

METODY PRO PROKLÁDÁNÍ KŘIVEK S POUŽITÍM NA REÁLNÝCH DATECH

Daniel HLUBINKA¹

MFF UK, KPMS

Abstract. In the paper an unique set of meteorological measurements, namely aerological measurements of gamma and beta radiation, is treated. The main purpose is to show possible statistical approaches to the data analysis when the physical model is unknown. We are looking, first of all, for a parametrical model best fitting the data. It turns out to be a derivative of the standard Richards growth model. Then we try to model the data using well known Poisson process with non-constant intensity. In the second part we consider non and semiparametrical statistics. The kernel regression estimator and its modifications trying to improve the estimation closest the borders of observed interval and/or deal with the data heteroscedasticity are considered. Then we show the possibility how to use local polynomial smoother. In the last section several nonparametrical estimators of derivatives of estimated curve are presented.

Резюме: В статье изучается единственная совокупность метеорологических измерений, точнее аэрологических измерений бета и гамма излучения. Главная проблема является показать возможные статистические процедуры анализа данных когда физикальная модель неизвестна. Прежде всего мы ищем параметрическую модель лучше всех аппроксимаций. Хорошей является модель Рычардсовы кривой возрастания. В следующем мы попробуем использовать модель процесса Пуассона с переменной напряженностью. Во второй части изучается непараметрическая и семипараметрическая статистика. Вместе с ядровой оценкой регрессии остоновлюсь ее модификации для крайних точек интервала и данных с переменной вариацией. После того изучается оценка регрессии методом локальных полиномов и в самом конце покажем три методы непараметрической оценки производной неизвестной функции.

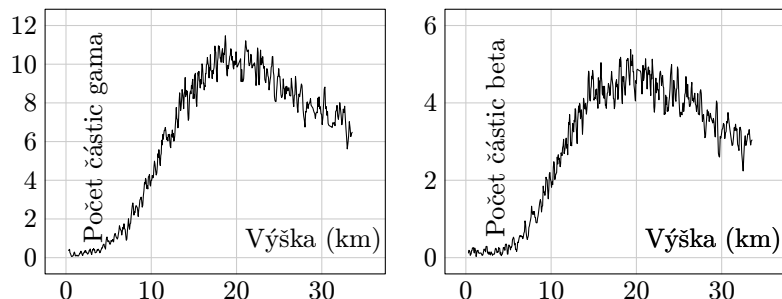
1. ÚVOD

Jednou měsíčně jsou na meteorologické observatoři v Praze – Libuši měřeny, kromě jiného, údaje o intenzitě gama a beta záření v atmosféře pomocí výškových balónů. Údaje jsou dostupné za období od léta 1994, přičemž v současné době jsou nám k dispozici data z 34 výstupů, další data máme slíbená. Pro další potřeby používáme slovo *pozorování*, budeme-li mluvit o jednom výstupu balónu, a *měření* pro označení matice hodnot daného pozorování. Měření radioaktivity, jež probíhá pomocí Geigerova – Müllerova počítače, je meteorology upraveno na „oficiální tvar“ přepočítáním na počet částic za vteřinu v desetivteřinovém časovém úseku. Tyto údaje jsou udány s přesností na dvě desetinná místa. Dále měření obsahuje údaj o teplotě a tlaku,

¹Autor děkuje Dr. Jaromíru Antochovi z KPMS MFF UK za nezištnou pomoc a cenné rady poskytnuté v průběhu vzniku tohoto článku.

Tato práce vznikla s podporou grantu GA UK 3051–10/716.

z tlaku se také odvozuje výška balonu. Měření probíhá po kalibraci sondy a jejím vypuštění od země. Řada se ukončuje po začátku pádu balónu, tedy při opětovném růstu tlaku. Takto jsou získány hodnoty o radioaktivitě až do výšky kolem třiceti kilometrů.



Obr. 1 a 2: Povšimněme si tvaru křivky (v podstatě shodné pro beta i gama záření) a zejména prudké změny rozptylu kolem desátého kilometru.

Naším základním úkolem je nalézt vhodnou metodu pro proložení naměřených dat křivkou. Omezíme se na jedno náhodně vybrané pozorování, jmenovitě ze dne 11. dubna 1995. Na tomto praktickém příkladě si ukážeme vhodné parametrické i neparametrické postupy při řešení tohoto úkolu.

Z fyzikálního hlediska je nejzajímavější závislost intenzity záření na výšce. Doposud však bohužel nejsou známy žádné fyzikální zákonitosti, jež by mohly být použity jako základ pro parametrický model. Proto jsme byli postaveni před další úkol – vybrat sami vhodný model. Této otázce je věnována první část o modelování pomocí růstových křivek, jejichž parametry odhadneme pomocí metody nejmenších čtverců.

Krátce se též zmíníme o možnosti použít k modelování náhodný proces, přičemž se zde přímo nabízí proces Poissonův. K odhadu intenzity Poissonova procesu použijeme model z první části, tentokrát však za použití metody maximální věrohodnosti. Tím získáme dva různé způsoby odhadu parametrů téhož modelu.

V dalších částech se obrátíme k takzvaným lokálním vyhlazovacím metodám. Přehledně si uvedeme odvození jádrového odhadu regresní křivky, lokálně polynomického vyhlazení a ukážeme možnou modifikaci jádrového odhadu tak, abychom i na okrajích měřeného intervalu získali méně vychýlený odhad. Dále si uvedeme tři možnosti odhadu derivace regresní křivky, a zmíníme možnost použití proměnné šířky okna při jádrovém odhadu s cílem snížit vliv heteroskedasticity vstupních dat na výsledný model.

Tato práce neobsahuje věty a důkazy. Důraz je spíše kladen na myšlenkové postupy potřebné pro odvození jednotlivých metod.

2. ANALÝZA POMOCÍ RŮSTOVÝCH KŘIVEK

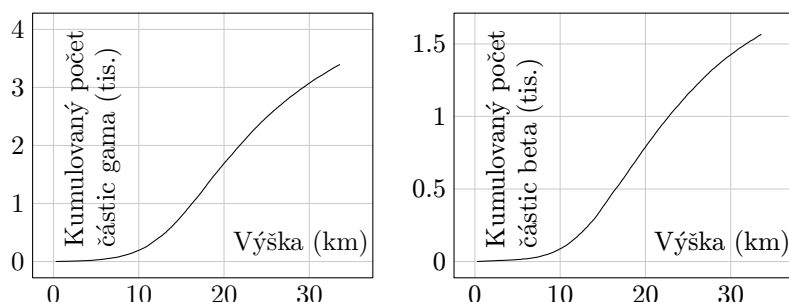
2.1. Volba modelu. Naší snahou v tomto oddíle je nalézt parametrickou křivku, která by aproximovala co nejlépe naměřená data, zde metodou nejmenších čtverců. Zavedme si jednotné značení, kterého se budeme v následujícím textu držet.

Řada měření je dána dvojicemi (X_i, Y_i) , kde X_i značí výšku a Y_i odpovídající hodnotu radioaktivity v i -tém časovém intervalu. Proměnné x a y použijeme opět pro výšku a radioaktivitu, nyní však jako spojité proměnné, nikoliv naměřené hodnoty. Pro čas jako proměnnou použijeme obvyklé značení t .

Původní řada měření tvoří poměrně zřejmou, avšak složitou křivku, kde nemáme příliš jasnou představu o možném tvaru aproximující funkce. Proto jsme se rozhodli studovat křivku částečných součtů, jinými slovy místo hodnotami radioaktivity v dané výšce (průměry v desetivteřinových intervalech) se budeme zabývat počtem částic do dané výšky (součtem udávaných průměrů). Do další analýzy tak vstupujeme s údaji

$$Z_k = \sum_{i=1}^k Y_i$$

Tyto hodnoty jsme si opět znázornili proti výšce a usoudili, že by je bylo možno aproximovat některou vhodnou růstovou křivkou. Ze známých křivek po kratším rozboru a studiu dat přicházely do úvahy v podstatě pouze tři křivky: logistická, Gompertzova a Richardsova. Po vyzkoušení všech tří jsme dospěli k názoru, že našemu účelu nejlépe poslouží Richardsova křivka. Hovořila pro ni především větší (byť stále ještě ne optimální) pružnost, nízká rezidua a jejich průběh v závislosti na vysvětlující proměnné. Navíc je, na rozdíl od předchozích dvou křivek čtyřparametrická. Další výhody se ukáží v pozdějším textu.



