$. Na základě druhé mocniny reziduí $\hat{r}_i = [Y_i - \hat{m}_n^{LL}(X_i)]^2$ spočteme XP odhad jako

$$\hat{\sigma}^2_{XP,n}(x) = \frac{\sum_{i=1}^n \hat{w}_{ni}(x) W(\frac{X_i - x}{h_{n,2}}) \hat{r}_i^2}{\sum_{i=1}^n \hat{w}_{ni}(x) W(\frac{X_i - x}{h_{n,2}})}, \tag{2.6}$$

kde $W(\cdot)$ je jádrová funkce, $h_{n,2}$ je druhý vyhlazovací parametr a $\hat{w}_{ni}(x)$ řeší optimalizační úlohu:

$$\{\hat{w}_{n1}(x), \ldots, \hat{w}_{nn}(x)\} = \arg \min_{\{w_{n1}(x), \ldots, w_{nn}(x)\}} -2 \sum_{i=1}^n \log(n w_{ni}(x)),$$

s podmínkami:

$$w_{ni}(x) \geq 0, \quad \sum_{i=1}^n w_{ni}(x) = 1, \tag{2.7}$$

$$\sum_{i=1}^n w_{ni}(x)(X_i - x) \frac{1}{h_{n,2}} W\left(\frac{X_i - x}{h_{n,2}}\right) = 0. \tag{2.8}$$

Poznámky:

- Podmínka (2.7) zaručuje nezápornost XP odhadu.

- Podmínka (2.8) je splněna váhami lokálně lineárního odhadu (viz (1.9)) a je považována za klíčovou podmínku pro redukci vychýlení (viz (1.10)) a pro dosažení eficientního odhadu (viz (1.14)).
- Naopak bez podmínky (2.8) jsou řešením optimalizační úlohy $w_{ni}(x) = 1/n$ pro všechna i (lze snadno nahlédnout z konvexnosti funkce: $-\log$ a podmínek (2.7)). XP odhad se pak redukuje na lokálně konstantní odhad.
- Váhy $\hat{w}_{ni}(x)$ v (2.6) můžeme získat použitím Lagrangeových multiplikátorů, tj.:

$$\hat{w}_{ni}(x) = \frac{1}{n[1 + \lambda(X_i - x)W_{h_{n,2}}(X_i - x)]}, \tag{2.9}$$

kde λ splňuje:

$$F(\lambda) = \sum_{i=1}^{n} \frac{(X_i - x)W_{h_{n,2}}(X_i - x)}{n[1 + \lambda(X_i - x)W_{h_{n,2}}(X_i - x)]} = 0, \tag{2.10}$$

$$1 + \lambda(X_i - x)W_{h_{n,2}}(X_i - x) > 0, \quad i = 1, \ldots, n. \tag{2.11}$$

Má-li $W(\cdot)$ omezený nosič $(-1, 1)$, pak zřejmě nutná podmínka řešitelnosti (2.10) za podmínky (2.11) je, aby existovaly $i, j \in \{1, \ldots, n\}$ t.ž.:

$$[X_i \in (x - h_{n,2}, x)] \wedge [X_j \in (x, x + h_{n,2})]\,. \tag{2.12}$$

V případě neomezeného nosiče $W(\cdot)$ stačí, aby existovaly $i, j \in \{1, \ldots, n\}$ t.ž.:

$$[X_i < x] \wedge [X_j > x]\,.$$

2. Podmínky:
   (a) Platí podmínky 1 uvedené u věty 7 o asymptotické normalitě FY odhadu.
   (b) Existuje konstanta $M < \infty$ taková, že $|g_{1,t}(y_1,\, y_t|x_1,\, x_t)| \leq M$ pro všechna $t \geq 2$, kde $g_{1,t}(y_1,\, y_t|x_1,\, x_t)$ je podmíněná hustota $(Y_1,\, Y_t)$ v bodě $(y_1,\, y_t)$ při daném $(X_1, X_t) = (x_1, x_t)$.
   (c) Jádrové funkce W a K mají omezený nosič $[-1, 1]$.

Věta 8. *Nechť platí podmínky 2.*

(a) Nechť navíc $f_X(x + h_n) > 0$ a $f_X(x - h_n) > 0$, pak pro $n \to \infty$ platí

$$(nh_{n,2})^{\frac{1}{2}}[\hat{\sigma}^2_{XP,n}(x) - \sigma^2(x) - \frac{h^2_{n,2}}{2}\sigma^2_W \ddot{\sigma}^2(x)]$$
$$\xrightarrow{d} \mathcal{N}(0, f_X^{-1}(x)\sigma^4(x)\lambda^2(x) \int W^2(t)\,dt),$$

kde

$$\lambda^2(x) = \mathsf{E}[(\epsilon_t^2 - 1)^2|X_t = x], \quad \epsilon_t = \frac{Y_t - m(X_t)}{\sigma(X_t)}, \quad \sigma^2_W = \int t^2 W(t)\,dt.$$

(b) Nechť navíc $f_X(\cdot)$ má omezený nosič $[a,b]$, c je konstanta splňující $0 < c < 1$. Označme:

$$\overline{W}_0 = \int_{-1}^{c} [1 - \overline{\lambda}_c t W(t)]^{-1} W(t)\, dt,$$
$$\overline{W}_1 = \int_{-1}^{c} [1 - \overline{\lambda}_c t W(t)]^{-1} t^2 W(t)\, dt,$$
$$\overline{W}_2 = \int_{-1}^{c} [W(t)/(1 - \overline{\lambda}_c t W(t))]^2\, dt$$

a $\overline{\lambda}_c$ splňuje $\overline{L}_c(\overline{\lambda}_c) = 0$, kde

$$\overline{L}_c(\lambda) = \int_{-1}^{c} tW(t)/(1 - \lambda t W(t))\, dt.$$

Analogicky:

$$\underline{W}_0 = \int_{-c}^{1} [1 - \underline{\lambda}_c t W(t)]^{-1} W(t)\, dt,$$
$$\underline{W}_1 = \int_{-c}^{1} [1 - \underline{\lambda}_c t W(t)]^{-1} t^2 W(t)\, dt,$$
$$\underline{W}_2 = \int_{-c}^{1} [W(t)/(1 - \underline{\lambda}_c t W(t))]^2\, dt$$

a $\underline{\lambda}_c$ splňuje $\underline{L}_c(\underline{\lambda}_c) = 0$, kde

$$\underline{L}_c(\lambda) = \int_{-c}^{1} tW(t)/(1 - \lambda t W(t))\, dt,$$

pak pro $n \to \infty$ platí

$$(nh_{n,2})^{\frac{1}{2}}[\hat{\sigma}^2_{XP,n}(a + ch_{n,2}) - \sigma^2(a + ch_{n,2}) - h^2_{n,2}\underline{W}_1 \ddot{\sigma}^2(a + ch_{n,2})/2\underline{W}_0] \xrightarrow{d} \mathcal{N}(0, \sigma^4(a)\lambda^2(a)\underline{W}_2/[f_X(a)\underline{W}_0^2]),$$

$$(nh_{n,2})^{\frac{1}{2}}[\hat{\sigma}^2_{XP,n}(b - ch_{n,2}) - \sigma^2(b - ch_{n,2}) - h^2_{n,2}\overline{W}_1 \ddot{\sigma}^2(b - ch_{n,2})/2\overline{W}_0] \xrightarrow{d} \mathcal{N}(0, \sigma^4(b)\lambda^2(b)\overline{W}_2/[f_X(b)\overline{W}_0^2]).$$

Důkaz. V Xu a Phillips (2012). □

Poznámky k větě 8:

- Všimněme si, že ze symetrie W plyne

$$\int_{-c}^{1} tW(t)/(1 - \lambda t W(t))\, dt \overset{!}{=} 0 \Leftrightarrow \int_{-1}^{c} tW(t)/(1 - (-\lambda) t W(t))\, dt \overset{!}{=} 0,$$

odsud dále plyne

$$\underline{\lambda}_c = -\overline{\lambda}_c \Rightarrow \underline{W}_j = \overline{W}_j, \quad j = 0, 1, 2.$$

- XP odhad je asymptoticky ekvivalentní FY odhadu pro vnitřní body, je tedy (100%) asymptoticky eficientní mezi lineárními odhady (ve smyslu věty 3) a téměř asymptoticky eficientní mezi všemi odhady (ve smyslu věty 2).
- Konstanty $\underline{\lambda}_c$ a $\overline{\lambda}_c$ klesají s rostoucím c a konvergují k nule pro $c \to 1$. Navíc část (b) platí i pro vnitřní body, neboť $\overline{W}_0 = \underline{W}_0 = 1$, $\overline{W}_1 = \underline{W}_1 = \sigma_W^2$ a $\overline{W}_2 = \underline{W}_2 = \int W^2(t)\,dt$ pro $c \geq 1$.
- Z Cai (2001) plyne, že XP odhad je adaptivní k neznámé regresní funkci $m(\cdot)$, jelikož je asymptoticky ekvivalentní svému oracle odhadu.
- XP odhad netrpí hraničními efekty, neboť řád konvergence vychýlení i rozptylu pro hraniční body je stejný jako pro vnitřní body. Protože však neznáme asymptotiku FY odhadu v hraničních bodech, nemůžeme ji přímo porovnat s asymptotikou XP odhadu v hraničních bodech.
- Asymptotický rozptyl $\hat{\sigma}^2_{XP,n}(x)$ může být konzistentně odhadnut pro vnitřní i hraniční body. To nám umožní konstrukci bodově konzistentních intervalů spolehlivosti pro $\sigma^2(x)$.

Věta 9. *Označme* $\hat{\Omega}(x) = \hat{f}_X^{-2}(x)\hat{V}(x)$, *kde*

$$\hat{V}(x) = nh_{n,2}^{-1}\sum_{i=1}^{n} W^2((X_i - x)/h_{n,2})[\hat{r}_i^2 - \hat{\sigma}^2_{XP,n}(x)]^2,$$
$$\hat{f}_X(x) = h_{n,2}^{-1}\sum_{i=1}^{n} W((X_i - x)/h_{n,2}).$$

Nechť platí předpoklady věty 8 (a) , pak pro $n \to \infty$,

$$\hat{\Omega}(x) \xrightarrow{P} f_X^{-1}(x)\sigma^4(x)\lambda^2(x)\int W^2(t)\,dt.$$

Nechť platí předpoklady věty 8 (b), pak pro $n \to \infty$,

$$\hat{\Omega}(a + ch) \xrightarrow{P} \sigma^4(a)\lambda^2(a)\underline{W}_2/[f_X(a)\underline{W}_0^2],$$
$$\hat{\Omega}(b - ch) \xrightarrow{P} \sigma^4(b)\lambda^2(b)\overline{W}_2/[f_X(b)\overline{W}_0^2].$$

Důkaz. V Xu a Phillips (2012).

□

2.4 Asymptotika odhadů v hraničních bodech

Chceme-li důkladně srovnat FY a XP odhad, měli bychom srovnat i jejich chování na hranici. Ovšem Fan a Yao ve svém článku asymptotiku pro hraniční body nenabídli. V této podkapitole si tedy sami odvodíme podmíněné asymptotické vychýlení a rozptyly obou odhadů v hraničních bodech.

2.4.1 FY odhad v hraničních bodech

Označme nejprve $\boldsymbol{S}_1 = \boldsymbol{S}_{1,h_{n,1}}$ vyhlazovací matici lokálně lineární regrese, která byla použita pro výpočet $\hat{\boldsymbol{m}}^{LL} = (\hat{m}_n^{LL}(X_1)^T, \ldots, \hat{m}_n^{LL}(X_n))$, tj.:

$$(\boldsymbol{S}_1)_{i,j} = \boldsymbol{e}_1^T(\boldsymbol{X}_1^T(X_i)\boldsymbol{K}_{\boldsymbol{h_{n,1}}}(X_i)\boldsymbol{X}_1(X_i))^{-1}\boldsymbol{X}_1^T(X_i)\boldsymbol{K}_{\boldsymbol{h_{n,1}}}(X_i)\boldsymbol{e}_j,$$

pak $\hat{\boldsymbol{m}}^{LL} = \boldsymbol{S}_1\boldsymbol{y}$. Označme vektor reziduí

$$\hat{\boldsymbol{r}} = (Y_1 - \hat{m}_n^{LL}(X_1), \ldots, Y_n - \hat{m}_n^{LL}(X_n))^T \Rightarrow \hat{\boldsymbol{r}} = (\boldsymbol{I} - \boldsymbol{S}_1)\boldsymbol{y}.$$

Nechť body $x_1, x_2, \ldots, x_n$ jsou předem pevně zvolené, označme:

$$\hat{\boldsymbol{\sigma}}^2_{FY} = (\hat{\sigma}^2_{FY,n}(x_1), \ldots, \hat{\sigma}^2_{FY,n}(x_n))^T, \quad \boldsymbol{\sigma}^2 = (\sigma^2(X_1), \ldots, \sigma^2(X_n))^T.$$

Je-li navíc $\boldsymbol{S}_2 = \boldsymbol{S}_{1,h_{n,2}}$ vyhlazovací matice lokálně lineární regrese pro výpočet $\hat{\boldsymbol{\sigma}}^2_{FY}$, tj.:

$$(\boldsymbol{S}_2)_{i,j} = \boldsymbol{e}_1^T(\boldsymbol{X}_1^T(x_i)\boldsymbol{W}_{\boldsymbol{h_{n,2}}}(x_i)\boldsymbol{X}_1(x_i))^{-1}\boldsymbol{X}_1^T(x_i)\boldsymbol{W}_{\boldsymbol{h_{n,2}}}(x_i)\boldsymbol{e}_j,$$

pak $\hat{\boldsymbol{\sigma}}^2_{FY} = \boldsymbol{S}_2\hat{\boldsymbol{r}}^2$.

Součinem dvou sloupcových vektorů dále myslíme součin po složkách. Součin matic po složkách značíme symbolem "$\circ$" (tzv. *Hadamardův součin*). Diagonální matici značíme diag a diagonálu matice pro změnu diagonal. Pro vektor $\hat{\boldsymbol{\sigma}}^2_{FY}$ nyní zformulujeme a dokážeme pomocnou větu o podmíněném vychýlení a kovarianci.

Lemma 10. *Nechť platí podmínka (a) z podkapitoly 1.3.2 o LL odhadu. Dále označme:*

$$\begin{aligned} q(x) = \mathsf{E}[\epsilon_i^3|X_i = x], \quad t(x) = \mathsf{E}[\epsilon_i^4|X_i = x], \\ \boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}^2), \quad \boldsymbol{Q} = \mathrm{diag}_{1\le i\le n}\{\sigma^3(X_i)q(X_i)\}, \\ \boldsymbol{T} = \mathrm{diag}_{1\le i\le n}\{\sigma^4(X_i)t(X_i)\}, \quad \boldsymbol{b}_1 = (\boldsymbol{S}_1 - \boldsymbol{I})\boldsymbol{m}, \end{aligned}$$

pak platí:

$$\mathsf{E}[\hat{\boldsymbol{\sigma}}^2_{FY}|\boldsymbol{X}] = \boldsymbol{S}_2\boldsymbol{\sigma}^2 + \boldsymbol{S}_2[\mathrm{diagonal}(\boldsymbol{b}_1\boldsymbol{b}_1^T + \boldsymbol{S}_1\boldsymbol{\Sigma}\boldsymbol{S}_1^T - 2\boldsymbol{S}_1\boldsymbol{\Sigma})]$$

a

$$\begin{aligned} var[\hat{\boldsymbol{\sigma}}^2_{FY}|\boldsymbol{X}] = \boldsymbol{S}_2[&\{(\boldsymbol{S}_1 - \boldsymbol{I}) \circ (\boldsymbol{S}_1 - \boldsymbol{I})\}(\boldsymbol{T} - 3\boldsymbol{\Sigma}^2)\{(\boldsymbol{S}_1 - \boldsymbol{I}) \circ (\boldsymbol{S}_1 - \boldsymbol{I})\}^T \\ &+ 2(\mathrm{diag}(\boldsymbol{b}_1))(\boldsymbol{S}_1 - \boldsymbol{I})\boldsymbol{Q}\{(\boldsymbol{S}_1 - \boldsymbol{I}) \circ (\boldsymbol{S}_1 - \boldsymbol{I})\}^T \\ &+ 2\{(\boldsymbol{S}_1 - \boldsymbol{I}) \circ (\boldsymbol{S}_1 - \boldsymbol{I})\}\boldsymbol{Q}(\boldsymbol{S}_1 - \boldsymbol{I})^T(\mathrm{diag}(\boldsymbol{b}_1)) \\ &+ 2\{(\boldsymbol{S}_1 - \boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{S}_1 - \boldsymbol{I})^T\} \circ \{(\boldsymbol{S}_1 - \boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{S}_1 - \boldsymbol{I})^T\} \\ &+ 4\{(\boldsymbol{S}_1 - \boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{S}_1 - \boldsymbol{I})^T\} \circ (\boldsymbol{b}_1\boldsymbol{b}_1^T)]\boldsymbol{S}_2^T. \end{aligned}$$

Důkaz. Připomeňme, že matice $\boldsymbol{S}_1, \boldsymbol{S}_2$ závisí pouze na $\boldsymbol{X}$, nikoliv na $\boldsymbol{y}$, a že z podmínky (a) plyne $\mathsf{E}[Y_iY_j|\boldsymbol{X}] = \delta_{i,j}\sigma^2(X_i) + m(X_i)m(X_j)$, kde $\delta_{i,j} = 1$ pro $i = j$, jinak 0. Dále můžeme psát

$$\hat{\boldsymbol{\sigma}}^2_{FY} = \boldsymbol{S}_2\hat{\boldsymbol{r}}^2 = \boldsymbol{S}_2\,\mathrm{diagonal}[(\boldsymbol{S}_1 - \boldsymbol{I})\boldsymbol{y}\boldsymbol{y}^T(\boldsymbol{S}_1 - \boldsymbol{I})^T] \Rightarrow$$

$$\mathsf{E}\,[\hat{\boldsymbol{\sigma}}^2_{FY}|\boldsymbol{X}] = \boldsymbol{S}_2\,\mathrm{diagonal}[(\boldsymbol{S}_1 - \boldsymbol{I})(\boldsymbol{\Sigma} + \boldsymbol{m}\boldsymbol{m}^T)(\boldsymbol{S}_1 - \boldsymbol{I})^T].$$

Odsud už snadno dostaneme požadovaný výsledek pro vychýlení. Důkaz kovariance by se provedl analogicky jako v článku Ruppert a kol. (1995b).

