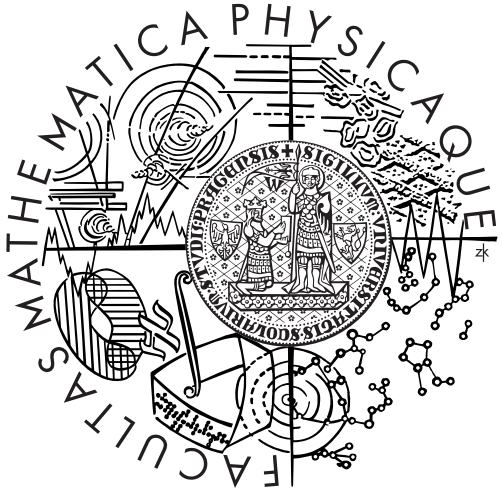
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Tomáš Jurczyk

Metody pro odhadování růstových křivek

Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Mgr. Michal Kulich, Ph.D.
Studijní program: Matematika, Matematická statistika

Poděkování

Na tomto místě bych rád poděkoval svému vedoucímu diplomové práce Mgr. Michalu Kulichovi, Ph.D. za cenné rady, návrhy a připomínky, které přispěly ke zlepšení této práce.

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

V Praze dne 16. dubna 2007 Tomáš Jurczyk

Obsah

Název práce: Metody pro odhadování růstových křivek
Autor: Tomáš Jurczyk
Katedra (Ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Mgr. Michal Kulich, Ph.D.
e-mail vedoucího: kulich@karlin.mff.cuni.cz
Abstrakt: Potřeba sestavování růstových křivek vedla ke vzniku velkého počtu metod, které se snaží růstové křivky odhadovat. Tato práce porovnává dvě dnes hojně používané metody: LMS metodu založenou na penalizované věrohodnosti a metodu kvantilové regrese s regresní maticí odpovídající vybrané B-splinové bázi. V rámci simulačních studií bylo vygenerováno 118 datových souborů různého typu a velikosti. Na odhadnutých růstových křivkách vzniklých na těchto souborech dat byly zkoumány základní vlastnosti obou metod, kvalita odhadnutých křivek a vhodnost použití metod pro různé typy dat. Metoda kvantilové regrese není oproti LMS metodě zatížena předpoklady na její použití, proto nás především zajímá chování LMS metody při nedodržení jejich předpokladů. Základním rysem této práce je, že pro určení kvality odhadnutých křivek bylo využito znalosti teoretického rozdělení dat. Znalost skutečných růstových křivek je přínosná především při hledání nejlepších modelů jednotlivých metod. Ve srovnání s metodami porovnávání odhadnutých křivek na reálných datech, dává tento přístup věcnější a objektivnější srovnání.
Klíčová slova: Růstové křivky, LMS metoda, kvantilová regrese, B-splinová báze.

Title: Growth Curve Estimation Methods
Author: Tomáš Jurczyk
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Michal Kulich, Ph.D.
Supervisor's e-mail address: kulich@karlin.mff.cuni.cz
Abstract: The need for constructing growth curves led to creation of many methods, which are estimating growth curves. A number of methods for estimation of growth curves have been proposed. This thesis compares two commonly used methods: the LMS method based on penalized likelihood and quantile regression on B-spline basis functions. In a simulation study, 118 data sets of different sample sizes were generated from several models. Growth curves were estimated from these data sets by both methods. The properties of the methods, the quality of the estimated curves and the performance of the methods under various circumstances were examined. Unlike quantile regression, the LMS method requires parametric assumptions, so we focused on the behaviour of the LMS method when the assumptions are violated. In this thesis, the knowledge of the true distribution of the data is used not only to evaluate the quality of the estimated growth curves but also to select the best model. In contrast to comparing estimated growth curves obtained from real data sets by competing methods, this approach provides a more relevant and more objective comparison.
Keywords: Growth curve, LMS method, quantile regression, B-spline basis.

Kapitola 1

Růstové křivky

1.1 Motivace

V medicíně, ale i v jiných oborech, vede řada experimentů k potřebě vytvoření rozmezí hodnot pro sledovanou veličinu (znak). Pokud už máme takové rozmezí, dá se pak jednoduše posoudit, jestli je napozorovaná veličina vně nebo uvnitř tohoto rozmezí, z čehož se pak může vyvodit nějaký závěr.

Potřeba vyjadřovat tato rozmezí v závislosti na nějaké spojité veličině (v medicíně je to většinou věk jedince) vede ke vzniku kvantilových křivek (křivky, vyjadřující kvantily veličiny závislé na nějaké spojité veličině). Růstové křivky (někdy také referenční křivky) se skládají ze skupiny kvantilových křivek, které tak ilustrují rozdělení nějakého znaku.

1.2 Použití růstových křivek v medicíně

Asi nejvíce je pojem růstových křivek znám z medicíny. Růstové křivky jsou používány především pediatry ke sledování růstu dětí. Slouží k porovnání nějakého tělesného znaku vzhledem k populaci dětí stejného věku (odděleně pro každé pohlaví). Sestavují se pozorováním velkého počtu dětí. Po sestavení pak křivky slouží jako nástroj k rozpoznání možné poruchy růstu dětí, přičemž rozhodnutí závisí na tom, jestli hodnota sledovaného znaku dítěte leží pod nebo nad určitou kvantilovou křivkou.

Křivky, které se u dětí standardně sledují: závislost výšky na věku, váhy na věku, váhy na výšce, obvod hlavy na věku, BMI (*body mass index*, BMI = $\frac{váha(kg)}{výška^2(m^2)}$) na věku. Většinou se křivky sestavují pro 3., 5., 10., 25., 50., 75., 90., 95. a 97. percentil.

Hlavní reprezentace křivek je v grafické podobě, kdy se do grafu vykreslí křivky základních kvantilů proti veličině, na které závisí.

1.3 Zaměření a struktura práce

Hlavním cílem této práce je porovnání metod pro sestavování růstových křivek na datech vytvořených v rámci simulačních studií. Těmto metodám je věnována kapitola 2. Je v ní uvedeno několik základních metod pro odhadování růstových křivek a obecná pravidla, která by měla být splněna jednotlivými metodami. Podrobněji jsou popsány dvě metody, které jsou v centru zájmu práce a to: LMS metoda (popsaná v části 2.3) a metoda kvantilové regrese s použitím B-splinové báze (popsaná v části 2.4).

Odhadnuté růstové křivky těchto dvou hlavních metod jsou porovnávány na množství datových souborů vytvořených v rámci simulačních studií. Kapitola 3 je zaměřena na problém generování dat. Také je v ní popsána hlavní myšlenka porovnávání modelů s využitím znalosti teoretického rozdělení nasimulovaných dat. Celá práce se věnuje porovnání metod na tomto základě. Ve srovnání s porovnáváním metod na reálných datech, by proto tato práce měla podat pohled na tuto problematiku z trochu jiného úhlu. Stejně tak i na úlohu, jak vybírat konkrétní parametry jednotlivých metod. V závěrečné části kapitoly 3 pak lze najít, jaké konkrétní datové soubory byly nagenerovány pro porovnání dvou hlavních metod.

Výsledky a vlastní srovnávání odhadnutých křivek lze najít v kapitole 4. Díky znalosti teoretického rozdělení dat (tedy i znalosti přesné polohy teoretických růstových křivek) se porovnávají odhadnuté křivky na základě toho, jak se blíží k křivkám teoretickým. Také se odhadnuté křivky srovnávají z hlediska, jak moc splňují požadavky, které jsou kladeny na tvar růstových křivek obecně.

Práce je zakončena shrnutím dosažených poznatků a stručným popisem toho, jak by se dalo na tuto práci navázat (kapitola 5).

Kapitola 2

Metody pro odhadování růstových křivek

2.1 Klasifikace metod a několik příkladů

Potřeba sestavování růstových křivek vedla ke vzniku velkého počtu metod, které se snaží tento problém řešit. Tyto metody se dělí do tří kategorií podle toho, jak se k problému přistupuje. Z velkého počtu metod uvedu alespoň některé základní. Metody se tedy dělí na

Parametrické:

Do této třídy patří například metody založené na transformacích, které se snaží převést data na normálně rozdělená. Po této transformaci stačí pro sestrojení podmíněných kvantilů (podmíněné spojitou veličinou) odhadnout podmíněnou střední hodnotu a podmíněný rozptyl. Tato střední hodnota a rozptyl jako funkce veličiny, na které závisí, se pak modelují pomocí lineární nebo obecnější polynomiální regrese.

Pozn.: Nevýhodou těchto metod je, že nemáme jistotu, zdali takováto normující transformace existuje a navíc předpoklady parametrického přístupu jsou vždy omezující.

Semiparametrické:

Hlavním zástupcem je LMS metoda [1, 2]. Růstové křivky odhadnuté touto metodou závisí na třech přirozených kubických splinech, z nichž jeden odhaduje podmíněný medián, druhý podmíněný variační koeficient veličiny, o kterou se zajímáme, a třetím splinem je pak křivka mocninné transformace, která má převést data na normálně rozdělená (tato metoda je v centru zájmu této práce, proto bude ještě podrobněji popsána).

Neparametrické:

Zde už nejsou kladeny žádné předpoklady na rozdělení dat. Patří sem metody založené na rozdělení nosiče spojité veličiny do podskupin, na nichž se pak spočítají empirické kvantily proměnné, o kterou se zajímáme. Pro vznik potřebných křivek se pak takto vypočtené skupiny kvantilů nakonec vyhladí. Další metodou je neparametrická kvantilová regrese založená na B-splinové bázi (podrobněji je popsána v části 2.4). A také sem patří například metody založené na jádrovém nebo lokálně konstantním odhadu podmíněných kvantilových křivek (více se o těchto metodách můžete dočíst v článku [8]).

2.2 Požadavky na metody pro sestavování růstových křivek

Následuje výčet některých základních požadavků, které by měly metody a kři/-vky jimi odhadnuté splňovat:

- Jednotlivé odhadnuté křivky pro jednotlivé kvantily se nesmí křížit.
- Křivky by měly být v rozmezí možných hodnot (např.: Jedná-li se o váhu jedince, pak by křivka neměla být v žádném úseku záporná).
- Výsledné křivky mají být hladké (hlavně kvůli tomu, že se předpokládá, že s malou změnou veličiny, na které znak závisí, se málo změní i vyšetřovaný znak).
- Křivky dobře odpovídají datům.
- Počet dat pod určitou křivkou by měl přibližně odpovídat očekávané hodnotě.
- Data mimo referenční křivky se neshlukují, ale jsou rozmístěny rovnoměrně podél celého rozsahu veličiny, na které závisí.
- Nejdůležitější je (vzhledem k použití) odhad krajních kvantilů.

2.3 LMS metoda

Tuto dnes hojně užívanou metodu pro odhadování růstových křivek navrhl T. J. Cole (1992) [2]. Je zařazována mezi semiparametrické metody.

Je založena na předpokladu, že po vhodné transformaci veličiny, o kterou se zajímáme (označíme ji jako Y), se dají data považovat za normálně rozdělená.

Box a Cox [3] navrhli mocninnou transformaci

$$\begin{aligned} X &= \frac{\left(\frac{Y}{\mu}\right)^{\lambda} - 1}{\lambda}, & \lambda \neq 0; \\ X &= \ln \frac{Y}{\mu}, & \lambda = 0. \end{aligned} \tag{2.1}$$

Předpokládá se tedy, že Y^{λ} ($\ln Y$ v případě $\lambda = 0$) má normální rozdělení, Y je kladná, μ je medián Y. Medián X je tedy 0. Díky symetrii normálního rozdělení je střední hodnota X také rovna 0.

Pro $\lambda = 1$ je normálně rozdělená i veličina Y a medián Y je roven střední hodnotě Y. Z toho plyne, že pro $\lambda = 1$ je směrodatná odchylka X rovna variačnímu koeficientu Y (směrodatná odchylka Y podělená střední hodnotou Y). Jak se píše v [2], toto přibližně platí i pro ostatní rozumná λ. Rovnici (2.1) podělíme směrodatnou odchylkou σ veličiny X (variačním koeficientem Y) a dostaneme

$$\begin{aligned} Z &= \frac{\left(\frac{Y}{\mu}\right)^{\lambda} - 1}{\lambda\sigma}, & \lambda \neq 0; \\ Z &= \frac{\ln \frac{Y}{\mu}}{\sigma}, & \lambda = 0. \end{aligned} \tag{2.2}$$

Potom Z už má normální rozdělení s nulovou střední hodnotou a jednotkovým rozptylem (to plyne z toho, že považujeme X za normálně rozdělené). Pokud teď budeme předpokládat, že rozdělení Y se mění v závislosti na t, dostáváme

$$\begin{aligned} Z &= \frac{\left(\frac{Y}{M(t)}\right)^{L(t)} - 1}{L(t)S(t)}, & L(t) \neq 0; \\ Z &= \frac{\ln \frac{Y}{M(t)}}{S(t)}, & L(t) = 0. \end{aligned} \tag{2.3}$$

Křivky $L(t)$, $M(t)$ a $S(t)$ odpovídají hodnotám λ, μ a σ pro daná t (díky těmto třem křivkám dostala metoda své označení). Z (2.3) vyplývají vztahy pro výpočet kvantilových křivek

$$\begin{aligned} Q_{\tau}(t) &= M(t)[1 + L(t)S(t)q_{\tau}]^{\frac{1}{L(t)}}, & L(t) \neq 0; \\ Q_{\tau}(t) &= M(t)\exp\{S(t)q_{\tau}\}, & L(t) = 0, \end{aligned} \tag{2.4}$$

kde q_{τ} je τ-kvantil normovaného normálního rozdělení.

2.3.1 Výpočet křivek $L(t)$, $M(t)$ a $S(t)$

Máme n nezávislých pozorování dat: $[t_i, Y_i], i = 1, \ldots, n$. Podle věty o substituci je hustota náhodného vektoru $(Y_1, \ldots, Y_n)^\top$ rovna

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{\left(\frac{y_i}{M(t_i)}\right)^{L(t_i)} - 1}{L(t_i)S(t_i)} \right)^2 \right\} \frac{\left(\frac{y_i}{M(t_i)}\right)^{L(t_i)-1}}{M(t_i)S(t_i)}. \quad (2.5)$$

Logaritmus věrohodnostní funkce odvozené z (2.5) vypadá až na konstantu takto

$$l(L, M, S) = \sum_{i=1}^{n} \left(L(t_i) \log \frac{Y_i}{M(t_i)} - \log S(t_i) - \frac{1}{2} Z_i^2 \right),$$

kde Z_i je hodnota příslušná Y_i spočtená z (2.3).