Obr. 3 a 4: Tvar křivky pro kumulované součty vede k použití růstových modelů, heteroskedasticita je vzhledem k velikosti součtů zanedbatelná.

Ukázalo se tedy, že máme vhodnou křivku pro aproximaci částečných součtů. Původní měření pak z částečných součtů získáme rozdílem dvou po sobě jdoucích hodnot, neboli $Y_i = Z_i - Z_{i-1}$. K proložení původní řady měření proto použijeme derivaci Richardsovy křivky.

Samotný tvar Richardsovy křivky je

$$(1) \quad R(x; a, b, c, d) = a \left(1 + (b-1)e^{-c(x-d)} \right)^{1/(1-b)}.$$

Naším prvním úkolem nyní je odhadnout parametry a, b, c, d tak, aby

$$(2) \quad \sum_i \left(Y_i - R(X_i; a, b, c, d) \right)^2 = \min.$$

Z tohoto zadání již je patrné, že odhad bude muset být spočten iteračně a nám tak přibývá úloha stanovit počáteční odhad pro spuštění numerické minimalizace.

2.2. Rozbor Richardsovy křivky. Abychom mohli dobře používat vybraný model, je dobré dozvědět se co nejvíce o úloze jednotlivých parametrů, typických vlastnostech zvolené křivky, stejně jako i odhalit případné nedostatky.

Růstové křivky jsou typicky řešením nějaké diferenciální rovnice popisující jistý děj. I z tohoto důvodu je možné považovat růstovou křivku za vhodnou volbu pro modelování daného fyzikálního procesu, ačkoliv nejde o popis tohoto procesu – ten je prozatím neznámý. Richardsova křivka je řešením rovnice

$$(3) \quad \frac{\partial R(x; a, b, c, d)}{\partial x} = c \frac{R(x; a, b, c, d) \cdot (a^{b-1} - R(x; a, b, c, d)^{b-1})}{b-1}.$$

Její derivace, kterou hodláme použít pro proložení původních měření, má tvar

$$(4) \quad \frac{\partial R(x; a, b, c, d)}{\partial x} = r(x; a, b, c, d) = ac(1 + (b-1)e^{-c(x-d)})^{b/(1-b)} e^{-c(x-d)}.$$

2.2.1. Význam parametrů. Nyní se pokusíme načrtnout význam jednotlivých parametrů Richardsovy křivky a jejich vliv na její výsledný tvar. Zaměříme se na limitní chování této křivky. Jelikož parametr d popisuje pouze posunutí v x a parametr a je parametrem měřítka celé křivky, stačí se zaměřit na čtyři význačné případy parametrů b a c . Vyloučíme prozatím situaci $c = 0$ a $b = 1$.

$$(1) \quad b > 1, c > 0$$

V tomto případě je $\lim_{x \rightarrow +\infty} R(x) = a$ a $\lim_{x \rightarrow -\infty} R(x) = 0$.

$$(2) \quad b > 1, c < 0$$

Nyní dostáváme naopak $\lim_{x \rightarrow +\infty} R(x) = 0$ a $\lim_{x \rightarrow -\infty} R(x) = a$.

$$(3) \quad b < 1, c > 0$$

Pro tento případ je $\lim_{x \rightarrow +\infty} R(x) = a$, zatímco limita pro $x \rightarrow$

$-\infty$ postrádá smyslu, neboť by docházelo k mocnění záporného čísla mocninou $1/(1-b)$, což obecně není přirozené číslo.

$$(4) \quad b < 1, c < 0$$

Zde obdobně $\lim_{x \rightarrow -\infty} R(x) = a$, zatímco limita pro $x \rightarrow +\infty$ nemá smysl.

Hranice těch x , pro něž má Richardsova křivka v případě 3 a 4 ještě smysl, bude, spolu s případem $b = 1$, probrána dále.

Předpokládejme, že pracujeme s nezápornými čísly, což u modelů růstu bývá obvyklý případ a proto můžeme předpokládat $a > 0$. Ptejme se na průběh křivky, tedy hledejme její derivace a jejich znaménka.

Již víme, že Richardsova křivka je definovaná na celém \mathbb{R}_1 pouze pro $b > 1$. Pro taková b je pak rozhodující pro směr křivky parametr c , kde $c > 0$ znamená, že křivka je všude rostoucí s dolní/horní uzavěrou $0/a$, zatímco pro $c < 0$ je klesající se stejnými uzavěrami. Je-li $c = 0$, pak dostaneme konstantní funkci $R(x) = ab^{1/(1-b)}$. Pro naše data je však zřejmě $a > 0$ a $c > 0$.

Je-li parametr $b < 1$, pak má křivka $R(x)$ omezený definiční obor a má prakticky smysl pouze tehdy, je-li $1 + (b-1)\exp\{-c(x-d)\} \geq 0$. Je zřejmé, že

$$(5) \quad 1 + (b-1)\exp\{-c(x-d)\} \geq 0 \Leftrightarrow \begin{cases} x \geq d + \frac{\log(1-b)}{c} & c > 0, \\ x \leq d - \frac{\log(1-b)}{c} & c < 0. \end{cases}$$

Dále platí, že je-li $c > 0$, je kladná i první derivace, zatímco je-li $c < 0$, je záporná i první derivace.

Ještě nám zbývá vyšetřit případ $b = 1$. Zde ovšem platí

$$(6) \quad \lim_{b \rightarrow 1+} R(x, b) = \lim_{b \rightarrow 1-} R(x, b) = a \cdot \exp[-\exp\{-c(x-d)\}] = G(x; a, c, d),$$

což je Gompertzova křivka.

Obraťme pozornost k další úloze parametrů b a d . Potřebujeme k tomu spočítat druhou derivaci Richardsovy křivky. Mějme $b \neq 1$, potom

$$(7) \quad \frac{\partial^2 R(x)}{\partial x^2} = ac^2 \cdot e^{-c(x-d)} \left[b \cdot e^{-c(x-d)} (1 + (b-1)e^{-c(x-d)})^{(2b-1)/(1-b)} - (1 + (b-1)e^{-c(x-d)})^{b/(1-b)} \right].$$

Položíme-li druhou derivaci rovnu nule, vyjde nám jako řešení $x = d$. Proto je d bodem inflexe Richardsovy křivky. Podle (5) je však pro $b < 0$ hodnota d mimo definiční obor křivky a daná křivka nemá inflexní bod. Je tudíž celá konkávní a popisuje tak procesy startující z jednoho bodu a mající za asymptotu a . Tímto máme vysvětlenou úlohu parametru d .

Položíme-li $x = d$ (za předpokladu $b > 0$), pak $R(d) = a \cdot b^{1/(1-b)}$. Jelikož je bod d bodem inflexe, je vidět, že hodnota $b^{1/(1-b)}$ udává poměr konvexní části růstu (pro $c > 0$) k celkovému růstu křivky, neboli jistou nesymetrii kolem bodu d . Poznamenejme, že funkce $b^{1/(1-b)}$ je monotónní a nabývá hodnoty 0.5 pro $b = 2$, z čehož plyne, že pro $b = 2$ je Richardsova křivka symetrická okolo bodu $(d, a/2)$. Podrobnější grafy znázorňující vliv parametrů viz Seber (1988).

