

Metoda nejmenších čtverců I

Zdeněk Mikulášek, Ústav teoretické fyziky a astrofyziky

1 Úvodem

Objekty s proměnnými charakteristikami jsou předmětem soustředěného zájmu astrofyziků, protože svou proměnností toho o sobě prozrazují více, než objekty neproměnné. Zjištění a matematické vyjádření povahy časové proměnnosti měřených veličin (jasnost, magnetické pole, intenzita spektrálních čar, polarizace apod.), hledání trendů, cyklických změn, periodicit apod. - to jsou nejčastější úkoly, které praktická astrofyzika řeší. Nejdůležitějším nástrojem pro matematické zpracování těchto závislostí je tzv. *regresní analýza* a zejména její nejstarší a nejpropracovanější disciplína – *metoda nejmenších čtverců* (MNČ, anglicky *least square method* - LSM).

Dříve než přistoupíte ke zpracování pomocí regresní analýzy, doporučuji abyste si celou situaci nejprve zevrubně obhlédli, což mj. znamená, že si do nejrůznějších grafů či schémat vynesete vzájemné závislosti všech možných veličin dotyčného objektu, ať už vámi naměřených nebo převzatých z literatury. Věřte, že tyto „obrázky“ vám o povaze vzájemných souvislostí mezi jednotlivými charakteristikami povědí více než sebedokonalejší číselné rozbory. Zjistíte-li, že zobrazené výsledky měření $\{y_i\}$ jeví jistou časovou závislost, zřejmě též pocítíte neodolatelné nutkání tuto závislost proložit (fit) nějakou elegantní hladkou křivkou. Proč? Nejspíš proto, abyste viděli, jak se daná veličina doopravdy mění, tedy jak by to asi vypadalo, pokud byste dotyčnou veličinu dokázali měřit nepřetržitě a přitom navíc absolutně přesně. K tomuto ideálu samozřejmě nedospějete nikdy, lze se mu však alespoň přiblížit. Metoda nejmenších čtverců přitom naznačuje osvědčenou cestu, jak toho dosáhnout.

Doporučuji vám, abyste ale předem zvážili, zda je vůbec třeba něco prokládat a počítat! Chceme-li totiž jen dokumentovat, že tu ona závislost existuje, tak je poctivější do grafu žádnou křivku nevkreslovat, stačí jen zvolit vhodná měřítka na osách a obrázků prezentovat v jeho originální podobě. Pouze tehdy, chceme-li s výsledky proložením dále pracovat a něco z nich vyvozovat, je záhodno pustit se do matematického zpracování.

1.1. Regresní model

Vyšetřujme nejprve časovou závislost vybrané měřené veličiny y na základě *časové řady*, což je soubor n trojic $\{t_i, y_i, \sigma_i\}$. Předpokládejme přitom, že čas měření t známe naprosto přesně, lze jej tedy pokládat za *nezávislou veličinu*, zatímco jednotlivá měření *závisle proměnné veličiny* y, y_i , jsou zatížena určitou nejistotou, řekněme σ_i .

Naším záměrem nyní bude najít takovou skalární funkci času t , $f(t)$, která optimálně prochází mezi mezi naměřenými body a co nejlépe vystihuje reálnou časovou závislost pozorované veličiny.

Triviálním řešením této úlohy v případě časové závislosti je pospojování všech po časově sobě následujících bodů lomenou čarou $\{t_i, y_i\}$, případně nějakou sice hladkou, ale dostatečně zvlněnou čarou (např. polynomem stupně $n-1$), která by procházela důsledně všemi naměřenými

body¹. Takovýto postup by měl své opodstatnění pouze tehdy, pokud bychom jak čas, tak závisle proměnnou veličinu znali absolutně přesně, což je nereálné. Mnohem hodnověrnější výsledky dává prostá grafická metoda, kdy mezi body vyneseny do grafu táhneme od ruky hladkou křivku, která dle našeho přesvědčení co nejlépe vyjadřuje pozorovanou závislost. Tento způsob proložení však není obecně reprodukovatelný (i vy sami nakreslíte tu svou optimální křivku pokaždé trochu jinak), navíc se s tímto grafickým řešením potom dosti špatně pracuje.