□

Lemma 11. *Nechť $\{Y_n\}_{n=1}^{\infty}$ je posloupnost náhodných veličin, $\{g_n\}_{n=1}^{\infty}$ posloupnost funkcí a X náhodná veličina. Nechť platí:*

$$var[Y_n|X] \xrightarrow{s.j.} 0, \quad \boldsymbol{E}\,[Y_n|X] - g_n(X) \xrightarrow{s.j.} 0, \quad n \to \infty,$$

pak

$$Y_n = g_n(X) + o_P(1), \quad n \to \infty.$$

Důkaz. Stačí ukázat, že pro všechna $\epsilon > 0$

$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - g_n(X)| > \epsilon]}\,|X] \xrightarrow{s.j.} 0, \quad n \to \infty,$$

neboť z Lebesgueovy věty o dominantní konvergenci plyne

$$\lim_{n\to\infty} P(|Y_n - g_n(X)| > \epsilon) = \lim_{n\to\infty} \mathsf{E}\,[\mathrm{I}_{[|Y_n - g_n(X)| > \epsilon]}] =$$
$$\lim_{n\to\infty} \mathsf{E}\,(\mathsf{E}\,[\mathrm{I}_{[|Y_n - g_n(X)| > \epsilon]}\,|X]) = \mathsf{E}\,(\lim_{n\to\infty} \mathsf{E}\,[\mathrm{I}_{[|Y_n - g_n(X)| > \epsilon]}\,|X]).$$

Podmínky Lebesgueovy věty jsou splněny, neboť s.j. platí

$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - g_n(X)| > \epsilon]}\,|X] \leq 1.$$

Dále z trojúhelníkové nerovnosti dostáváme

$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - g_n(X)| > \epsilon]}\,|X] \leq \mathsf{E}\,[\mathrm{I}_{[|Y_n - \mathsf{E}\,[Y_n|X]| + |\,\mathsf{E}\,[Y_n|X] - g_n(X)| > \epsilon]}\,|X] =$$
$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - \mathsf{E}\,[Y_n|X]| + |\,\mathsf{E}\,[Y_n|X] - g_n(X)| > \epsilon, |\,\mathsf{E}\,[Y_n|X] - g_n(X)| < \epsilon/2]}\,|X] + \tag{2.13}$$
$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - \mathsf{E}\,[Y_n|X]| + |\,\mathsf{E}\,[Y_n|X] - g_n(X)| > \epsilon, |\,\mathsf{E}\,[Y_n|X] - g_n(X)| \geq \epsilon/2]}\,|X]. \tag{2.14}$$

Nyní ukážeme, že výraz (2.14) konverguje skoro jistě k 0, nejprve ho odhadneme shora:

$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - \mathsf{E}\,[Y_n|X]| + |\,\mathsf{E}\,[Y_n|X] - g_n(X)| > \epsilon, |\,\mathsf{E}\,[Y_n|X] - g_n(X)| \geq \epsilon/2]}\,|X] \leq \mathsf{E}\,[\mathrm{I}_{[|\,\mathsf{E}\,[Y_n|X] - g_n(X)| \geq \epsilon/2]}\,|X],$$

a protože funkce $\mathrm{I}_{[|\,\mathsf{E}\,[Y_n|X] - g_n(X)| \geq \epsilon/2]}$ je $\sigma(X)$ měřitelná, dostáváme

$$\mathsf{E}\,[\mathrm{I}_{[|\,\mathsf{E}\,[Y_n|X] - g_n(X)| \geq \epsilon/2]}\,|X] = \mathrm{I}_{[|\,\mathsf{E}\,[Y_n|X] - g_n(X)| \geq \epsilon/2]} \xrightarrow{s.j.} 0, \quad n \to \infty.$$

Výraz (2.13) s pomocí Čebyševovy nerovnosti odhadneme shora:

$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - \mathsf{E}\,[Y_n|X]| + |\,\mathsf{E}\,[Y_n|X] - g_n(X)| > \epsilon, |\,\mathsf{E}\,[Y_n|X] - g_n(X)| < \epsilon/2]}\,|X] \leq$$
$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - \mathsf{E}\,[Y_n|X]| > \epsilon/2]}\,|X] \leq \frac{\mathrm{var}[Y_n|X]}{\epsilon^2/4} \xrightarrow{s.j.} 0, \quad n \to \infty.$$

Tedy celkově

$$\mathsf{E}\,[\mathrm{I}_{[|Y_n - g_n(X)| > \epsilon]}\,|X] \xrightarrow{s.j.} 0 \Rightarrow Y_n - g_n(X) \xrightarrow{P} 0, \quad n \to \infty.$$

□

Lemma 12. *Nechť platí předpoklady lemmatu 10, nechť jádrová funkce K je omezená a má omezený nosič. Nechť hustota $f_X(\cdot)$ má nosič $[0,1]$ a je na něm spojitá. Nechť navíc $0 < h_n \to 0$ a $nh_n \to \infty$, pro $n \to \infty$, dále označme*

$$S_{n,l}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{X_i - x}{h_n}\right)\left(\frac{X_i - x}{h_n}\right)^l,$$

pak pro $j = 1, \ldots, n$ platí

$$S_{n,l}(X_j) = g_{n,l}(X_j) + o_P(1) = \begin{cases} f_X(X_j)\int K(t)t^l\,dt + o_P(1), & l = 2k, \\ o_P(1), & l = 2k+1. \end{cases}$$

Důkaz. BÚNO $j = 1$, označme nejprve

$$Y_i(X_1) = \frac{1}{h_n} K\left(\frac{X_i - X_1}{h_n}\right)\left(\frac{X_i - X_1}{h_n}\right)^l, \quad i = 1, \ldots, n,$$

pak

$$S_{n,l}(X_1) = \frac{1}{n}\sum_{i=2}^{n} Y_i(X_1).$$

Dle lemmatu 11 stačí ukázat, že:

$$\operatorname{var}[S_{n,l}(X_1)|X_1] \overset{s.j.}{\to} 0, \quad \mathsf{E}\,[S_{n,l}(X_1)|X_1] - g_{n,l}(X_1) \overset{s.j.}{\to} 0, \quad n \to \infty.$$

Protože $Y_2(X_1), \ldots, Y_n(X_1)$ jsou X_1-podmíněně i.i.d., s použitím vhodné substituce dostáváme

$$\mathsf{E}\,[S_{n,l}(X_1)|X_1] = \frac{n-1}{n}\,\mathsf{E}\,[Y_2(X_1)|X_1] = \frac{n-1}{n}\int K(t)t^l f_X(X_1 + th_n)\,dt.$$

Z Lebesgueovy věty (o záměně limity a integrálu) plyne

$$\begin{aligned} P\left(w : \lim_{n\to\infty} \int \frac{n-1}{n} K(t)t^l f_X(X_1(w) + th_n)\,dt = \int K(t)t^l f_X(X_1(w) + th_n)\,dt\right) \\ = P\left(w : \int K(t)t^l f_X(X_1(w))\,dt = \int K(t)t^l f_X(X_1(w))\,dt\right) = 1. \end{aligned}$$

Podmínka Lebesgueovy věty je splněna, neboť ze spojitosti $f_X(\cdot)$ plyne, že existuje $C < \infty$ t.ž. $|f_X(X_1 + th_n)| \le C$, a z omezenosti $K(\cdot)$ a jejího nosiče plyne, že integrál $\int |K(t)t^l|\,dt$ existuje a je konečný. Celkově tedy dostáváme

$$\mathsf{E}\,[S_{n,l}(X_1)|X_1] = f_X(X_1)\int K(t)t^l\,dt + o_{s.j.}(1) = g_{n,l}(X_1) + o_{s.j.}(1).$$

Výsledek pro l liché plyne ze symetrie K. Nyní rozptyl.

$$\begin{aligned} \operatorname{var}[S_{n,l}(X_1)|X_1] &= \frac{n-1}{n^2}\operatorname{var}[Y_2(X_1)|X_1] \\ &= \frac{n-1}{n^2}(\mathsf{E}\,[Y_2^2(X_1)|X_1] - (\mathsf{E}\,[Y_2(X_1)|X_1])^2) \\ &= \frac{n-1}{n^2}[\frac{1}{h_n}O_{s.j.}(1) - O_{s.j.}(1)] = o_{s.j.}(1). \end{aligned}$$

Odtud již plyne požadovaný výsledek.

□

Lemmata 11 a 12 mají následující důsledek.

Lemma 13. *Nechť platí předpoklady lemmatu 12, nechť $m(\cdot)$ je dvakrát diferencovatelná na intervalu $[0,1]$ a $m''(\cdot)$, $\sigma^2(\cdot)$ jsou spojité na intervalu $[0,1]$. Nechť navíc vyhlazovací parametr splňuje $0 < h_{n,1} \to 0$ a $nh_{n,1} \to \infty$, pro $n \to \infty$, pak pro $j = 1, \ldots, n$ platí:*

$$\text{bias}[\hat{m}_n^{LL}(X_j)|\boldsymbol{X}] = O_P(h_{n,1}^2),$$

$$var[\hat{m}_n^{LL}(X_j)|\boldsymbol{X}] = O_P\left(\frac{1}{nh_{n,1}}\right),$$

tedy celkem

$$\text{MSE}[\hat{m}_n^{LL}(X_j)|\boldsymbol{X}] = O_P\left(h_{n,1}^4 + \frac{1}{nh_{n,1}}\right).$$

Důkaz. Analogicky jako ve větě 4, s použitím lemmat 11 a 12.

□

Lemma 14. *Nechť platí předpoklady lemmatu 13, jádrová funkce $W(\cdot)$ je omezená a má nosič $[-1,1]$, x_1 je levý hraniční bod, tj. $x_1 = ch_{n,2}$, kde $0 \leq c < 1$. Nechť navíc vyhlazovací parametr splňuje $0 < h_{n,2} \to 0$ a $nh_{n,2} \to \infty$, pro $n \to \infty$, pak pro funkci $g(\cdot)$ spojitou na intervalu $[0,1]$, jejíž druhá derivace je zprava spojitá v 0, platí:*

1. $\boldsymbol{e}_1^T \boldsymbol{S}_2 \boldsymbol{g} = g(ch_{n,2}) + \frac{h_{n,2}^2}{2} g''(0+) b_L^W(c) + o_P(h_{n,2}^2)$,
   kde $\boldsymbol{g} = (g(X_1), \ldots, g(X_n))^T$, $b_L^W(c) = \frac{\mu_{2,c}^2 - \mu_{1,c}\mu_{3,c}}{\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2}$ *a* $\mu_{j,c} = \int_{-c}^{\infty} u^j W(u)\, du$.

2. $\boldsymbol{e}_1^T \,\text{diagonal}(\boldsymbol{S}_2 \,\text{diag}(\boldsymbol{g}) \boldsymbol{S}_2^T) = v_L^W(c) \frac{g(0+)}{f_X(0+) nh_{n,2}} + o_P(\frac{1}{nh_{n,2}})$,
   $\text{diagonal}(\boldsymbol{S}_2 \,\text{diag}(\boldsymbol{g}) \boldsymbol{S}_2^T) = O_P(\frac{1}{nh_{n,2}})_{n\times 1}$,
   kde $v_L^W(c) = \frac{\int_{-c}^{\infty} (\mu_{2,c} - u\mu_{1,c})^2 W^2(u)\, du}{(\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2)^2}$.

3. $\text{diagonal}(\boldsymbol{S}_1 \,\text{diag}(\boldsymbol{g}) \boldsymbol{S}_1^T) = O_P(\frac{1}{nh_{n,1}})_{n\times 1}$.

4. $\boldsymbol{S}_1 = O_P(\frac{1}{nh_{n,1}})_{n\times n}$.

5. $\boldsymbol{S}_1 \,\text{diag}(\boldsymbol{g}) \boldsymbol{S}_1^T = O_P(\frac{1}{nh_{n,1}})_{n\times n}$.

Důkaz. Připomeňme, že vyhlazovací matice (LL odhadu) $\boldsymbol{S}_1$ odhaduje regresní funkci v náhodných bodech $(X_1, \ldots, X_n)$ s použitím vyhlazovacího parametru $h_{n,1}$ a jádrové funkce K. Vyhlazovací matice (LL odhadu) $\boldsymbol{S}_2$ odhaduje regresní funkci v pevných bodech $(x_1, \ldots, x_n)$ s použitím vyhlazovacího parametru $h_{n,2}$ a jádrové funkce W, kde $0 < x_i < 1$ pro $i = 2, \ldots, n$ a navíc x_1 je levým hraničním bodem, tj. $x_1 = ch_{n,2}$.

1. Podmíněná střední hodnota LL odhadu v bodě x_1 s použitím vyhlazovacího parametru $h_{n,2}$ a jádrové funkce W lze vyjádřit

$$\mathsf{E}\,[\hat{m}_n^{LL}(ch_{n,2})|\boldsymbol{X}] = \boldsymbol{e}_1^T \boldsymbol{S}_2 \boldsymbol{m},$$

a protože jsou splněny předpoklady věty 5, tak zároveň platí

$$\mathsf{E}\left[\hat{m}_n^{LL}(ch_{n,2})|\boldsymbol{X}\right] = m(ch_{n,2}) + \frac{h_n^2}{2}m''(0+)b_L^W(c) + o_P(h_{n,2}^2).$$

Z faktu, že $m(\cdot)$ je libovolná funkce, jejíž druhá derivace je zprava spojitá v 0, plyne požadovaný výsledek.

2. Podmíněný rozptyl LL odhadu v bodě x_i s použitím vyhlazovacího parametru $h_{n,2}$ a jádrové funkce W lze vyjádřit

$$\begin{aligned}\mathrm{var}[\hat{m}_n^{LL}(x_i)|\boldsymbol{X}] &= \boldsymbol{e}_i^T\,\mathsf{E}\left[\mathrm{diagonal}(\boldsymbol{S}_2(\boldsymbol{y}-\boldsymbol{m})(\boldsymbol{y}-\boldsymbol{m})^T\boldsymbol{S}_2^T)|\boldsymbol{X}\right]\\ &\overset{(a)}{=} \boldsymbol{e}_i^T\,\mathrm{diagonal}(\boldsymbol{S}_2\,\mathrm{diag}(\boldsymbol{\sigma}^2)\boldsymbol{S}_2^T),\end{aligned}$$

a protože jsou splněny předpoklady vět 4 a 5, tak zároveň platí

$$\mathrm{var}[\hat{m}_n^{LL}(x_i)|\boldsymbol{X}] = \begin{cases} v_L^W(c)\frac{\sigma^2(0+)}{f_X(0+)nh_{n,2}} + o_P(\frac{1}{nh_{n,2}}) = O_P(\frac{1}{nh_{n,2}}), & i=1,\\ \frac{\sigma^2(x_i)\int K^2(u)\,du}{f_X(x_i)nh_{n,2}} + o_P(\frac{1}{nh_{n,2}}) = O_P(\frac{1}{nh_{n,2}}), & i=2,\ldots,n.\end{cases}$$

Uvědomme si, že pro každou spojitou funkci $g(\cdot)$ na $[0,1]$ existují spojité a nezáporné funkce $\sigma_1^2(\cdot)$, $\sigma_2^2(\cdot)$ na $[0,1]$ takové, že:

$$g(x) = \sigma_1^2(x) - \sigma_2^2(x).$$

Z faktu, že $\sigma^2(\cdot)$ je libovolná funkce, která je spojitá a nezáporná na intervalu $[0,1]$, plyne požadovaný výsledek.

3. Přímo z lemmatu 13 plyne

$$\begin{aligned}\mathrm{diagonal}(\boldsymbol{S}_1\,\mathrm{diag}(\boldsymbol{\sigma}^2)\boldsymbol{S}_1^T) &= \mathsf{E}\left[\mathrm{diagonal}(\boldsymbol{S}_1(\boldsymbol{y}-\boldsymbol{m})(\boldsymbol{y}-\boldsymbol{m})^T\boldsymbol{S}_1^T)|\boldsymbol{X}\right]\\ &= \mathrm{diagonal}(\mathrm{var}[\hat{\boldsymbol{m}}^{LL}|\boldsymbol{X}]) = O_P\left(\frac{1}{nh_{n,1}}\right)_{n\times 1}.\end{aligned}$$

Z faktu, že $\sigma^2(\cdot)$ je libovolná funkce, která je spojitá a nezáporná na intervalu $[0,1]$, plyne požadovaný výsledek.

4. Z (1.6) plyne

$$\hat{m}_n^{LL}(X_i) = \sum_{j=1}^n Y_j W_{nj}^{LL}(X_i) = \sum_{j=1}^n Y_j(\boldsymbol{S}_1)_{ij} \Rightarrow (\boldsymbol{S}_1)_{ij} = W_{nj}^{LL}(X_i),$$

kde

$$W_{nj}^{LL}(X_i) = \frac{\frac{1}{nh_{n,1}}K(\frac{X_j-X_i}{h_{n,1}})[S_{n,2}(X_i) - \frac{X_j-X_i}{h_{n,1}}S_{n,1}(X_i)]}{S_{n,0}(X_i)S_{n,2}(X_i) - S_{n,1}^2(X_i)}, i,j = 1,\ldots,n.$$

Z lemmatu 12 dostáváme

$$W_{nj}^{LL}(X_i) = \frac{\frac{1}{nh_n}K(\frac{X_j-X_i}{h_{n,1}})[f_X(X_i)\sigma_K^2 + o_P(1) - \frac{X_j-X_i}{h_{n,1}}o_P(1)]}{f_X^2(X_i)\sigma_K^2 + o_P(1)}.$$

Jádrová funkce $K(\cdot)$ je omezená a má omezený nosič, proto $K(\frac{X_j-X_i}{h_{n,1}})$, $K(\frac{X_j-X_i}{h_{n,1}})(\frac{X_j-X_i}{h_{n,1}})$ a σ_K^2 jsou omezené shora skoro jistě. Hustota $f_X(\cdot)$ je spojitá a kladná na $[0,1]$, proto existují $\epsilon > 0$ a $C < \infty$ t.ž.: $\epsilon < f_X(X_i) < C$ s.j. pro všechna i. Odtud plyne požadovaný výsledek.