Zajímavou otázkou může být též vliv změny měřítka na odhadované parametry. Změna měřítka totiž může při numerických odhadech parametrů značně ovlivnit numerickou stabilitu odhadu, proto je tato otázka hodna zájmu. Změna měřítka na ose x , v praxi zejména záměna kilometrů a metrů, vede ke změně v parametrech c a d . Pokud vynásobíme údaje na ose x konstantou k , musíme stejnou konstantou vynásobit parametr d a vydělit parametr c . Změna měřítka na ose y , obvykle normalizace tak, aby horní uzávěra byla rovna 1, ovlivní pouze parametr a . Vynásobíme-li údaje na ose y konstantou k , musíme i parametr a touto konstantou vynásobit.

2.2.2. *Podmodely Richardsovy křivky.* Při předcházejícím studiu limitních vlastností Richardsovy křivky pro $b \rightarrow 1$ jsme si uvedli, že vhodnou volbou b získáme Gompertzovu křivku. V Richardsově modelu jsou však obsaženy i jiné známé modely.

- *Logistickou křivku* dostaneme volbou $b = 2$. Získáme tak model

$$(8) \quad L(x; a, c, d) = R(x; a, 2, c, d) = \frac{a}{1 + e^{-c(x-d)}} = a \frac{e^{c(x-d)}}{1 + e^{c(x-d)}}, \forall x \in \mathbb{R}.$$

- Připomeňme pro úplnost již zmíněný *Gompertzův model*, jenž je tvaru

$$(9) \quad G(x; a, c, d) = R(x; a, 1, c, d) = ae^{-\exp\{-c(x-d)\}}, \forall x \in \mathbb{R}.$$

- Volbou $b = 0$ získáme pro $x \geq d$ *monomolekulární model*

$$(10) \quad M(x; a, c, d) = R(x; a, 0, c, d) = a(1 - e^{-c(x-d)}).$$

- Naposledy zmiňme *von Bertalanffyho model*, který je pro $x \geq d - \log(3)/c$ definován rovnicí

$$(11) \quad B(x; a, c, d) = R(x; a, 2/3, c, d) = a\left(1 - \frac{1}{3}e^{-c(x-d)}\right)^3.$$

Jak jsme již uvedli, v našich počátečních úvahách jsme se zabývali jako možnými kandidáty pro modelování logistickou, Gompertzovou a Richardsovou křivkou. Nyní jsme ukázali, že Richardsova křivka v sobě zbývající dvě obsahuje, je tedy obecnější. Blíže o těchto modelech viz Seber (1988), či Ratkowski (1983).

2.3. Odhadování parametrů. Po důkladné analýze Richardsovy křivky můžeme přistoupit k vlastnímu provedení odhadu parametrů. Z dostupných údajů musíme nejprve stanovit počáteční odhad, který pak budeme iteračně vylepšovat (necháme to za nás udělat počítač). Teoreticky sice můžeme stanovit celkem optimální počáteční odhad, v praxi však pro naše data narazíme na některé zásadní překážky.

2.3.1. Parametrizace částečných součtů. Počáteční odhad a_0 parametru a stanovíme jako $a_0 = \sum_{i=1}^n Y_i = Z_n$, neboť jde o horní asymptotu Richardsovy křivky.

Další z parametrů, parametr d , je celkem snadno odhadnutelný. Jde o bod inflexe křivky, což je zde bod s největší derivací. Proto můžeme použít výšku X_i takovou, že $Y_i/(X_i - X_{i-1}) = \max$. Pokud bychom měli ekvidistantní sít bodů X_i (balón by stoupal rovnoměrně), stačilo by zřejmě hledat výšku X_i , v níž je naměřena maximální hodnota Y_i . Pro naše data však můžeme použít i výšku odpovídající modusu měření. Nerovnoměrnost stoupání balónu je zanedbatelná zejména s ohledem na rozptyl naměřených hodnot.

K odhadu parametru b_0 uijeme rovnosti $R(d) = ab^{1/(1-b)}$. Položíme-li $\sum_{i: X_i \leq d_0} Y_i = a_0 b_0^{1/(1-b_0)}$, pak tato rovnice má, vzhledem k monotonii funkce $b^{1/(1-b)}$, jednoznačné řešení. Nemůžeme jej však spočítat analyticky, je nutno tak učinit numericky.

Abychom mohli odhadnout parametr c , musíme se ještě jednou podívat na jeho úlohu v Richardsově křivce, neboť samotné zjištění, že Richardsova křivka je pro $a > 0$ rostoucí právě když $c > 0$, nepostačuje. Opět nám pomůže bod inflexe, neboť pro derivaci Richardsovy křivky platí

$$(12) \quad r(d; a, b, c, d) = acb^{b/(1-b)}.$$

Odtud již s pomocí počátečních odhadů a_0 , b_0 a d_0 můžeme stanovit odhad $c_0 = Y(X_{d_0})/[a_0 b_0^{b_0/(1-b_0)}]$.

Shrňme si tedy navržený počáteční vektor odhadů:

$$(13) \quad \begin{aligned} a_0 &= \sum_{i=1}^n Y_i = Z_n \\ d_0 &= \{X_i : Y_i = \max_j Y_j\} \\ b_0 &: \sum_{\{i: X_i \leq d_0\}} Y_i = a_0 b_0^{1/(1-b_0)} \\ c_0 &= \frac{Y(X_{d_0})}{(X_{d_0} - X_{d_0-1}) a_0 b_0^{b_0/(1-b_0)}}. \end{aligned}$$

Nyní si všimněme obtíží provázejících tento počáteční odhad. První, a zásadní, je „neukončenost“ pozorované Richardsovy křivky. Z grafů je zřejmé, že naměřená křivka nedosahuje horní asymptoty a navržený odhad a_0 se tak

stává nevhodným. Jelikož na něm velice podstatně závisejí odhady b_0 a c_0 , dostáváme se do obtíží, a navíc vzhledem k nesymetrii Richardsovy křivky nejsme ani schopni odhadnout ze známého průběhu neznámý zbytek. Další výhrady jsou již sice menšího, nikoliv však zanedbatelného, významu.

Naměřená křivka nezačíná v nule, ale vzhledem k přirozenému pozadí radioaktivity na povrchu země začíná od nějaké nezáporné hodnoty. Pokud bychom však chtěli tuto hodnotu odečíst, může se stát, že některá měření pak budou záporná, což je v rozporu s podstatou růstových modelů. Přesto jsme tak učinili, výsledný odhad však byl z hlediska celé křivky zanedbatelně odlišný, zlepšení lze zaznamenat především v přízemní části měření. Chceme-li lépe vyrovnat přízemní hodnoty, můžeme doporučit následující postup; od původních měření odečteme odhadnutou přízemní hodnotu (např. 0,2) a poté spočteme částečné součty a nalezneme odhad parametrů pro Richardsovu křivku. K takto odhadnuté křivce zpětně připočteme částečné součty konstantní posloupnosti 0,2 a tím získáme křivku aproximující skutečně naměřené hodnoty.

Odhad parametru d_0 může být vzhledem k velkému rozptylu naměřených hodnot též zavádějící, většinou ale padne do blízkosti později odhadnutého parametru d . Pro počáteční odhad musíme opět dbát na to, aby $x_1 \geq d_0 + \log(1 - b_0)/c_0$ jestliže $b_0 < 1$

2.3.2. *Parametrizace původních měření.* Hlavním úkolem od samého počátku bylo proložení vhodné křivky naměřenými hodnotami radioaktivity. Pomocí modelování částečných součtů jsme našli vhodného kandidáta v derivaci Richardsovy křivky, jejíž vlastnosti jsme rozebrali výše. Nyní máme za úkol stanovit odhad parametrů její derivace.