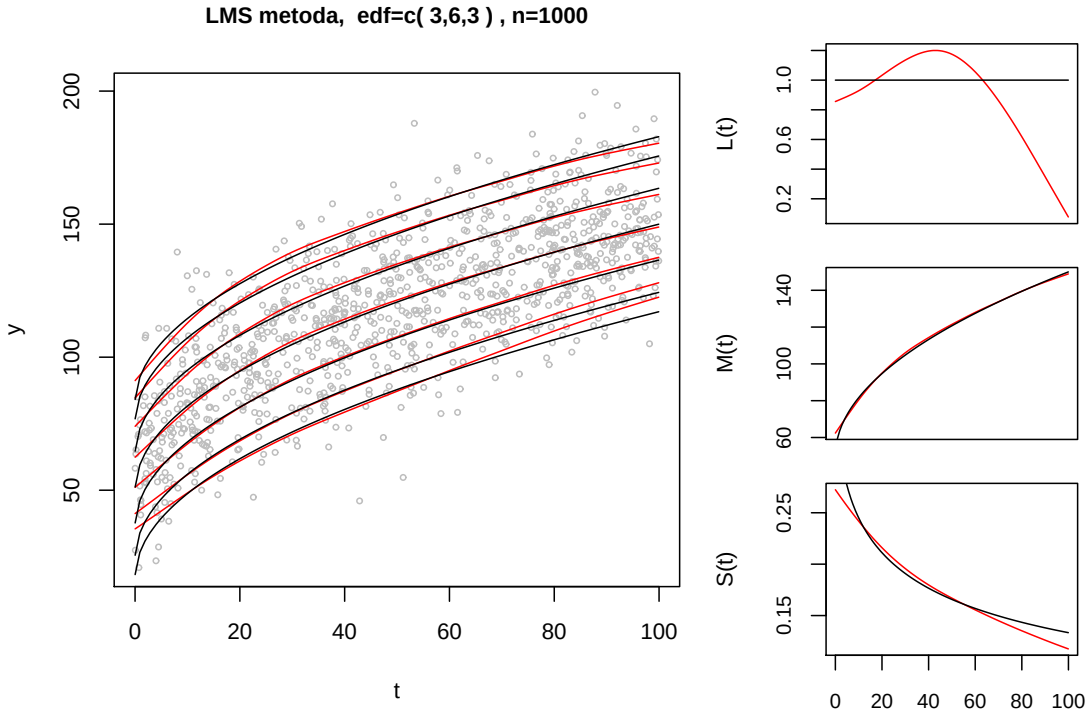
Řešení v podobě tří křivek L, M a S získáme maximalizací penalizované věrohodnosti

$$l(L, M, S) - \frac{1}{2}\alpha_\lambda \int (L''(t))^2 dt - \frac{1}{2}\alpha_\mu \int (M''(t))^2 dt - \frac{1}{2}\alpha_\sigma \int (S''(t))^2 dt, \quad (2.6)$$

kde α_λ, α_μ a α_σ jsou vyhlazovací parametry. Penalizovaná věrohodnost se používá kvůli požadavku na hladkost růstových křivek. Křivka L (resp. M, S), která se moc vlní, je penalizována v závislosti na velikosti α_λ (resp. α_μ, α_σ). Dá se ukázat, že řešením této maximalizace jsou přirozené kubické spliny s uzly v každé z navzájem různých hodnot t_i (možno najít v [5]). K určení L, M a S tedy stačí zvolit hodnoty parametrů α_λ, α_μ a α_σ.

V celé práci se místo s parametry α_λ, α_μ a α_σ pracuje s tzv. odpovídajícími stupni volnosti (*equivalent degrees of freedom - e.d.f.*), což jsou jakési míry složitosti křivek L, M a S. Dá se říci, že je to jakási analogie stupňů volnosti u polynomiálních křivek. Spodní hranice e.d.f. pro každou z křivek je 2, to odpovídá nejhladší možnosti a tou je přímka. S rostoucím e.d.f. se pak jedná o složitější křivky (e.d.f. nabývá reálné hodnoty). Přesnou početní interpretaci najdete v [2]. Je možno také definovat křivku s e.d.f. rovno 1, což je přímka se směrnicí 0, odhaduje se pouze posunutí této přímky. Za křivku s e.d.f. rovno 0 se pak považuje přímka s nulovou směrnicí a pevně daným posunutím. To má smysl definovat především pro křivku $L(t)$, $L(t) = 1$ totiž odpovídá tomu, že jsou data normálně rozdělená i bez mocninné transformace.

V praxi si tedy zvolíme trojici odpovídajících stupňů volnosti (tím řekneme, jak chceme, aby L, M a S byly hladké), k nim pak existuje příslušná trojice (α_λ, α_μ, α_σ), tak aby výsledkem maximalizace (2.6) byly právě takto hladké funkce. Růstové křivky se pak jednoduše sestaví z (2.4).



Obrázek 2.1: LMS metoda: Na levém obrázku jsou data s odhadnutými křivkami (zobrazené červeně) spolu s teoretickými kvantilovými křivkami (zobrazené černě). Vpravo jsou výsledné spliny $L(t)$, $M(t)$ a $S(t)$ (červeně), k nim jsou přikresleny (černě) teoretické křivky $L(t)$, $M(t)$ a $S(t)$, které by vedly k vytvoření teoretických kvantilových křivek.

Tato metoda je implementována v softwaru R v knihovně `gamlss` [6] jako součást funkce *gamlss*, kde za rodinu rozdělení musíme zvolit BCCG, formule pro parametry *nu*, *sigma* a *y* musíme zvolit jako kubické spliny příslušného stupně.

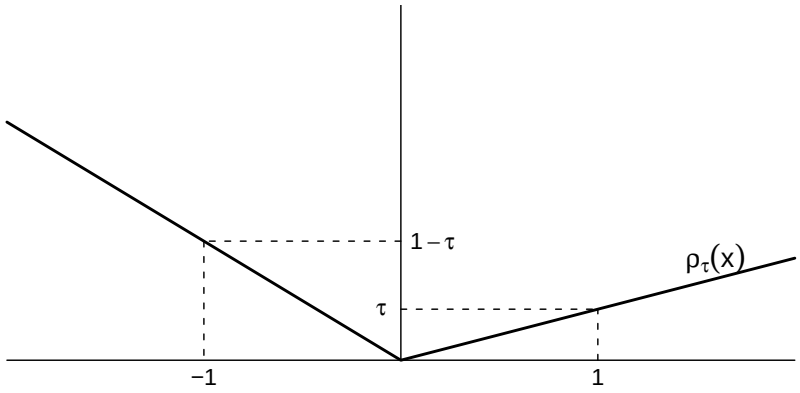
Jak vypadá příklad jedné z realizací této metody je ukázáno na obrázku 2.1. Na obrázku je vidět, jak proložení daty, tak i kubické spliny $L(t)$, $M(t)$ a $S(t)$.

2.4 Metoda založená na kvantilové regresi

Myšlenka kvantilové regrese vychází z následující jednoduché optimalizační úlohy: Mějme náhodnou veličinu Y s distribuční funkcí F. Funkce

$$\rho_\tau(x) = x(\tau - I(x < 0)) \tag{2.7}$$

je ztrátová funkce, kde $\tau \in (0, 1)$ (tato funkce je znázorněna na obrázku 2.2). Úlohou je najít $\hat{\xi}$, které by minimalizovalo očekávanou ztrátu. Chceme tedy najít



Obrázek 2.2: Ztrátová funkce ρ_τ.

řešení minimalizace

$$\min_{\xi \in \mathbb{R}} E\rho_\tau(Y - \xi). \tag{2.8}$$

Věta 1. *τ-kvantil náhodné veličiny Y s absolutně spojitou distribuční funkcí F je řešením minimalizace (2.8).*

Důkaz.

$$\begin{aligned} E\rho_\tau(Y-\xi) &= (\tau-1)\int_{-\infty}^{\xi}(y-\xi)dF(y) + \tau\int_{\xi}^{\infty}(y-\xi)dF(y) \\ &= \tau\int_{-\infty}^{\infty}(y-\xi)dF(y) - \int_{-\infty}^{\xi}(y-\xi)dF(y) \\ &= \tau EY - \tau\xi + \xi F(\xi) - \int_{-\infty}^{\xi} y dF(y). \end{aligned} \tag{2.9}$$

Zderivujeme (2.9) podle ξ a položíme rovno 0

$$\begin{aligned} 0 &= -\tau + F(\xi) + \xi F'(\xi) - \xi F'(\xi) \\ &= -\tau + F(\xi). \end{aligned}$$

Protože F je monotónní, minimalizuje očekávanou ztrátu každý prvek z $\{\xi : F(\xi) = \tau\}$. Pokud je řešení $F(\xi) = \tau$ jediné, pak $\hat{\xi} = F^{-1}(\tau)$, v opačném případě je řešením interval. Nejmenší hodnota tohoto intervalu pak přesně vyhovuje definici τ-kvantilu. τ-kvantil veličiny Y je tedy řešením minimalizace (2.8).

□

Pokud nahradíme F empirickou distribuční funkcí

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y),$$

pak se změní úloha (2.8) na úlohu

$$\min_{\xi\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(Y_i - \xi). \tag{2.10}$$

Řešením této minimalizace je nepodmíněný výběrový τ-kvantil (toto bude dokázáno níže).

Koenker a Bassett [4] rozšířili tuto ideu. Navrhli při hledání odhadu podmíněné kvantilové funkce Y za podmínky x řešit minimalizaci

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i^\top \beta) = \min_{\beta\in\mathbb{R}^p} \left[\tau \sum_{i:e_i>0} e_i + (\tau - 1) \sum_{i:e_i<0} e_i \right], \tag{2.11}$$

kde β je p-dimenzionální parametr, $e_i = Y_i - x_i^\top \beta$ a $Q_Y(\tau|x) = x^\top \beta(\tau)$ je podmíněná kvantilová funkce.

Definice 1 (Regresní kvantily). *Regresním kvantilem nazveme řešení $\hat{\beta}(\tau)$ minimalizace (2.11). Obor zabývající se regresními kvantily se nazývá kvantilová regrese.*

Odhadem podmíněné kvantilové funkce $x^\top \beta(\tau)$ je tedy $x^\top \hat{\beta}(\tau)$.

2.4.1 Metoda pro růstové křivky založená na B-splinové bázi

Definice 2 (Normované B-spline funkce). *Nechť $a_1 \equiv u_{-k} \leq \cdots \leq u_0 \equiv a \leq u_1 \leq \cdots \leq u_N \equiv b \leq \cdots \leq u_{N+k} \equiv b_1$ je množina uzlů, $N \geq 1$. Definujme rekurentně i-tou normovanou B-spline funkci stupně k předpisem:*

$$\phi_k^i(t) = \frac{t - u_i}{u_{i+k} - u_i} \phi_{k-1}^i(t) + \frac{u_{i+k+1} - t}{u_{i+k+1} - u_{i+1}} \phi_{k-1}^{i+1}(t)$$

pro $i = -k, \ldots, N-1$ a $t \in \langle a_1, b_1 \rangle$, přičemž výraz v rekurenci obsahující 0 ve jmenovateli se vynechává.

$$\phi_0^i(t) = \begin{cases} 1 & \text{pro } u_i \leq t < u_{i+1} \\ 0 & \text{jinak.} \end{cases}$$

Některé důležité vlastnosti normovaných B-spline funkcí:

1. Nosič $\phi_k^i(t)$ je $\langle u_i, u_{i+k+1} \rangle$.
2. $$\sum_{i=-k}^{N-1} \phi_k^i(t) = 1 \quad \text{pro } t \in \langle a, b \rangle.$$

3. $\phi_k^i(t) > 0$ pro $x \in (u_i, u_{i+k+1})$.

4. $\{\phi_n^i\}_{i=-k}^{N-1}$ tvoří bázi prostoru $S_n(U) = \mathcal{L}\{\phi_n^i\}_{i=-k}^{N-1}$, $U = \{u_i\}_{i=-k}^{N+k}$ je množina uzlů. Do tohoto prostoru tedy patří všechny funkce S tvaru

$$S(t) = \sum_{i=-k}^{N-1} c_i \phi_n^i(t).$$

V dalším je tedy množina funkcí $\{\phi_n^i\}_{i=-k}^{N-1}$ označována jako B-splinová báze.

Pozn.: Více o této problematice najdete například v knize Karla Najzara [5].

Pro odhadování růstových křivek lze použít metodu kvantilové regrese založenou na těchto B-spline funkcích. Ta spočívá v tom, že ve vztahu (2.11) zvolíme vektor x_i takto: $x_{ij} = \phi_3^j(t_i)$, t_i jsou hodnoty podmiňující veličiny napozorované spolu s Y_i, ϕ_3^j je j-tá normovaná B-spline funkce stupně 3 (vektor x_i má tedy stejnou délku jako je počet prvků B-splinové báze). Uzly pro funkce ϕ_3^j jsou navíc zvoleny tak, že $u_{-3} = u_{-2} = u_{-1} = u_0 = a$ a $u_N = u_{N+1} = u_{N+2} = u_{N+3} = b$. Toto je určeno pevně.

Pro určení konkrétní báze tedy zbývá určit parametr N (počet prvků báze je potom $N+3$) a umístění $N-1$ uzlů (to znamená uzlů mezi a a b – interval $\langle a, b\rangle$ je v praxi roven intervalu, na kterém chceme růstové křivky sestavovat). Těchto $N-1$ uzlů nazýváme vnitřními uzly. Nejmenší možný počet složek báze je 4, což odpovídá tomu, že nemáme žádný vnitřní uzel a $N = 1$.

B-splinové báze, přesně tak, jak zde byly popsány, jsou implementovány v knihovně `splines` softwaru R, přesněji – jedná se o funkci *bs*. Kvantilovou regresi pak najdeme v knihovně `quantreg` Rogera Koenkera [7].

Základem metody kvantilové regrese založené na B-splinové bázi (tuto metodu budu někdy zkráceně označovat jako RQ metodu) je tedy to, že podmíněná kvantilová funkce je ve formě lineární kombinace složek kubické B-splinové báze. Příklad použití kvantilové regrese i s B-splinovou bází, podle které růstové křivky vznikly, je na obrázku 2.3.