5. Analogicky jako v 3 vyjádříme

$$\mathrm{var}[\hat{\boldsymbol{m}}^{LL}|\boldsymbol{X}] = \mathsf{E}\left[\boldsymbol{S}_1(\boldsymbol{y}-\boldsymbol{m})(\boldsymbol{y}-\boldsymbol{m})^T\boldsymbol{S}_1^T|\boldsymbol{X}\right] = \boldsymbol{S}_1\,\mathrm{diag}(\boldsymbol{\sigma}^2)\boldsymbol{S}_1^T.$$

Z Cauchy-Schwarzovy nerovnosti navíc plyne, že

$$(\boldsymbol{S}_1\,\mathrm{diag}(\boldsymbol{\sigma}^2)\boldsymbol{S}_1^T)_{i,j} = \mathrm{cov}[\hat{m}_n^{LL}(X_i), \hat{m}_n^{LL}(X_j)|\boldsymbol{X}] \leq$$
$$\leq \max\{\mathrm{var}[\hat{m}_n^{LL}(X_i)|\boldsymbol{X}], \mathrm{var}[\hat{m}_n^{LL}(X_j)|\boldsymbol{X}]\} \overset{lemma13}{=} O_P\left(\frac{1}{nh_{n,1}}\right).$$

Z faktu, že $\sigma^2(\cdot)$ je libovolná funkce, která je spojitá na intervalu $[0,1]$, plyne požadovaný výsledek.

Konec důkazu.

□

Poznámky k lemmatu 14:

- Lemma 14 lze analogicky zformulovat a dokázat, i pokud je jeden z regresorů pravým krajním bodem, tj. $x_1 = 1 - ch_{n,2}$. Stačí předpokládat spojitost $g''(\cdot)$ v 1 zprava (místo v 0 zleva). Limity funkcí g, f jsou nyní v bodě 1 zleva a místo $b_L^W(c)$, $v_L^W(c)$ jsou ve výsledku konstanty $b_P^W(c)$, $v_P^W(c)$ definované analogicky jako ve větě 5 pro jádrovou funkci $W(\cdot)$.
- Pro platnost tvrzení 1 je spojitost $g''(\cdot)$ v 0 zprava postačují předpoklad o funkci $g(\cdot)$, naopak pro tvrzení 2, 3 a 5 je spojitost $g(\cdot)$ na $[0,1]$ postačují předpoklad o funkci $g(\cdot)$.

Nyní se můžeme vrátit k důkazu věty 6.

Důkaz. Z lemmatu 13 a ze 4 v lemmatu 14 víme:

$$\mathrm{MSE}[\hat{m}_n^{LL}(X_i)|\boldsymbol{X}] = O_P\left(h_{n,1}^4 + \frac{1}{nh_{n,1}}\right), \quad W_{ni}^{LL}(X_i) = O_P\left(\frac{1}{nh_{n,1}}\right).$$

Dokážeme

$$\mathrm{bias}[P(h_{n,1})|\boldsymbol{X}] = \sum_{i=1}^{n}\sigma^2(X_i) + o_P\left(h_{n,1}^4 + \frac{1}{nh_{n,1}}\right).$$

Postupně dosadíme za $Y_i = m(X_i) + \epsilon_i\sigma(X_i)$, poté za $\hat{m}_n^{LL}(X_i) = \sum_{j=1}^n (m(X_j) + \sigma(X_j)\epsilon_j)W_{nj}^{LL}(X_i)$ a využijeme vlastností náhodné složky $\boldsymbol{\epsilon}$ a nakonec uděláme Taylorův rozvoj 1. řádu se středem v nule penalizační funkce $\psi(\cdot)$ a dostáváme

$$\begin{aligned}
\mathsf{E}\left[P(h_{n,1})|\boldsymbol{X}\right] =& \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\left[\left(\hat{m}_n^{LL}(X_i) - Y_i\right)^2\psi(W_{ni}^{LL}(X_i))|\boldsymbol{X}\right] \\
=& \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\left[\left(\hat{m}_n^{LL}(X_i) - m(X_i) - \epsilon_i\sigma(X_i)\right)^2\psi(W_{ni}^{LL}(X_i))|\boldsymbol{X}\right] \\
=& \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2(X_i)[1 - 2W_{ni}^{LL}(X_i)] + \mathrm{MSE}[\hat{m}_n^{LL}(X_i)|\boldsymbol{X}]\right)\psi(W_{ni}^{LL}(X_i))
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2(X_i)[1-2W_{ni}^{LL}(X_i)]\right)\left(1+2W_{ni}^{LL}(X_i)+O_P\left(\frac{1}{n^2h_{n,1}^2}\right)\right)\\
&+\frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}[\hat{m}_n^{LL}(X_i)|\boldsymbol{X}]\left(1+2W_{ni}^{LL}(X_i)+O_P\left(\frac{1}{n^2h_{n,1}^2}\right)\right)\\
=&\,\mathrm{MASE}(h_{n,1})+\frac{1}{n}\sum_{i=1}^{n}\sigma^2(X_i)+o_P\left(h_{n,1}^4+\frac{1}{n^2h_{n,1}^2}\right).
\end{aligned}
$$

□

Věta 15. *Nechť platí předpoklady lemmatu 14, funkce $q(\cdot)$ a $t(\cdot)$ definované v lemmatu 10 jsou spojité na $[0,1]$, nechť $\ddot{\sigma}^2(\cdot)$ je zprava spojitá v 0 a nechť navíc vyhlazovací parametry splňují:*

$$
nh_{n,1}^2\to\infty, \tag{2.15}
$$

$$
\left\{h_{n,1}^4+\frac{1}{nh_{n,1}}\right\}=o(h_{n,2}^2), \tag{2.16}
$$

pro $n\to\infty$, pak

$$
\mathrm{bias}[\hat{\sigma}_{FY,n}^2(ch_{n,2})|\boldsymbol{X}]=\frac{h_{n,2}^2}{2}\ddot{\sigma}^2(0+)b_L^W(c)+o_P(h_{n,2}^2),
$$

$$
var[\hat{\sigma}_{FY,n}^2(ch_{n,2})|\boldsymbol{X}]=v_L^W(c)\frac{\sigma^4(0+)\lambda^2(0+)}{f_X(0+)nh_{n,2}}+o_P\left(\frac{1}{nh_{n,2}}\right),
$$

kde $b_L^W(c)$ a $v_L^W(c)$ jsou definované v lemmatu 14.

Důkaz. Začněme s vychýlením. Z lemmatu 10 plyne

$$
\mathsf{E}\left[\hat{\sigma}_{FY,n}^2(ch_{n,2})|\boldsymbol{X}\right]=\boldsymbol{e}_1^T\,\mathsf{E}\left[\hat{\boldsymbol{\sigma}}_{FY}^2|\boldsymbol{X}\right]
$$

$$
=\boldsymbol{e}_1^T\boldsymbol{S}_2\boldsymbol{\sigma}^2+\boldsymbol{e}_1^T\boldsymbol{S}_2[\mathrm{diagonal}(\boldsymbol{b}_1\boldsymbol{b}_1^T)+\mathrm{diagonal}(\boldsymbol{S}_1\boldsymbol{\Sigma}\boldsymbol{S}_1^T-2\boldsymbol{S}_1\boldsymbol{\Sigma})]. \tag{2.17}
$$

Víme, že lokálně lineární regrese zachovává konstantní vektor, tj. $\boldsymbol{S}_1\mathbf{1}=\boldsymbol{S}_2\mathbf{1}=\mathbf{1}$. Z lemmatu 13 víme, že $\mathrm{bias}[\hat{m}_n^{LL}(X_i)|\boldsymbol{X}]=O_P(h_{n,1}^2)$ pro $i=1,\ldots,n$, odtud plyne

$$
\mathrm{diagonal}(\boldsymbol{b}_1\boldsymbol{b}_1^T)=O_P(h_{n,1}^4)_{n\times1}\overset{(2.16)}{=}o_P(h_{n,2}^2)_{n\times1}
$$

$$
\Rightarrow\boldsymbol{S}_2\,\mathrm{diagonal}(\boldsymbol{b}_1\boldsymbol{b}_1^T)=o_P(h_{n,2}^2)_{n\times1}.
$$

Ze 3 v lemmatu 14 plyne

$$
\begin{aligned}
\mathrm{diagonal}(\boldsymbol{S}_1\boldsymbol{\Sigma}\boldsymbol{S}_1^T)=O_P\left(\frac{1}{nh_{n,1}}\right)_{n\times1}&\overset{(2.16)}{=}o_P(h_{n,2}^2)_{n\times1}\\
\Rightarrow\boldsymbol{S}_2\,\mathrm{diagonal}(\boldsymbol{S}_1\boldsymbol{\Sigma}\boldsymbol{S}_1^T)&=o_P(h_{n,2}^2)_{n\times1}.
\end{aligned}
$$

Ze 4 v lemmatu 14 plyne

$$\text{diagonal}(\boldsymbol{S}_1\boldsymbol{\Sigma}) = \text{diagonal}(\boldsymbol{S}_1)\boldsymbol{\sigma}^2 = O_P\left(\frac{1}{nh_{n,1}}\right)_{n\times 1} \overset{(2.16)}{=} o_P(h_{n,2}^2)_{n\times 1}$$
$$\Rightarrow \boldsymbol{S}_2\,\text{diagonal}(\boldsymbol{S}_1\boldsymbol{\Sigma}) = o_P(h_{n,2}^2)_{n\times 1}.$$

Tedy dominantní člen v (2.17) je $\boldsymbol{e}_1^T\boldsymbol{S}_2\boldsymbol{\sigma}^2$, z 1 v lemmatu 14 dostáváme požadovaný výsledek pro vychýlení. Nyní rozptyl.

$$\begin{aligned}\text{var}[\hat{\sigma}^2_{FY,n}(ch_{n,2})|\boldsymbol{X}] &= \boldsymbol{e}_1^T\,\text{diagonal}(\text{var}[\hat{\boldsymbol{\sigma}}^2_{FY}|\boldsymbol{X}]) \\ &= \boldsymbol{e}_1^T\,\text{diagonal}(\boldsymbol{S}_2(\text{var}[\hat{\boldsymbol{r}}^2|\boldsymbol{X}])\boldsymbol{S}_2^T).\end{aligned}$$

Ukážeme, že dominantními členy $\text{var}[\hat{\boldsymbol{r}}^2|\boldsymbol{X}]$, který je rozepsaný v lemmatu 10, jsou:

$$(\boldsymbol{T} - 3\boldsymbol{\Sigma}^2) + 2\boldsymbol{\Sigma}^2.$$

Označme $\boldsymbol{D} = \text{diag}_{1\le i\le n}\{g(X_i)\}$ typ diagonální matice, kde g je nějaká spojitá funkce na $[0,1]$. Ze spojitosti $\sigma^2(\cdot)$, $q(\cdot)$ a $t(\cdot)$ na $[0,1]$ plyne, že matice $(\boldsymbol{T}-3\boldsymbol{\Sigma}^2)$, $\boldsymbol{Q}$ a $\boldsymbol{\Sigma}$ jsou typu $\boldsymbol{D}$.

Stačí ukázat, že zbylé členy ve $\text{var}[\hat{\boldsymbol{r}}^2|\boldsymbol{X}]$ jsou buďto matice typu $\boldsymbol{D}$, jejíchž diagonální prvky konvergují v pravděpodobnosti k nule, tj. $o_P(1)\boldsymbol{D}$, a nebo matice $o_P(\frac{1}{nh_{n,2}})_{n\times n} = o_P(\frac{1}{nh_{n,2}})\boldsymbol{1}_{n\times n}$, neboť platí:

$$\text{diagonal}(\boldsymbol{S}_2 o_P(1)\boldsymbol{D}\boldsymbol{S}_2^T) \overset{(2)}{=} o_P(1)O_P\left(\frac{1}{nh_{n,2}}\right)_{n\times 1} = o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times 1},$$

$$\text{diagonal}\left(\boldsymbol{S}_2 o_P\left(\frac{1}{nh_{n,2}}\right)\boldsymbol{1}_{nxn}\boldsymbol{S}_2^T\right) = o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times 1}.$$

Ze 4 v lemmatu 14 plyne:

$$(\boldsymbol{S}_1 \circ \boldsymbol{I})\boldsymbol{D} = o_P(1)\boldsymbol{D}, \quad (\boldsymbol{S}_1\boldsymbol{D}) \circ \boldsymbol{D} = o_P(1)\boldsymbol{D},$$

$$(\boldsymbol{S}_1 \circ \boldsymbol{S}_1)\boldsymbol{D} = O_P\left(\frac{1}{n^2h_{n,1}^2}\right)_{n\times n} \overset{(2.15)}{=} o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n},$$

$$(\boldsymbol{S}_1 \circ \boldsymbol{S}_1)\boldsymbol{D}(\boldsymbol{S}_1 \circ \boldsymbol{S}_1)^T = O_P\left(\frac{1}{n^3h_{n,1}^4}\right)_{n\times n} \overset{2.15}{=} o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n},$$

$$(\boldsymbol{S}_1\boldsymbol{D}) \circ (\boldsymbol{S}_1\boldsymbol{D}) = O_P\left(\frac{1}{n^2h_{n,1}^2}\right)_{n\times n} \overset{2.15}{=} o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n}.$$

Z 5 v lemmatu 14 plyne

$$\boldsymbol{S}_1\boldsymbol{D}\boldsymbol{S}_1^T = O_P\left(\frac{1}{nh_{n,1}}\right)_{n\times n},$$

odtud plyne:

$$(\boldsymbol{S}_1\boldsymbol{D}\boldsymbol{S}_1^T)\circ(\boldsymbol{S}_1\boldsymbol{D}\boldsymbol{S}_1^T) = O_P\left(\frac{1}{n^2h_{n,1}^2}\right)_{n\times n} \overset{2.15}{=} o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n},$$

$$(\boldsymbol{S}_1\boldsymbol{D}\boldsymbol{S}_1^T)\circ(\boldsymbol{S}_1\boldsymbol{D}) = O_P\left(\frac{1}{n^2h_{n,1}^2}\right)_{n\times n} \overset{2.15}{=} o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n}.$$

Z lemmatu 13 navíc plyne:

$$(\mathrm{diag}(\boldsymbol{b}_1))\boldsymbol{D} = o_P(1)\boldsymbol{D}, \quad \boldsymbol{D}\circ(\boldsymbol{b}_1\boldsymbol{b}_1^T) = o_P(1)\boldsymbol{D},$$

$$(\mathrm{diag}(\boldsymbol{b}_1))\boldsymbol{S}_1\boldsymbol{D} = O_P(h_{n,1}^2)O_P\left(\frac{1}{nh_{n,1}}\right)_{n\times n} = o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n},$$

$$(\boldsymbol{S}_1\boldsymbol{D})\circ(\boldsymbol{b}_1\boldsymbol{b}_1^T) = O_P\left(\frac{1}{nh_{n,1}}\right)_{n\times n} O_P(h_{n,1}^4) = o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n},$$

$$(\boldsymbol{S}_1\boldsymbol{D}\boldsymbol{S}_1^T)\circ(\boldsymbol{b}_1\boldsymbol{b}_1^T) = O_P\left(\frac{1}{nh_{n,1}}\right)_{n\times n} O_P(h_{n,1}^4) = o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n}.$$

Tedy celkově dostáváme

$$\begin{aligned}\mathrm{var}[\hat{\sigma}^2_{FY,n}(ch_{n,2})|\boldsymbol{X}] &= \boldsymbol{e}_1^T\,\mathrm{diagonal}(\boldsymbol{S}_2[\boldsymbol{T}-\boldsymbol{\Sigma}^2 + o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n} + o_P(1)\boldsymbol{D}]\boldsymbol{S}_2^T)\\ &= \boldsymbol{e}_1^T\,\mathrm{diagonal}(\boldsymbol{S}_2\,\mathrm{diag}(\boldsymbol{\eta})\boldsymbol{S}_2^T) + o_P\left(\frac{1}{nh_{n,2}}\right),\end{aligned}$$

kde $\boldsymbol{\eta} = (\sigma^4(X_1)\lambda^2(X_1),\ldots,\sigma^4(X_n)\lambda^2(X_n))$. Zřejmě funkce $\eta(x) = \sigma^4(x)\lambda^2(x) = \sigma^4(x)(t(x)-1)$ je spojitá na $[0,1]$, stačí tedy aplikovat 2 z lemmatu 14 a dostáváme požadovaný výsledek.

□

Poznámky k větě 15:

- Větu 15 lze analogicky zformulovat a dokázat pro pravý hraniční bod. Stačí předpokládat, že $\ddot{\sigma}^2(\cdot)$ je zleva spojitá v 1 (místo spojitosti zprava v 0), pak platí:

$$\mathrm{bias}[\hat{\sigma}^2_{FY,n}(1-ch_{n,2})|\boldsymbol{X}] = \frac{h_{n,2}^2}{2}\ddot{\sigma}^2(1-)b_P^W(c) + o_P(h_{n,2}^2),$$

$$\mathrm{var}[\hat{\sigma}^2_{FY,n}(1-ch_{n,2})|\boldsymbol{X}] = v_P^W(c)\frac{\sigma^4(1-)\lambda^2(1-)}{f_X(1-)nh_{n,2}} + o_P\left(\frac{1}{nh_{n,2}}\right),$$

kde konstanty $b_P^W(c)$, $v_P^W(c)$ jsou definované analogicky jako ve větě 5 pro jádrovou funkci $W(\cdot)$.

- Analogicky může být formulována a dokázána věta o podmíněném asymptotickém vychýlení a rozptylu FY odhadu pro vnitřní body $x \in (0,1)$. Stačí předpokládat, že $\ddot{\sigma}^2(\cdot)$ je spojitá na okolí bodu x (místo spojitosti v 0 zprava), pak platí:

$$\operatorname{bias}[\hat{\sigma}^2_{FY,n}(x)|\boldsymbol{X}] = \frac{h^2_{n,2}}{2}\ddot{\sigma}^2(x)\int u^2 W(u)\,du + o_P(h^2_{n,2}),$$

$$\operatorname{var}[\hat{\sigma}^2_{FY,n}(x)|\boldsymbol{X}] = \frac{\sigma^4(x)\lambda^2(x)\int W^2(u)\,du}{f_X(x)nh_{n,2}} + o_P\left(\frac{1}{nh_{n,2}}\right).$$

- Z vlastností jádrové funkce snadno nahlédneme, že pro konstantní faktory platí:

$$\lim_{c\to 1} b^W_L(c) = \lim_{c\to 1} b^W_P(c) = \int u^2 W(u)\,du,$$

$$\lim_{c\to 1} v^W_L(c) = \lim_{c\to 1} v^W_P(c) = \int W^2(u)\,du,$$

  což jsou přesně konstantní faktory v podmíněném asymptotickém vychýlení a rozptylu ve vnitřních bodech. Tento výsledek odpovídá našim očekáváním, neboť pro $c \geq 1$ jsou body $ch_{n,2}$ a $1 - ch_{n,2}$ vnitřními body a platí pro ně předcházející poznámka.