Všimněme si, že z rovnice $Y_i = Z_i - Z_{i-1}$ můžeme odvodit pomocí Taylova rozvoje vztah

$$(14) \quad Y_i = R(X_i) - R(X_{i-1}) \approx r(X_i)(X_i - X_{i-1}).$$

Pokud by platilo $X_i - X_{i-1} = \Delta$, pak z vlastností Richardsovy křivky a její derivace můžeme odvodit parametrizaci $a_r = \Delta a_R$, $b_r = b_R$, $c_r = c_R$ a $d_r = d_R$, kde index R znamená odhad parametru pro Richardsovu křivku a r pro její derivaci. Jelikož však tento předpoklad je porušen, musíme opět nalézt odhad pomocí iterací, kde počáteční odhad je nyní dán rovnicemi

$$(15) \quad \begin{aligned} a_0 &= \sum_{i=1}^n Y_i (X_i - X_{i-1}) \\ d_0 &= \{X_i : Y_i = \max_j Y_j\} \\ b_0 &: \sum_{\{i: x_i \leq d_0\}} Y_i = a_0 b_0^{1/(1-b_0)} \end{aligned}$$

$$c_0 = \frac{Y(X_{d_0})}{(X_{d_0} - X_{d_0-1})a_0b_0^{b_0/(1-b_0)}}.$$

Výhrady vyslovené pro počáteční odhad parametrů Richardsovy křivky jsou platné i zde.

2.3.3. Vlastní výpočet odhadu. Naměřená data načteme do formátu S-plus a stanovíme počáteční odhad parametrů, kde si odhadneme (pokusně) $a_0 = 4/3 \sum Y_i$. Poté odhadneme i další parametry uvedeným způsobem. Zkontrolujeme odhad tak, aby $x_1 \geq d_0 + \log(1 - b_0)/c_0$. Poté přistoupíme k odhadu pomocí vestavěné funkce `nls`, *non-linear least squares*. Jelikož náš odhad počátečních parametrů není možno provést přesně podle vzorců (13) a (15), je možné, že procedura `nls` nedokonzervuje, případně ukončí běh z jiných důvodů ($x_1 < d_n + \log(1 - b_n)/c_n$ v n -tém kroku) a my budeme nuceni počáteční odhad pozměnit.

Dobré výsledky jsme dosáhli též použitím programu NCSS, který jako jediný z nám dostupných programů má Richardsovu křivku přímo zabudovanou.

Pro zvýšení numerické stability odhadu je vhodné též zkusit normalizovat součet měření radioaktivity na hodnotu 1. Lepší by bylo normalizovat hodnotu a , ovšem vzhledem potížím spojeným s tímto parametrem, zejména jeho počátečním odhadem, je normalizace součtu měření vhodným kompromisem. Pro osu x postačí používat hodnoty v kilometrech, výška udávaná v metrech totiž dosahuje až desetitisícových hodnot a parametr c naopak pouze statisícin.

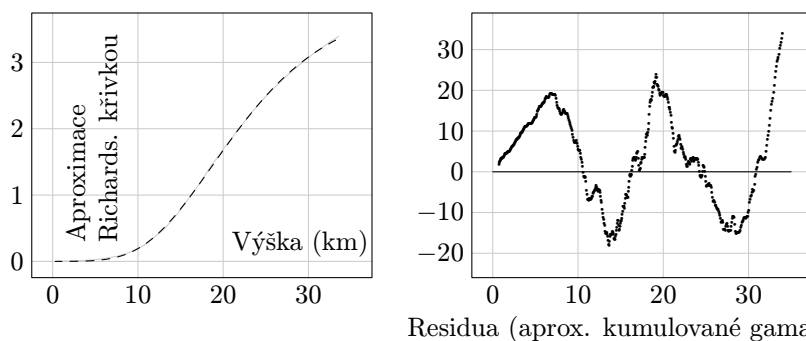
2.4. Výsledky. Pro dokreslení předchozího výkladu jsme použili vybrané pozorování (z dubna 1995). Na grafech v obrázcích 5–8 si můžeme dobře povšimnout některých zásadních a neodstranitelných nedostatků. Předně, v blízkosti země křivka podhodnocuje skutečné měření, což souvisí s jak již zmíněným přirozeným pozadím radioaktivity, tak faktem, že Richardsova křivka a její derivace začínají prakticky v nule. oznamenejme dále, že Richardsova křivka, jejíž derivace je hladká a jednovrcholová, není z principu dostatečně pružná na to, aby postihla lokální chování měřených údajů. Totéž samozřejmě platí i pro její derivaci. Tuto vlastnost můžeme dobře vidět především na residuích pro Richardsovu křivku.

Naopak nespornou výhodou je zde jasná interpretovatelnost parametrů, krásný hladký průběh obou parametrických křivek a v neposlední řadě i nízká hodnota residuů. Obzvláště u derivace Richardsovy křivky je porovnání residuů s residuemi neparametrických odhadů velmi lichotivé.

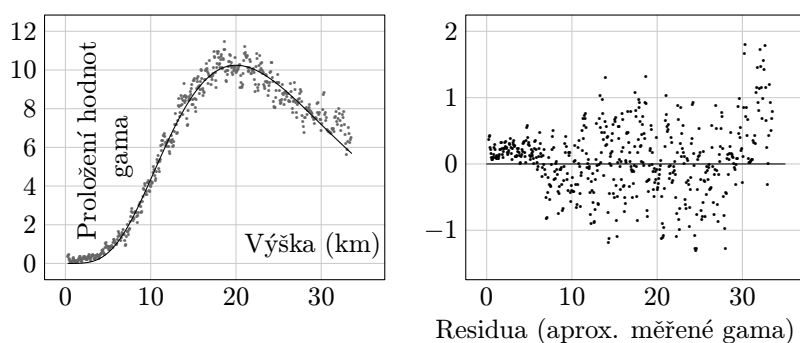
Tabulka konečných odhadů parametrů pro aproximaci měření počtu částic gama v závislosti na výšce je pro pozorování z 11. dubna 1995 následující:

Řada	Odhadnutá hodnota parametru			
	a	b	c	d
Původních měření	276,479	0,85494	0,09331	20,011
Částečných součtů	3914,56	1,00073	0,12658	18,7147

Všimněme si zejména rozdílů v odhadech parametrů b , c , a d , které by při rovnoměrné síti výšek pro jednotlivá měření měly být stejné.



Obr. 5 a 6: Aproximace Richardsovou křivkou je velmi slibná, při naměřených hodnotách až 3 000 částic nepřesahují residua hodnotu 30. Průběh residuí je dán pružností Richardsovy křivky



Obr. 7 a 8: Derivace Richardsovy křivky velmi dobře aproximuje střední část pozorování, oba kraje však jsou podhodnoceny.

Na grafu residuí pro původní měření si všimněme, že ke konci měření již odhad pomocí derivace Richardsovy křivky začíná naměřené údaje (relativně dost) podceňovat. Naše domněnka je, že se pokles intenzity radioaktivity zpomaluje, či dokonce zastavuje, což však již Richardsova křivka nemůže ze své podstaty postihnout. K naší lítosti se tak děje ve výšce nejvyššího dostupného balónu, takže si nemůžeme tuto hypotézu ověřit. Jinak by zřejmě

bylo nutné v těchto výškách model pozměnit z Richardsovy křivky na jiný a pokusit se o hladký přechod mezi těmito modely.

3. MODELOVÁNÍ POMOCÍ POISSONOVA PROCESU

3.1. Poissonův proces. Skutečnost, že předmětem našeho zájmu je intenzita radioaktivního záření, přímo vnucuje myšlenku využití Poissonova procesu, který je v oblasti radioaktivního záření velmi oblíbeným a přes svou jednoduchost velmi dobrým modelem.

Poissonův proces lze ve stručnosti popsat následovně. V daném čase máme danou intenzitu procesu $\lambda(t)$. Pravděpodobnost, že v malém časovém úseku $[t_1, t_2]$ dojde ke k impulsům je dána Poissonovým rozdělením

$$(16) \quad P(N_{[t_1, t_2]} = k) = \exp \left\{ - \int_{t_1}^{t_2} \lambda(t) dt \right\} \frac{\left(\int_{t_1}^{t_2} \lambda(t) dt \right)^k}{k!}, \quad k \in \{0, \dots, \infty\},$$

kde $\Lambda[t_1, t_2] = \int_{t_1}^{t_2} \lambda(t) dt$ je parametr tohoto Poissonova rozdělení a $N_{[t_1, t_2]}$ počet impulsů v intervalu $[t_1, t_2]$.