2.4.2 Výpočet regresních kvantilů

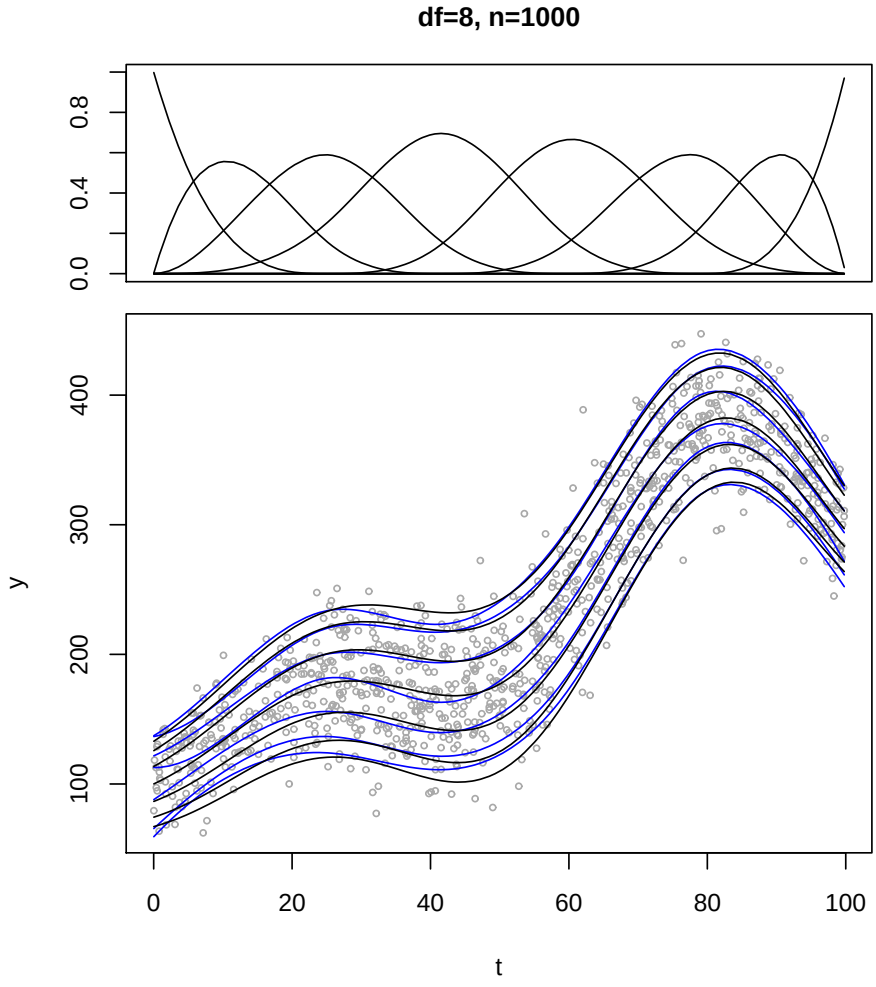
Úloha (2.11) se dá přepsat jako úloha lineárního programování:

$$\begin{aligned} \text{minimalizujeme:} \quad & \tau \sum_{i=1}^{n} e_i^+ + (1-\tau) \sum_{i=1}^{n} e_i^- \\ \text{za podmínky:} \quad & \sum_{j=1}^{p} x_{ij}\beta_j + e_i^+ - e_i^- = Y_i, \;\; i = 1, \ldots, n, \end{aligned} \tag{2.12}$$

$$\beta_j \in \mathbb{R},\ j = 1, \ldots, p,\ e_i^+, e_i^- \geq 0,\ i = 1, \ldots, n,\ 0 < \tau < 1,$$

kde $e_i^+ = \max(0, e_i)$, $e_i^- = \max(0, -e_i)$. $\hat{\beta}(\tau)$ je pak složka β optimálního řešení (β, e^+, e^-) úlohy (2.12) s pevným τ.

Optimální řešení pak umožňuje vypočítat simplexová metoda.



Obrázek 2.3: Metoda kvantilové regrese: V horní části obrázku je kubická B-splinová báze s ekvidistantními uzly a 8 prvky báze, v dolní části data s odhadnutými (modrými) a teoretickými (černými) růstovými křivkami.

2.4.3 Vlastnosti kvantilové regrese

Metoda kvantilové regrese se začala používat pro odhadování růstových křivek především díky následující vlastnosti:

Věta 2. *Nechť N_k, N_z a N_n je počet kladných, záporných a nulových složek vektoru reziduí $Y - X\hat{\beta}(\tau)$, $Y = (Y_1, \ldots, Y_n)^\top$, X je regresní matice. Pokud existuje $\alpha \in \mathbb{R}^p$ takové, že $X\alpha = 1$, pak pro každý regresní kvantil $\hat{\beta}(\tau)$ platí*

$$N_z \leq n\tau \leq N_z + N_n$$

a

$$N_k \leq n(1-\tau) \leq N_k + N_n.$$

Důkaz. Nejdříve odvodíme podmínku, kterou musí splňovat optimální řešení úlohy minimalizace funkce

$$T(\beta) = \sum_{i=1}^{n} \rho_\tau(Y_i - x_i^\top \beta).$$

Funkce $T(\beta)$ je po částech lineární a spojitá. To znamená, že v každém bodě existují derivace T ve všech směrech. Derivace T ve směru v se dá napsat jako

$$\begin{aligned}
\nabla T(\beta, v) &\equiv \frac{d}{dt} T(\beta + tv) \mid_{t=0} \\
&= \frac{d}{dt} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i^\top \beta - x_i^\top tv)[\tau - I(Y_i - x_i^\top \beta - x_i^\top tv < 0)] \mid_{t=0} \\
&= -\sum_{i=1}^{n} \psi_\tau(Y_i - x_i^\top \beta, -x_i^\top v) x_i^\top v,
\end{aligned}$$

kde

$$\psi_\tau(a, b) = \begin{cases} \tau - I(a < 0) & \text{pokud } a \neq 0 \\ \tau - I(b < 0) & \text{pokud } a = 0. \end{cases}$$

Pokud je $\hat{\beta}(\tau)$ regresní kvantil, pak je řešením minimalizace a tedy musí být v $\hat{\beta}(\tau)$ derivace ve všech směrech nezáporné. Platí tedy

$$-\sum_{i=1}^{n} \psi_\tau(Y_i - x_i^\top \hat{\beta}(\tau), -x_i^\top v) x_i^\top v \geq 0$$

pro všechny směry $v \in \mathbb{R}^p$. Speciální volbou $v = \alpha$ $(X\alpha = 1)$ dostáváme

$$\begin{aligned}
-\sum_{i=1}^{n} \psi_\tau(Y_i - x_i^\top \hat{\beta}(\tau), -1) &\geq 0 \\
-\tau N_k - \tau N_z + N_z - \tau N_n + N_n &\geq 0 \\
N_z + N_n &\geq \tau(N_k + N_z + N_n).
\end{aligned}$$

Analogicky pro $v = -\alpha$ dostaneme

$$\begin{aligned} \tau N_k + \tau N_z - N_z + \tau N_n &\geq 0 \\ \tau(N_k + N_z + N_n) &\geq N_z. \end{aligned}$$

Zbytek plyne z těchto nerovností a faktu, že $n = N_k + N_z + N_n$.

□

V případě kvantilové regrese s B-splinovou bází existuje α takové, že $X\alpha = 1$. A to díky 2. z vlastností normovaných B-spline funkcí na straně 13.

Z teorie lineárního programování víme, že množina přípustných řešení úlohy (2.12) je konvexní polyedrická množina. Řešení je pak buď ve vrcholu této množiny (jediné řešení) nebo ve více vrcholech, tehdy je řešením hrana nebo celá stěna množiny přípustných řešení. Tyto vrcholy odpovídají bodům v parametrickém prostoru, pro které je p pozorování interpolováno (tedy případ, kdy je $N_n = p$). Pokud mají $Y_i, i = 1, \ldots, n$ spojité rozdělení, pak to, že by bylo pro nějaký vrchol interpolováno více než p pozorování, nastane s pravděpodobností 0.

Důsledek 1. *Přímým důsledkem věty 2 je, že pokud $N_n = p$, pak poměr záporných složek vektoru reziduí je přibližně τ:*

$$\frac{N_z}{n} \leq \tau \leq \frac{N_z + p}{n}$$

a poměr kladných přibližně $1 - \tau$:

$$\frac{N_k}{n} \leq 1 - \tau \leq \frac{N_k + p}{n}.$$

Důsledek 2. *Ve speciálním případě, kdy $X = 1_n$, pak řešením minimalizace (2.11) je výběrový τ-kvantil. Pokud $n\tau$ není přirozené číslo, pak je výběrový τ-kvantil dán jednoznačně. Pokud $n\tau$ je přirozené číslo, pak řešením je interval mezi dvěmi po sobě jdoucími pořádkovými statistikami.*

Kapitola 3

Simulační studie

Tato kapitola se věnuje vlastním generováním dat a popisem postupu analýzy využívajícím znalosti přesného rozdělení těchto dat.

Nejprve je zde popsána hlavní myšlenka, jak využít znalosti teoretického rozdělení dat k určení, pro který model jsou odhadnuté křivky nejlepší (pro daná data). Dále je zde plán celé studie – tedy postup, který byl zvolen pro základní analýzu. V závěru kapitoly je popsáno, jak byla data konstruována a pro jaké typy dat byly obě metody vyzkoušeny.

3.1 Míra kvality modelu

V praxi neznáme přesné rozdělení dat a tedy ani umístění jejich kvantilů, tudíž nemůžeme přesně vědět, který model je nejlepší pro daná data. V rámci simulačních studií si ale data sami generujeme, což znamená, že známe přesně teoretické rozdělení dat. To nám dává možnost porovnat odhadnuté křivky se „skutečnými“ křivkami. Nutno poznamenat, že tento přístup je naprosto odlišný od toho, jak se kvalita křivek určuje v praxi, kde neznáme teoretické rozdělení dat (pro srovnání například článek [9]).

Snahou bylo určit kvalitu odhadnutých křivek jedním číslem a to s využitím znalosti teoretického (skutečného) rozdělení. Jako kritérium bylo zvoleno

$$\begin{aligned} V &= \frac{1}{k}\sum_{j=1}^{k} w_j V_j, \\ V_j &= \int_{\chi} \left(Q_{\tau_j}(t) - \hat{Q}_{\tau_j}(t)\right)^2 dt, \end{aligned} \tag{3.1}$$

kde:

χ je interval, na kterém vytváříme růstové křivky.

$\tau = (\tau_1, \tau_2, \ldots, \tau_k)$ je vektor délky k určující, o které kvantily se zajímáme.

Q_{τ_j} je teoretická (známá) růstová křivka pro kvantil τ_j.

$\hat{Q}_{\tau_j}$ je odhadnutá křivka pro kvantil τ_j.
w_j je váha pro τ_j-tou kvantilovou křivku, $\sum_{j=1}^{k} w_j = k$.

Jako lepší by potom byl shledán model s menšími hodnotami V. V je tedy spíše jakási ztrátová funkce, která odhadnutým křivkám přiřadí číslo podle toho, jak moc se odlišují od teoretických křivek.

Důvod pro takovouto volbu ztrátové funkce: Potřebujeme, aby odhadnuté křivky co nejlépe kopírovaly teoretické křivky a to po celé její délce. Přičemž je důležité, aby se nevyskytovaly nějaké lokální odskoky (proto upřednostňuji tuto volbu oproti například integrálu s absolutní hodnotou). Volbou vah w_j pak můžeme přidat důležitost některému kvantilu.

To, že je tato volba dobrá, se ukázalo na četných simulacích, které byly provedeny. Nejlépe je to vidět, pokud si vykreslíme teoretické a odhadnuté křivky několika modelů (obrázek 3.1 nebo obrázky v příloze). Modely s menšími hodnotami V bychom opticky opravdu vyhodnotili jako lépe kopírující teoretické křivky (ve smyslu předešlého odstavce).

3.2 Plán simulační studie

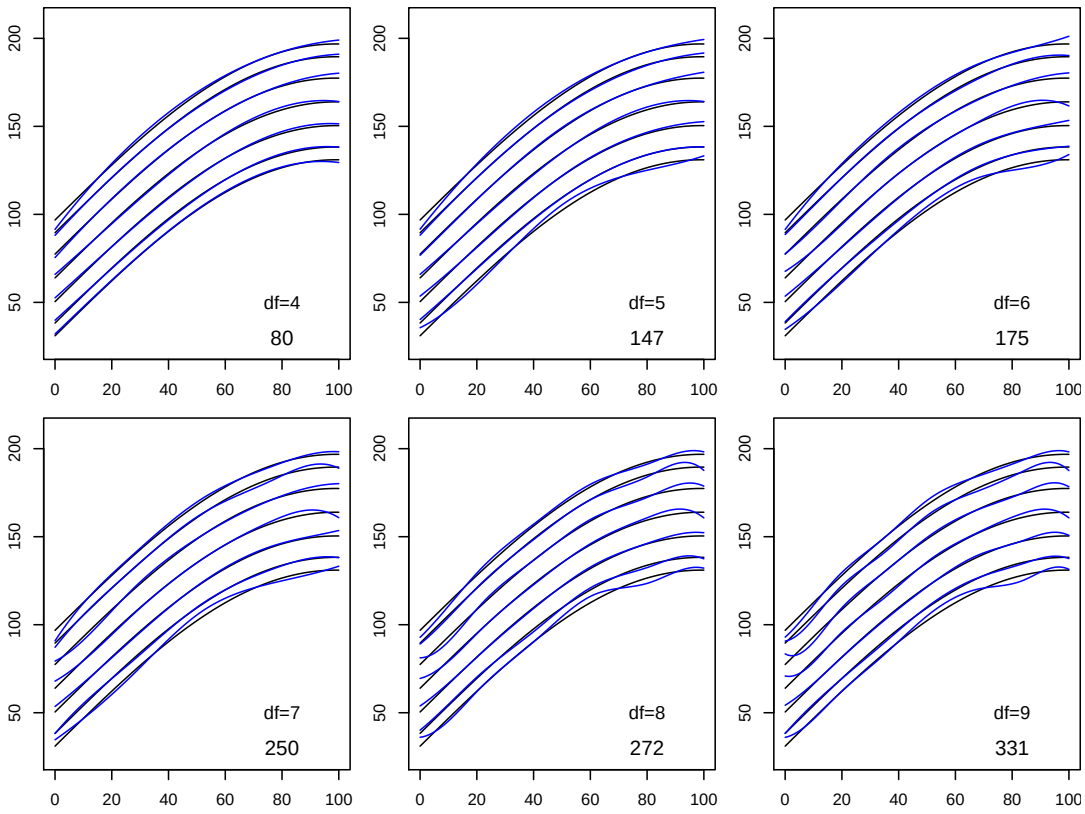
V této části je popsána strategie postupu základní analýzy.

Generuji jeden soubor dat $\{[t_i, Y_i] : i = 1, \ldots, n\}$ pro zvolené schéma generování používající známé rozdělení a pro danou velikost výběru n. Tím se zaručí, že přesné teoretické rozdělení dat je známé a tedy jsou známy i jeho teoretické kvantilové křivky (jako funkce t).

Na vygenerovaný soubor dat zkouším postupně modely LMS metody s různými hodnotami trojice stupňů volnosti a modely kvantilové regrese s různou velikostí dimenze B-splinové báze. Ze zkoušených modelů potom vyberu podle kritéria kvality modelu (vzorec (3.1)) nejlepší model pro LMS metodu (model s nejmenší hodnotou V) a nejlepší model pro metodu kvantilové regrese. Porovnáním dosažených hodnot V na těchto nejlepších modelech zjistím, která metoda si pro tato konkrétní data vedla lépe.

Jde tedy o myšlenku zjistit, jak blízko se mohou k teoretickým kvantilům dostat jednotlivé metody.

Celý tento postup můžeme opakovat mnohokrát pro mnoho různých datových souborů – díky tomu, že si data sami generujeme. To je jistá výhoda a odlišnost tohoto přístupu oproti porovnávání metod na reálných datech. Při zkoumání reálných dat je totiž většinou k dispozici jen jeden datový soubor.



Obrázek 3.1: Porovnání šesti nejlepších modelů metody kvantilové regrese na základě dosažených hodnot V na jednom souboru dat rozsahu $n = 4000$. Tyto hodnoty V ($w_i = 1, i = 1, \ldots, 7$) jsou v pravých dolních rozích spolu s uvedením počtu prvků B-splinové báze použité pro sestavení křivek. Černě jsou přikresleny teoretické kvantilové křivky. S rostoucím V je vidět zhoršování kvality odhadnutých křivek.