  Navíc pro rovnoměrné, Epanechnikovo a Gaussovo jádro autoři článku Fan a Gijbels (1992) ukázali, že:

$$\left(b^W_L(c)\right)^2 < \left(\int u^2 W(u)\,du\right)^2, \quad v^W_L(c) > \int W^2(u)\,du,$$

  pro všechna $c < 1$. Odsud plyne, že druhá mocnina podmíněného vychýlení FY odhadu je pro hraniční body nižší než pro vnitřní body, tedy minimálně za předpokladu stejných hodnot $\ddot{\sigma}^2(\cdot)$. Podmíněný rozptyl je pro hraniční body vyšší než pro vnitřní body, tedy minimálně za předpokladu stejných hodnot $\sigma^4(\cdot)$, $\lambda^2(\cdot)$ a $f_X(\cdot)$. Tyto výsledky odpovídají našim očekáváním, protože v hraničních bodech je komplexnost odhadu menší (používá se menší počet pozorování), a proto je obecně menší vychýlení a větší rozptyl.

- Z věty 5 přímo plyne, že FY odhad má stejné podmíněné asymptotické vychýlení a rozptyl v hraničních bodech jako jeho příslušný oracle odhad. Tedy FY odhad má 94,64% podmíněnou asymptotickou eficienci v krajních bodech nosiče $f_X(\cdot)$ (viz 1.3.2).

- Věta může být formulována i pro obecný lokálně polynomický odhad rozptylové funkce (přesněji řečeno lokálně polynomický odhad stupně p_2 rozptylové funkce založený na reziduích lokálně polynomického odhadu stupně p_1 regresní funkce). Nás ovšem zajímal především jeho speciální případ, FY odhad.

2.4.2 XP odhad v hraničních bodech

Odvodíme podmíněnou asymptotiky XP odhadu v hraničním bodě $ch_{n,2}$.

Věta 16. *Nechť platí předpoklady věty 15 a podmínky 2. Pak platí*

$$\text{bias}[\hat{\sigma}^2_{XP,n}(ch_{n,2})|\boldsymbol{X}] = \frac{h^2_{n,2}\underline{W}_1\ddot{\sigma}^2(0+)}{2\underline{W}_0} + o_P(h^2_{n,2}),$$

$$var[\hat{\sigma}^2_{XP,n}(ch_{n,2})|\boldsymbol{X}] = \frac{\underline{W}_2\sigma^4(0+)\lambda^2(0+)}{\underline{W}^2_0 f_X(0+)nh_{n,2}} + o_P\left(\frac{1}{nh_{n,2}}\right),$$

kde $\underline{W}_0$, $\underline{W}_1$ *a* $\underline{W}_2$ *jsou definované ve větě 8.*

Důkaz. Začneme vychýlením.

$$\mathsf{E}\left[\hat{\sigma}_{XP}(ch_{n,2})|\boldsymbol{X}\right] = \frac{\sum_{i=1}^n \hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})\,\mathsf{E}\left[\hat{r}^2_i|\boldsymbol{X}\right]}{\sum_{i=1}^n \hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})}.$$

Rezidua lze psát

$$\hat{r}_i = [m(X_i) - \hat{m}^{LL}_n(X_i)] + \sigma(X_i)\epsilon_i \Rightarrow$$

$$\hat{r}^2_i = \sigma^2(X_i)\epsilon^2_i + 2\sigma(X_i)\epsilon_i[m(X_i) - \hat{m}^{LL}_n(X_i)] + [m(X_i) - \hat{m}^{LL}_n(X_i)]^2.$$

Dosadíme $\hat{m}^{LL}_n(X_i) = \sum_{j=1}^n (m(X_j) + \sigma(X_j)\epsilon_j)W^{LL}_{nj}(X_i)$, využijeme vlastností náhodné složky $\boldsymbol{\epsilon}$ a dostáváme

$$\mathsf{E}\left[\hat{r}^2_i|\boldsymbol{X}\right] = \sigma^2(X_i)[1 - 2W^{LL}_{ni}(X_i)] + \text{MSE}[\hat{m}^{LL}_n(X_i)|\boldsymbol{X}].$$

Z lemmatu 13 a předpokladů na vyhlazovací parametry plyne

$$\text{MSE}[\hat{m}^{LL}_n(X_i)|\boldsymbol{X}] = O_P\left(h^4_{n,1} + \frac{1}{nh_{n,1}}\right) = o_P(h^2_{n,2}),$$

tedy tento člen můžeme zanedbat. Ukážeme, že

$$\frac{\sum_{i=1}^n \hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})\sigma^2(X_i)}{\sum_{i=1}^n \hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})} = \sigma^2(ch_{n,2}) + \frac{h^2_{n,2}\underline{W}_1\ddot{\sigma}^2(0+)}{2\underline{W}_0} + o_P(h^2_{n,2}).$$

Tedy $2W^{LL}_{ni}(X_i)$ můžeme také zanedbat, protože ze 4 v lemmatu 14 plyne

$$W^{LL}_{ni}(X_i) = O_P\left(\frac{1}{nh_{n,1}}\right) = o_P(h^2_{n,2}).$$

Provedeme Taylorův rozvoj $\sigma^2(X_i)$ v bodě $ch_{n,2}$ a dostáváme

$$\mathsf{E}\left[\hat{\sigma}_{XP}(ch_{n,2})|\boldsymbol{X}\right] = \sum_{i=1}^n \frac{\hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})}{\sum_{i=1}^n \hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})} \times$$

$$\left[\sigma^2(ch_{n,2}) + \dot{\sigma}^2(ch_{n,2})(X_i - ch_{n,2}) + \frac{\ddot{\sigma}^2(ch_{n,2})}{2}(X_i - ch_{n,2})^2 + Z_2(X_i)\right],$$

kde $Z_2(X_i) = R_2(X_i)(X_i - ch_{n,2})^2$ a $R_2(\cdot)$ je Peanův tvar zbytku. Výraz můžeme dále zjednodušit, protože $|X_i - ch_{n,2}| < h_{n,2}$, jinak by $W(\cdot)$ byla nulová. Odtud dostáváme

$$|R(X_i)| \leq \sup_{z:|z-ch_{n,2}|<h_{n,2}} |R(z)| = o(1) \Rightarrow R(X_i)(X_i - ch_{n,2})^2 = o_P(h_{n,2}^2).$$

Z podmínky (2.8) kladené na váhy $\hat{w}_{ni}(ch_{n,2})$ se výraz dále zjednodušuje:

$$\text{bias}[\hat{\sigma}_{XP}(ch_{n,2})|\boldsymbol{X}] = \sum_{i=1}^{n} \frac{\hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})\frac{\ddot{\sigma}^2(ch_{n,2})}{2}(X_i - ch_{n,2})^2}{\sum_{i=1}^{n}\hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})}.$$

Z lemmatu A.3. v článku Cai (2001) plyne:

$$\hat{w}_{ni}(ch_{n,2}) = \frac{1}{n(1 - \underline{\lambda}_c(X_i - ch_{n,2})W_{h_{n,2}}(X_i - ch_{n,2}))}[1 + o_P(1)],$$

kde $\underline{\lambda}_c$ je kořen rovnice $\underline{L}_c(\lambda) = 0$ a $\underline{L}_c(\cdot)$ je definovaná ve větě 8. Označme:

$$Z_i = \frac{W(\frac{X_i - ch_{n,2}}{h_{n,2}})}{(1 - \underline{\lambda}_c(X_i - ch_{n,2})W_{h_{n,2}}(X_i - ch_{n,2}))},$$

$$A_n = \frac{1}{nh_{n,2}^3}\sum_{i=1}^{n} Z_i\ddot{\sigma}^2(ch_{n,2})(X_i - ch_{n,2})^2/2, \quad B_n = \frac{1}{nh_{n,2}}\sum_{i=1}^{n} Z_i.$$

Z faktu, že $X_1, \ldots, X_n$ jsou i.i.d. pro každé pevné $n \in \mathbb{N}$, a s použitím vhodné substituce a Lebesgueovy věty (o záměně limity a integrálu), dostáváme:

$$\left[\mathsf{E}\,[A_n] \to \underline{W}_1 f_X(0+)\ddot{\sigma}^2(0+)/2 \wedge \text{var}[A_n] \to 0\right] \Rightarrow A_n \xrightarrow{P} \underline{W}_1 f_X(0+)\frac{\ddot{\sigma}^2(0+)}{2},$$

$$[\mathsf{E}\,[B_n] \to \underline{W}_0 f_X(0+) \wedge \text{var}[B_n] \to 0] \Rightarrow B_n \xrightarrow{P} \underline{W}_0 f_X(0+).$$

Požadovaný výsledek pak plyne přímo ze vztahu

$$\text{bias}[\hat{\sigma}_{XP}(ch_{n,2})|\boldsymbol{X}] = \frac{h_{n,2}^2 A_n[1 + o_P(1)]}{B_n[1 + o_P(1)]}.$$

Nyní rozptyl.

$$\text{var}[\hat{\sigma}_{XP}(ch_{n,2})|\boldsymbol{X}] = \text{var}\left[\frac{\sum_{i=1}^{n}\hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})\hat{r}_i^2}{\sum_{i=1}^{n}\hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})}\middle|\boldsymbol{X}\right] =$$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}_{ni}(ch_{n,2})\hat{w}_{nj}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}})W(\frac{X_j - ch_{n,2}}{h_{n,2}})\,\text{cov}[\hat{r}_i^2, \hat{r}_j^2|\boldsymbol{X}]}{(\sum_{i=1}^{n}\hat{w}_{ni}(ch_{n,2})W(\frac{X_i - ch_{n,2}}{h_{n,2}}))^2} = (*).$$

Z důkazu věty 15 plyne

$$\mathrm{var}[\hat{\boldsymbol{r}}^2|\boldsymbol{X}] = \boldsymbol{T} - \boldsymbol{\Sigma}^2 + o_P(1)\boldsymbol{D} + o_P\left(\frac{1}{nh_{n,2}}\right)_{n\times n} \Rightarrow$$

$$\mathrm{cov}[\hat{r}_i^2, \hat{r}_j^2|\boldsymbol{X}] = \delta_{i,j}[\sigma^4(X_i)\lambda^2(X_i) + o_P(1)] + (1-\delta_{i,j})o_P\left(\frac{1}{nh_{n,2}}\right),$$

tedy aplikací lemmatu A.3. z Cai (2001), dostáváme

$$(*) = \frac{C_n[1+o_P(1)]o_P(\frac{1}{nh_{n,2}}) + \frac{1}{nh_{n,2}}D_n[1+o_P(1)] + \frac{1}{nh_{n,2}}E_n o_P(1)}{\left(B_n[1+o_P(1)]\right)^2},$$

kde

$$C_n = \frac{1}{n^2h_{n,2}^2}\sum_{i\neq j} Z_iZ_j, \quad D_n = \frac{1}{nh_{n,2}}\sum_{i=1}^n Z_i^2[\sigma^4(X_i)\lambda^2(X_i)], \quad E_n = \frac{1}{nh_{n,2}}\sum_{i=1}^n Z_i^2.$$

Již víme, že $\left(B_n[1+o_P(1)]\right)^2 = \underline{W}_0^2 f_X^2(0+) + o_P(1)$. Opět výpočtem střední hodnoty a rozptylu lze ukázat:

$$C_n = O_P(1), \quad E_n = O_P(1), \quad D_n = \underline{W}_2 f_X(0+)\sigma^4(0+)\lambda^2(0+) + o_P(1).$$

Odsud již plyne požadovaný výsledek. Ukážeme výpočet pouze pro D_n, pro ostatní je postup analogický. Využíváme opět toho, že sčítance jsou i.i.d. Po vhodné substituci dostáváme

$$\mathsf{E}\,[D_n] = \int_{-c}^{1} \frac{W^2(t)f_X(h_{n,2}t + h_{n,2}c)\sigma^4(h_{n,2}t + h_{n,2}c)\lambda^2(h_{n,2}t + h_{n,2}c)}{(1-\underline{\lambda}_c tW(t))^2}\,dt.$$

Ověříme, že jsou splněny předpoklady Lebesgueovy věty (o záměně limity a integrálu). Funkce $f_X(\cdot)$, $\sigma^4(\cdot)$ a $\lambda^2(\cdot)$ jsou spojité na $[0,1]$, tedy existuje konstanta C tak, že $|f_X(h_{n,2}t + h_{n,2}c)\sigma^4(h_{n,2}t + h_{n,2}c)\lambda^2(h_{n,2}t + h_{n,2}c)| \leq C$ pro všechna t a n. Navíc integrál

$$\int_{-c}^{1} \frac{|W^2(t)|C}{|1-\underline{\lambda}_c tW(t)|^2}\,dt = C\underline{W}_2$$

existuje a je konečný, aplikací Lebesgueovy věty tedy dostáváme

$$\mathsf{E}\,[D_n] \to \underline{W}_2 f_X(0+)\sigma^4(0+)\lambda^2(0+).$$

$$\begin{aligned}\mathrm{var}[D_n] &= \frac{1}{nh_{n,2}}\left[\mathsf{E}\left[\frac{Z_1^4(\sigma^4(X_1)\lambda^2(X_1))^2}{h_{n,2}}\right] - h_{n,2}\left(\mathsf{E}\left[\frac{Z_1^2(\sigma^4(X_1)\lambda^2(X_1))}{h_{n,2}}\right]\right)^2\right] \\ &= \frac{1}{nh_{n,2}}\left[O(1) - O(h_{n,2})\right] = o(1).\end{aligned}$$

□

Poznámky ke větě 16:

- Podmínky lze zjednodušit. Protože $\{(X_i, Y_i)\}_{i\in\mathbb{N}}$ jsou nezávislé, podmínka 1d platí pro $\delta = 0$ a podmínka 1b se reguluje na $\mathsf{E}[Y^4] < \infty$. Některé další podmínky mohou být významně zmírněny a jsou uvedeny pro jednoduchost důkazu. Podmínky 2 nám umožňují použít lemmata A.2. a A.3. z článku Cai (2001).

- Větu 16 lze analogicky zformulovat a dokázat pro pravý hraniční bod. Stačí předpokládat, že $\ddot{\sigma}^2(\cdot)$ je zleva spojitá v 1 (místo spojitosti zprava v 0), pak platí:

$$\operatorname{bias}[\hat{\sigma}^2_{XP,n}(1 - ch_{n,2})|\boldsymbol{X}] = \frac{h^2_{n,2}\overline{W}_1\ddot{\sigma}^2(1-)}{2\overline{W}_0} + o_P(h^2_{n,2}),$$

$$\operatorname{var}[\hat{\sigma}^2_{XP,n}(1 - ch_{n,2})|\boldsymbol{X}] = \frac{\overline{W}_2\sigma^4(1-)\lambda^2(1-)}{\overline{W}^2_0 f_X(1-)nh_{n,2}} + o_P\left(\frac{1}{nh_{n,2}}\right),$$

  kde $\overline{W}_0$, $\overline{W}_1$ a $\overline{W}_2$ jsou definované ve větě 8.

- Analogicky, za použití lemmatu A.2. z Cai (2001), může být formulována a dokázána věta o podmíněném asymptotickém vychýlení a rozptylu XP odhadu pro vnitřní body $x \in (0, 1)$. Stačí předpokládat, že $\ddot{\sigma}^2(\cdot)$ je spojitá na okolí bodu x (namísto pravého okolí nuly) a ostatní předpoklady zachovat, pak platí:

$$\operatorname{bias}[\hat{\sigma}^2_{XP,n}(x)|\boldsymbol{X}] = \frac{h^2_{n,2}}{2}\ddot{\sigma}^2(x)\int u^2 W(u)\,du + o_P(h^2_{n,2}),$$

$$\operatorname{var}[\hat{\sigma}^2_{XP,n}(x)|\boldsymbol{X}] = \frac{\sigma^4(x)\lambda^2(x)\int W^2(u)\,du}{f_X(x)nh_{n,2}} + o_P\left(\frac{1}{nh_{n,2}}\right).$$

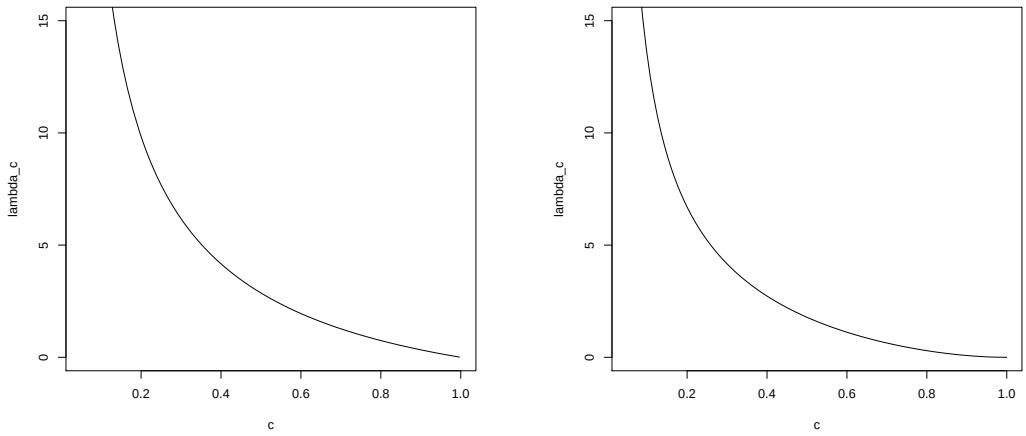
- Označme konstantní faktory vychýlení a rozptylu pro levé a pravé hraniční body:

$$B_L(c) = \frac{\underline{W}_1}{\underline{W}_0}, \quad V_L(c) = \frac{\underline{W}_2}{\underline{W}^2_0}, \quad B_P(c) = \frac{\overline{W}_1}{\overline{W}_0}, \quad V_P(c) = \frac{\overline{W}_2}{\overline{W}^2_0}.$$

  Z poznámky pod větou 8 plyne, že asymptotické chování FY odhadu nezáleží na tom, zda se jedná o levý nebo pravý hraniční bod, tj. $B_L(c) = B_P(c)$ a $V_L(c) = V_P(c)$ pro všechna c.