Běžně se proto dává přednost takovým metodám, které vedou k analytickému vyjádření prokládané funkce a k objektivnímu, reprodukovatelnému stanovení kritéria nejlepší shody. Obvykle si hned na počátku definujeme tzv. *regresní model* (regression model). Regresním modelem si z nekonečného množství funkcí, jimiž by bylo možno pozorovanou závislost proložit, vybereme jen jistou omezenou množinu funkcí, přičemž každá z funkcí této zvolené množiny modelových funkcí bude plně definována g předem neznámými volnými parametry, které si pracovně označíme $\beta_1, \beta_2, \beta_3, \dots, \beta_g$. Veličina g pak vyjadřuje *počet stupňů volnosti* (degree of freedom) zvoleného modelu. Na tom, jak si dokážeme zvolit ten správný regresní model, který v sobě obsahuje funkce co nejpodobnější reálné závislosti $y(t)$ a použít přitom co nejmenší počet volných parametrů, pak závisí úspěch celého našeho dalšího počínání.

Pokud nevíme o fyzikální podstatě závislosti jedné z pozorovaných veličin na druhé vůbec nic, pak jako regresní model volíme soubor co nejjednodušších funkcí - polynomy, harmonické funkce - s nimiž lze snadno pracovat. Pokud však již předem víme, jakou modelovou funkcí by měla být pozorovaná závislost popsána, měli bychom jí dát přednost, protože jinak si způsobilíme zbytečné problémy při interpretaci zjištěné závislosti. Správnou a citlivou volbou regresního modelu lze ze souboru dat vytěžit spoustu informací, naopak zvolením neadekvátního modelu, lze snadno dospět i ke zcela mylným a falešným vývodům.

Regresní model představuje množinu podobných funkcí, které se od sebe liší jen jinými hodnotami volných parametrů $\beta_1, \beta_2, \dots, \beta_g$: $f(t) = f(\beta_1, \beta_2, \dots, \beta_g, t)$. Uspořádanou g -ticí parametrů β_j je výhodné zapisovat jako g -rozměrný vektor nebo sloupcovou matici $\boldsymbol{\beta}$ o rozměrech $g \times 1$ (g řádků a 1 sloupec): $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_g)^T$.

Předpokládejme nyní, že jsme v rámci regresního modelu zvolili nějakou konkrétní hodnotu vektoru parametrů pro i -té měření $\{t_i, y_i\}$ pak lze vyjádřit odchylku e_i tohoto měření od dané závislosti vztahem

$$e_i = y_i - f(t_i, \boldsymbol{\beta}). \quad (1)$$

Je zjevné, že čím menší budou odchylky měření od modelové předpovědi, tím lepší bude proložení.

Je však třeba navíc uvážit, že jednotlivá měření mají různou kvalitu, či chcete-li váhu, která bude nějak souviset s nejistotou jejich určení σ_i . Je užitečné zavést si tzv. *modifikovanou odchylku* \tilde{e}_i , kde $\tilde{e}_i = e_i/\sigma_i$, a tu pak brát jako rozhodující při posuzování úspěšnosti modelování nějakých pozorovaných závislostí, tedy:

$$\tilde{e}_i = \frac{e_i}{\sigma_i} = \frac{y_i - f(t_i, \boldsymbol{\beta})}{\sigma_i}. \quad (2)$$

Naším úkolem nyní bude vybrat z množiny funkcí, které připouští zvolený regresní model, $f(t, \boldsymbol{\beta})$ popsaných vektorem $\boldsymbol{\beta}$, takový vektor $\boldsymbol{\beta} = \mathbf{b}$, pro nějž budou modifikované

¹Tímto polynomem stupně $n - 1$ může být třeba Lagrangeův nebo Newtonův interpolační polynom.

odchylky $\{\tilde{e}_i\}$ minimální. Onu podmínku minimálnosti je ovšem třeba nejprve matematicky precizovat. Nejčastěji používanou, a z mnoha důvodů nejoblíbenější (nikoli však jedinou²), je podmínka, aby součet čtverců modifikovaných odchylek pro všechna měření, označovaný běžně jako veličina χ^2 , tedy

$$\chi^2 = \sum_{i=1}^n \tilde{e}_i^2 = \sum_{i=1}^n \left(\frac{e_i}{\sigma_i} \right)^2 \quad (3)$$

byl minimální. Z této podmínky pak vychází moderní varianta, jinak již letité metody nejmenších čtverců, které se budeme nadále věnovat.