Převedeno na naše data, předpokláme-li na chvíli spojitý čas píšeme Z_t , získáváme

$$(17) \quad P(Z_t = k) = \exp \left\{ - \int_0^t \lambda(s) ds \right\} \frac{\left(\int_0^t \lambda(s) ds \right)^k}{k!}.$$

Naším úkolem je odhad intenzity $\lambda(t)$. Pro tento typ dat bohužel nelze předpokládat homogenost Poissonova procesu, takže nelze provést zjednodušující předpoklad $\lambda(t) = \lambda$ a je třeba přikročit k modelování intenzity $\lambda(t)$ nějakou funkcí.

Jelikož víme, že střední hodnota náhodné veličiny s Poissonovým rozdělením je rovna parametru tohoto rozdělení, lze vlastně pro odhad intenzity procesu použít přiblížení naměřených hodnot Y_t a Z_t odpovídajícími hodnotami intenzity, neboli modelovat

$$(18) \quad Z_t = \Lambda(t) = \int_0^t \lambda(s) ds$$

Jak si můžeme povšimnout, Richardsova křivka je jedním takovým modelem, ve kterém se používá pro parametrizaci intenzity $\lambda(t)$ derivace Richardsovy funkce. Jestliže tedy budeme spokojeni s proložením údajů Richardsovou křivkou, můžeme její derivaci použít pro (bodový) odhad intenzity Poissonova procesu a zpětně tuto intenzitu použít pro stanovení intervalu spolehlivosti pro libovolný časový bod. To je velice důležitý krok, neboť v předchozí části jsme takový interval spolehlivosti nemohli snadno ustanovit.

Předchozí odvození jsme učinili pro závislost intenzity Poissonova procesu na čase. Nás ovšem zajímá hlavně závislost na výšce. Naštěstí je zde jednoduché řešení v nahrazení proměnné čas proměnnou výška, což jsme pro Richardsovu křivku v předchozích oddílech již učinili.

3.2. Odhad maximální věrohodnosti. Parametry Poissonova procesu jsou obvykle odhadovány metodou maximální věrohodnosti. To ovšem znamená, že pro modelování intenzity derivací Richardsovy křivky dostaneme odhad jejích parametrů maximálně věrohodnou metodou.

Zaveďme si nyní značení $\Lambda_i(\boldsymbol{\theta}) = \int_{X_{i-1}}^{X_i} r(x; \boldsymbol{\theta}) dx = R(X_i; \boldsymbol{\theta}) - R(X_{i-1}; \boldsymbol{\theta})$, kde X_i značí výšku v i -tém časovém intervalu a $\boldsymbol{\theta} = (a, b, c, d)$. Pak maximálně věrohodný odhad parametrů a, b, c, d Richardsovy křivky lze nalézt jako libovolné řešení úlohy

$$(19) \quad \max_{\boldsymbol{\theta}} \prod_{i=1}^n e^{-\Lambda_i} \frac{\Lambda_i^{Y_i}}{Y_i!},$$

odkud po troše algebry dojdeme k výrazu

$$(20) \quad \max_{\boldsymbol{\theta}} \sum_{i=1}^n \left(Y_i \log(R(X_i; \boldsymbol{\theta}) - R(X_{i-1}; \boldsymbol{\theta})) - (R(X_i; \boldsymbol{\theta}) - R(X_{i-1}; \boldsymbol{\theta})) \right).$$

Zde je na místě připomenout, že naměřené údaje jsou údané jako *průměrný počet částic za vteřinu v úseku deseti vteřin*, což znamená, že nejde o celočíselné hodnoty. Pro Poissonův proces však lze použít pouze celočíselné údaje a proto je nutno použít hodnotu *počet částic za deset vteřin* namísto průměru (vynásobíme průměr deseti). Tím samozřejmě změníme i odhad parametru a Richardsovy křivky na desetinásobek, což je ovšem snadno napravitelná změna.

Poznámka: Pomocí programu Mathematica jsme se, žel bezúspěšně, pokoušeli nalézt řešení této maximalizační úlohy. Po mnoha hodinách počítání jsme však nedostávali stále žádný výsledek a proto jsme výpočet ukončili.

4. NEPARAMETRICKÉ VYHLAZOVÁNÍ

Jak jsme viděli výše, při pokusu o parametrizaci křivky prokládající naměřená data se stětváváme s některými zásadními praktickými problémy, zejména při výpočtu parametrů a při modelování lokálního chování sledovaného procesu. Tuto nesnáz nám pomůže překonat výpočetně „průhledný“ způsob tzv. neparametrického vyhlazování.

Myšlenka jádrového regresního vyhlazení vychází, podobně jako regresogram z histogramu, z jádrového odhadu hustoty, daného rovnicí

$$(21) \quad \hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i),$$

kde body X_i jsou měření (o nichž předpokládáme, že jsou *iid* s hustotou $f(x)$) a x bod ve kterém odhadujeme. Funkce K se nazývá jádro. Většinou jde o nějakou nezápornou funkci, symetrickou kolem nuly, $\int K(x) dx = 1$, a obvykle $\int xK(x) dx = 0$. Z tohoto důvodu se používá nejčastěji nějaká hustota. Popis a stručné porovnání vlastností nejčastěji používaných jader je v tabulce 1.

4.1. Od odhadu hustoty k odhadu regresní křivky. Vyjděme z rovnice

$$(22) \quad m(x) = E(Y|X = x) = \int yf(y|x) dy = \int y \frac{f(x, y)}{f_X(x)} dy.$$

Chceme-li najít jádrový odhad $\hat{m}(x)$ regresní funkce $m(x)$, použijeme tuto rovnici, kde hustoty f nahradíme jejich jádrovými odhady. Je tedy zřejmé, že musíme najít jádrový odhad $\hat{f}(x, y)$ sdružené hustoty a za odhad $f_X(x)$ použijeme (21).

Sdruženou hustotu můžeme odhadnout dvourozměrným jádrem způsobem popsaným např. v Härdle (1997)

$$(23) \quad \hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\det(\mathbf{H})} \mathbb{K} \left[\mathbf{H}^{-1} \begin{pmatrix} x - X_i \\ y - Y_i \end{pmatrix} \right].$$

Zde matice \mathbf{H} reprezentuje kovarianční strukturu rozdělení vektoru (X, Y) .

V praxi se však zpravidla používá diagonální matice s prvky h_x a h_y na diagonále. Takto dojdeme k výrazu

$$(24) \quad \hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_x h_y} \mathbb{K} \left(\frac{x - X_i}{h_x}, \frac{y - Y_i}{h_y} \right).$$

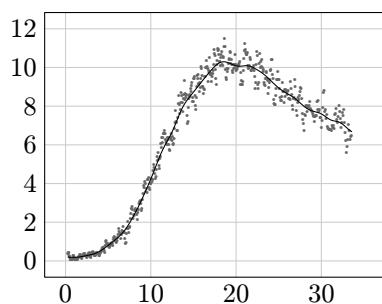
Dvourozměrné jádro se nakonec nahrazuje součinem dvou jednorozměrných jader, $\mathbb{K} = K^X \cdot K^Y$, čímž dochází k dalšímu podstatnému zjednodušení

$$(25) \quad \hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_x h_y} K^X \left(\frac{x - X_i}{h_x} \right) K^Y \left(\frac{y - Y_i}{h_y} \right).$$

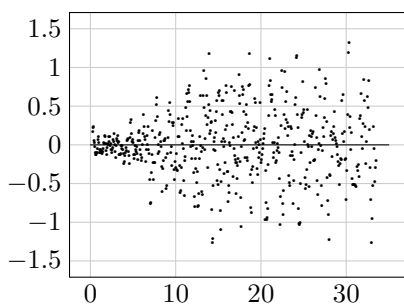
Toto zjednodušení se může zdát neopodstatněné v případě regrese, jak však uvidíme, vede k žádoucímu výsledku. Na druhou stranu se mnohdy v regresi hodnoty X za náhodné nepovažují, a v tom případě není zcela korektní zápis $(Y|X)$.