2.4.3 Porovnání FY a XP odhadu v hraničních bodech

Z vět 15 a 16 plyne, že podmíněný asymptotický rozptyl a vychýlení FY a XP odhadu v hraničních bodech se liší pouze v konstantních faktorech. Jejich srovnání pro rovnoměrné a Epanechnikovo jádro a vypočtené hodnoty $\overline{\lambda}_c$ nabízí následující obrázky.



(a) Rovnoměrné jádro (b) Epanechnikovo jádro

Obrázek 2.1: Vypočtené $\overline{\lambda}_c$ v závislosti na c u XP odhadu.

Poznámky k obrázkům 2.1:

- Pro vykreslení konstantních faktorů $B_L(c), V_L(c)$ u rovnoměrného jádra bylo třeba vyjádřit vztah $\overline{\lambda}_c$ a c z rovnice

$$\int_{-1}^{c} uW(u)/[1 - \lambda uW(u)]\, du \stackrel{!}{=} 0,$$

  což vedlo na transcendentní rovnici

$$2\log\left(\frac{2+\lambda}{2-c\lambda}\right) - \lambda(c+1) \stackrel{!}{=} 0,$$

  jejíž řešení jsme vyjádřili pomocí *Lambertovy* W funkce jako

$$c = \frac{2(W(z_c)+1)}{\overline{\lambda}_c}, \text{kde } z_c = -\frac{1}{2}(\overline{\lambda}_c + 2)\sqrt{e^{-\overline{\lambda}_c - 2}}$$

  a dopočítali pro různé hodnoty $\overline{\lambda}_c$. Lambertova W funkce je implementována v balíčku gsl (Hankin, 2006).

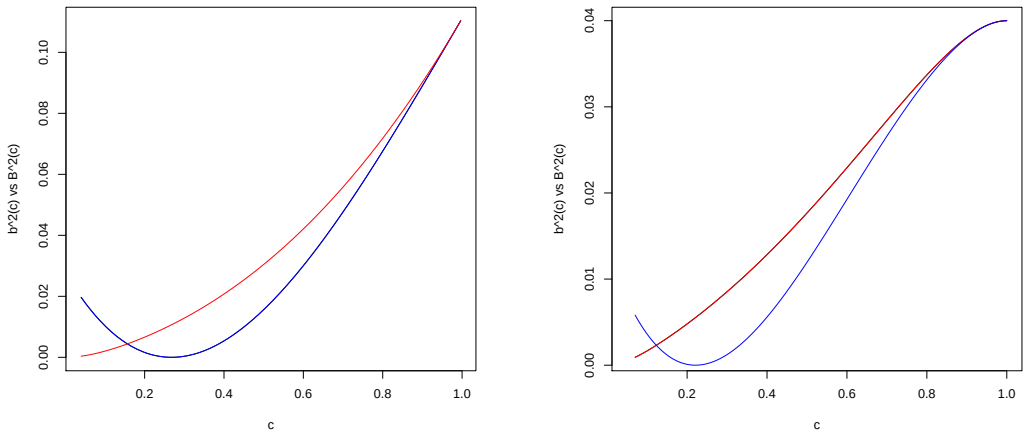
- U Epanechnikova jádra jsme $\underline{\lambda}_c$ získali pomocí metody půlení intervalů jako řešení

$$\underline{\lambda}_c = \underline{\lambda}(c) = \arg_\lambda F(\lambda, c) \stackrel{!}{=} 0,$$

  kde

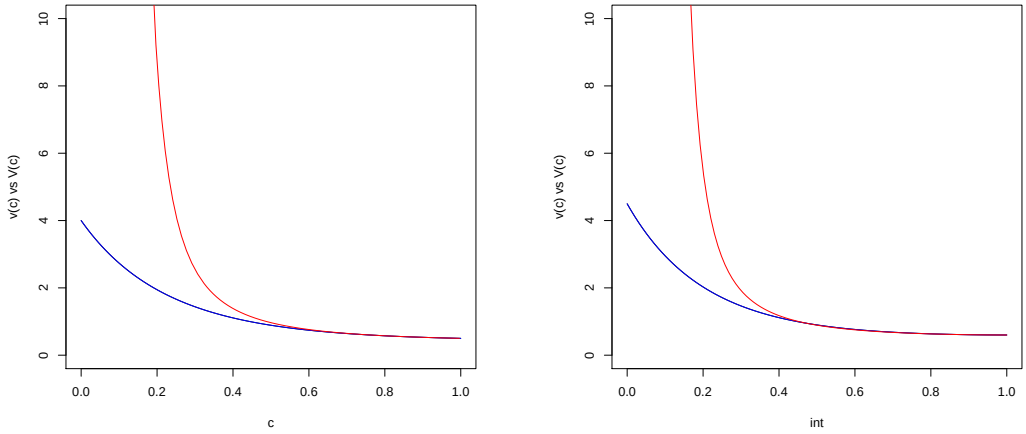
$$F(\lambda, c) = \int_{-c}^{1} \frac{W(u)u}{1 - \lambda uW(u)}.$$

  Pro konvergenci metody bylo třeba volit krajní body a, b intervalu hledání tak, aby $F(a,c)F(b,c) < 0$. Připomeňme, že hodnoty $\underline{\lambda}_c$ plynou z identity $\underline{\lambda}_c = -\overline{\lambda}_c$. Metoda půlení intervalů je implementována v balíčku rootSolve (Soetaert, 2015). Zdrojový kód je k nahlédnutí v příloze A.1.



(a) Rovnoměrné jádro (b) Epanechnikovo jádro

Obrázek 2.2: Srovnání druhých mocnin konstantních faktorů vychýlení: $b_L^2(c)$ vs $B_L^2(c)$. Modře FY odhad, červeně XP odhad.



(a) Rovnoměrné jádro (b) Epanechnikovo jádro

Obrázek 2.3: Srovnání konstantních faktorů rozptylu: $v_L(c)$ vs $V_L(c)$. Modře FY odhad, červeně XP odhad.

Poznámky k obrázkům 2.2 a 2.3:

- Protože oba odhady mají stejné podmíněné asymptotické vychýlení a rozptyl pro vnitřní body (viz poznámky pod větami 15 a 16), rozdíly konstantních faktorů pro vychýlení i rozptyl jsou limitně nulové pro $c \to 1$.
- Navíc z obrázků 2.2a, 2.2b je patrné, že $b_L^2(c) < B_L^2(c)$ pro $c > 0{,}16$ u rovnoměrného jádra a pro $c > 0{,}12$ u Epanechnikova jádra, jinde $b_L^2(c) > B_L^2(c)$. Srovnání konstantních faktorů rozptylu vychází jednoznačně ve prospěch FY odhadu. Z obrázků 2.3a a 2.3b plyne, že $v_L(c) < V_L(c)$ pro $c < 0{,}45$ a $v_L(c) \approx V_L(c)$ pro $c > 0.45$ u obou jádrových funkcí. Navíc $V_L(c)$ roste velmi rychle pro $c \to 0$. Můžeme tedy říct, že celkově lepší výsledky pro hraniční body vykazuje FY odhad.

- Z našich výpočtů mimo jiné vyplývají vztahy pro znaménko podmíněného asymptotického vychýlení obou odhadů v hraničních bodech. Pro XP odhad při použití obou jader platí

$$\operatorname{sgn}\left(\operatorname{abias}[\hat{\sigma}^2_{XP,n}(ch_{n,2})|\boldsymbol{X}]\right) = \operatorname{sgn}\left(\ddot{\sigma}^2(0+)\right), \quad c \in [0,1]; \tag{2.18}$$

pro FY odhad s Epanechnikovou jádrovou funkcí platí

$$\operatorname{sgn}\left(\operatorname{abias}[\hat{\sigma}^2_{FY,n}(ch_{n,2})|\boldsymbol{X}]\right) = \begin{cases} -\operatorname{sgn}\left(\ddot{\sigma}^2(0+)\right), & c \leq 0{,}22; \\ +\operatorname{sgn}\left(\ddot{\sigma}^2(0+)\right), & c > 0{,}22; \end{cases} \tag{2.19}$$

a pro FY odhad s rovnoměrnou jádrovou funkcí

$$\operatorname{sgn}\left(\operatorname{abias}[\hat{\sigma}^2_{FY,n}(ch_{n,2})|\boldsymbol{X}]\right) = \begin{cases} -\operatorname{sgn}\left(\ddot{\sigma}^2(0+)\right), & c \leq 0{,}27; \\ +\operatorname{sgn}\left(\ddot{\sigma}^2(0+)\right), & c > 0{,}27. \end{cases} \tag{2.20}$$

Poznamenejme ještě, že pro oba odhady ve vnitřních bodech $x \in (0,1)$ platí

$$\operatorname{sgn}\left(\operatorname{abias}[\hat{\sigma}^2_{n}(x)|\boldsymbol{X}]\right) = \operatorname{sgn}\left(\ddot{\sigma}^2(x)\right), \text{neboť } \sigma^2_W > 0.$$

2.5 Úvaha: modifikovaný FY odhad

V předcházejících podkapitolách jsme ukázali, že FY odhad má ideální asymptotické vlastnosti. Tedy je (téměř) asymptoticky eficientní (ve smyslu vět 2 a 3), adaptivní k neznámé regresní funkci a netrpí hraničními efekty. Jediným jeho nedostatkem je, že může nabývat záporných hodnot. Proto jsme se zabývali XP odhadem, který je pro vnitřní body asymptoticky ekvivalentní FY odhadu (tedy eficientní ve stejném smyslu), netrpí hraničními efekty a navíc má zaručenu nezápornost. Nevýhodou ovšem je jeho vyšší výpočetní náročnost a především horší chování na hranici, jak jsme ukázali v 2.4.3. Tyto nedostatky nás motivují k definovaní modifikovaného FY odhadu (dále také FYM odhadu):

$$\hat{\sigma}^2_{FYM,n}(x) = \max\{\hat{\sigma}^2_{FY,n}(x), 0\}.$$

Odhad má pak zřejmě zaručenu nezápornost. Odvodíme asymptotiku odhadu.

Věta 17. *Nechť platí předpoklady věty 7, pak $\hat{\sigma}^2_{FYM,n}(\cdot)$ má (pro vnitřní body) stejné asymptotické rozdělení jako FY odhad.*

Důkaz. Z věty 7 víme

$$X_n := (nh_{n,2})^{\frac{1}{2}}[\hat{\sigma}^2_{FY,n}(x) - \sigma^2(x) - \theta_n] \xrightarrow{d} Z,$$

kde $Z \sim \mathcal{N}(0, \tau^2)$ a $\tau^2 = f_X^{-1}(x)\sigma^4(x)\lambda^2(x)\int W^2(t)\,dt$. Chceme dokázat, že

$$Y_n := (nh_{n,2})^{\frac{1}{2}}[\hat{\sigma}^2_{FYM,n}(x) - \sigma^2(x) - \theta_n] \xrightarrow{d} Z.$$

Označme $F_n(\cdot)$ distribuční funkci Y_n, $G_n(\cdot)$ distribuční funkci X_n a konečně $\Phi_\tau(\cdot)$ distribuční funkci Z. Z věty 18.4. v Jacod a Protter (2004) a spojitosti $\Phi_\tau(\cdot)$ ve všech bodech plyne, že nám stačí ukázat, že pro všechna $y \in \mathbb{R}$ platí

$$F_n(y) = \Phi_\tau(y) + o(1),$$

a zároveň víme, že pro všechna $y \in \mathbb{R}$ platí

$$G_n(y) = \Phi_\tau(y) + o(1).$$

Z trojúhelníkové nerovnosti dále plyne

$$|F_n(y) - \Phi_\tau(y)| \leq |F_n(y) - G_n(y)| + |G_n(y) - \Phi_\tau(y)| = |F_n(y) - G_n(y)| + o(1).$$

Dále můžeme psát

$$|F_n(y) - G_n(y)| = |P(\hat{\sigma}^2_{FYM,n}(x) \leq z_n) - P(\hat{\sigma}^2_{FY,n}(x) \leq z_n)|,$$

kde $z_n = \frac{y}{\sqrt{nh_{n,2}}} + \sigma^2(x) + \theta_n$ a z věty 7 plyne, že $\lim_{n\to\infty} z_n = \sigma^2(x) > 0$. Odtud již plyne požadovaný výsledek, neboť $P(\hat{\sigma}^2_{FYM,n}(x) \leq z) = P(\hat{\sigma}^2_{FY,n}(x) \leq z)$ pro všechna $z > 0$.

□

Poznámky k větě 17:

- V souladu s očekáváními jsme ukázali, že modifikovaný FY odhad je asymptoticky ekvivaletní (pro vnitřní body) FY odhadu. Tedy je (100%) asymptoticky eficientní mezi lineárními odhady (ve smyslu věty 3) a téměř asymptoticky eficientní mezi všemi odhady (ve smyslu věty 2).
- Analogicky se dá ukázat, že oracle odhad příslušný FYM odhadu, tj. :

  $$\hat{\sigma}^2_{FYMO,n}(x) = \max\{\hat{\sigma}^2_{FYO,n}(x), 0\},$$

  je asymptoticky ekvivalentní FYO odhadu, tedy i FY odhadu, tedy i FYM odhadu. Tedy FYM odhad je adaptivní.

Jak se však chová odhad, máme-li pouze konečný počet pozorování? Porovnejme rozptyl, vychýlení a střední čtvercovou chybu FYM odhadu s FY odhadem.

Věta 18. *Pro libovolné $n \in \mathbb{N}$ platí:*

$$var[\hat{\sigma}^2_{FYM,n}(x)] \leq var[\hat{\sigma}^2_{FY,n}(x)],$$

$$\mathrm{MSE}[\hat{\sigma}^2_{FYM,n}(x)] \leq \mathrm{MSE}[\hat{\sigma}^2_{FY,n}(x)].$$

Důkaz. Označme nejprve $X = \hat{\sigma}^2_{FY,n}(x)$, $Z = \hat{\sigma}^2_{FYM,n}(x)$, $a_1 = \mathsf{E}[X|X \geq 0] \geq 0$, $a_2 = \mathsf{E}[X|X < 0] \leq 0$, $b_1 = \mathsf{E}[X^2|X \geq 0] \geq a_1^2 \geq 0$, $b_2 = \mathsf{E}[X^2|X < 0] \geq a_2^2 \geq 0$, $p_1 = P(X \geq 0) \geq 0$, $p_2 = P(X < 0) \geq 0$, pak platí:

$$\begin{aligned}\mathrm{var}[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 &= b_1p_1 + b_2p_2 - (a_1p_1 + a_2p_2)^2 \\ &= b_1p_1 + b_2p_2 - a_1^2p_1^2 - 2a_1p_1a_2p_2 - a_2^2p_2^2,\end{aligned}$$

$$\mathrm{var}[Z] = \mathsf{E}[Z^2] - (\mathsf{E}[Z])^2 = b_1p_1 - a_1^2p_1^2,$$

$$\begin{aligned}\mathrm{var}[X] - \mathrm{var}[Z] = b_2p_2 - 2a_1p_1a_2p_2 - a_2^2p_2^2 &= b_2p_2 - 2a_1p_1a_2p_2 - a_2^2p_2(1-p_1) \\ &= p_2(b_2 - a_2^2) - 2a_1p_1a_2p_2 + p_1p_2a_2^2.\end{aligned}$$

Jelikož všechny sčítance na pravé straně jsou nezáporné, $\text{var}[X] - \text{var}[Z] \geq 0 \Rightarrow \text{var}[\hat{\sigma}^2_{FYM,n}(x)] \leq \text{var}[\hat{\sigma}^2_{FY,n}(x)]$. Nyní střední čtvercová chyba. Z faktu $\sigma^2(x) > 0$ plyne

$$|\hat{\sigma}^2_{FYM,n}(x) - \sigma^2(x)| \leq |\hat{\sigma}^2_{FY,n}(x) - \sigma^2(x)| \Rightarrow \text{MSE}[\hat{\sigma}^2_{FYM,n}(x)] \leq \text{MSE}[\hat{\sigma}^2_{FY,n}(x)].$$

□

Poznámky k větě 18:

- Pokud navíc platí $P(\hat{\sigma}^2_{FY,n}(x) < 0) > 0$, pak platí ostré nerovnosti, tj.:

$$\text{var}[\hat{\sigma}^2_{FYM,n}(x)] < \text{var}[\hat{\sigma}^2_{FY,n}(x)],$$

$$\text{MSE}[\hat{\sigma}^2_{FYM,n}(x)] < \text{MSE}[\hat{\sigma}^2_{FY,n}(x)].$$

- Je-li x_0 krajní bod nosiče $f_X(\cdot)$, pak analogicky platí

$$\text{MSE}[\hat{\sigma}^2_{FYM,n}(x_0)|\boldsymbol{X}] \leq \text{MSE}[\hat{\sigma}^2_{FY,n}(x_0)|\boldsymbol{X}], \quad n \in \mathbb{N},$$

  tedy FYM odhad má alespoň 94,64% podmíněnou asymptotickou eficienci v krajních bodech nosiče $f_X(\cdot)$ (viz 1.3.2).

- Analogická nerovnost platí např. i pro *průměrnou absolutní odchylku odhadu* ve vyhodnocovaných bodech $x_1, \ldots, x_m$ (zkráceně MAD), tj.:

$$\begin{aligned}\text{MAD}[\hat{\sigma}^2_{FYM,n}] &= \sum_{i=1}^{m} |\hat{\sigma}^2_{FYM,n}(x_i) - \sigma^2(x_i)| \\ &\leq \sum_{i=1}^{m} |\hat{\sigma}^2_{FY,n}(x_i) - \sigma^2(x_i)| = \text{MAD}[\hat{\sigma}^2_{FY,n}].\end{aligned}$$

- Vychýlení odhadů nemůžeme přímo porovnat, nicméně snadno nahlédneme, že $\mathsf{E}\,[\hat{\sigma}^2_{FYM,n}(x)] > \mathsf{E}\,[\hat{\sigma}^2_{FY,n}(x)]$. Tedy speciálně, pokud FY odhad nadhodnocuje skutečnou hodnotu rozptylové funkce (tj. $\mathsf{E}\,[\hat{\sigma}^2_{FY,n}(x)] > \sigma^2(x)$, viz (2.19) a (2.20)), je nový odhad více vychýlený. Pokud FY odhad podhodnocuje skutečnou hodnotu $\sigma^2(x)$, může být nový odhad více nebo i méně vychýlený.