Metoda nejmenších čtverců je nástroj, pomocí něhož lze poměrně jednoduše stanovit hodnoty parametrů zvoleného regresního modelu tak, aby tento model co nejlépe souhlasil s tím, co jsme napozorovali. Pokud jsme měli šťastnou ruku při výběru modelu, budeme moci i předpovědět, jak se zkoumaný objekt choval, a to i v době, když jsme jej neměli pod dohledem. Budeme moci předpovědět, co by se s ním mělo dít v budoucnosti. Všechny tyto předpovědi známe i jistou dávkou nepřesnosti, která je dána jednak tím, že zvolený model nemusí úplně přesně odpovídat realitě, ale zejména proto, že všechna pozorovací data jsou zatížena jistou nepřesností danou způsobem měření a řadou neznámých faktorů, které výsledky pozorování ovlivňují. Velkou předností MNC je, že umožňuje nejen předpovídat, ale i odhadnout nejistotu těchto předpovědi

2 Metoda nejmenších čtverců

2.1. Hledání řešení metodou nejmenších čtverců

Suma $\chi^2(\boldsymbol{\beta})$ je bezrozměrná skalární funkce vektoru parametrů $\boldsymbol{\beta}$:

$$\chi^2(\boldsymbol{\beta}) = \sum_{i=1}^n \left[\frac{y_i - f(t_i, \boldsymbol{\beta})}{\sigma_i} \right]^2 = \sum_{i=1}^n \frac{e_i^2}{\sigma_i^2} = \sum_{i=1}^n e_i^2 w_i^2 = \sum_{i=1}^n [y_i - f(t_i, \boldsymbol{\beta})]^2 w_i, \quad (4)$$

jež je úměrná záporně vzatému logaritmu pravděpodobnosti daného řešení. Místo individuálních nejistot σ_i lze z výpočetních důvodů použít i *individuální váhy*³ dané vztahem: $w_i = \sigma_i^{-2}$.

Hledejme nyní takový vektor $\boldsymbol{\beta}$, ($\boldsymbol{\beta} = \mathbf{b}$) pro nějž je tato suma $\chi^2 = \chi^2(\boldsymbol{\beta} = \mathbf{b})$ minimální. Funkci $\chi^2(\boldsymbol{\beta})$ si lze představit jako zprohýbanou plochu v $(g + 1)$ rozměrném prostoru, kde g rozměrů je vyhrazeno pro složky vektoru $\boldsymbol{\beta}$ a g plus první rozměr je rezervován pro funkční hodnotu $\chi^2(\boldsymbol{\beta})$. Obecně může mít taková plocha dosti komplikovaný vzhled. Nicméně většinou na ní můžeme najít jedno nebo i více lokálních minim, z nichž ovšem jen některá budou mít nějaký dobrý fyzikální smysl.

²Jinou takovou podmínkou může být minimálnost součtu absolutních hodnot modifikovaných odchylek nebo jejich čtvrtých mocnin. Nicméně takto definované podmínky se používají jen zřídka, a ve zcela odůvodněných případech. Naopak často se používají jisté modifikace MNC, které dokáží eliminovat hrubé chyby. Těmto modifikacím se pak říká *robustní regrese*.

³U těchto vah je však třeba mít na paměti, že to nejsou bezrozměrné veličiny, ale že mají individuální rozměr $\dim(w_i) = [\dim(y_i)]^{-2}$.