Vlastní odvození je poměrně snadné. Pro jednoduchost položíme $h_x = h_y$, což však není těžké zobecnit. Potom

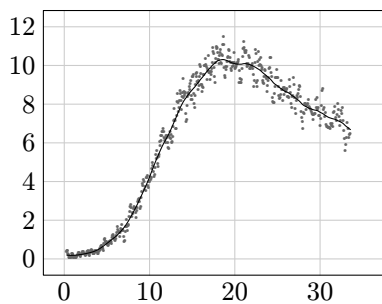
$$\begin{aligned}
 \hat{m}(x) &= \int y \frac{\hat{f}(x, y)}{\hat{f}_X(x)} dy = \int y \frac{\sum_{i=1}^n K_h(y - Y_i) K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)} dy \\
 &= \frac{\sum_{i=1}^n \int y K_h(y - Y_i) K_h(x - X_i) dy}{\sum_{i=1}^n K_h(x - X_i)} \\
 (26) \quad &= \frac{\sum_{i=1}^n K_h(x - X_i) \int (zh + Y_i) K(z) dz}{\sum_{i=1}^n K_h(x - X_i)} \\
 &= \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)} = \frac{\frac{1}{nh} \sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)}.
 \end{aligned}$$



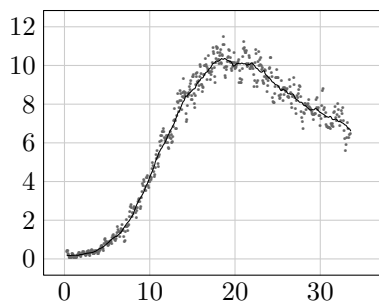
Jádrové vyhlazení
norm. jádro $h = 20$



Residua (norm. jádro $h = 20$)



Jádrové vyhlazení
trojúh. jádro $h = 20$



Jádrové vyhlazení
rovnom. jádro $h = 20$

Obr. 9 – 12: Normální a trojúhelníkové jádro dávají takřka nerozlišitelné odhady. Pro rovnoměrné jádro při stejné šířce okna dostáváme méně hladký odhad.

Z výsledného zápisu můžeme vidět, že jde o vážený průměr naměřených hodnot, kde váhy jsou dány hodnotou $w_i(x) = K_h(x - X_i) / \sum K_h(x - X_i)$.

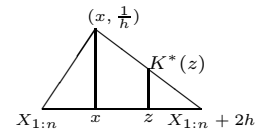
Jádro	$K(z)$	Základní vlastnost
Rovnoměrné	$\frac{1}{2}\mathbf{I}(z \leq 1)$	Jednoduché, nespojitý odhad
Trojúhelníkové	$(1 - z)\mathbf{I}(z \leq 1)$	Jednoduché, spojitý odhad
Epanečnikovovo	$\frac{3}{4}(1 - z^2)\mathbf{I}(z \leq 1)$	Parabola, spojitý odhad
Normální	$\frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$	Spojitý odhad, obor celé \mathbb{R}_1

TABULKA 1. Doporučovaná symetrická jádra

4.2. Hraniční efekt. Z definice jádrového odhadu je zřejmé, že v blízkosti okrajů definičního oboru pro X bude značná část hmoty jádra zasahovat do oblastí bez pozorování. Takové snížení počtu pozorování zpravidla má značný vliv na chování a kvalitu odhadu na těchto oblastech, především na vychýlení. Pokusíme se proto navrhnout postup, který by hraniční efekt částečně odstranil.

Naše idea blízkosti krajních bodů je měnit postupně jádro tak, aby pro bod $X_{1:n}$ mělo definiční obor zleva tímto bodem omezený (obdobně pro bod $X_{n:n}$). Pro trojúhelníkové jádro budeme pozměňovat tvar trojúhelníku, zatímco normální jádro na krajích zkombinujeme s hustotou exponenciálního rozdělení. U rovnoměrného jádra se v blízkosti krajních bodů obvykle odhad ponechá konstantní, stejně jako u odhadu pomocí k nejbližších sousedů.

4.2.1. Trojúhelníkové jádro. Pokud pro trojúhelníkové jádro bude pro bod x , pro který hledáme odhad, platit $(x - X_{1:n})/h \leq 1$, „upevníme“ levý krajní bod základny trojúhelníku do bodu $X_{1:n}$ při současném zachování základny o délce $2h$ a vrcholu v bodě, pro který hledáme odhad. Ukažme si přesný tvar takto modifikovaného jádra pro bod $x \in [X_{1:n}, X_{1:n} + h]$:

$$(27) \quad K^*(z) = \begin{cases} \frac{[1-z/(2h+X_{1:n}-x)]^+}{h} & \text{pro } z \geq 0 \\ \frac{[1-z/(X_{1:n}-x)]^+}{h} & \text{pro } z \leq 0 \end{cases}$$


4.2.2. Normální jádro. Jak jsme již uvedli výše, zde je myšlenkou postupně nahrazovat u kraje normální jádro směsí normálního a exponenciálního jádra tak, aby pro odhad v bodě $X_{1:n}$ bylo použito čistě exponenciální jádro. Zatímco normální jádro je určeno tak, že jeho střední hodnota je vždy v bodě, pro který hledáme odhad, exponenciální jádro posuneme tak, že počátek má v bodě $X_{1:n}$. Zvolíme hodnotu δ a pro pevný bod $x \in [X_{1:n}, X_{1:n} + \delta]$ použijeme jádro tvaru:

$$(28) \quad K(z) = \frac{x - X_{1:n}}{\delta} \cdot \frac{1}{h} \phi\left(\frac{z}{h}\right) + \frac{\delta - x + X_{1:n}}{\delta} \cdot \frac{1}{h} e^{-(x - X_{1:n} - z)/h}.$$

Pokud se nám zdá, že takto dáváme příliš velkou váhu bodu $X_{1:n}$, můžeme též místo exponenciálního jádra použít „deformovaný trojúhelník“ z předchozího postupu.

Zůstává stanovit volbu hodnoty δ . Vhodným odhadem může být taková vzdálenost od bodu $X_{1:n}$, při níž předem daná hmota jádra (například 2,5%) padne mimo interval pozorování. Tedy

$$(29) \quad \frac{\delta}{h} = u(0,975) \approx 1,96 \implies \delta \approx 2h.$$

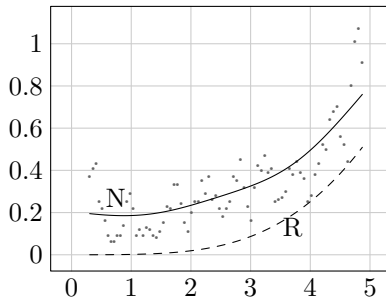
Ještě jednou možností pro normální jádro je jeho přenásobení konstantou tak, aby

$$(30) \quad C \int_{x-X_{1:n}}^{X_{n:n}-x} \frac{1}{h} \phi\left(\frac{z}{h}\right) dz = 1,$$

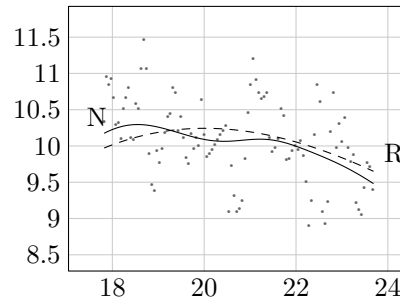
čehož snadno dosáhneme volbou $C = \Phi((X_{n:n} - x)/h) - \Phi((x - X_{1:n})/h)$. Tím získáváme pro zvolený bod x jádro, jež má uvnitř intervalu $[X_{1:n}, X_{n:n}]$ hmotu jedna.

Celý nápad s řešením krajového efektu pochází z proložení pomocí k nejbližších sousedů. Můžeme říci, že při „skoro rovnoměrném“ rozdělení bodů X dokáže skloubit výhody obou postupů. Uvnitř intervalu s dostatečným počtem měření je počet bodů zahrnutých do odhadu přibližně stejný pro všechny body, zatímco na okrajích vhodným pozměněním jádra tohoto stavu dosáhneme též. Nevýhodou vyhlazení pomocí k nejbližších sousedů je totiž využití stejných vah $1/k$ pro všechny body mezi k nejbližšími, a tedy z toho plynoucí nespojitost, na krajích pak zůstává odhad konstantní.