- MSE je pro nás hlavní ukazatel kvality odhadu, jestliže odhadujeme $\sigma^2(x)$ pouze jednou (např. aplikace na empirická data). Naopak pokud odhady opakujeme za pomoci simulovaných dat, jako výsledný odhad $\sigma^2(x)$ obvykle bereme průměr odhadů. V tomto případě je podstatnější, aby byl odhad co nejméně vychýlen. Proto námi navržená modifikace je vylepšením FY odhadu v empirických příkladech, ale při simulačních studiích s l opakováním, pokud bychom $\sigma^2(x)$ odhadovali jako průměr odhadů $\{\hat{\sigma}^2_{FYM,ni}(x)\}_{i=1,\ldots,l}$, může vykazovat horší výsledky (speciálně pokud FY odhad nadhodnocuje skutečnou hodnotu rozptylové funkce).

2.6 Volba vyhlazovacího parametru při odhadu rozptylové funkce

Protože FY, FYM a XP odhad mají stejnou podmíněnou asymptotickou střední čtvercovou chybu jako lokálně lineární odhad FYO (podmíněná asymptotika FY a XP odhadu je popsána v poznámkách pod větami 15 a 16 a podmíněná asymptotika FYO odhadu plyne přímo z věty 4), můžeme použít metody volby vyhlazovacího parametru pro LL odhad regresní funkce popsané v 1.4. Označme $\hat{h}(X_1, \ldots, X_n; Y_1, \ldots, Y_n)$ některou z metod volby vyhlazovacího parametru založenou na datech $(X_1, Y_1), \ldots, (X_n, Y_n)$ jako například metodu křížového ověřování, popsanou v 1.4.5, či některou z klasických nebo plug-in metod. Pro výpočet odhadů rozptylové funkce používáme následující algoritmus:

1. Použijeme $h_{n,1} = \hat{h}(X_1, \ldots, X_n; Y_1, \ldots, Y_n)$ k výpočtu $\hat{m}_n^{LL}(X_i)$, pro $i = 1, \ldots, n$.
2. Spočítáme $\hat{r}_i^2 = (Y_i - \hat{m}_n^{LL}(X_i))^2$, $i = 1, \ldots, n$.
3. Použijeme $h_{n,2} = \hat{h}(X_1, \ldots, X_n; \hat{r}_1^2, \ldots, \hat{r}_n^2)$ k výpočtu odhadů rozptylové funkce.

V praktických výpočtech budeme používat metody doporučené ve Fan a Yao (1998). Z těch, které jsme popsali v 1.4, doporučují autoři článku buď metodu křížového ověřování, nebo plug-in metodu prezentovanou v článku Ruppert a kol. (1995a), přičemž preferují druhou zmíněnou metodu. Metoda křížového ověřování je implementována v balíčku locpol (Cabrera, 2012) pod názvem regCVBwSelC. Pro výpočet plug-in vyhlazovacího parametru LL odhadu s Gaussovým jádrem je v balíčku KernSmooth (Wand, 2015) implementována funkce dpill. V simulačních příkladech budeme pro srovnání také počítat asymptoticky optimální globální vyhlazovací parametr pro odhady FY, FYM a XP. Ve všech případech se jedná o stejný vyhlazovací parametr a z jejich podmíněné asymptotiky plyne

$$h_{n,2}^{opt} = \arg\min_{h_{n,2}>0} \int \mathrm{AMSE}[\hat{\sigma}_n^2(x)|\boldsymbol{X}] f_X(x) w(x)\, dx =$$

$$\arg\min_{h_{n,2}>0} \frac{h_{n,2}^4}{4} (\sigma_W^2)^2 \int (\ddot{\sigma}^2(x))^2 f_X(x) w(x)\, dx + \frac{\int W^2(t)\, dt}{n h_{n,2}} \int \sigma^4(x) \lambda^2(x) w(x)\, dx.$$

Odsud snadno spočítáme, že

$$h_{n,2}^{opt} = \left[\frac{\int W^2(t)\, dt \int \sigma^4(x) \lambda^2(x) w(x)\, dx}{n(\sigma_W^2)^2 \int (\ddot{\sigma}^2(x))^2 f_X(x) w(x)\, dx} \right]^{1/5}.$$

Pokud navíc zvolíme za $W(\cdot)$ Epanechnikovu jádrovou funkci a $w(x) = 1$, dostáváme

$$h_{n,2}^{opt} = \left[\frac{15 \int \sigma^4(x) \lambda^2(x)\, dx}{n \int (\ddot{\sigma}^2(x))^2 f_X(x)\, dx} \right]^{1/5}. \tag{2.21}$$

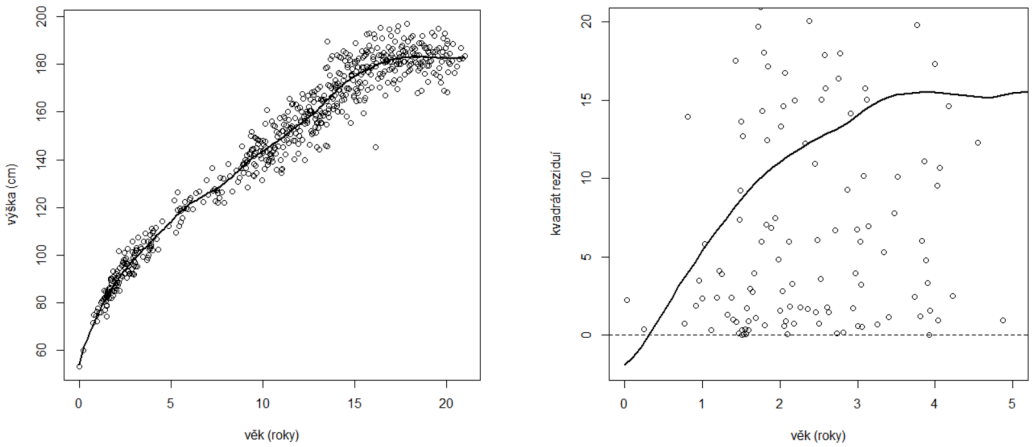
Pro výpočet plug-in vyhlazovacího parametru pro LL odhad s Epanechnikovou jádrovou funkcí vynásobíme parametr získaný funkcí dpill konstantou:

$$\left[\frac{\int W_{EPA}^2(t)\, dt (\sigma_{W_{GAUSS}}^2)^2}{(\sigma_{W_{EPA}}^2)^2 \int W_{GAUSS}^2(t)\, dt} \right]^{1/5} = (30\sqrt{\pi})^{1/5}. \tag{2.22}$$

2.7 Motivační příklad

Abychom ilustrovali riziko zápornosti FY odhadu, bylo potřeba najít vhodná data. Rozhodli jsme se modelovat výšku chlapců v závislosti na věku. Datový vzorek nazvaný boys7482 je dostupný v balíčku AGD (van Buuren, 2015).

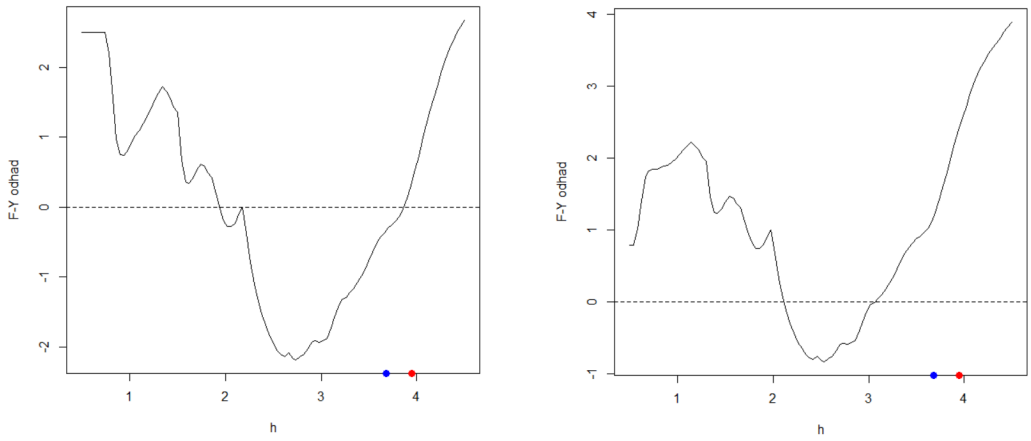
Negativní odhady rozptylové funkce nastávají typicky v bodech, v jejichž blízkém okolí je četnost pozorování relativně nízká, proto jsme vzorek dat „naředili“ tak, abychom snížili relativní četnost pozorování pro x blízké nule. Výsledný datový vzorek obsahuje 592 pozorování výšky chlapců v závislosti na věku, který se pohybuje od 0 do 21 let. Všechny datové úpravy a výpočty byly provedeny v programu R Core Team (2015) a jsou k nahlédnutí v přiloženém souboru. Pro výpočet LL odhadu jsme použili funkci locLinSmootherC z balíčku locpol (Cabrera, 2012). Implementaci XP odhadu jsme provedli sami pomocí funkce multiroot z balíčku rootSolve (Soetaert, 2015) a je uvedena v příloze A.2.



(a) LL odhad $m(x)$ (b) FY odhad

Obrázek 2.4: Výška chlapců v závislosti na věku: (a) LL odhad regresní funkce při použití $h_{n,1}^{CV} = 1{,}40$; (b) FY odhad rozptylové funkce založený na druhých mocninách reziduí při použití $h_{n,2} = 2{,}50$.

Na obrázku 2.4a vidíme LL odhad regresní funkce. Vyhlazovací parametr $h_{n,1}$ byl vybrán pomocí metody křížového ověřování popsané v sekci 1.4.5. Na základě tohoto odhadu jsme získali druhé mocniny reziduí a FY odhad rozptylové funkce pro 211 hodnot věků chlapců ekvidistantně rozdělených od $x = 0$ do $x = 21$. Na obrázku 2.4b je FY odhad znázorněn speciálně pro $x = 0$ až $x = 5$. Z obrázku můžeme nahlédnout, že pro malé hodnoty x je FY odhad záporný. Pro ilustraci jsme zvolili $h_{n,2} = 2{,}50$, nicméně z obrázku 2.5a, znázorňující FY odhad v bodě 0 v závislosti na $h_{n,2}$, je patrné, že FY odhad vychází záporně i pro vyhlazovací parametr získaný metodou křížového ověřování, ten je na ose x zaznačen modře, červeně je pak zaznačen plug-in vyhlazovací parametr (Ruppert a kol., 1995a). Pro porovnání jsme na obrázku 2.5b ještě vykreslili hodnotu FY odhadu v bodě 0,2 v závislosti na $h_{n,2}$. V tomto bodě by již ani jedna ze zmíněných metod nevedla k zápornému odhadu.

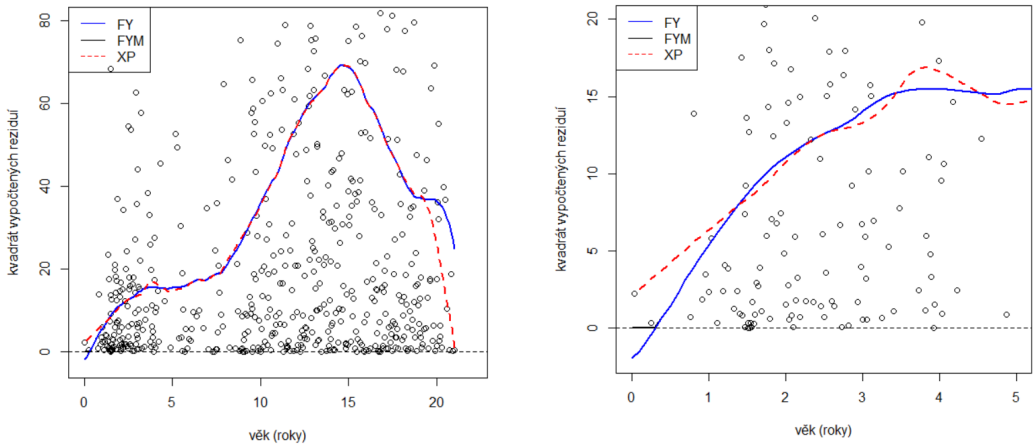


(a) $x = 0$ (b) $x = 0{,}2$

Obrázek 2.5: FY odhad v závislosti na vyhlazovacím parametru: (a) v bodě $x = 0$; (b) v bodě $x = 0{,}2$. Modře je označen vyhlazovací parametr získaný metodou křížového ověřování, červeně pak plug-in metodou.

Srovnání s XP odhadem

Riziko zápornosti odhadu nás motivovalo k popsání XP odhadu a posléze definovaní modifikovaného FY odhadu. Jak dopadly na stejných datech (s použitím stejných reziduí), se stejnou jádrovou funkcí (Epanechnikovou) a se stejným vyhlazovacím parametrem $h_{n,2} = 2{,}50$, můžeme nahlédnout na obrázku 2.6.

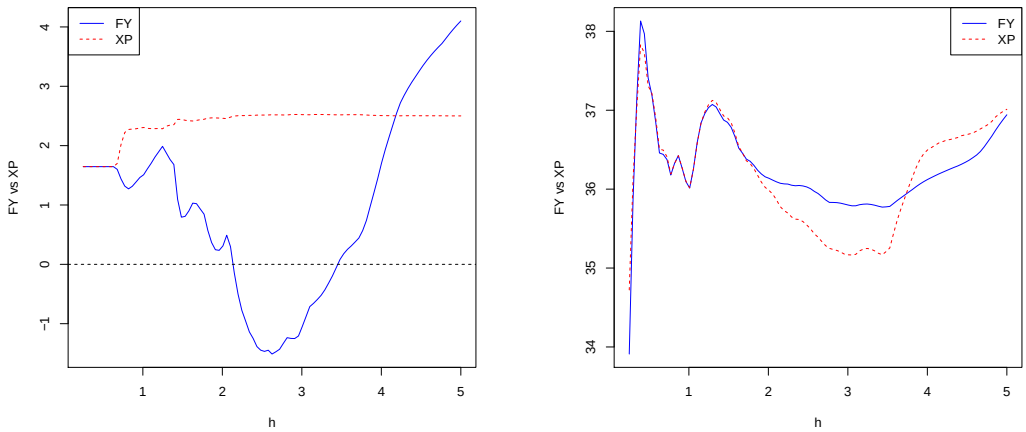


(a) (b)

Obrázek 2.6: Srovnání FY, FYM a XP odhadu na motivačních datech při použití vyhlazovacího parametru $h_{n,2} = 2{,}50$. Modře je označen FY odhad, modře a černě FYM odhad a červeně XP odhad.

Vykreslené křivky jsou spočteny opět v 211 evaluačních bodech ekvidistantně rozdělených od $x = 0$ do $x = 21$. Obrázek 2.6b je pouze přiblížením sousedního obrázku pro $x = 0$ až $x = 5$. Nepřekvapí nás, že XP odhad a modifikovaný FY odhad vycházejí nezáporně, neboť mají nezápornost garantovanou již z definice.

Z obrázku 2.6a můžeme dále nahlédnout, že pro vnitřní body se křivky FY a XP odhadu téměř překrývají, což nás také nepřekvapí, neboť odhady jsou (pro vnitřní body) asymptoticky ekvivalentní. Zajímavější je jejich srovnání v hraničních bodech, zejména vůči rozptylové funkci, čemuž se budeme věnovat v simulačních příkladech. Nejprve však srovnejme citlivost odhadů na změnu vyhlazovacího parametru. Na obrázku 2.7a je znázorněna závislost odhadů v hraničním bodě $x = 0{,}1$ pro vyhlazovací parametr $h_{n,2} = 0{,}25$ až $h_{n,2} = 5$. Je zde patrný velký rozdíl, XP je evidentně poměrně stabilní, pro $h_{n,2} > 1$ téměř konstantní. Naopak FY odhad je velmi citlivý a navíc, jak jsme ukázali dříve, pro některé vyhlazovací parametry nabývá záporných hodnot. Vysvětlením stability XP odhadu je podmínka (2.8). Díky ní je pozorováním napravo od bodu x s nenulovou jádrovou funkcí přidělována velmi malá váha. Tohle již neplatí u vnitřních bodů, což můžeme vidět na obrázku 2.7b.



(a) $x = 0{,}1$ (b) $x = 10$

Obrázek 2.7: FY a XP odhad v závislosti na vyhlazovacím parametru při použití Epanechnikovy jádrové funkce v bodě: (a) $x = 0{,}1$; (b) $x = 10$. Modře je označen FY odhad, červeně XP odhad.

Chování odhadů je zde velmi podobné. Celkově můžeme konstatovat, že u FY odhadu (především v hraničních bodech) je nutné klást větší důraz na volbu vyhlazovacího parametru i s ohledem na možnou zápornost odhadu. Poznamenejme ještě, že citlivost odhadů na volbu vyhlazovacího parametru je do velké míry dána také volbou jádrové funkce. Např. u rovnoměrného jádra bychom očekávali vyšší citlivost a u Gaussova jádra naopak nižší.

Výpočet vah u XP odhadu

Pro výpočet vah $\hat{w}_{ni}(x)$ u XP odhadu jsme vycházeli z poznámky (2.9). λ jsme nalezli pomocí Newton-Raphsonova iteračního algoritmu pro hledání kořene funkce (též zvaného metoda tečen), který jsme aplikovali na funkci $F(\cdot)$ defino-

vanou v (2.10). Kořen jsme hledali na intervalu $(\mathrm{DMEZ}, \mathrm{HMEZ})$, kde:

$$\mathrm{DMEZ} = \begin{cases} \frac{-1}{\max\limits_{1\le i\le n}\{(X_i-x)_+ W_{h_{n,2}}(X_i-x)\}}, & \max\limits_{1\le i\le n}\{(X_i-x)_+ W_{h_{n,2}}(X_i-x)\} > 0, \\ -\infty, & \max\limits_{1\le i\le n}\{(X_i-x)_+ W_{h_{n,2}}(X_i-x)\} = 0, \end{cases}$$

$$\mathrm{HMEZ} = \begin{cases} \frac{1}{\max\limits_{1\le i\le n}\{(x-X_i)_+ W_{h_{n,2}}(X_i-x)\}}, & \max\limits_{1\le i\le n}\{(x-X_i)_+ W_{h_{n,2}}(X_i-x)\} > 0, \\ \infty, & \max\limits_{1\le i\le n}\{(x-X_i)_+ W_{h_{n,2}}(X_i-x)\} = 0. \end{cases}$$

Uvedený interval zaručuje nezápornost vah. Zdrojový kód výpočtu XP odhadu je v příloze A.2.