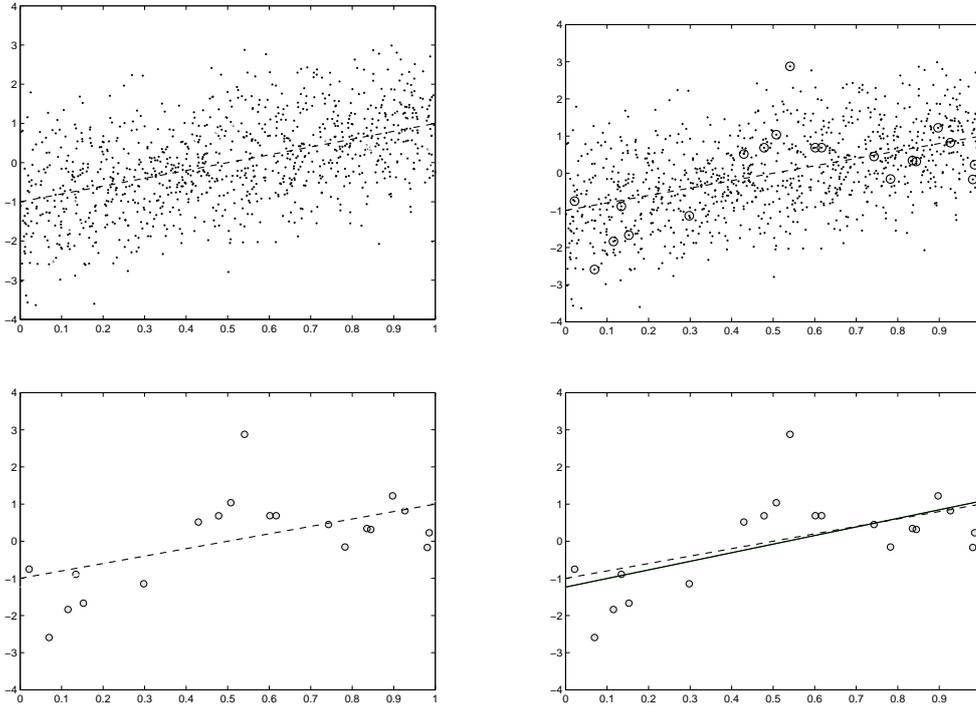


Fig. 1. Na těchto obrázcích si můžete ověřit sílu MNČ. Předpokládejme, že y je lineárně závislé na veličině x (typicky na čase). Každý z 1000 naměřených bodů necht' je zatížen stejnou nejistotou σ_i . Nyní si z těchto 1000 bodů náhodně vybereme 20, které jsou na druhém obrázku zvýrazněny kroužky. Z nich vypočteme odhad závislosti $y(x)$ a znázorníme si ji. V grafu je pro informaci vynesena i výsledná závislost, ovšem s vědomím, že tuto závislost v té chvíli ještě neznáme. Nyní je třeba zvolit správný model pro tuto závislost. I když by v těchto 20 bodech bylo možné vidět i úsek paraboly, dostačujícím modelem závislosti tu bude přímka definovaná dvěma parametry. Tato přímka se zjevně dobře shoduje se skutečnou závislostí definovanou padesátkrát více body, než kolik jich máme k dispozici.

Při hledání extrémů (minima nebo maxima) skalární funkce je vhodné si zavést pojem *gradient funkce*. Gradient v daném bodě je vektor orientovaný v opačném směru než spádnicí, přičemž délka vektoru je tím větší, čím strměji v daném bodě funkce probíhá. Číselně jsou složky vektoru gradientu funkce χ^2 , která je funkcí g proměnných parametrů, rovny parciálními derivacím podle těchto parametrů

$$\vec{\nabla}\chi^2(\mathbf{b}) = \left(\frac{\partial\chi^2}{\partial\beta_1}, \frac{\partial\chi^2}{\partial\beta_2}, \dots, \frac{\partial\chi^2}{\partial\beta_g} \right). \quad (5)$$

Gradient lze takto podle potřeby chápat jako buď jako vektor o g složkách nebo řádkovou matici s g sloupci. Pomocí gradientu součtu čtverců odchylek lze podmínku pro nalezení extrému funkce nebo jeho sedlového bodu lze pak elegantně zapsat

$$\vec{\nabla}\chi^2(\mathbf{b}) = \mathbf{0}, \quad (6)$$

kde $\mathbf{0}$ je řádkový vektor o g složkách, jež jsou všechny rovny nule. Podmínka tak říká, že extrém (sedlový bod) skalární funkce nastává v takovém bodě, kde všechny složky gradientu funkce jsou rovny nule. Nás ovšem zajímají právě jen minima této funkce. Velikost vektoru gradientu je v minimu nulová, jsme totiž na dně - hlouběji se okolí tohoto bodu dostat nelze. Popisované metodě hledání minima skalární funkce se proto říká též *gradientní metoda* (gradient method).