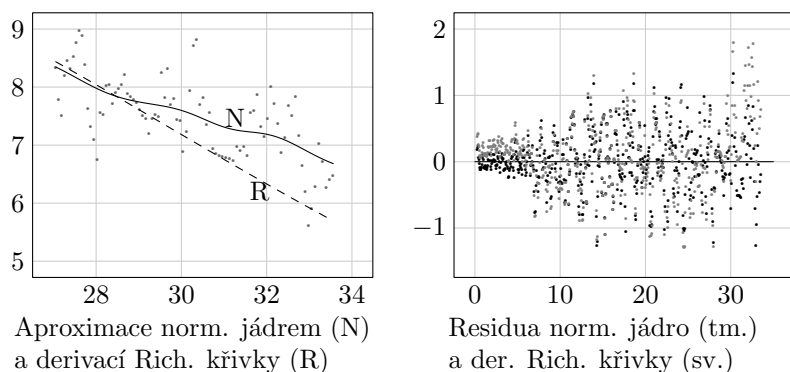
Pro normální jádro jsou sice teoreticky do okna zahrnuty všechny body, ale jejich váhy rychle klesají prakticky na zanedbatelnou hodnotu se zvětšující se vzdáleností od bodu v němž odhadujeme. To je důvod k použití exponenciálního rozdělení, které má těžší chvost.



Aproximace norm. jádrem (N)
a derivací Rich. křivky (R)



Aproximace norm. jádrem (N)
a derivací Rich. křivky (R)



Obr. 13 – 16: Na krajích měřeného intervalu je jádrový odhad výrazně lepší než aproximace derivací Richardsovy křivky. Fyzikální interpretace jádrového odhadu je však výrazně složitější.

4.3. Heteroskedasticita. Nestejné rozptyly jednotlivých měření sice nejsou pro vlastní jádrové vyhlazení nebezpečné, ale ovlivňují rozptyl odhadu $\hat{m}(x)$. Při splnění některých podmínek však můžeme dopady nestejnosti rozptylů zmírnit.

Předpokládejme, že jsme schopni rozdělit měření na intervaly, kde měření mají shodný rozptyl, řekněme σ_i^2 pro interval I_i , $i = 1, \dots, K$. Naší snahou je nalézt pro tento interval takovou hodnotu h_i , abychom pro všechny tyto intervaly získali shodnou hodnotu střední čtvercové chyby, a tuto konstantu chceme mít pochopitelně co nejmenší. Jde tedy o modifikaci hledání hodnoty h metodou minimalizace střední čtvercové chyby.

Dosáhnout tohoto stavu samozřejmě není snadné (pokud je to vůbec možné), proto chceme stanovit jednodušší postup. Využijeme předpokladu o homoskedastických intervalech. Požadujeme, aby

$$(31) \quad \frac{\sigma_i^2}{|I_i|} \int_{I_i} \frac{\sum_{i=1}^n K_{h_i}^2(x - X_i)}{(\sum_{i=1}^n K_{h_i}(x - X_i))^2} dx = const.$$

Takto můžeme pro všechny intervaly získat optimální hodnoty h_i pro všechna $i = 1, \dots, K$. Pro hodnoty x na hranicích dvou intervalů použijeme interpolovanou hodnotu h . Hranice je zde stanovena podobně, jako v oddíle o hraničním efektu. při použití trojúhelníkového jádra tedy začneme používat interpolaci

$$(32) \quad h_{int} = h_{i+1} + \frac{b_{i+1} - x}{b_{i+1} - b_i} (h_i - h_{i+1}),$$

kde body b_i jsou voleny tak, že jejich vzdálenost od konce intervalu I_i je právě h_i . Pro normální jádro použijeme opět přibližného kvantilu a bod b_i stanovíme ve vzdálenosti $2h_i$ od okraje intervalu.

Je zřejmé, že nás tento postup může dovést na scestí. Jestliže například odhadnuté h_1 povede k takovému střednímu rozptylu, že i při zahrnutí všech pozorování druhého intervalu (tj. když $2h_2 > |I_2|$, případně $4h_2 > |I_2|$), musíme upravit (zmenšit) následně volbu h_1 . Celý algoritmus pak připomíná volbu h pomocí minimální čtvercové chyby, musíme však dát pozor na simultánnost minimalizace – sledovat více intervalů v každém kroku této optimalizace.

5. LOKÁLNÍ POLYNOMICKÁ REGRESE

Jednou z dalších možností vyhlazení křivky je použití takzvané *lokální polynomické regrese*. Omezíme se na jednu vysvětlující proměnnou, stejně jako v předchozím výkladu využijeme jako regresor výšku. Lokální polynomická regrese se, podobně jako jádrové vyhlazování, provádí v jednotlivých bodech. Myšlenka je velmi jednoduchá, neboť jde o přímé použití Taylorova rozvoje v daném bodě, respektive o prvních k členů tohoto rozvoje.

Vezměme bod x pro který chceme nalézt odhad. Vytvoříme si regresní matici

$$(33) \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 - x & (X_1 - x)^2 & \dots & (X_1 - x)^k \\ 1 & X_2 - x & (X_2 - x)^2 & \dots & (X_2 - x)^k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_n - x & (X_n - x)^2 & \dots & (X_n - x)^k \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix},$$

kde k je řád polynomu, kterým chceme prokládat a (X_1, \dots, X_n) je vektor hodnot nezávisle proměnné. Regresní model je zřejmý, jde o

$$(34) \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, i = 1, \dots, n).$$

5.1. Odhad parametrů a predikce. Přikročme k odhadu vektoru $\boldsymbol{\beta}$. Provádí se minimalizací součtu čtverců *vážených* residuí, tj. minimalizuje se

$$(35) \quad \frac{\sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}.$$

Pro účely názorného zápisu nyní předpokládejme, že máme zvolenu kvadratickou regresi. Pak v okolí bodu x_0 používáme model

$$(36) \quad \hat{l}(x) = \hat{\beta}_0 + \hat{\beta}_1(x - x_0) + \hat{\beta}_2(x - x_0)^2,$$

takže $\hat{l}(x_0) = \hat{\beta}_0$ je pak odhad vlastní hodnoty v bodě x . Tento postup provedeme v „každém“ bodě a dostaneme tím proložení křivky metodou lokálních polynomů. Zvolený postup je výhodný nejen při požadavku odhadování neznámé křivky, ale též její derivace, jak uvidíme dále.

Tento postup je vlastně semiparametrický, neboť v daném bodě použijeme parametrické (polynomické) vyjádření regresní závislosti. V porovnání s jádrovými odhady jsme dospěli k názoru, že lokálně polynomické vyhlazení citlivěji reaguje na změny hodnoty h . Srovnatelnou „hladkost“ odhadu

s jádrovým vyhlazením (normálním s $h_J = 20$) jsme dosáhli při nižší hodnotě (zde $h_{LP} = 10$). Na rozdíl od jádrových odhadů zde způsobuje potíže heteroskedasticita našich dat. Možné řešení uvádí Holst (1996).

6. ODHADY DERIVACE

Neparametrický odhad regresní křivky je někdy vhodné doplnit odhadem derivace/derivací „skutečné“ křivky. Zatímco v parametrických modelech můžeme jednoduše derivovat odhadnutou funkci, pro jádrové odhady je nezbytné spočítat nový odhad. Zde si ukážeme tři možné postupy.