2.8 Simulace: srovnání odhadů

Nyní provedeme několik simulací s různou rozptylovou funkcí $\sigma^2(\cdot)$, abychom otestovali vlastnosti odhadů na konečných výběrech. Ve všech případech simulujeme 200 náhodných výběrů délky $n = 201$ z modelu

$$Y_i = \exp\{x_i\} + \sigma(x_i)\epsilon_i, \tag{2.23}$$

kde nezávislou proměnnou předem volíme v ekvidistantně rozdělených bodech $x_1 = 0, x_2 = 0{,}005, \ldots, x_{201} = 1$ a náhodné chyby $\{\epsilon_i\}_{i=1}^n$ generujeme z normovaného normálního rozdělení. Vyhlazovací parametry $h_{n,1}$ a $h_{n,2}$ volíme pomocí plug-in metody popsané v článku Ruppert a Wand (1997) a používáme Epanechnikovu jádrovou funkci.

Vzhledem ke zvolenému designu nemůžeme příliš očekávat zápornost FY odhadu, tedy FY a FYM odhad budou téměř vždy stejné. Nicméně u rozdělení s velkými shluky a velkou rozptylovou funkcí může být FY odhad relativně často záporný, v takových případech je FYM odhad jeho významným vylepšením. Cílem simulační studie je empiricky potvrdit teoretické výsledky ze sekce 2.4.3 o horším chování XP odhadu v hraničních bodech.

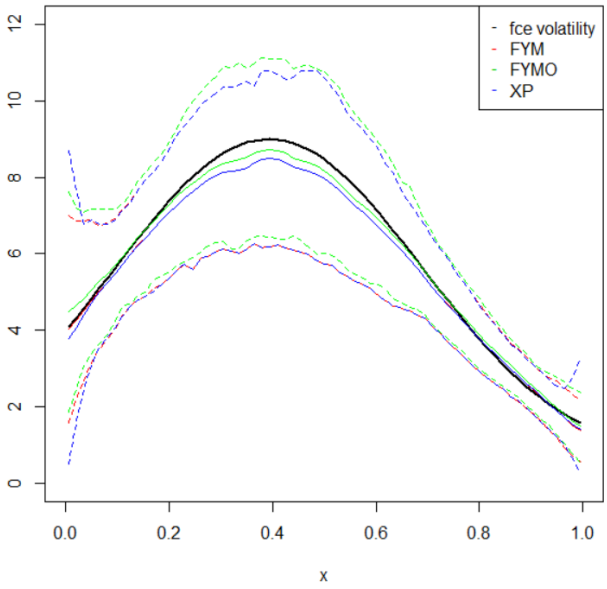
Simulace 1

Rozptylovou funkci volíme

$$\sigma^2(x) = (\sin(4x) + 2)^2.$$

Průměrná hodnota vyhlazovacího parametru pro odhad rozptylové funkce vychází $\mathsf{E}\,[h_{n,2}^{PI}] = 0{,}21$ a jeho směrodatná odchylka $\mathrm{sd}[h_{n,2}^{PI}] = 0{,}07$. Pro porovnání jsme také spočítali, dle vzorce (2.21), optimální vyhlazovací parametr $h_{n,2}^{opt} = 0{,}29$. Můžeme konstatovat, že pro většinu výběrů došlo k podhlazení odhadu. Pro každý vygenerovaný výběr dat spočítáme FY, FYM, FYMO a XP odhad v ekvidistantně rozdělených evaluačních bodech $x_1 = 0{,}005, x_2 = 0{,}020, \ldots, x_{67} = 0{,}995$. Jejich průměry, 10% a 90% kvantily můžeme porovnat na obrázku 2.8. Zřejmě nejlepších výsledků dosahuje FYMO odhad, jeví se nejméně vychýlený. FYM a XP odhady jsou pro vnitřní body téměř totožné, ovšem v hraničních bodech je rozpětí

kvantilů XP odhadu zřetelně větší a výrazně narůstá směrem ke krajním bodům intervalu, což odpovídá našim očekáváním (viz konstantní faktory podmíněného asymptotického rozptylu na obr. 2.3). Zda znaménka výběrových vychýlení odhadů odpovídají těm teoretickým (znaménkům podmíněnému asymptotickému vychýlení), si může čtenář snadno ověřit pomocí (2.18) a (2.19).



Obrázek 2.8: Průměry, 10% kvantily a 90% kvantily FYM, FYMO a XP odhadu rozptylové funkce $\sigma^2(x) = (\sin(4x) + 2)^2$ v modelu (2.23) na 200 opakováních a při použití plug-in vyhlazovacího parametru a Epanechnikovy jádrové funkce.

Porovnání průměrných absolutních odchylek odhadů

Nyní srovnáme průměrné (přes evaluační body) absolutní odchylky odhadů pro každý náhodný výběr, tj:

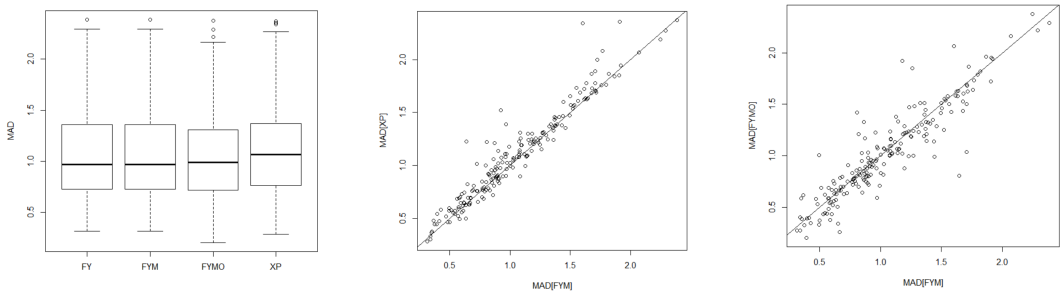
$$\mathrm{MAD}[\hat{\sigma}_n^2] = \frac{1}{67}\sum_{i=1}^{67} |\hat{\sigma}_n^2(x_i) - \sigma^2(x_i)|.$$

Boxploty $\mathrm{MAD}[\hat{\sigma}^2_{FY,n}]$, $\mathrm{MAD}[\hat{\sigma}^2_{FYM,n}]$, $\mathrm{MAD}[\hat{\sigma}^2_{FYMO,n}]$ a $\mathrm{MAD}[\hat{\sigma}^2_{XP,n}]$ spočtené na 200 simulacích jsou zobrazené na obrázku 2.9a. Vidíme, že (ve smyslu průměrných absolutních odchylek) nejhorších výsledků dosahuje XP odhad. FY, FYM a FYMO dosahují srovnatelných výsledků. Seřadíme-li průměrné hodnoty (přes všechny simulace) průměrných absolutních odchylek, dostaneme vzestupně:

$$\frac{1}{200}\sum_{s=1}^{200} \mathrm{MAD}[\hat{\sigma}^2_{FYMO,n,s}] = 1{,}0358 \quad \frac{1}{200}\sum_{s=1}^{200} \mathrm{MAD}[\hat{\sigma}^2_{FYM,n,s}] = 1{,}0546,$$

$$\frac{1}{200}\sum_{s=1}^{200} \mathrm{MAD}[\hat{\sigma}^2_{FY,n,s}] = 1{,}0547, \quad \frac{1}{200}\sum_{s=1}^{200} \mathrm{MAD}[\hat{\sigma}^2_{XP,n,s}] = 1{,}1075.$$

Nejlepší hodnotu tohoto ukazatele má FYMO odhad. Bylo předem jasné, že $\frac{1}{200}\sum_{s=1}^{200} \mathrm{MAD}[\hat{\sigma}^2_{FYM,n,s}] \leq \frac{1}{200}\sum_{s=1}^{200} \mathrm{MAD}[\hat{\sigma}^2_{FY,n,s}]$. V našem případě nastává dokonce ostrá nerovnost, tedy FY odhad nabývá v některých bodech záporných hodnot. Přímé srovnání FYM odhadu vůči FYMO odhadu a také FYM odhadu vůči XP odhadu nabízí obrázky 2.9b a 2.9c.



(a) Boxploty (b) FYM vs XP (c) FYM vs FYMO

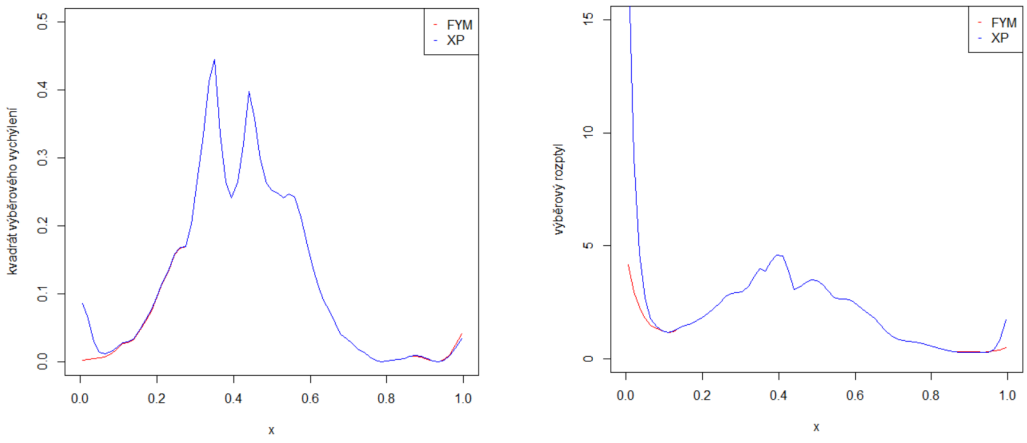
Obrázek 2.9: (a) Boxploty MAD FY, FYM, FYMO a XP odhadu; srovnání průměrných absolutních odchylek: (b) FYM vs XP; (c) FYM vs FYMO.

Je mírně vyšší šance, že FYM odhad „porazí“ XP odhad (ve smyslu nižší průměrné absolutní odchylky), nežli opačně a nepatrně vyšší šanci na to, že FYMO „porazí“ FYM odhad. Kvantifikujeme-li to přesně, pak situace

$$\mathrm{MAD}[\hat{\sigma}^2_{FYM,n,s}] < \mathrm{MAD}[\hat{\sigma}^2_{XP,n,s}]$$

nastává ve 138 případech ze 200 a ve 82 případech ze 200:

$$\mathrm{MAD}[\hat{\sigma}^2_{FYM,n,s}] < \mathrm{MAD}[\hat{\sigma}^2_{FYMO,n,s}].$$



(a) Kvadrát výběrového vychýlení (b) Výběrový rozptyl

Obrázek 2.10: Srovnání FY a XP odhadu: (a) druhá mocnina výběrového vychýlení; (b) výběrový rozptyl.

Na závěr můžeme porovnat druhou mocninu výběrového vychýlení a výběrový rozptyl FYM a XP odhadu. Z obrázku 2.10a je patrné, že kvadrát výběrového vychýlení XP odhadu je větší než u FYM odhadu v hraničních bodech, ve vnitřních

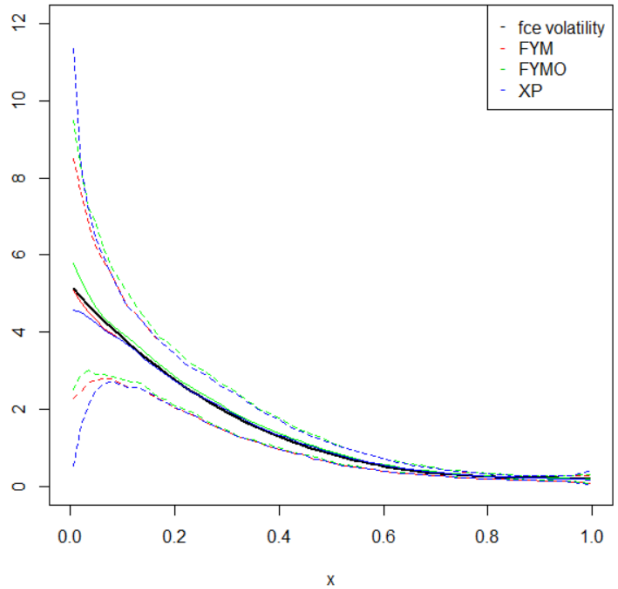
bodech je přibližně stejný. U výběrového rozptylu je situace obdobná, z obrázku 2.10b vidíme, že v hraničních bodech je větší výběrový rozptyl XP odhadu a rozdíl narůstá směrem ke krajním bodům nosiče $f_X(\cdot)$, jinde jsou výběrové rozptyly přibližně stejné. Poznamenejme ještě, že výběrovou MSE bychom získali prostým sečtením kvadrátu výběrového vychýlení a výběrového rozptylu.

Simulace 2

Rozptylovou funkci volíme

$$\sigma^2(x) = \frac{1}{5} + 5(1-x)^3.$$

Průměrná hodnota vyhlazovacího parametru pro odhad rozptylové funkce vychází $\mathsf{E}\,[h_{n,2}^{PI}] = 0{,}18$ a jeho směrodatná odchylka $\mathrm{sd}[h_{n,2}^{PI}] = 0{,}08$. Pro porovnání jsme také spočítali, dle vzorce (2.21), optimální vyhlazovací parametr $h_{n,2}^{opt} = 0{,}29$. Můžeme konstatovat, že pro většinu výběrů opět došlo k podhlazení odhadu. Odhady počítáme opět v evaluačních bodech $x_1 = 0{,}005, x_2 = 0{,}020, \ldots, x_{67} = 0{,}995$ a jejich kvantily jsou k nahlédnutí na obrázku 2.11. Můžeme si všimnout, že průměry a kvantily všech vykreslených odhadů jsou si velmi podobné s výjimkou levých hraničních bodů. Zde je rozpětí kvantilů XP odhadu zřetelně větší a významně narůstá směrem k nule. V těchto bodech je nejméně vychýlen FYM odhad, FYMO odhad (v průměru) nadhodnocuje a XP odhad podhodnocuje rozptylovou funkci.



Obrázek 2.11: Průměry, 10% kvantily a 90% kvantily FYM, FYMO a XP odhadu rozptylové funkce $\sigma^2(x) = \frac{1}{5} + 5(1-x)^3$ v modelu (2.23) na 200 opakováních a při použití plug-in vyhlazovacího parametru a Epanechnikovy jádrové funkce.

Porovnání průměrných absolutních odchylek odhadů

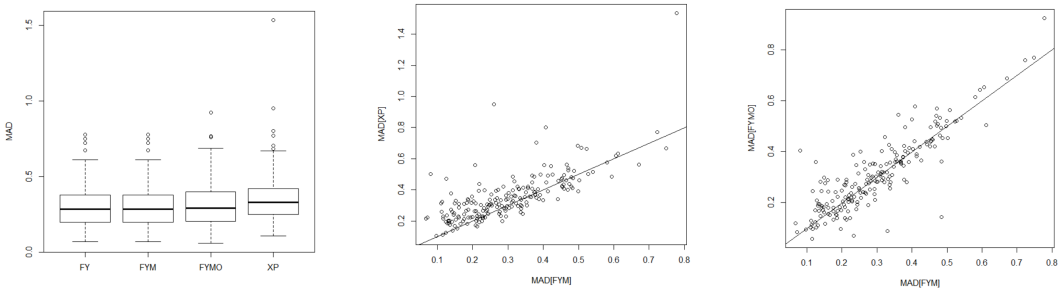
Boxploty $\mathrm{MAD}[\hat{\sigma}^2_{FY,n}]$, $\mathrm{MAD}[\hat{\sigma}^2_{FYM,n}]$, $\mathrm{MAD}[\hat{\sigma}^2_{FYMO,n}]$ a $\mathrm{MAD}[\hat{\sigma}^2_{XP,n}]$ spočtené na 200 simulacích jsou zobrazené na obrázku 2.12a. Viditelně nejhorších výsledků

dosahuje XP odhad, a to zřejmě v důsledku velmi vysokého rozptylu v levých hraničních bodech. Modifikovaný FY odhad vykazuje srovnatelné výsledky jako jeho oracle odhad. Pokud navíc zprůměrujeme hodnoty průměrných absolutních odchylek přes všechny simulace, dostaneme vzestupně:

$$\frac{1}{200}\sum_{s=1}^{200}\mathrm{MAD}[\hat{\sigma}^2_{FYM,n,s}] = 0{,}3003, \quad \frac{1}{200}\sum_{s=1}^{200}\mathrm{MAD}[\hat{\sigma}^2_{FYMO,n,s}] = 0{,}3111,$$

$$\frac{1}{200}\sum_{s=1}^{200}\mathrm{MAD}[\hat{\sigma}^2_{XP,n,s}] = 0{,}3553.$$

Tentokrát nejlepší hodnotu tohoto ukazatele má FYM odhad. Přímé srovnání FYM odhadu vůči FYMO odhadu a také FYM odhadu vůči XP odhadu nabízí obrázky 2.12b a 2.12c.



(a) Boxploty (b) FYM vs FYMO (c) FYM vs XP

Obrázek 2.12: (a) Boxploty MAD FY, FYM, FYMO a XP odhadu; srovnání průměrných absolutních odchylek: (b) FYM vs XP; (c) FYM vs FYMO.