Dosadíme-li nyní výraz pro váhovanou sumu čtverců odchylek do (6) po krátkých úpravách dojdeme k jediné vektorové podmínce

$$\sum_{i=1}^n \frac{\mathbf{x}_i f(t_i, \mathbf{b})}{\sigma_i^2} = \sum_{i=1}^n \frac{\mathbf{x}_i y_i}{\sigma_i^2},$$

$$\text{nebo} \quad \sum_{i=1}^n \mathbf{x}_i f(t_i, \mathbf{b}) w_i = \sum_{i=1}^n \mathbf{x}_i y_i w_i, \quad (7)$$

$$\mathbf{x}_i = \vec{\nabla} f(t_i, \mathbf{b}) = \left(\frac{\partial f(t_i, \mathbf{b})}{\partial \beta_1}, \frac{\partial f(t_i, \mathbf{b})}{\partial \beta_2}, \dots, \frac{\partial f(t_i, \mathbf{b})}{\partial \beta_g} \right). \quad (8)$$

Vektor příslušný k i -tému měření \mathbf{x}_i s g složkami je tedy gradientem podle složek parametrů prokládané funkce v daném bodě. Složky tohoto vektoru tak lze pokládat za nezávislé proměnné. Soustavu g obecně nelineárních rovnic o g neznámých, složek parametru \mathbf{b} pak řešíme běžným způsobem.⁴

2.2. Odhad nejistot jednotlivých měření

V praxi se občas stává, že nemáme vždy spolehlivou informaci o nejistotách $\{\sigma_i\}$ pro jeden každý bod měření. Přitom většinou jde o měření provedená v minulosti, tedy neopakovatelná a tudíž unikátní. Někdy o nejistotách vstupních údajů nevíme zhora nic. Jenže ony nejistoty k výpočtu χ^2 nutně potřebujeme. Nebylo by poctivější oprášit starou dobrou *prostou metodu nejmenších čtverců* se sumou čtverců odchylek v podobě: $S(\mathbf{b}) = \sum [y_i - f(t_i, \mathbf{b})]^2$, v níž není ani nejistoty σ_i ani váhy w_i zapotřebí? Lze to ale vůbec takto udělat?

Lze to učinit, ale jen v tom případě, kdy máme co do činění s daty stejného druhu, o nichž víme, že všechna mají zaručeně stejnou nejistotu $\sigma_i = \sigma$. Pokud by tato podmínka splněna nebyla, neměli bychom MNČ používat nebo alespoň bychom neměli tvrdit, že jsme k nějakým závěrům dospěli pomocí této metody. Výsledky, které bychom dostali, by byly nutně zkreslené, zejména by nebylo možné se spolehnout na odhady nejistot.

Připustíme-li, že v souboru zpracovávaných dat se nacházejí data nebo skupiny dat s rozdílným rozptylem, s rozdílnou kvalitou⁵, je naší povinností vše udělat pro to, abyste

⁴Triviálním příkladem regrese řešené pomocí MNČ je nalezení střední hodnoty n naměřených hodnot $\{y_i\}$ se stejnou nejistotou σ . Model regresní funkce $f(t) = \beta$, $\mathbf{x}_i = \vec{\nabla} f_i = \partial f_i / \partial \beta = 1$, $\chi^2(\beta) = \sigma^{-2} \sum (y_i - \beta)^2$.

Minimum funkce $\chi^2(\beta)$ nastává v bodě $\beta = b$, v němž platí, že $\partial \chi^2 / \partial \beta = -2\sigma^{-2} \sum (y_i - b) = 0$, tedy $b = \frac{1}{n} \sum y_i = \bar{y}$ hledaným středem je aritmetický průměr. Suma kvadrátů modifikovaných odchylek \tilde{e}_i^2 pro $b = \bar{y}$, $\chi^2(\beta = \bar{y}) = \sigma^{-2} \sum (y_i - \bar{y})^2 = \sigma^{-2} \sum y_i^2 - 2y_i \bar{y} + \bar{y}^2 = n \sigma^{-2} (\bar{y}^2 - \bar{y}^2)$.

Poučný je i průběh funkce $= \sigma^{-2} \sum (y_i - \beta)^2 = \sigma^{-2} \sum y_i^2 - 2\beta \sum y_i + \beta^2 = \chi^2(b) + n \sigma^{-2} (\beta - \bar{y})^2$ - jde o parabolu, křivku s minimem v $\beta = b = \bar{y}$ s minimální hodnotou $\chi^2(\beta)_{\min} = \chi^2(b)$.

⁵Zde úplně stačí, když používáme data od různých pozorovatelů, získaná různou pozorovací technikou, v různých fotometrických filtrech, v různých klimatických podmínkách atp.

ony nejistoty či váhy nějak odhadli a použili vztahy zohledňující rozdílné nejistoty, respektive váhy jednotlivých měření.