3.3 Parametrizace problému

Před jakýmkoli dalším postupem je potřeba udělat si představu o tom, co všechno si musíme zvolit, abychom získali v rámci simulačních studií datový soubor a také abychom posléze na tomto souboru dokázali porovnat jednotlivé metody postupem popsaným v předešlé části kapitoly.

Je třeba poznamenat, že celá tato část textu představuje jednu z možných parametrizací celého problému.

Nejprve popíšu, co je potřeba volit, aby bylo možno nagenerovat jeden konkrétní datový soubor $\{[t_i, Y_i] : i = 1, \ldots, n\}$. Pro získání složky dat t_i si musíme určit interval $\chi = [a, b]$, na kterém budeme tuto složku dat generovat, velikost datového souboru n a rozdělení t_i na χ. Pro dogenerování složky dat Y_i potřebujeme

nějaké schéma, podle kterého budeme generovat Y_i v závislosti na již vygenerovaném t_i (příklady takových schémat jsou popsány níže).

Pro sestavení modelů pro nasimulovaná data je nutné zvolit skupinu trojic odpovídajících stupňů volnosti v případě LMS metody a skupinu B-splinových bází pro metodu kvantilové regrese (pro určení báze je potřeba znát počet prvků báze a rozmístění vnitřních uzlů).

Abychom mohli ze vzniklých modelů vybrat nejlepší modely metodou popsanou v předchozí části kapitoly, zbývá pro porovnání výsledných modelů určit kvantily $\tau_1, \tau_2, \ldots, \tau_k$, které nás zajímají, a váhy $w_j, j = 1, \ldots, k$.

3.3.1 Příklady schémat pro generování Y_i

Tato kapitolka ukazuje některé možnosti, jak lze generovat Y_i. Protože Y_i závisí na t_i, předpokládáme, že již složku dat $\{t_i : i = 1, \ldots, n\}$ známe. Jako příklady schémat pro generování Y_i uvedu 2 možnosti:

Schéma 1

Podle vzoru LMS metody:

$$Y_i = M(t_i)[1 + L(t_i)S(t_i)R_i]^{\frac{1}{L(t_i)}},$$

kde L, M, S jsou funkce na χ (musí být zvoleny tak, aby Y_i bylo kladné pro všechna i) a R_i jsou náhodné veličiny s rozdělením $\mathcal{R}$, které je pevně dáno.

Vzorec teoretických křivek je pak

$$Q_\tau(t) = M(t)[1 + L(t)S(t)q_\tau]^{\frac{1}{L(t)}},$$

kde q_τ je kvantilová funkce příslušná $\mathcal{R}$.

Pokud za $\mathcal{R}$ zvolíme normované normální rozdělení, pak data vygenerovaná podle tohoto schématu splňují předpoklady LMS metody.

Schéma 2

$$\begin{aligned} g^{-1}(Y_i - c) &= f(t_i) + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{R}_{t_i}, \end{aligned} \tag{3.2}$$

kde f a g jsou funkce, ε_i jsou nezávislé náhodné veličiny s rozdělením $\mathcal{R}_{t_i}$, toto rozdělení můžeme volit libovolně, například to může být rozdělení nějaké transformované veličiny, kterou umíme v rámci simulačních studií generovat. Navíc parametry tohoto rozdělení mohou záviset na t_i (například $\varepsilon_i \sim N(0, t_i + 5)$), c je konstanta – ta je zde kvůli předpokladu LMS metody na kladné hodnoty veličiny,

která nás zajímá (v praxi např.: váha, věk), a také proto, aby se nestalo, že odhady křivek budou v nějakém úseku záporné – z důvodů softwarové implementace LMS metody by se totiž takovéto odhadnuté křivky nedaly porovnávat.

Vzorec teoretických křivek je pak vzhledem k volbě Y_i jako (3.2) takový:

$$Q_{\tau_j}(t) \;=\; g\left(f(t) + q_t(\tau_j)\right) + c, \quad j = 1, \ldots, k,$$

kde q_t je kvantilová funkce příslušná volbě rozdělení $\mathcal{R}_t$.

Pro generování konkrétních dat je tedy potřeba všechny tyto funkce a parametry zvolit.

3.4 Konkrétní volby pro základní analýzu

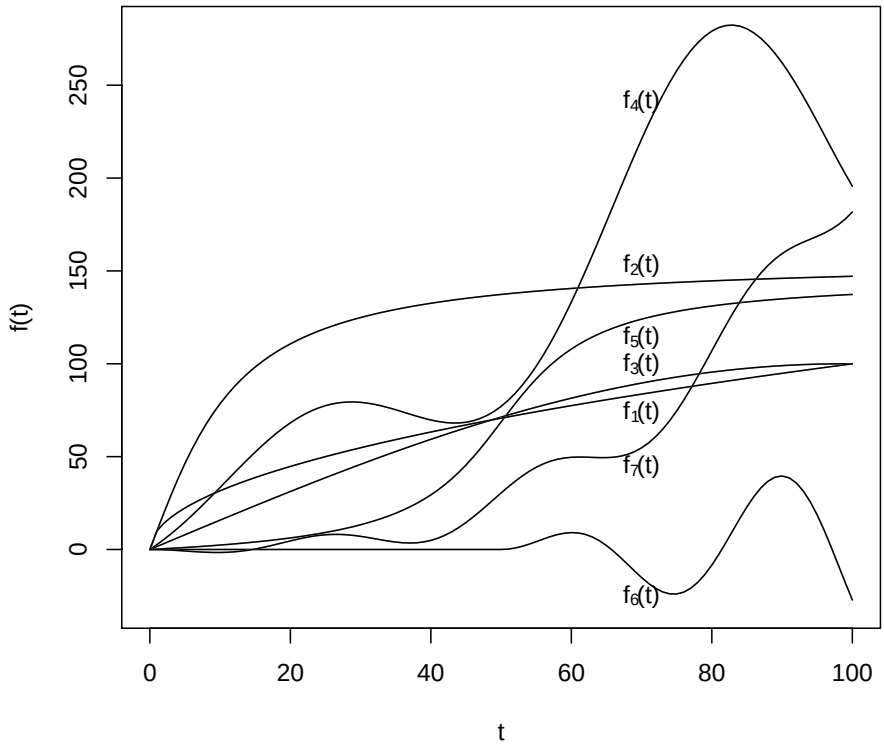
Následuje přehled toho, jak byly konkrétně voleny všechny parametry a funkce popsané v předchozí části kapitoly.

Je zřejmé, že už jen prostor možností, jak generovat data je obrovský. Proto je potřeba sepsat, jak bylo vše voleno a jaké omezení z toho plynou.

Data: Pro všechny analýzy je interval χ (bez újmy na obecnosti) roven $[0, 100]$. Pro určení hodnot t_i se v celé práci omezíme na případ, kdy t_i je vždy generováno z rovnoměrného rozdělení na intervalu $[0, 100]$. Počet dat n pro jedno generování je pak roven jedné z hodnot v množině $\{500, 1000, 2000, 4000, 8000\}$ (vyšší hodnoty nebylo možno použít kvůli paměťové náročnosti při výpočtu LMS metody pomocí [6]).

Za funkce (použité v simulacích) modelující podmíněný medián Y v závislosti na t bylo zvoleno 7 funkcí:

$$\begin{aligned}
f_1(t) &= 100\sqrt{\frac{t}{100}},\\
f_2(t) &= 100 \arctan \frac{t}{100},\\
f_3(t) &= 100 \sin \frac{t}{63},\\
f_4(t) &= 2.5t + t \sin \frac{t}{10},\\
f_5(t) &= 50 \arctan \frac{t-50}{10} + 50 \arctan 5,\\
f_6(t) &= -\max(0, t-50) \sin\left(\frac{\max(0, t-50)}{5} + \pi\right),\\
f_7(t) &= \frac{1}{5}\left(\left(\frac{t}{10}\right)^3 - t \sin \frac{t}{5}\right).
\end{aligned}$$



Obrázek 3.2: Přehled funkcí použitých pro modelování podmíněného mediánu Y v závislosti na t.

Graficky jsou tyto funkce znázorněny na obrázku 3.2.

Pro základní analýzu bylo nagenerováno celkem 118 datových souborů. Snahou bylo vytvořit data s různými tvary M křivky a různého rozsahu dat, které by nebyly příliš složité. Bylo vytvořeno 58 souborů splňujících předpoklady LMS metody a 60 souborů porušujících tyto předpoklady, ty byly generovány za pomocí jiného než normálního rozdělení.

Konkrétní volby, jak byla v jednotlivých případech data generována, jsou vidět v tabulkách (4.1 až 4.4) v následující kapitole. Všechna data jsou generována podle schémat popsaných výše a jsou uložena na přiloženém médiu.

Volba skupin trojic stupňů volnosti, z kterých vybíráme nejlepší model, byla určena v závislosti na teoretickém tvaru křivek L, M a S (při existující normální transformaci). Při neznámé transformaci nebo při jiném rozdělení dat, pak bylo použito více modelů z větším rozsahem parametrů. V průměru bylo pro jedna data vyzkoušeno více než 10 modelů.

Volba jednotlivých B-splinových bází, pomocí kterých budeme modelovat,

je takováto: vnitřní uzly jsou zvoleny pro jednoduchost vždy ekvidistantně a dimenze bází je postupně $4, 5, \ldots, 25$, což je plně dostačující rozsah pro všechny naše nagenerovaná data. Přesněji: Nebyly zkoušeny vždy všechny možnosti dimenze 4, 5, až 25, ale vždy vhodná podmnožina $\{4, 5, \ldots, 25\}$ (s po sobě jdoucími prvky). *Pozn.*: Volba B-splinových bází s ekvidistantními uzly je omezující. V praxi by se volil větší počet uzlů v oblastech, kde se mají data v závislosti na t tendenci vlnit nebo prudce měnit, a naopak menší počet uzlů tam, kde se data se změnou t příliš nemění. Uzly byly nakonec zvoleny ekvidistantně z toho důvodu, že v opačném případě by se muselo určovat rozmístění uzlů ke každému souboru dat individuálně a to by bylo vzhledem k počtu provedených simulací velmi pracné. Nehledě na další komplikace, které s tím souvisí (například obtížné porovnávání modelů).

Vektor τ zvolíme pevně jako $(0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95)^\top$, což je většinou rozmezí kvantilů, které nás zajímá.

Jako ztrátové funkce modelu byly zvoleny dvě možnosti výpočtu V. První s vahami $w_j = 1,\ j = 1, \ldots, k$ (tuto míru kvality označíme jako V^u) a druhá s vahami $w = \frac{1}{1+2(2+3+4)}(4, 3, 2, 1, 2, 3, 4)$ pro zvýšení důležitosti krajních kvantilů (označení V^w).

Kapitola 4

Výsledky

V této kapitole jsou sepsány výsledky na základě porovnání metod na 118 datových souborech vytvořených v rámci simulačních studií pomocí postupů popsaných výše.

4.1 Porovnání metod na základě dosažených hodnot V

Nejprve se zaměříme na porovnání metod s využitím znalosti teoretického rozdělení dat a to na základě dosažených hodnot V. Tyto výsledky jsou rozděleny na dvě části – v první jsou případy, kdy jsou pro data splněny předpoklady LMS metody, v druhé jsou pak případy, kdy tyto předpoklady splněny nejsou.

4.1.1 Podmínky LMS metody splněny

Nejdříve se tedy podíváme na situaci, kdy máme zaručeno splnění požadavků LMS metody. Tím se myslí situace, kdy jsou data $[t_i, Y_i]$ vytvořena tak, že je generování Y_i ekvivalentní generování podle následujícího vztahu

$$\begin{aligned} Y_i &= M(t_i)[1 + L(t_i)S(t_i)Z_i]^{\frac{1}{L(t_i)}}, \qquad L(t_i) \neq 0; \\ Y_i &= M(t_i) \exp S(t_i)Z_i, \qquad\qquad\quad L(t_i) = 0, \end{aligned} \tag{4.1}$$

kde Z_i je generováno z normovaného normálního rozdělení a funkce L, M a S přesně známe.

Výsledky simulací pro tento případ, jsou vidět v tabulkách 4.1 a 4.2.

Přesná interpretace konkrétního výsledku (to znamená řádku) z tabulky 4.1 (podobně i tabulek 4.2, 4.3 a 4.4) je takováto: pokud je V u LMS menší než V u kvantilové regrese, pak to znamená, že pro daná data existuje LMS model lepší (ve smyslu V) než modely kvantilové regrese s B-splinovou bází s ekvidistantními uzly. Je potřeba upozornit na to, že výsledek V v konkrétním řádku se váže pouze

k jednomu datovému souboru. Při generování jiných dat stejným způsobem mohou být hodnoty V odlišné, stejně jako může být jiný i nejlepší model pro tu kterou metodu. To je potřeba brát v úvahu.

Z jednoduchých, ale i složitějších případů dat, kdy jsou splněny předpoklady LMS metody, se dá z výsledků simulací shrnutých v tabulce 4.1 a 4.2 jednoznačně říci, že LMS metoda má lepší výsledky než kvantilová regrese. Nejsme si sice jisti u konkrétního výsledku, že na jiných datech generovaných stejným způsobem se nestane naopak lepší metoda kvantilové regrese, ale pokud se na výsledky díváme jako na celek: tedy uvažujeme každý řádek (jednotlivý výsledek) jako jednu realizaci simulace se splněnými předpoklady LMS metody, pak můžeme opravdu říci, že LMS metoda je v těchto případech lepší.

Je také vidět, že V^u a V^w se chovají velmi podobně, dokonce ve většině případů je vybrán i stejný nejlepší model. Stejné je to u všech zkoumaných případů, proto jsou hodnoty V^w uvedeny pouze v tabulce 4.1, v ostatních tabulkách jsou již hodnoty V^w vynechány.

4.1.2 Data založená na jiných rozděleních

Pokud teď budeme uvažovat případ, kdy nejsou splněny předpoklady LMS metody (pro data neplatí vztah (4.1)). Uvažuje se zde případ, kdy nejsou splněny tím, že Z_i v (4.1) generujeme z jiného rozdělení než z normálního (přesněji spojení „nejsou splněny“ platí, neexistuje-li žádná transformace, která by takto generovaná data na normální převedla. Nebo tato transformace existuje, ale data se po jejím použití nedají přepsat na tvar v (4.1)).

Výsledky takovéhoto simulování jsou v tabulce 4.3 a 4.4. Zde je na zkoušených případech vidět, že u některých rozdělení už si lépe vede metoda kvantilové regrese. V tabulkách jsou většinou jednoduché případy, kdy jsou data tvaru $Y_i = f_k(t_i) + W_i$, $i = 1, \ldots, n$, kde W_i je generováno z jiného rozdělení než normálního, všechny W_i a tedy i Y_i mají stejný konstantní rozptyl.

Pokud se podíváme na příklady, kde kvantilová regrese vítězí nad LMS metodou, zjistíme, že mají jeden společný znak. Především při větších rozsazích výběru je to vidět při vykreslení teoretických a odhadnutých křivek (viz obrázek 4.1). LMS metoda dobře vystihne tvar křivek, ale odhadnuté křivky některých kvantilů jsou trochu posunuty vzhledem k těm teoretickým ve směru svislé osy. V některých případech to vede až k tomu, že pod některými odhadnutými kvantilovými křivkami je rozdíl mezi očekávaným a skutečným procentem dat pod křivkou až 8%.