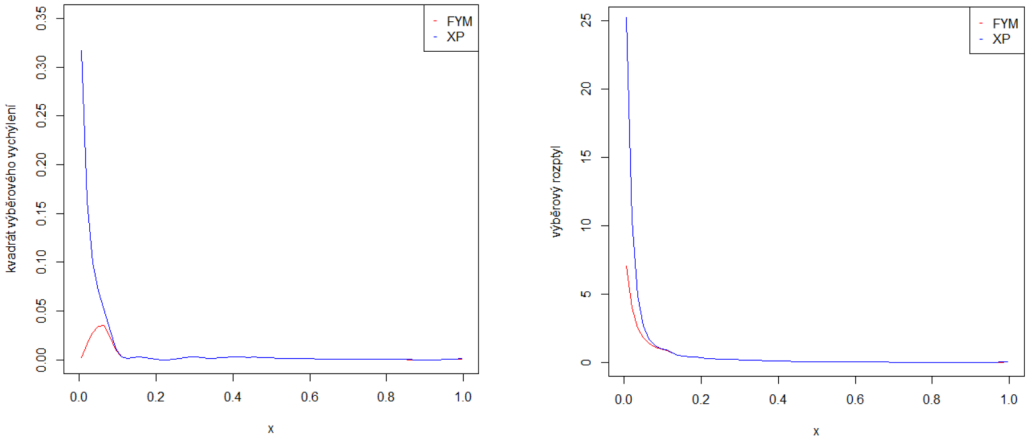
Zřejmě je mírně vyšší šance, že FYM odhad „porazí“ XP odhad (ve smyslu nižší průměrné absolutní odchylky) nežli opačně. FYM a FYMO odhad mají přibližně stejné šance na to, že „porazí“ jeden druhého. Kvantifikujeme-li to přesně, pak situace

$$\mathrm{MAD}[\hat{\sigma}^2_{FYM,n,s}] < \mathrm{MAD}[\hat{\sigma}^2_{XP,n,s}]$$

nastává ve 138 případech ze 200 a ve 105 případech ze 200:

$$\mathrm{MAD}[\hat{\sigma}^2_{FYM,n,s}] < \mathrm{MAD}[\hat{\sigma}^2_{FYMO,n,s}]$$

Na závěr můžeme porovnat druhou mocninu výběrového vychýlení a výběrový rozptyl FYM a XP odhadu. Z obrázku 2.13a je patrné, že kvadrát výběrového vychýlení XP odhadu je vyšší v levých hraničních bodech, jinde přibližně stejný. Na obrázku 2.13b vidíme, že výběrové rozptyly obou odhadů jsou srovnatelné až na levé hraniční body, kde výběrový rozptyl XP odhadu velmi rychle roste pro $x \to 0+$.



(a) Kvadrát výběrového vychýlení (b) Výběrový rozptyl

Obrázek 2.13: Srovnání FYM a XP odhadu: (a) druhá mocnina výběrového vychýlení; (b) výběrový rozptyl.

Tyto výsledky odpovídají našemu očekávání plynoucímu z teoretických výsledků v sekci 2.4.3. V obou simulačních případech jsme ukázali, že FYM odhad vykazuje lepší výsledky než XP odhad (ve smyslu průměrné absolutní odchylky). To je dáno především jeho podstatně menším rozptylem v hraničních bodech. Mimo jiné jsme také empiricky potvrdili adaptivitu FYM odhadu, neboť vycházel velmi blízko svému oracle odhadu.

S použitím přiloženého kódu si může samotný čtenář vyzkoušet chování odhadů i pro jiné rozptylové funkce, případně i jiný design nezávislé proměnné.

Závěr

Tato diplomová práce se zaměřuje na problematiku lokálně polynomických odhadů funkce podmíněného rozptylu v heteroskedastickém regresním modelu. V první kapitole, věnované odhadům regresní funkce, jsme důkladně popsali skvělé asymptotické vlastnosti lokálně lineárního odhadu regresní funkce. V druhé kapitole, věnované odhadům rozptylové funkce, jsme ukázali, že se tyto vlastnosti přenášejí i do lokálně lineárního odhadu rozptylu založeném na reziduích. Nicméně Xu a Phillips upozornili na riziko zápornosti lokálně lineárního odhadu rozptylu a navrhli vlastní odhad založený na reziduích, který má zaručenu nezápornost. Ačkoliv ukázali, že oba odhady jsou asymptoticky ekvivalentní pro vnitřní body, nesrovnali teoreticky jejich chování v hraničních bodech. To byl jeden z cílů této práce. Podařilo se nám ukázat, že FY odhad má podstatně lepší asymptotické chování v hraničních bodech, což se potvrdilo i v simulační studii. To nás motivovalo k triviální modifikaci FY odhadu, která zaručuje jeho nezápornost a nezhoršuje jeho kvalitu (ve smyslu střední čtvercové chyby). Navíc jsme ukázali, že modifikovaný odhad je asymptoticky ekvivalentní FY odhadu pro vnitřní body. Kromě odhadů založených na reziduích jsme se také zamysleli nad přímočarým odhadem a ukázali jeho nevhodnost kvůli vyššímu asymptotickému vychýlení.

Mezi další populární neparametrické metody, jejíchž popis by byl nad rámec této práce, patří vyhlazovací spliny. Vztah mezi jádrovou regresí a vyhlazovací spliny byl popsán v článku Silverman (1984), obsáhlý popis těchto metod včetně jejich aplikace na odhad rozptylové funkce najdeme v knize Wang (2011).

Seznam použité literatury

Bowman, A. W. a Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. First edition. Oxford University Press, New York.

Cabrera, J. L. O. (2012). *locpol: Kernel local polynomial regression*. URL `http://CRAN.R-project.org/package=locpol`.

Cai, Z. (2001). Weighted Nadaraya-Watson regression estimation. *Statistics and Probability Letters*, **51**(3), 307–318.

Cheng, M. Y., Fan, J. a Marron, J. S. (1993). *Minimax Efficiency of Local Polynomial Fit Estimators at Boundaries*. Institute of Statistics MIMEO series. University of North Carolina at Chapel Hill.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, **21**(1), 196–216.

Fan, J. a Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, **20**(4), 2008–2036.

Fan, J. a Gijbels, I. (1995). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, **4**(1), 213–217.

Fan, J. a Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. First edition. Chapman and Hall, London, New York.

Fan, J. a Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, **3**(1), 35–56.

Fan, J. a Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**(3), 645–660.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M. a Engel, J. (1995). On nonparametric estimation via local polynomial regression.

Fan, J., Gijbels, I., Hu, T.-C. a Huang, L.-S. (1996). An asymptotic study of variable bandwidth selection for local polynomial regression with application to density estimation. *Statistica Sinica*, **6**(1), 113–127.

Hall, P. a Marron, J. S. (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Th. Rel.*, **90**(2), 149–173.

Hankin, R. K. S. (2006). Special functions in R: introducing the gsl package. *R News*, **6**.

Hastie, T. J. a Tibshirani, R. J. (1990). *Generalized Additive Models*. First edition. Chapman and Hall/CRC, London.

Härdle, W. (1990). *Applied Nonparametric Regression*. First edition. Cambridge University Press.

Härdle, W., Hall, P. a Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, **83**(401), 86–95.

Härdle, W. a Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, **81**(1), 223–242.

Hurvich, C. M., Simonoff, J. S. a Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Sociaty*, **60**(2), 271–293.

Jacod, J. a Protter, P. (2004). *Probability Essentials*. Second edition. Springer-Verlag, Paris.

Koláček, J. (2004). *Jádrové odhady regresní funkce*. PhD thesis, Masarykova univerzita, Brno.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, **9**, 141–142.

Omelka, M. (2015). Zápisy z přednášky Moderní statistické metody. URL `https://www.karlin.mff.cuni.cz/~omelka/Vyuka_nmst434_1617.php`.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Ruppert, D. a Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, **22**(3), 1346–1370.

Ruppert, D. a Wand, M. P. (1997). Local polynomial variance-function estimation. *Technometrics*, **39**(3), 262–273.

Ruppert, D., Sheather, S. J. a Wand, M. P. (1995a). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**(432), 1257–1270.

Ruppert, D., Wand, M. P., Holst, M. P. a Hőssjer, O. (1995b). Local polynomial variance function estimation. *Technometrics*, **39**(3), 262–273.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics*, **12**, 898–916.

Soetaert, K. (2015). *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*. URL `http://CRAN.R-project.org/package=rootSolve`.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, **5**(4), 595–645.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, **8**(6), 1348–1360.

Tibshirani, R. a Hastie, T. J. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**, 559–567.

van Buuren, S. (2015). *AGD: Analysis of Growth Data*. URL `http://CRAN.R-project.org/package=AGD`.

Wand, M. (2015). *KernSmooth: Functions for Kernel Smoothing*. URL `http://CRAN.R-project.org/package=KernSmooth`.

Wang, Y. (2011). *Smoothing splines: methods and applications*. CRC Press.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics*, **26**(4), 359–372.

Xu, K.-L. a Phillips, P. C. B. (2012). Tilted nonparametric estimation of volatility functions with empirical applications. *Journal of Business and Economic Statistics*, **29**(4), 518–528.

Yao, Q. a Tong, H. (1994). Quantifying the influence of initial values on nonlinear prediction. *Journal of the Royal Statistical Society*, **56**(4), 701–725.

Seznam obrázků

Seznam použitého značení

AMISE	asymptotická střední integrovaná čtvercová chyba
AMSE	asymptotická střední čtvercová chyba
$\approx$	přibližně rovno
ASE	průměrná čtvercová chyba
$\asymp$	stejný asymptotický řád
$\boldsymbol{S}_h$	vyhlazovací matice jádrového odhadu
$\circ$	Hadamardův součin matic
$\delta_{i,j}$	Kroneckerovo delta
$\hat{\sigma}_n^2(x)$	odhad rozptylové funkce
$\hat{m}_n(x)$	odhad regresní funkce
$\mathrm{I}_{[x \in A]}$	indikační funkce, 1 pokud $x \in A$, nula jinak
ISE	integrovaná čtvercová chyba
MAD	průměrná absolutní odchylka
MASE	střední průměrná čtvercová chyba
MISE	střední integrovaná čtvercová chyba
MSE	střední čtvercová chyba
$\mathbb{N}$	množina všech přirozených čísel
$\mathbb{R}$	množina všech reálných čísel
RMS	reziduální střední čtverec
sgn	znaménková funkce
$\sigma^2(x)$	rozptylová funkce
$f_X(x)$	hustota X
$K(x)$	jádrová funkce u odhadu regresní funkce
$m(x)$	regresní funkce
$W(x)$	jádrová funkce u odhadu rozptylové funkce
$W_{ni}(x)$	váhová funkce
FY	Fan-Yaoův odhad

FYM	modifikovaný FY odhad
FYMO	oracle odhad příslušný FYM odhadu
FYO	oracle odhad příslušný FY odhadu
LL	lokálně lineární odhad
NW	Nadaraya-Watsonův odhad
XP	Xu-Phillipsův odhad

Přílohy

A. Zdrojový kód programu

V této příloze uvádíme části zdrojového kódu, které jsme programovali sami. Celý zdrojový kód je k dispozici na přiloženém CD. Výpočet konstantních faktorů b_L, v_L a B_L, V_L pro Epanechnikovu jádrovou funkci je v podkapitole A.1. Pro rovnoměrnou jádrovou funkci byl postup analogický s tím rozdílem, že hodnoty $\overline{\lambda}_c$ jsme spočítali přímo pomocí Lambertovy W funkce. Implementace XP odhadu je v podkapitole A.2.

A.1 Výpočet konstantních faktorů

```
# knihovna metod hledání kořene
library(rootSolve)

# Spočítáme lambda_cD (dolní) pomocí metody půlení intervalu.
# a značí lambda_cD.

FE <- function(a){
   f0E <- function(u){
       (u*(3/4)*(1-u^2))/(1-a*u*(3/4)*(1-u^2))
   }
   return(f0E)
}

KE <- function (a,c) {
   # Funkce F(a,c) pro Epa jádro
   integrate(FE(a),  lower=-c, upper=1)$value
}

# Analyzovali jsme průběh funkce u*(3/4)*(1-u^2).
# Na intervalu (-1, 1) má minimum -sqrt(1/12) v bodě -sqrt(1/3) .
# Na intervalu (-sqrt(1/3), 0) je funkce rostoucí.
# Tedy hledáme lambda_cD na intervalech:
# a) (-2 *sqrt(3), 0), je-li c >= sqrt(1/3),
# b) (-4 / (3c(1-c^2)), 0), je-li c < sqrt(1/3).

koren <- function(c){
   # c >= sqrt(1/3)
   uniroot(function(a) KE(a, c), c(-2 *sqrt(3)+0.00001, 0))$root
}
koren2 <- function(c){
   # c < sqrt(1/3)
   uniroot(function(a) KE(a,c),c(-4/(3*c*(1-c^2))+0.00001,0))$root
```

```
}

# Spočítáme lambda pro ekvid. posloupnost bodů c od 0.07 do 1.
cc <- seq(0.07, 1, length.out = 101)
lambda <- rep(1:101)
for (i in 1:55) {
   #c < sqrt(1/3)
   lambda[i]=koren2(cc[i])
}
for (i in 1:46) {
   # c >= sqrt(1/3)
   lambda[i+55]=koren(cc[i+55])
}

# KONSTANTNÍ FAKTORY XP ODHADU

E0 <- function(i){
   e0 <- function(u){
      ((3/4)*(1-u^2))/(1-lambda[i]*u*(3/4)*(1-u^2))
   }
   return(e0)
}

WE0 <- function (i) {
   # W0 - Epanechnikovo jádro
   integrate(E0 (i),  lower=-cc[i], upper=1)$value
}

E1 <- function(i){
   e1 <- function(u){
      ((u^2)*(3/4)*(1-u^2))/(1-lambda[i]*u*(3/4)*(1-u^2))
   }
   return(e1)
}

WE1 <- function (i) {
   # W1 - Epanechnikovo jádro
   integrate(E1(i),  lower=-cc[i], upper=1)$value
}

# Konstantní faktor vychýlení XP odhadu B(c)
biasXPE <- rep(1:101)
for (i in 1:101) {
   biasXPE[i]=WE1(i)/WE0(i)
}

E2 <- function(i){
   e2 <- function(u){
```

```
      (((3/4)*(1-u^2))/(1-lambda[i]*u*(3/4)*(1-u^2)))^2
   }
   return(e2)
}

WE2 <- function (i) {
   # W2 - Epanechnikovo jádro
   integrate(E2(i),  lower=-cc[i], upper=1)$value
}

# Konstantní faktor rozptylu XP odhadu V(c)
rozptylXPE <- rep(1:101)
for (i in 1:101) {
   rozptylXPE[i]=WE2(i)/(WE0(i))^2
}

# KONSTANTNÍ FAKTORY FY ODHADU

IE <- function(j){
   integrandE <- function(u){
      (u^j)*(3/4)*(1-u^2)
   }
   return(integrandE)
}


UE <- function(j, i){
   # mi(j, c)
   integrate(IE(j), lower=-cc[i], upper=1)$value
}

# Konstantní faktor b(c) v posloupnosti cc
biasFYE <- rep(1:101)
for (i in 1:101){
   biasFYE[i]=( (UE(2, i)^(2))-UE(1, i)*UE(3, i) )/
   ( UE(2, i)*UE(0, i)-(UE(1, i)^(2)) )
}

# Rozptyl FY spočítáme pro ekvid. posl. bodů c od 0 do 1.
int <- seq(0, 1, length.out = 101)
UEint<- function(j, i){
   integrate(IE(j), lower=-int[i], upper=1)$value
}

VE <- function(i){
   vnitrekE <- function(u){
      ((9*(1-u^2)^2)*(UEint(2, i)-u*UEint(1, i))^2)/
      (16*(UEint(2, i)*UEint(0,i)-UEint(1, i)^2)^2)
```

```
   }
   return(vnitrekE)
}

rozptylFYE <- c(1:101)
for (i in 1:101){
   rozptylFYE[i]=integrate(VE(i), lower=-int[i],  upper=1)$value
}
```

A.2 Implementace XP odhadu

```
# knihovna metod hledání kořene
library(rootSolve)


Epanechnik=function(x, h, z){
   # Epanechnikova jádrová funkce
   ((3/h)*(1-((x-z)/h)^2)/4)*(ifelse(abs(((x-z)/h))<=1, 1, 0))
}

# Dolní a horní mez intervalu, kde hledáme kořen.
# Zaručují nezápornost vah.
dmez <- function(x, h, z){
   ifelse(max((x-z)*(x-z>0)*Epanechnik(x, h, z))>0,
   -1/( max((x-z)*(x-z>0)*Epanechnik(x, h, z)) ),-Inf)
}
hmez <- function(x, h, z){
   ifelse(max((z-x)*(z-x>0)*Epanechnik(x,h,z))>0,
   1/( max((z-x)*(z-x>0)*Epanechnik(x, h, z)) ), Inf)
}


F <- function(a, x, h, z){
   # Mimo interval (dmez, hmez) předefinujeme na log(abs(a)).
   # Uvnitř intervalu (dmez, hmez) se jedná o funkci F.
   # F je definovaná v rovnice (2.10).
   ifelse((dmez(x, h, z)<a)&(a<hmez(x, h, z)),
   sum(((1+a*(x-z)*Epanechnik(x, h, z))^(-1))*
   (x-z)*Epanechnik(x, h, z)), log(abs(a)))
}

# Výpočet lambda pomocí multiroot (Newton-Raphsonova metoda).
# Hledáme kořen nejvýše ze tří startovních bodů:
# 0, dmez+0.01 (pokud je konečná) a hmez-0.01 (pokud je konečná)

lambdaL <- function(x, h, z){
   multiroot(function(a) F(a, x, h, z),
```

```
   ifelse((dmez(x,h,z)>-Inf),dmez(x,h,z)+0.01,0),rtol=1e-10)$root
}
lambda0 <- function(x, h, z){
   multiroot(function(a) F(a, x, h, z),0,rtol = 1e-10)$root
}
lambdaP <- function(x, h, z){
   multiroot(function(a) F(a, x, h, z),
  ifelse((hmez(x,h,z)<Inf),hmez(x,h,z)-0.01,0),rtol = 1e-10)$root
}


minn <- function(x, h, z){
   which.min(c(abs(F(lambdaL(x, h, z),x, h, z)),
   abs(F(lambda0(x, h, z),x, h, z)),
   abs(F(lambdaP(x, h, z),x, h, z))))
}

lambda <- function(x, h, z) {
   # Zvolíme ten bod, který nejlépe splňuje definici kořene.
   ifelse(minn(x, h, z)==1,lambdaL(x, h, z),
   ifelse(minn(x, h, z)==2,lambda0(x, h, z),lambdaP(x, h, z)))
}

# kontrola kořene
F(lambda(x, h, z), x, h, z)


w <- function(x, i, h, z){
   # vahy XP odhad
   1/(length(x)*(1+lambda(x,h,z)*(x[i]-z)*Epanechnik(x[i], h, z)))
}

# kontrola vah
sum( w(x, seq(1:length(x)), h, z) )


XPodhad <- function(x, r, h, z){
   # XP odhad
   XP <- (sum(w(x,seq(1:length(x)),h, z)*Epanechnik(x, h, z)*r) )/
   (sum(w(x, seq(1:length(x)), h, z)*Epanechnik(x, h, z)))
   return(XP)
}
```