6.1. Derivace pomocí rozvoju. Hlavním nástrojem pro odvození nám bude rozvoj $f(x + \delta) = f(x) + \delta f'(x) + \delta^2 f''(x)/2! + \dots$, takže nám opět pomůže odhad derivace hustoty.

$$\begin{aligned}
 \hat{f}'(x) &\approx \frac{\hat{f}(x + \delta) - \hat{f}(x)}{\delta} = \frac{1}{n\delta} \left(\sum_{i=1}^n K_h(x + \delta - X_i) - \sum_{i=1}^n K_h(x - X_i) \right) \\
 (37) \quad &= \frac{1}{n} \sum_{i=1}^n \frac{K_h(x + \delta - X_i) - K_h(x - X_i)}{\delta} \\
 &\approx \frac{1}{n} \sum_{i=1}^n \frac{K' \left(\frac{x + \delta - X_i}{h} \right) - K \left(\frac{x - X_i}{h} \right)}{h\delta} \approx \frac{1}{nh} \sum_{i=1}^n K'_h(x - X_i).
 \end{aligned}$$

Tento výraz budeme nyní často používat. Pusťme se do odvození odhadu derivace $m'(x) = \lim_{\delta \rightarrow 0} (m(x + \delta) - m(x))/\delta$.

$$\begin{aligned}
 m'(x) &\approx \frac{m(x + \delta) - m(x)}{\delta} = \int \left[y \frac{f(x + \delta, y)}{f(x + \delta)} - y \frac{f(x, y)}{f(x)} \right] \frac{1}{\delta} dy \\
 (38) \quad &= \int [yf(x)f(x + \delta, y) - yf(x + \delta)f(x, y)] \frac{1}{\delta f(x + \delta)f(x)} dy \\
 &\approx \int \frac{[yf(x)(f(x + \delta, y) - f(x, y)) - y\delta f'(x)f(x, y)]}{\delta f(x + \delta)f(x)} dy \\
 (39) \quad &\rightarrow \int y \frac{f'(x, y)}{f(x)} - y \frac{f(x, y)}{f(x)} \cdot \frac{f'(x)}{f(x)} dy \\
 &= \int y \frac{f'(x, y)}{f(x)} dy - m(x) \log' f(x),
 \end{aligned}$$

kde všechny derivace jsou myšleny podle x a konvergence platí při $\delta \rightarrow 0$. Využijeme-li součinnové jádro, dostaneme hledaný odhad tvaru

$$(40) \quad \widehat{m}'(x) = \frac{1}{h} \cdot \frac{\sum_{i=1}^n Y_i K'_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)} - \widehat{m}(x) \frac{\hat{f}'(x)}{\hat{f}(x)}.$$

6.2. Proložení pomocí směrnic. Druhou možností odhadu derivace je jádrové vyhlazení naměřených směrnic $T_i = (Y_i - Y_{i-1})/(X_i - X_{i-1})$. Z naměřených údajů si spočítáme směrnice úseček spojujících dvě po sobě následující měření, derivaci (směrnici tečny) pak odhadneme jádrovým vyhlazením těchto hodnot. Dostáváme tak postupně

(41)

$$\begin{aligned} \widehat{m}'(x) &= \frac{\sum_{i=2}^n \frac{Y_i - Y_{i-1}}{X_i - X_{i-1}} K_h(x - X_i)}{\sum_{i=2}^n K_h(x - X_i)} \\ &= \frac{\sum_{i=2}^n \frac{Y_i}{X_i - X_{i-1}} K_h(x - X_i) - \frac{Y_{i-1}}{X_i - X_{i-1}} K_h(x - X_{i-1} + X_{i-1} - X_i)}{\sum_{i=2}^n K_h(x - X_i)} \\ &\approx \frac{\sum_{i=2}^n \frac{Y_i}{X_i - X_{i-1}} K_h(x - X_i)}{\sum_{i=2}^n K_h(x - X_i)} - \\ &\quad - \frac{\sum_{i=2}^n \frac{Y_{i-1}}{X_i - X_{i-1}} \left[K_h(x - X_{i-1}) + \frac{X_{i-1} - X_i}{h} K_h'(x - X_{i-1}) \right]}{\sum_{i=2}^n K_h(x - X_i)} \\ &= \frac{\sum_{i=2}^n \frac{Y_i}{X_i - X_{i-1}} K_h(x - X_i) - \frac{Y_{i-1}}{X_i - X_{i-1}} K_h(x - X_{i-1})}{\sum_{i=2}^n K_h(x - X_i)} + \\ &\quad + \frac{1}{h} \cdot \frac{\sum_{i=2}^n Y_{i-1} K_h'(x - X_{i-1})}{\sum_{i=2}^n K_h(x - X_i)} \end{aligned}$$

Všimněme si, že pokud platí $X_i - X_{i-1} = \Delta$, pak se (41) zjednoduší na

$$(42) \quad \frac{\frac{Y_n}{\Delta} K_h(x - X_n) - \frac{Y_1}{\Delta} K_h(x - X_1)}{\sum_{i=1}^{n-1} K_h(x - X_i)} + \frac{\sum_{i=1}^{n-1} Y_i K_h'(x - X_i)}{h \sum_{i=1}^{n-1} K_h(x - X_i)},$$

takže dostáváme přibližně

$$(43) \quad \widehat{m}'(x) = \frac{\sum_{i=1}^{n-1} Y_i K_h'(x - X_i)}{h \sum_{i=1}^{n-1} K_h(x - X_i)}.$$

Tento výraz je až na člen $\widehat{m}(x)\widehat{f}'(x)/\widehat{f}(x)$ shodný s odhadem odvozeným v předešlém odstavci.

6.3. Odhad parametry lokální polynomicke regrese. Připomeňme tvar lokálně polynomicke regrese:

$$(44) \quad l(x) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \dots,$$

a zároveň Taylorův rozvoj regresní funkce:

$$(45) \quad l(x) = l(x_0) + (x - x_0)l'(x_0) + \frac{(x - x_0)^2}{2!}l''(x_0) + \dots$$

Již víme, že odhad $\hat{\beta}_0 = \hat{l}(x_0)$. Z předchozích rovnic pak můžeme odvodit odhady $\hat{\beta}_1 = \hat{l}'(x_0)$ a $2!\hat{\beta}_2 = \hat{l}''(x_0)$ atd., takže vhodnou volbou řádu prokládaného polynomu lze stanovit odhady i pro vyšší derivace.

6.4. Obecná poznámka o odhadech derivací. K odhadu vyšších řádů derivací lze využít i prvních dvou postupů. Zde pak pro odhad n -té derivace dostáváme první členy ve tvaru $\sum Y_i K_h^{(n)} / h^n$.

Ze všech tří navržených postupů je zřejmé, ačkoliv možná ne na první pohled, že chceme-li odhadovat derivace vyšších řádů dostatečně hladce, musíme zvětšovat šířku okna h . Jelikož však máme pouze omezený počet dat na omezeném intervalu, nelze doporučit odhadování více než druhé derivace.

LITERATURA

- [1] Antoch J., Odhady hustoty, *ROBUST'82*, JČSMF Praha, 1982, 1–9.
- [2] Antoch J., Neparаметrické odhady regresních křivek, *ROBUST'86*, JČSMF Praha, 1986, 1–21.
- [3] Antoch J., Neparаметrické odhady hustot a regresních křivek, *Acta Oeconomica Pragensia* 13 (Robust 1980–1990), 1991.
- [4] Boularan J. et. al., Growth curves: a two stage nonparametric approach, *Journal of Statistical Planning and Inference* 38, 1994, 327–350.
- [5] Budíková M., Horová I., Různé přístupy k jádrovému odhadu regresní funkce, *Robust '94*, 13–23.
- [6] Härdle W. et. al., *A Course on Non- and Semiparametric Modelling*, 1997, Humbolt-Universität zu Berlin.
- [7] Holst U. et. al., Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurement, *Environmetrics*, vol. 7, 1996, 401–416.
- [8] Michálek J., Jádrové odhady: Základní vlastnosti a optimální výběr parametrů odhadů, *Robust '94*, 102–113.
- [9] Ratkowski D.A., *Nonlinear Regression Modeling*, 1983, M. Dekker, New York.
- [10] Seber G.A.F., Wild C.J., *Nonlinear Regression*, 1988, J. Wiley, New York.
- [11] Veselý V., Jádrové odhady – implementace, ilustrativní příklady, aplikace, *Robust '94*, 160–171.