Lepší představu o této situaci si uděláme, pokud se zaměříme na jednotlivé V_i pro všechny zkoušené modely a to jak u LMS metody, tak i u metody kvantilové regrese. Konkrétní případ na datech generovaných za pomocí dvojitě exponenciálního rozdělení ukazuje obrázek 4.2. Na obrázku je vidět, že chování obou metod je odlišné. LMS metoda má v tomto případě největší problémy s odhadnutím křivek pro 1. a 3. kvartil, zatímco kvantilová regrese se chová „přirozeně“

Tabulka 4.1: Výsledky simulování (předpoklad LMS splněn – jednodušší případy): $V^u_{(\text{model})}$ znamená, že nejlepšího výsledku V^u bylo dosaženo pro model v závorce. $Z_i \sim N(0,1), i = 1,\ldots,n$, f_j jsou funkce zmíněné v závěru kapitoly 3, c je konstanta pro zaručení kladnosti všech Y_i i odhadnutých křivek.

Schéma generování dat		LMS	RQ	LMS	RQ
předpis	n	$V^u_{(\text{model})}$	$V^u_{(\text{model})}$	$V^w_{(\text{model})}$	$V^w_{(\text{model})}$
$Y_i = f_1(t_i) + 20Z_i + c$	500	$805_{(0,6,3)}$	$1540_{(5)}$	$953_{(0,6,3)}$	$1921_{(5)}$
	1000	$449_{(0,6,3)}$	$479_{(4)}$	$467_{(0,6,3)}$	$520_{(4)}$
	2000	$138_{(3,9,3)}$	$283_{(5)}$	$153_{(3,9,3)}$	$288_{(5)}$
	4000	$110_{(0,9,6)^\star}$	$129_{(6)}$	$119_{(1,9,6)^\star}$	$152_{(6)}$
$Y_i = f_3(t_i) + 20Z_i + c$	500	$783_{(0,6,3)}$	$1524_{(4)}$	$894_{(0,6,3)}$	$1901_{(4)}$
	1000	$406_{(0,6,3)}$	$660_{(5)}$	$454_{(0,6,3)}$	$800_{(6)}$
	2000	$94_{(3,6,6)^\star}$	$210_{(4)}$	$103_{(3,6,6)^\star}$	$244_{(4)}$
	4000	$70_{(1,6,6)^\star}$	$80_{(4)}$	$76_{(1,6,6)^\star}$	$85_{(4)}$
$Y_i = f_3(t_i) + 10Z_i + c$	500	$165_{(4,4,4)}$	$269_{(4)}$	$182_{(4,4,4)}$	$311_{(4)}$
	1000	$85_{(0,6,3)}$	$141_{(4)}$	$98_{(0,6,3)}$	$175_{(4)}$
	2000	$28_{(0,6,3)}$	$30_{(4)}$	$29_{(0,6,3)}$	$34_{(4)}$
	4000	$32_{(6,7,6)^\star}$	$22_{(4)}$	$33_{(6,7,6)^\star}$	$22_{(4)}$
	8000	$10_{(0,12,12)^\star}$	$15_{(4)}$	$11_{(0,12,12)^\star}$	$17_{(4)}$
$Y_i = f_4(t_i) + 20Z_i + c$	500	$1909_{(0,12,9)}$	$2441_{(10)}$	$2169_{(0,12,9)}$	$2714_{(8)}$
	1000	$974_{(0,12,9)}$	$1844_{(9)}$	$1038_{(0,12,6)}$	$2006_{(9)}$
	2000	$165_{(1,12,9)}$	$893_{(8)}$	$190_{(1,12,9)}$	$963_{(8)}$
	4000	$181_{(0,15,9)^\star}$	$440_{(12)}$	$202_{(0,15,9)^\star}$	$521_{(12)}$
	8000	$86_{(1,15,11)^\star}$	$195_{(10)}$	$92_{(1,15,11)^\star}$	$201_{(10)}$
$Y_i = f_5(t_i) + 20Z_i + c$	500	$812_{(6,9,9)}$	$1574_{(8)}$	$800_{(6,9,9)}$	$1586_{(8)}$
	1000	$428_{(3,9,6)}$	$1021_{(8)}$	$457_{(3,9,6)}$	$1198_{(8)}$
	2000	$227_{(1,12,9)}$	$922_{(8)}$	$241_{(1,12,9)}$	$975_{(8)}$
	4000	$126_{(1,12,9)^\star}$	$252_{(10)}$	$129_{(1,12,9)^\star}$	$258_{(10)}$
	8000	$84_{(0,15,11)^\star}$	$142_{(10)}$	$91_{(0,15,11)^\star}$	$168_{(10)}$
$Y_i = f_6(t_i) + 20Z_i + c$	500	$2387_{(0,15,9)}$	$3257_{(11)}$	$2544_{(0,15,9)}$	$3605_{(11)}$
	1000	$1035_{(0,18,12)}$	$3596_{(8)}$	$1131_{(0,18,12)}$	$3605_{(8)}$
	2000	$368_{(0,18,12)^\star}$	$887_{(12)}$	$397_{(0,18,12)^\star}$	$922_{(12)}$
	4000	$314_{(0,18,15)}$	$531_{(14)}$	$337_{(0,18,15)}$	$574_{(14)}$
	8000	$199_{(0,18,15)^\star}$	$294_{(14)}$	$208_{(0,18,15)^\star}$	$332_{(14)}$
$Y_i = f_7(t_i) + 20Z_i + c$	500	$1183_{(3,12,6)}$	$3449_{(10)}$	$1260_{(3,12,6)}$	$3899_{(10)}$
	1000	$629_{(1,15,9)}$	$1772_{(12)}$	$676_{(1,15,9)}$	$1914_{(12)}$
	2000	$358_{(3,15,6)}$	$545_{(10)}$	$390_{(3,15,6)}$	$608_{(10)}$
	4000	$129_{(1,15,9)^\star}$	$290_{(10)}$	$143_{(1,15,9)^\star}$	$314_{(10)}$
	8000	$70_{(0,18,12)^\star}$	$247_{(10)}$	$80_{(0,18,12)^\star}$	$278_{(10)}$

* Při větších hodnotách n se může stát, že si software pro výpočet LMS křivek sám zvýší počet stupňů volnosti oproti těm, které zadáme. Touto značkou jsou označeny případy, kdy si funkce v nějakém ze zkoušených modelů sama upravila počet e.d.f. V závorkách u V je pak upravený počet e.d.f.

Tabulka 4.2: Výsledky simulování (předpoklad LMS splněn – složitější případy): $V^u_{(\text{model})}$ znamená, že nejlepšího výsledku V^u bylo dosaženo pro model v závorce. $Z_i \sim N(0,1), i = 1,\ldots,n$, f_j jsou funkce zmíněné v závěru kapitoly 3, c je konstanta pro zaručení kladnosti všech Y_i i odhadnutých křivek.

Schéma generování dat		LMS	RQ
předpis	n	$V^u_{(\text{model})}$	$V^u_{(\text{model})}$
$Y_i = f_1(t_i) + f_{sd}(t_i)Z_i + c$	500	$596_{(1,6,3)}$	$899_{(4)}$
$f_{sd}(t) = \frac{1}{5}t + 10$	1000	$256_{(0,6,3)}$	$633_{(4)}$
	2000	$134_{(0,6,3)}$	$194_{(5)}$
	4000	$194_{(0,9,6)^\star}$	$219_{(5)}$
	8000	$68_{(1,10,11)^\star}$	$95_{(6)}$
$Y_i = f_2(t_i) + f_{sd}(t_i)Z_i + c$	500	$2793_{(6,12,6)}$	$1904_{(6)}$
$f_{sd}(t) = \frac{1}{100}(t-80)^2 + 10$	1000	$903_{(3,12,6)}$	$2271_{(5)}$
	2000	$706_{(1,20,9)}$	$953_{(6)}$
	4000	$511_{(6,16,9)^\star}$	$562_{(8)}$
	8000	$303_{(0,20,12)^\star}$	$257_{(7)}$
$Y_i = f_3(t_i) + (\frac{1}{5}t_i + 10)Z_i + c$	500	$429_{(3,4,3)}$	$1374_{(4)}$
$f_{sd}(t) = \frac{1}{5}t + 10$	1000	$451_{(3,4,3)}$	$555_{(4)}$
	2000	$331_{(0,9,6)}$	$594_{(4)}$
	4000	$77_{(0,6,6)^\star}$	$113_{(4)}$
	8000	$83_{(0,11,11)^\star}$	$113_{(5)}$
$Y_i = M(t_i)[1 + L(t_i)S(t_i)Z_i]^{\frac{1}{L(t_i)}}$	500	$3231_{(3,9,6)}$	$8331_{(8)}$
$L(t) = 1 + 0.1\sin(\frac{1}{100}t\pi)$	1000	$1696_{(6,12,9)}$	$4453_{(8)}$
$M(t) = f_4(t) + c$	2000	$1198_{(4,9,6)^\star}$	$2000_{(9)}$
$S(t) = \frac{1}{M(t)}(20 + 20\sin(\frac{1}{100}t\pi))$	4000	$921_{(6,15,12)^\star}$	$1497_{(10)}$
	8000	$209_{(10,11,11)^\star}$	$376_{(10)}$
$Y_i = M(t_i)[1 + L(t_i)S(t_i)Z_i]^{\frac{1}{L(t_i)}}$	500	$1567_{(3,9,6)}$	$3126_{(6)}$
$L(t) = 1 + 0.1\sin(\frac{1}{50}t\pi + \pi)$	1000	$1664_{(6,9,6)}$	$3300_{(6)}$
$M(t) = f_5(t) + c$	2000	$532_{(4,9,6)^\star}$	$1960_{(8)}$
$S(t) = \frac{1}{M(t)}(20 + 20\sin(\frac{1}{100}t\pi))$	4000	$358_{(6,12,9)^\star}$	$860_{(10)}$
	8000	$176_{(11,12,12)^\star}$	$426_{(10)}$

* Při velkých hodnotách n se může stát, že si software pro výpočet LMS křivek sám zvýší počet stupňů volnosti oproti těm, které zadáme. Touto značkou jsou označeny případy, kdy si funkce v nějakém ze zkoušených modelů sama upravila počet e.d.f. V závorkách u V^u je pak upravený počet e.d.f.

Tabulka 4.3: Výsledky simulování (předpoklad LMS nesplněn)(1.část): $V^u_{(\text{model})}$ znamená, že nejlepšího výsledku V^u bylo dosaženo pro model v závorce, f_j jsou funkce zmíněné v závěru kapitoly 3, c je konstanta pro zaručení kladnosti všech Y_i i odhadnutých křivek.

Schéma generování dat předpis	n	LMS $V^u_{(\text{model})}$	RQ $V^u_{(\text{model})}$	
$Y_i = f_5(t_i) + W_i + c$	500	$1957_{(0,9,6)}$	$2690_{(6)}$	
W_i má rovnoměrné rozdělení	1000	$1859_{(3,9,6)}$	$2165_{(6)}$	
na intervalu [-50,50]	2000	$1718_{(0,9,6)}$	$420_{(8)}$	
	4000	$1354_{(0,9,7)\star}$	$370_{(8)}$	◦
	8000	$1420_{(0,11,11)\star}$	$198_{(10)}$	
$Y_i = f_1(t_i) + W_i + c$	500	$1811_{(4,4,4)}$	$2693_{(5)}$	◦
W_i – dvojitě exponenciální	1000	$873_{(0,6,3)}$	$927_{(7)}$	◦
s hustotou	2000	$790_{(6,12,6)}$	$239_{(5)}$	◦
$p(x) = \frac{1}{20\sqrt{2}} \exp\{-\frac{\lvert x\rvert}{10\sqrt{2}}\}$	4000	$881_{(6,9,6)\star}$	$249_{(6)}$	◦
	8000	$900_{(11,15,11)\star}$	$120_{(6)}$	◦
$Y_i = f_5(t_i) + W_i + c$	500	$3665_{(0,12,6)}$	$5710_{(4)}$	
W_i – dvojitě exponenciální	1000	$2415_{(0,12,6)}$	$2705_{(6)}$	
s hustotou	2000	$1711_{(1,12,6)}$	$1295_{(10)}$	
$p(x) = \frac{1}{30\sqrt{2}} \exp\{-\frac{\lvert x\rvert}{15\sqrt{2}}\}$	4000	$1807_{(9,12,9)}$	$761_{(10)}$	◦
	8000	$1865_{(11,12,11)\star}$	$571_{(8)}$	
$Y_i = f_7(t_i) + W_i + c$	500	$5892_{(0,12,6)}$	$8796_{(10)}$	
W_i – dvojitě exponenciální	1000	$3812_{(0,12,6)}$	$6198_{(10)}$	
s hustotou	2000	$1823_{(0,12,6)}$	$1886_{(10)}$	
$p(x) = \frac{1}{30\sqrt{2}} \exp\{-\frac{\lvert x\rvert}{15\sqrt{2}}\}$	4000	$1602_{(0,15,9)}$	$885_{(10)}$	
	8000	$1711_{(12,15,12)\star}$	$549_{(10)}$	
$Y_i = f_4(t_i) + 20W_i + c$	500	$1771_{(0,12,6)}$	$11353_{(5)}$	
W_i – t-rozdělení	1000	$2185_{(6,12,6)}$	$2194_{(8)}$	
o 5 stupních volnosti	2000	$762_{(0,12,6)}$	$1767_{(8)}$	
	4000	$927_{(9,12,9)}$	$1463_{(9)}$	
	8000	$1096_{(18,18,12)\star}$	$520_{(9)}$	
$Y_i = f_5(t_i) + 10W_i + c$	500	$845_{(9,9,9)}$	$1533_{(6)}$	
W_i – t-rozdělení	1000	$360_{(0,12,6)}$	$382_{(8)}$	
o 5 stupních volnosti	2000	$214_{(6,12,6)}$	$295_{(8)}$	
	4000	$209_{(6,12,6)}$	$971_{(6)}$	
	8000	$223_{(10,18,10)\star}$	$146_{(10)}$	

* Při větších hodnotách n se může stát, že si software pro výpočet LMS křivek sám zvýší počet stupňů volnosti oproti těm, které zadáme. Touto značkou jsou označeny případy, kdy si funkce v nějakém ze zkoušených modelů sama upravila počet e.d.f. V závorkách u V^u je pak upravený počet e.d.f.

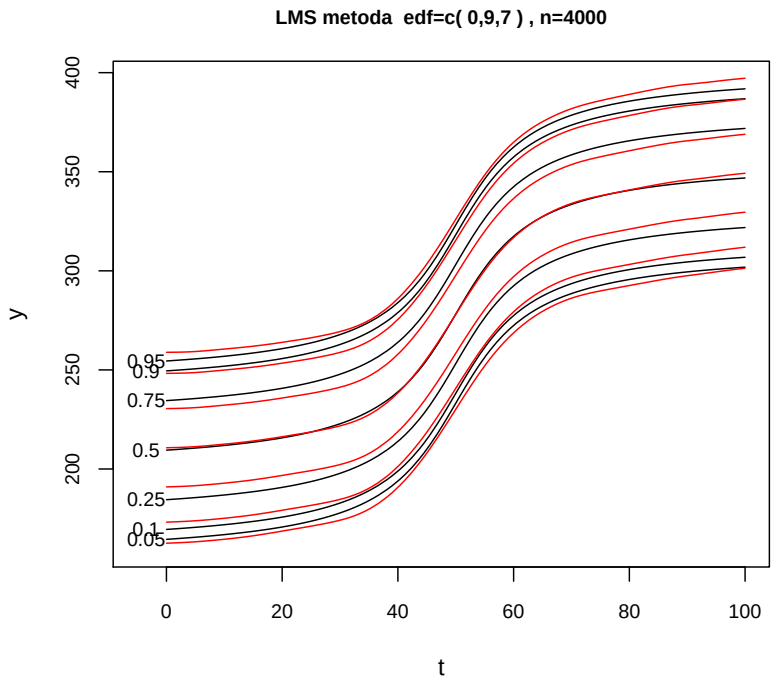
$^\circ$ Označuje, že tato data jsou použita v některém z obrázků této kapitoly.

Tabulka 4.4: Výsledky simulování (předpoklad LMS nesplněn)(2.část): $V^u_{(\text{model})}$ znamená, že nejlepšího výsledku V^u bylo dosaženo pro model v závorce, f_j jsou funkce z obrázku 3.2, c je konstanta pro zaručení kladnosti všech Y_i.

Schéma generování dat		LMS	RQ	
předpis	n	$V^u_{(\text{model})}$	$V^u_{(\text{model})}$	
$Y_i = f_2(t_i) + W_i + c$	500	$2259_{(6,9,6)}$	$3645_{(5)}$	
W_i – logistické	1000	$1379_{(9,9,6)}$	$1319_{(6)}$	
s distribuční funkcí	2000	$559_{(0,12,6)}$	$831_{(6)}$	
$F(x) = \frac{1}{15}\exp\{-\frac{x-10}{15}\}$	4000	$486_{(1,12,6)}$	$489_{(7)}$	
	8000	$306_{(0,12,10)^\star}$	$206_{(7)}$	
$Y_i = f_3(t_i) + (W_i)^3 + c$	500	$19363_{(0,6,3)}$	$13986_{(4)}$	
W_i – rovnoměrné	1000	$13515_{(0,6,3)}$	$3706_{(4)}$	
na [-5,5]	2000	$13575_{(3,6,3)}$	$1093_{(4)}$	
	4000	$13714_{(1,6,6)^\star}$	$1369_{(4)}$	
	8000	$13514_{(0,11,11)^\star}$	$375_{(4)}$	
$Y_i = f_4(t_i) + (W_i)^3 + c$	500	$17636_{(0,12,6)}$	$9871_{(7)}$	
W_i – rovnoměrné	1000	$17878_{(0,12,6)}$	$12673_{(5)}$	
na [-5,5]	2000	$15468_{(0,12,6)}$	$4494_{(8)}$	
	4000	$13944_{(0,12,6)}$	$3025_{(9)}$	
	8000	$13817_{(0,12,12)^\star}$	$2369_{(7)}$	
$Y_i = f_7(t_i) + f_e(W_i) + c$	500	$827_{(6,12,6)}$	$982_{(10)}$	
$f_e(x) = \text{sgn}(x)\sqrt{\lvert x\rvert}$	1000	$520_{(1,15,15)}$	$1413_{(10)}$	
W_i – dvojitě exponenciální	2000	$526_{(1,15,15)}$	$673_{(10)}$	
$p(x) = \frac{1}{400\sqrt{2}}\exp\{-\frac{\lvert x\rvert}{200\sqrt{2}}\}$	4000	$421_{(1,15,15)}$	$234_{(10)}$	
	8000	$414_{(0,18,15)^\star}$	$179_{(10)}$	
$Y_i = M(t_i)[1 + L(t_i)S(t_i)W_i]^{\frac{1}{L(t_i)}}$	500	$12563_{(3,9,6)}$	$16467_{(8)}$	
$L(t) = 1 + 0.1\sin(\frac{1}{100}t\pi)$	1000	$3759_{(3,12,6)}$	$2951_{(8)}$	
$M(t) = f_4(t) + c$	2000	$6664_{(9,12,6)}$	$4749_{(7)}$	
$S(t) = \frac{1}{M(t)}(20 + 20\sin(\frac{1}{100}t\pi))$	4000	$3455_{(5,9,6)^\star}$	$3925_{(7)}$	
$W_i : p(x) = \frac{1}{2}\exp\{-\lvert x\rvert\}$	8000	$3918_{(12,15,12)^\star}$	$781_{(10)}$	
$Y_i = f_4(t_i) + 50W_i + c$	500	$10020_{(1,9,6)}$	$15600_{(10)}$	†
W_i – logaritmicko	1000	$4531_{(6,12,6)}$	$7287_{(9)}$	†
normální s hustotou	2000	$3688_{(9,12,6)}$	$5528_{(8)}$	†
$p(x) = \frac{2}{x\sqrt{2\pi}}\exp\{-2(\ln x - \frac{1}{2})^2\}$	4000	$1378_{(6,9,6)}$	$1884_{(8)}$	†
	8000	$813_{(12,12,11)^\star}$	$1192_{(9)}$	†

* Při větších hodnotách n se může stát, že si software pro výpočet LMS křivek sám zvýší počet stupňů volnosti oproti těm, které zadáme. Touto značkou jsou označeny případy, kdy si funkce v nějakém ze zkoušených modelů sama upravila počet e.d.f. V závorkách u V^u je pak upravený počet e.d.f.

† Pro W_i v tomto případě existuje transformace převádějící data na normální. V kategorii nesplňující předpoklady LMS metody jsou tyto výsledky umístěny kvůli tomu, že se tato data nedají vyjádřit ve tvaru (4.1).

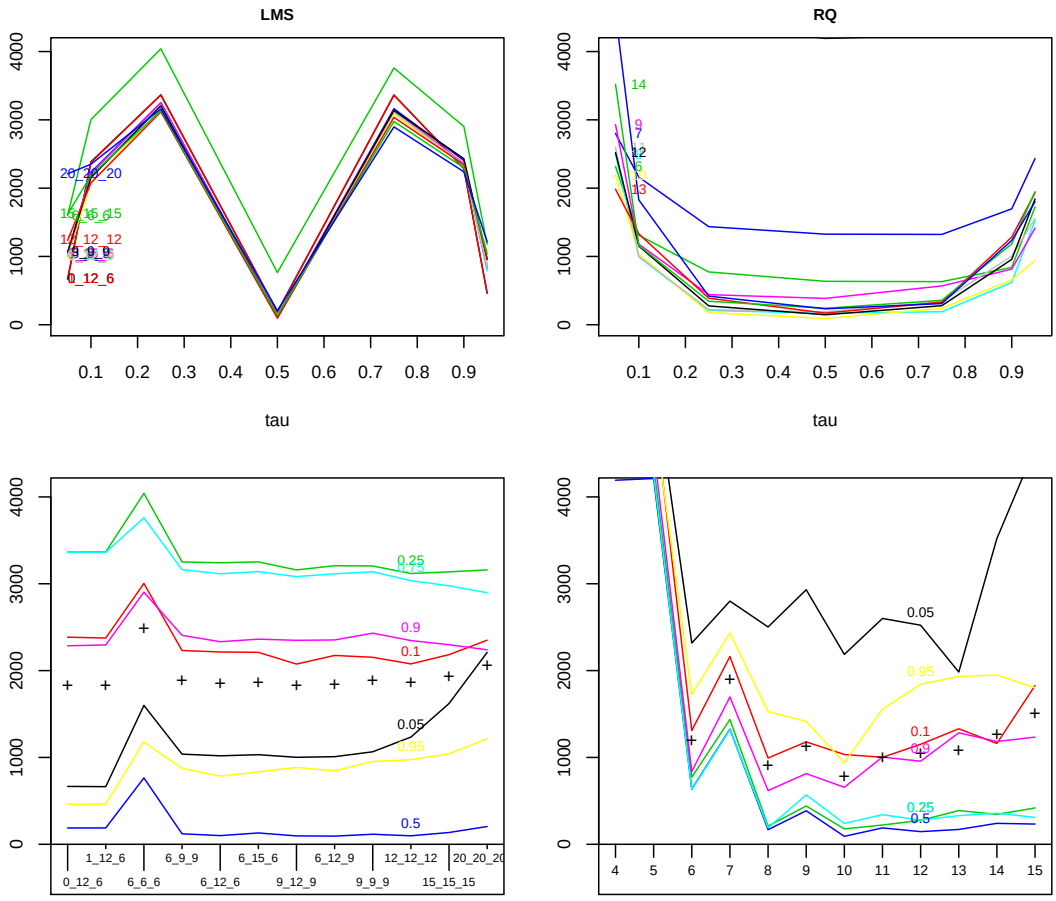


Obrázek 4.1: Nejlepší model pro LMS metodu při datech generovaných jako $Y_i = \ f_5(t_i) + 209.5 + W_i$, $i = 1, \ldots, 4000$, kde W_i je generováno z rovnoměrného rozdělení na $[-50, 50]$. Je zřetelně vidět jisté posunutí mezi teoretickými a odhadnutými kvantily.

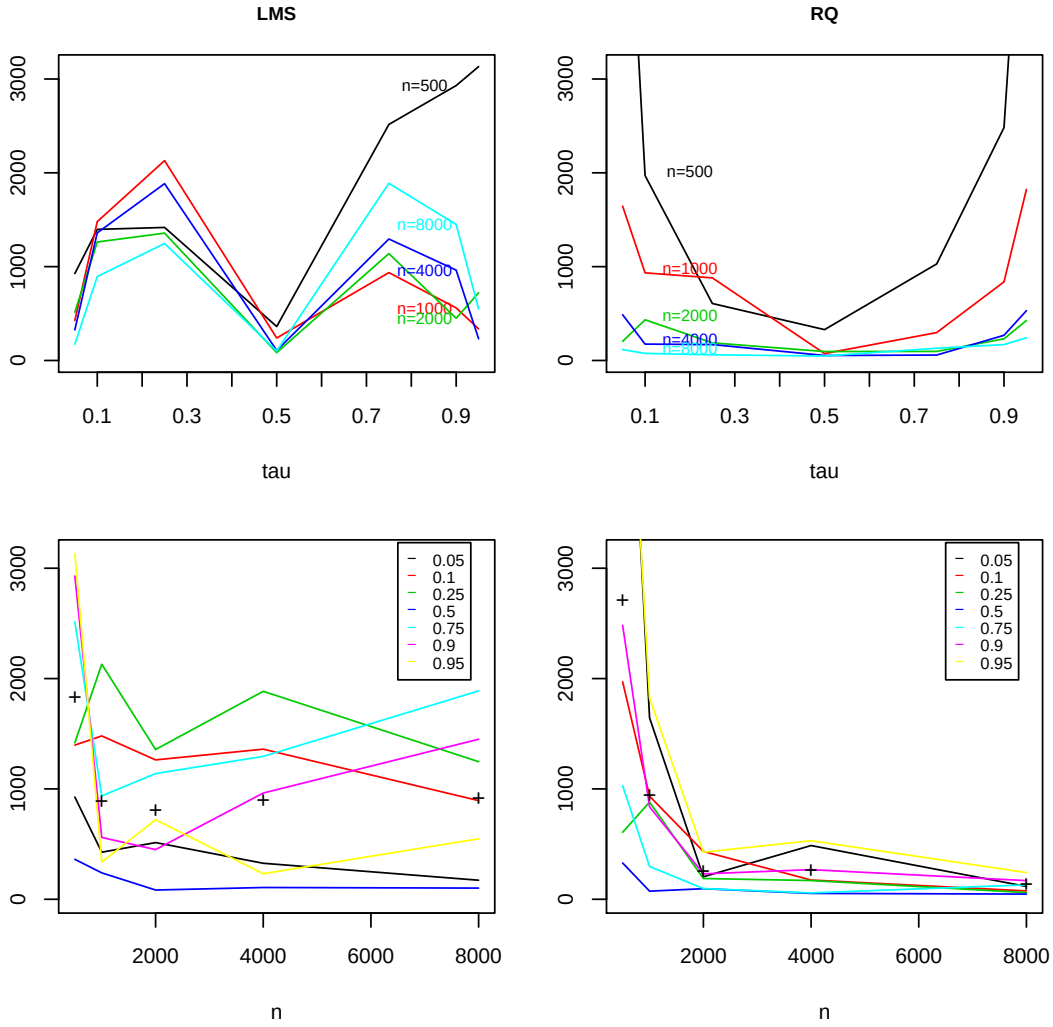
a největší hodnoty V_i jsou u krajních kvantilů.

Abych vyloučil možnost, že se jedná jen o špatnou volbu odpovídajících stupňů volnosti, provedl jsem pro data z obrázku 4.2 dodatečné výpočty pro mnohem rozsáhlejší skupinu trojic e.d.f. (celkem 30 modelů). Jednotlivé složky trojice e.d.f. byly voleny v rozmezí od 0 až do 40, také byly vyzkoušeny různé varianty toho, která složka této trojice bude největší, atd. Model, který by se k datům blížil lépe, ale nebyl nalezen. Navíc jako další poznatek se při této dodatečné analýze ukázalo, že všechny modely mají podobné chování jako na obrázku 4.2 bez ohledu na konkrétní e.d.f.

Zajímavé bude i srovnání metod, co se týče toho, jak se odhadnuté křivky v těchto případech blíží k teoretickým křivkám s rostoucí velikostí datového souboru (obrázek 4.3). Z obrázků i výsledků v tabulce 4.3 to vypadá, že s rostoucím rozsahem výběru se odhadnuté křivky u LMS metody blíží (konvergují) k trochu posunutým teoretickým křivkám (což by byl zásadní problém). Metoda kvantilové regrese tuto vlastnost nemá a „konverguje“ ke správným křivkám. Toto je také důvod, proč jsou od určitého rozsahu výběru hodnoty V pro metodu kvantilové regrese menší než hodnoty V pro LMS metodu (obrázek 4.3).



Obrázek 4.2: Porovnání metod na datovém souboru nesplňujícím předpoklady LMS metody. Konkrétně na datech $Y_i = f_5(t_i) + 286 + W_i$, $i = 1, \ldots, 4000$, kde W_i má hustotu $p(x) = \frac{1}{30\sqrt{2}} \exp\{-\frac{|x|}{15\sqrt{2}}\}$. Horní obrázky: každému modelu přísluší jedna lomená čára složená z hodnot $V_j, j = 1, \ldots, 7$ příslušných tomuto modelu. Vodorovná osa určuje, ke kterému kvantilu V_j patří. Na dolních obrázcích jsou pak vykresleny stejné V_j, ale lomená čára jedné barvy teď představuje hodnoty V_j pro konkrétní kvantil τ_j měnící se v závislosti na modelu (jednotka vodorovné osy tedy představuje jeden model). + ukazuje hodnotu V^u (průměr V_j) pro model, jehož označení je v popisku vodorovné osy. Z toho plyne, že jako nejlepší model pro tato data je vyhodnocen ten, u kterého je značka + nejníže. Obrázky na levé straně přísluší LMS metodě, na pravé straně RQ metodě. Na první pohled je vidět rozdílné chování LMS a RQ metody. Také je vidět, že modely LMS metody se chovají všechny v odhadování křivek podobně – největší problémy jsou s odhadnutím 1. a 3. kvartilu.



Obrázek 4.3: Hodnoty V_j závislé na velikosti dat – porovnání metod na datech nesplňujících předpoklady LMS metody. Na horní dvojici obrázků jsou zakresleny výsledky $V_j, j = 1, \ldots, 7$ nejlepších modelů pro různé rozsahy dat generovaných stejným způsobem. Jedna lomená čára značí hodnoty $V_j, j = 1, \ldots, 7$ nejlepšího modelu pro daný rozsah dat n. Na dolních obrázcích představuje lomená čára jedné barvy změnu V_j (pro daný kvantil τ_j) v závislosti na n pro nejlepší modely. Byla použita data, jejichž výsledky jsou v 6.–10. řádku tabulky 4.3. Je jasně vidět, že se v tomto případě chová výrazně lépe RQ metoda. Navíc nic neukazuje na to, že by se kvalita odhadnutých křivek u LMS metody s rostoucím rozsahem dat nějak vylepšovala.

Všechny výsledky ukazují na to, že LMS metoda v těchto případech nedokáže zcela správně popsat data a má vážné problémy s odhadem některých kvantilových křivek. Problém se nevyřeší ani při větších datových souborech. Konkrétně měla LMS metoda největší problémy s vystihnutím některých kvantilů u všech dat generovaných pomocí rovnoměrného rozdělení (obrázek 4.1 a tabulky 4.3, 4.4), dvojitě exponenciálního (obrázky 4.2, 4.3 a tabulky 4.3, 4.4) a jejich transformací. U logistického a t-rozdělení se tento jev také projevil, je ale méně patrný.

4.2 Chování odhadnutých růstových křivek

Nyní je potřeba se zmínit i o dalších vlastnostech sledovaných metod. Tyto vlastnosti jsou odvozeny především z výsledných odhadnutých křivek na vygenerovaných datech pro jednotlivé metody. Nejlépe se dá o následujících vlastnostech přesvědčit shlédnutím obrázků na přiloženém médiu.

Poznatky jsou shrnuty ve formě výhod a nevýhod jednotlivých metod.

Malé rozsahy dat a křížení odhadnutých křivek

Při menším rozsahu dat má RQ metoda problémy. Odhadnuté křivky se odlišují od teoretických mnohem více než křivky sestavené LMS metodou.

V několika případech dokonce dochází u odhadnutých křivek k nežádoucímu křížení některých kvantilů.

Jev křížení odhadnutých křivek se u metody kvantilové regrese může vyskytnout. To se dá ilustrovat na následujícím jednoduchém příkladu: Máme úlohu, kde je podmíněná kvantilová funkce rovna $Q_Y(\tau|x) = \beta_0(\tau) + \beta_1(\tau)x$, pak výsledným odhadem jednotlivých podmíněných kvantilů jsou přímky procházející každá dvěma datovými body. Každé dvě tyto přímky se někde protínají. Jde jen o to, jestli se tak stane uvnitř nebo mimo prostor, na kterém se podmíněnými kvantily zabýváme.

To je způsobeno tím, že kvantilová regrese využívá jen lokální informace okolo určitého kvantilu, není citlivá na velikosti odchylek dat nad nebo pod křivkou.

Problém křížení se týká krajů intervalů, na kterých křivky sestavujeme. Při rozsahu dat větším než 2000 pozorování došlo ke křížení sledovaných kvantilů na našich datech už jen jedenkrát. S rostoucí velikostí dat přestává být křížení kvantilových křivek problémem.

Chování na krajích intervalů

Z výsledných odhadnutých křivek je také patrné, že RQ metoda má největší potíže s odhadem křivek právě na krajích intervalů, na kterých křivky sestavujeme. Tyto potíže jsou větší než u LMS metody (to dokládá i obrázek 4.6, který byl vytvořen v rámci dodatečných analýz, které teprve budou popsány).

Počet dat pod křivkami

Nespornou výhodou RQ metody je již zmíněná vlastnost správného poměru dat pod křivkami. LMS metoda tuto vlastnost nezaručuje a jak jsme se mohli přesvědčit na příkladech dat nesplňujících předpoklady LMS metody, v některých případech je poměr dat pod křivkou až příliš odlišný od očekávaného (obrázek 4.5).

Hladkost nejlepších modelů

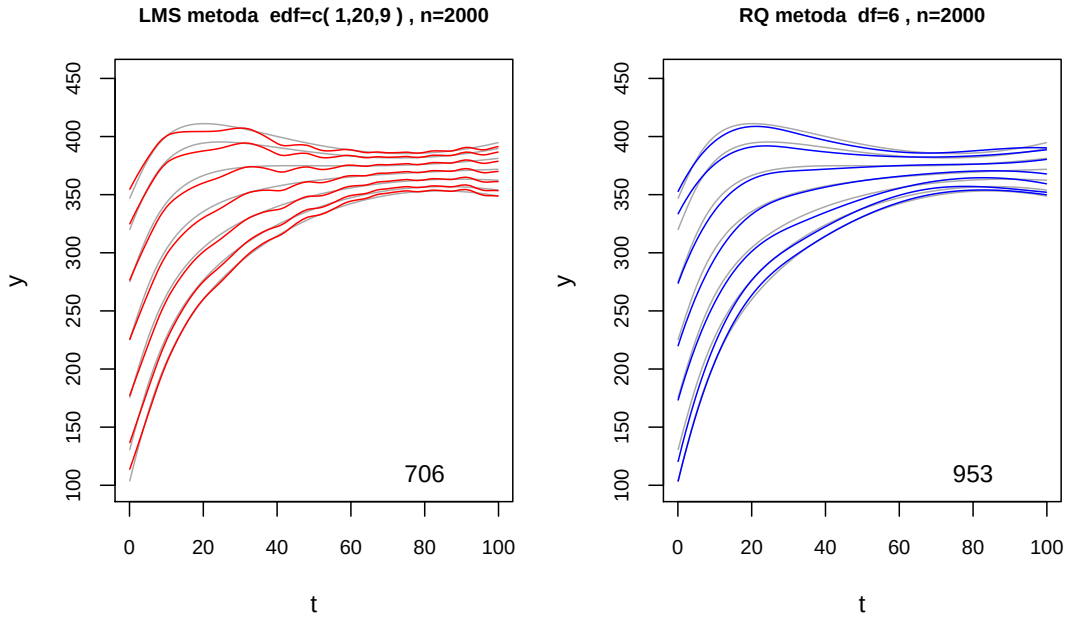
Z literatury je znám fakt, že při určování e.d.f. pro reálná data pro LMS metodu je největší problém ve vyvážení hladkosti a kvality modelu.

Při příliš velkých hodnotách e.d.f. nejsou výsledné křivky hladké (odhadnuté křivky se vlní okolo teoretických křivek, při pohledu na takovouto odhadnutou křivku se to pozná tak, že křivky mají mnoho více či méně patrných hrbolků, které nesouvisí s tvarem teoretického rozdělení). U RQ metody také platí, že při velkém počtu prvků báze se křivky příliš vlní (tento jev je vidět například na obrázku 3.1).

Pokud vezmeme jako měřítko kvality modelu zde používanou hodnotu V^u a porovnáme nejlepší modely pro jednotlivé metody, zjistíme, že nejlepší modely LMS metody ve většině případů nejsou tak hladké jako nejlepší modely RQ metody (jeden z takových případů je vidět na obrázku 4.4). Potvrzuje to tedy to, že model LMS metody blížící se nejlépe datům, má v některých případech problémy s hladkostí odhadnutých křivek. Naopak RQ metoda tento problém prakticky nemá – se zvýšením kvality modelu nevzniká problém s hladkostí. Na základě výsledných odhadnutých křivek se dá konstatovat, že u RQ metody není problém s vyvážením kvality a hladkosti.

Odpovídající stupně volnosti

Další vlastností LMS metody je to, že si míru složitosti e.d.f. pro některou z křivek $L(t)$, $M(t)$ a $S(t)$ volíme pro celý interval, na kterém křivky odhadujeme. To znamená, že každý úsek odhadnutého splinu má stejnou míru složitosti (úseky křivky se vlní „podobně“). Z toho plyne, že pokud je jedna část teoretických křivek výrazně hladší než druhá, pak bude mít LMS metoda v jedné z těchto částí problémy správně odhadnout teoretické rozdělení. To se potvrdilo především u dat, kde je křivka $M(t)$ volena jako $f_6(t) + c$ (viz obrázky v příloze).
Pozn.: V literatuře lze najít modifikaci LMS metody, která tento problém částečně řeší. Tato modifikace je popsána v [10]. Jedná se o to, že se nejdříve transformuje t na základě předběžného odhadu křivky M, poté se odhadne LMS model pomocí penalizované věrohodnosti a nakonec se převede transformované t zpět a sestaví se růstové křivky. Tato úprava nebyla na našich datech vyzkoušena.



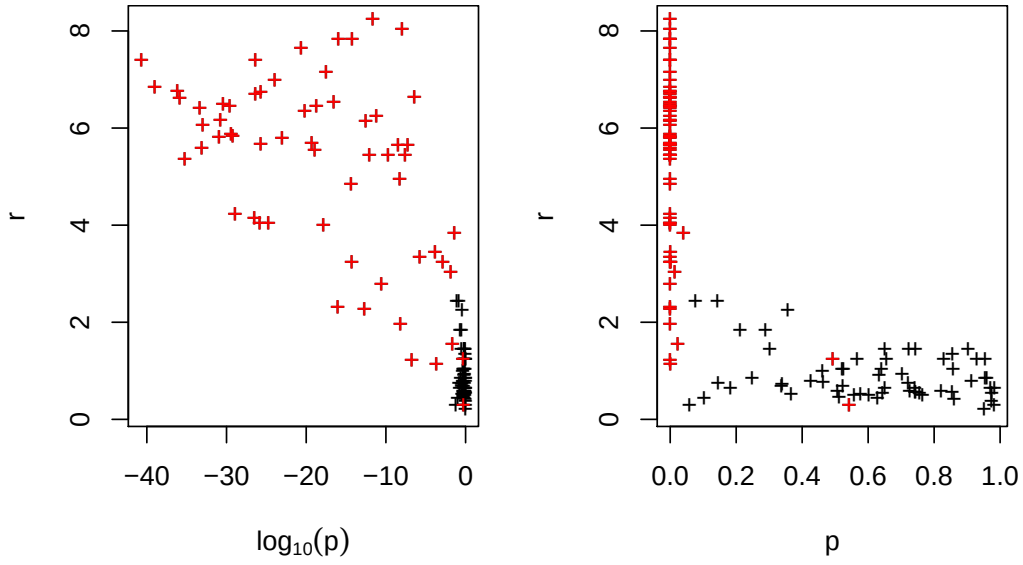
Obrázek 4.4: Odhadnuté křivky nejlepších modelů pro data $Y_i = f_2(t_i) + \frac{1}{100}(t_i - 80)^2 Z_i + 224$, $Z_i \sim N(0,1)$, $i = 1, \ldots, n$. Do obrázků jsou přikresleny šedou barvou teoretické křivky. V pravých dolních rozích jsou dosažené hodnoty V^u. Je vidět, že LMS metoda má problémy s hladkostí výsledných křivek.

4.3 Dodatečné analýzy

4.3.1 Normalita výsledných veličin Z_i

Jednou z možností, jak zjistit, jak dobře se LMS metodě daří data převést na normální, je otestování normality veličin $Z_i, i = 1, \ldots, n$. Z_i se dají jednoduše spočítat z konkrétních dat a z hodnot odhadnutých křivek L, M a S v bodech t_i podle vzorce (2.3). K testování použijeme Shapirův-Wilkův test normality. Tím budeme testovat normalitu veličin $Z_i, i = 1, \ldots, n$ nejlepšího modelu LMS metody zvlášť pro každý datový soubor.

Pro datové soubory, pro které jsou splněny předpoklady LMS metody, test ve všech případech nezamítl hypotézu normality na 5 procentní hladině významnosti. Naopak u dat, pro která nejsou splněny předpoklady LMS metody, test zamítl hypotézu normality na 5 procentní hladině ve všech případech kromě dvou (a to právě u dat generovaných z logaritmicko normálního rozdělení, které je svým postavením v kategorii nesplněných předpokladů LMS metody vyjímečné). To ukazuje, že LMS metoda v případě nesplnění předpokladů opravdu nedokáže převést data na normálně rozdělená.



Obrázek 4.5: Hodnoty rozdílu očekávaného a pozorovaného počtu dat pod odhadnutou křivkou v závislosti na výsledné p-hodnotě Shapirova-Wilkova testu normality výsledných Z_i. Hodnoty r svislé osy (v procentech) značí maximální odchylku (maximum přes sledované kvantily) očekávaného a pozorovaného počtu dat pod určitou odhadnutou křivkou nejlepšího modelu LMS metody. Červeně jsou vyznačeny případy, kdy data nesplňují předpoklady LMS metody. Oba obrázky zachycují stejná data, jen vodorovná osa má jiné měřítko.

Na obrázku 4.5 je zobrazena závislost výsledné p-hodnoty testu normality a rozdílu mezi očekávaným a pozorovaným počtem dat pod odhadnutou křivkou nejlepšího modelu. Je vidět, že s větším porušením normality roste i tento rozdíl. Je také vidět, jak až velký tento rozdíl může být pro jednu kvantilovou křivku. Pro srovnání: u RQ metody nebyl tento rozdíl pro jednu křivku nikdy větší než 1.2% (to je maximum pro všechny nejlepší modely a pro všechny datové soubory).

4.3.2 Data generovaná za pomoci t-rozdělení

V této části se podrobněji podíváme na vlastnosti dat generovaných jako

$$Y_i = f_5(t_i) + 10W_i + c, i = 1, \ldots, 4000, \tag{4.2}$$

kde W_i má t-rozdělení o 5 stupních volnosti. A to především kvůli výsledné hodnotě V^u pro RQ metodu (viz tabulka 4.3). Tato hodnota je nečekaně velká, proto nás může zajímat, jaký to má důvod.

Tabulka 4.5: Statistiky shrnující vypočtené velikosti hodnot V^u pro jednotlivé metody na 20 různých datových souborech generovaných podle (4.2).

metoda	minimum	1.kvartil	medián	průměr	3.kvartil	maximum
LMS	139.4	191.2	214.5	236.4	258.2	432.5
RQ	141.1	164.2	209.0	251.6	250.2	838.2

Za tímto účelem bylo dogenerováno 20 souborů dat generovaných stejně jako v (4.2). Na nich pak byly nalezeny nejlepší modely LMS a RQ metody a spočteny hodnoty V^u. Získané výsledky jsou v tabulce 4.5.

Hodnota $V^u = 971$ u RQ metody v tabulce 4.3 se nám zdála velká, protože se čekalo (vzhledem k ostatním výsledkům tabulky 4.3), že na tomto případě dat budou metody srovnatelné (v rámci dosaženého V^u). Z výsledků v tabulce 4.5 lze říci, že metody opravdu srovnatelné jsou. Hodnota $V^u = 971$ u RQ metody sice není typická pro tento typ dat, ale může se vyskytnout.

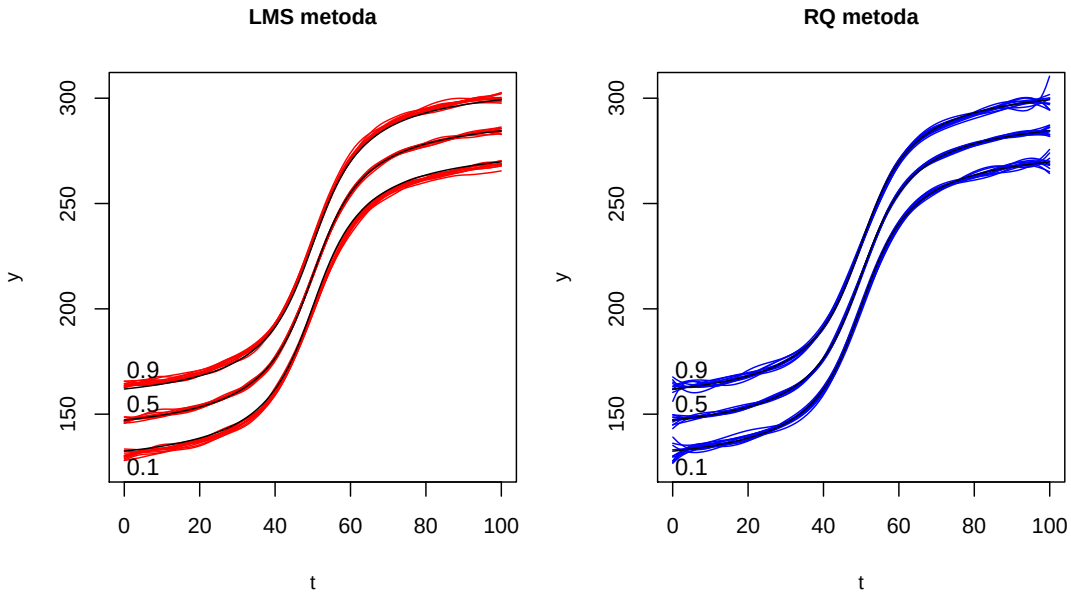
Na tomto příkladě je vidět, že hodnotu V je opravdu nutné spojovat s konkrétním datovým souborem a ne s určitým typem datového souboru.

Pro ilustraci vlastností obou metod bylo na základě této analýzy vykresleno do jednoho obrázku (obrázek 4.4) 10 souborů odhadnutých křivek, kde každý soubor křivek je vytvořen z jednoho datového souboru generovaného podle (4.2). Na obrázku je možné sledovat hned několik vlastností sledovaných metod. Je například vidět, že u LMS metody se odhadnuté křivky 10. percentilu odchylují od teoretické křivky 10. percentilu směrem dolů (podobně pro křivky pro 90. percentil), což je znovu dříve popisovaný problém odhadování křivek při nesplnění předpokladů LMS metody. Naopak u RQ metody tomu tak není. Dále jsou také na obrázku vidět problémy RQ metody na krajích intervalu, na kterém křivky sestavujeme.

4.3.3 Uzly B-splinové báze

Jak již bylo zmíněno, to že byly vnitřní uzly pro vytváření B-splinových bází voleny ekvidistantně, může být v některých případech omezující.

To, že to opravdu omezující může být, ukazuje následující příklad: Na datech $Y_i = f_6(t_i) + 20Z_i + 136$, $i = 1, \ldots, 4000$, $Z_i \sim N(0, 1)$, byla dodatečně provedena analýza, kdy se vnitřní uzly B-splinové báze nevolily ekvidistantně. Tato data byla pro tuto analýzu vybrána především kvůli tvaru funkce $f_6(t)$, která je konstantní na intervalu $[0, 50]$ a pak se začíná vlnit, ekvidistantní vnitřní uzly tedy nejsou příliš vhodné. Na obrázku 4.7 je vidět porovnání nejlepšího modelu pro metodu kvantilové regrese s B-splinovou bází s ekvidistantními vnitřními uzly a modelu s vnitřními uzly zvolenými sice ekvidistantně, ale první vnitřní uzel je až v bodě 45. Je zde jasně vidět vylepšení odhadnutých křivek.



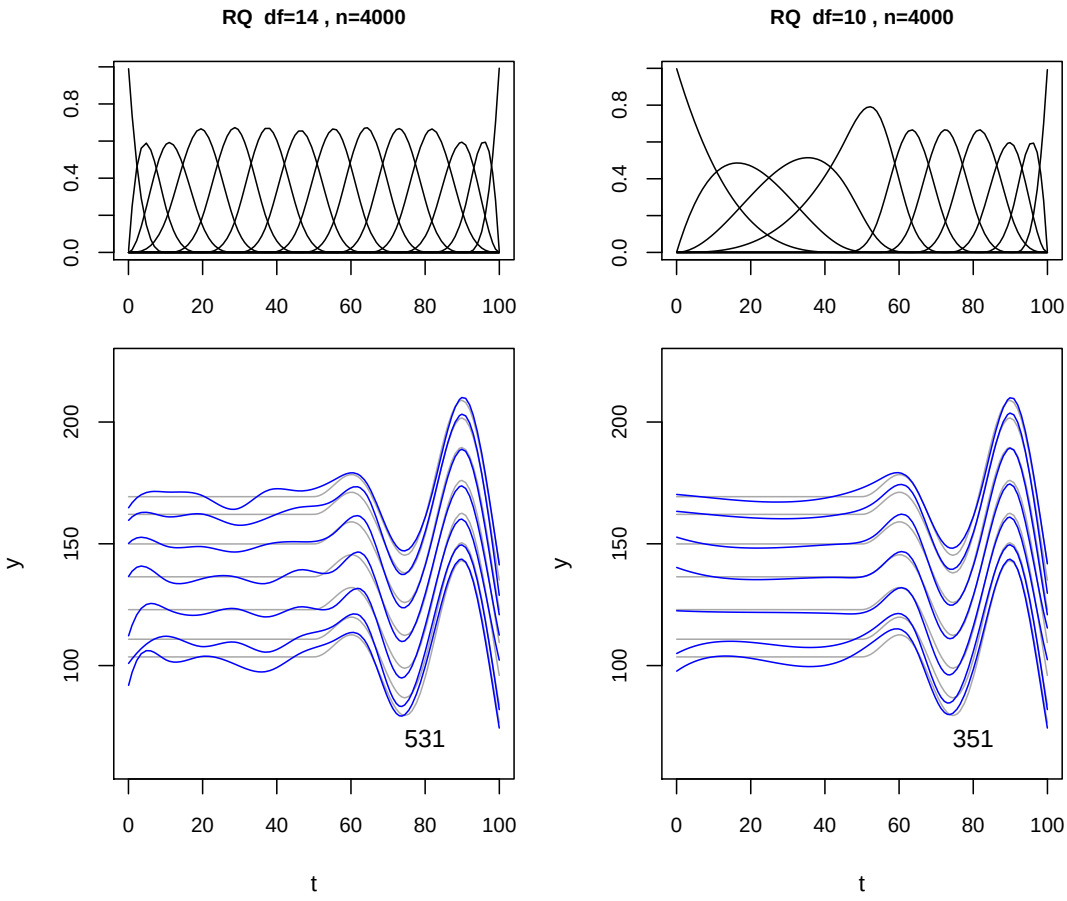
Obrázek 4.6: 10 souborů odhadnutých křivek pro nejlepší modely LMS a RQ metody pro 10 datových souborů generovaných jako $Y_i = f_5(t_i) + 10W_i + c, i = 1, \ldots, 4000$, kde W_i má t-rozdělení o 5 stupních volnosti. Pro přehlednost jsou vyznačeny jen odhadnuté křivky pro 10. a 90. percentil a pro medián. Černě jsou vyznačeny teoretické křivky.

Dá se tedy říci, že metoda kvantilové regrese s ekvidistantními vnitřními uzly má v tomto ohledu ještě jisté rezervy. Při volbě neekvidistantních uzlů je ale potřeba postupovat opatrně, nestačí se dívat na jednotlivé umístění uzlů, ale je třeba vybírat bázi jako celek.

4.4 Zmínka o obecnějších typech dat

V celé práci se zabýváme jen daty, kdy je sledovaný znak závislý jen na jedné spojité veličině a jednotlivá pozorování $[t_i, Y_i]$ jsou navzájem nezávislá. Ačkoli nejpoužívanější typ růstových křivek vzniká na základě takovýchto dat, můžeme si představit i obecnější varianty. V tomto kontextu je potřeba zmínit ještě jednu důležitou výhodu přístupu kvantilové regrese. Přidání další proměnné, na které by závisel sledovaný znak, je relativně jednoduché – stačí rozšířit regresní matici o další sloupce. Stejně tak není problém s daty, kde se vyskytují jisté závislosti. U reálných dat to jsou hlavně případy, kdy sledujeme znak u jednoho jedince delší časové období a máme tedy pro tohoto jedince více pozorování.

Zobecnění LMS metody jistě nebude tak jednoduché (článek [11]).



Obrázek 4.7: Příklad vylepšení RQ metody volbou neekvidistantních vnitřních uzlů B-splinové báze. Data generována jako $Y_i = f_6(t_i) + 20Z_i + 136$, $i = 1, \ldots, 4000$, $Z_i \sim N(0, 1)$. Vnitřní uzly vylepšující B-splinové báze (obrázek vpravo nahoře) jsou $(45, 54.17, 63.33, 72.5, 81.67, 90.83)$. Horní obrázky ukazují použité B-splinové báze, dolní obrázky ukazují odhadnuté růstové křivky (vykreslené modře) vzniklé RQ metodou pro tyto báze. V pravém dolním rohu dolních obrázků jsou dosažené hodnoty V^u. Šedou barvou jsou přikresleny teoretické křivky.

Kapitola 5

Shrnutí

Na základě analýz na nagenerovaných datových souborech lze vyvodit pro použití LMS metody a metody kvantilové regrese založené na B-splinové bázi několik následujících doporučení.

Použití RQ metody není vhodné pro rozsahy dat menší než 1500 pozorování – LMS metoda má v těchto případech (dokonce, i když nejsou splněny její předpoklady) kvalitnější odhadnuté křivky a navíc u RQ metody dochází v těchto případech nejčastěji k nežádoucímu křížení odhadnutých křivek pro sledované kvantily.

Nejdůležitějším výsledkem, který podává tato práce, je to, že pokud data nesplňují předpoklady LMS metody, pak se mohou výsledné odhadnuté růstové křivky zásadně lišit od teoretických křivek a to pro všechny zkoušené modely na těchto datech. Problémem u takových křivek je, že pro některé kvantily neodpovídá počet dat pod křivkou očekávané hodnotě, což je nevýhodné vzhledem k účelu, pro který jsou křivky sestavovány. Situace se nezlepší ani při větším rozsahu dat. Navíc je nepříjemné, že se toto nepozná na tvaru odhadnutých křivek.

LMS metodu bychom tedy měli použít, jen pokud nebudou zásadně porušeny její předpoklady (nebo, jak bylo řečeno, při malých rozsazích dat). Nesplnění předpokladů (tedy především případ neexistující transformace dat, která by převedla data na normálně rozdělená) se dá, jak se ukázalo, celkem snadno ověřit testem normality.

Pokud jsou splněny předpoklady LMS metody, pak LMS metoda lépe popisuje data, v některých případech je to ale na úkor hladkosti křivek. S rostoucím rozsahem dat se kvalita odhadnutých křivek RQ metody blíží výsledkům LMS metody. Navíc hladkost výsledných nejlepších modelů ukázala, že odhadnuté křivky RQ metody jsou hladší a také, že RQ metoda téměř nemá problémy s vyvážením hladkosti a kvality odhadnutých křivek.

Při větších rozsazích dat se tedy dá RQ metoda výhodně použít. Metoda není zatížena žádnými předpoklady, není problém s vyvážením hladkosti a kvality odhadnutých křivek a navíc zaručuje správný počet dat pod jednotlivými odhadnutými křivkami. Jedinou nevýhodou jsou možné problémy v chování křivek na

krajích intervalů.

5.1 Možnosti navázání na práci

Dalším možným krokem v práci by bylo navázat na získané výsledky ve snaze nalézt postup, který by řešil odhadování křivek na reálných datech. V literatuře je popsáno mnoho kritérií, jak určit kvalitu modelu na reálných datech (viz například [9]). Cílem by tedy bylo vyzkoušet takováto kritéria na souborech našich nagenerovaných dat a najít takové kritérium, které by vybíralo modely podobné těm modelům, které byly v této práci vyhodnoceny jako dobré. Toto kritérium (nebo kombinace kritérií) by bylo následně použito pro vybrání nejlepšího modelu na reálných datech. Přičemž by se upřednostňovalo použití RQ metody především vzhledem k tomu, že se nezdá být problémem ve vyvážení hladkosti a kvality odhadnutých křivek. To samozřejmě jen v případě, že by se jednalo o větší datový soubor.

Při soustředění na RQ metodu by se pak dalo podrobněji zabývat problematikou umístění vnitřních uzlů, což by mohlo vést k dalšímu vylepšení odhadnutých křivek.

Dodatek A

Popis přiloženého média a použitý software

K práci je přiloženo DVD, na kterém jsou nejdůležitější zdrojové kódy, řada obrázků a složky s daty. V každé ze složek jsou konkrétní nagenerovaná data, zkoušené modely LMS a RQ metody a výsledky porovnání metod. Vše ve formě souborů typu .Rdata.

Všechny výpočty byly provedeny programem R, verze 2.4.0 pro Windows. Pro konstrukci modelů LMS metody byla použita knihovna `gamlss` verze 1.4-0. Pro modely RQ metody knihovna `quantreg` verze 4.02 a pro ni potřebná knihovna `SparseM` verze 0.71.

Internetová adresa na stránky softwaru R je http://www.r-project.org.

Literatura

[1] Cole T. J., (1988): Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society, Series A* **151**, 385–428.

[2] Cole T. J., Green P.J. (1992): Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine* **11**, 1305–1319.

[3] Box G. E. P., Cox D. R. (1964): An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**, 211–252.

[4] Koenker R. (2005): Quantile Regression. Cambridge University Press, Cambridge.

[5] Najzar K. (2006): Základy teorie splinů. Karolinum, nakladatelství Univerzity Karlovy Praha.

[6] Stasinopoulos M., Rigby B., Akantziliotou C. (2004): GAMLSS: An R package for generalized additive models for location, scale and shape. http://www.londonmet.ac.uk/gamlss/

[7] Koenker R. (2004): Quantreg: An R package for quantile regression and related methods. http://cran.r-project.org

[8] Gannoun A., Girard S., Guinot Ch., Saracco J. (2002): Reference curves based on non-parametric quantile regression. *Statistics in Medicine* **20**, 3119–3135.

[9] Pan H., Cole T. J. (2004): A comparison of goodness of fit tests for age-related reference ranges. *Statistics in Medicine* **23**, 1749–1765.

[10] Cole T. J., Freeman J.V., Preece M.A. (1998): British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Statistics in Medicine* **17**, 407–429.

[11] Cole T. J. (1994): Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine* **13**, 2477–2492.