Jak tedy onu nejistotu měření veličiny σ_i odhadnout? Předně je třeba se smířit se skutečností, že onu nejistotu individuálního měření nikdy nedokážeme určit přesně: každé měření je jedinečné, neopakovatelné a nikdy zpětně nebudeme znát všechny okolnosti, které v tu chvíli mohly vlastní měření ovlivnit. Jistým vodítkem nám sice může být udávaná vnitřní nejistota (chyba), která ovšem zpravidla představuje jen dolní odhad skutečné nejistoty. Zde je třeba si uvědomit, že ona nejistota by se měla vztahovat k právě použitému regresnímu modelu, který nemusí realitu popisovat ideálně.

Východiskem tu může být použití prosté metody nejmenších čtverců s jednotkovými váhami a s následnou analýzou kvality proložení jednotlivými podskupinami v celém datovém souboru. Zlepšený odhad nejistot pak lze učinit za předpokladu, že přesnost měření v rámci určité relativně homogenní podskupiny dat bude nejspíš zhruba stejná (např. měření z určité noci v určitém filtru atp.). Tato nejistota pro j -tou podskupinu měření $-\sigma_j$ je pak dána rozptylem měření podskupiny vzhledem k modelové předpovědi. Platí tedy: $\sigma_{ji} = \sigma_j$. Takto lze upřesnit váhy všech měření ve zpracovávaném souboru a celou regresi zopakovat. Po několika iteracích dojdeme k ustálenému stavu, kdy se již výsledky nebudou dále měnit.

Odhadujeme-li nejistoty jednotlivých pozorování takto, musíme se smířit s tím, že se vážou na daný regresní model. Při volbě jiného modelu, můžeme dostat poněkud odlišné hodnoty odhadů $\sigma_{ji} = \sigma_j$ a tím i vah jednotlivých měření. Zkušenost však ukazuje, že tyto rozdíly povedou jen k marginálním změnám ve výsledku, takže je můžeme zanedbat.

3 Lineární regrese

Řešení soustavy rovnic (7) v jejich obecnosti bývá dosti komplikované, takže není divu, že se vyhledají takové regresní modely, s nimiž by se dalo zacházet jednodušeji. Příjemná práce je s tzv. *lineárními regresními funkcemi* $f(t, \boldsymbol{\beta})$, které je možné vyjádřit jako lineární kombinaci g funkcí času $\{x_1(t), x_2(t), \dots, x_g(t)\}$, které tvoří vektorovou funkci $\mathbf{x}(t) = (x_1, x_2, \dots, x_g)$. Hovoříme pak o lineární regresní funkci nebo o lineárním regresním modelu. Platí tedy

$$f(t, \boldsymbol{\beta}) = \beta_1 x_1(t) + \beta_2 x_2(t) + \dots + \beta_g x_g(t) = \sum_{j=1}^g \beta_j x_j(t) = \boldsymbol{\beta} \mathbf{x}(t) \quad (9)$$

$$\Rightarrow \vec{\nabla} f(t, \boldsymbol{\beta}) = \left(\frac{\partial f}{\partial \beta_1}, \frac{\partial f}{\partial \beta_2}, \dots, \frac{\partial f}{\partial \beta_g} \right) = \mathbf{x}(t). \quad (10)$$

Dosadíme-li nyní do rovnice (7) za $f(t, \boldsymbol{\beta})$ dostaneme

$$\sum_{i=1}^n \mathbf{x}(t_i) w_i \sum_{j=1}^g b_j x_j(t_i) = \sum_{i=1}^n \mathbf{x}(t_i) y_i w_i, \quad (11)$$

kde váha $w_i = \sigma_i^{-2}$. k -tou složku předchozí soustavy rovnic lze po roznásobení sum přepsat do tvaru

$$\sum_{j=1}^g b_j \sum_{i=1}^n x_k(t_i) x_j(t_i) w_i = \sum_{i=1}^n y_i x_k(t_i) w_i. \quad (12)$$

Celou soustavu g lineárních rovnic o g neznámých, jimiž jsou složky hledaného vektoru \mathbf{b} lze zapsat takto:

$$\begin{aligned} V_{11}b_1 + V_{12}b_2 + \dots + V_{1g}b_g &= U_1 \\ V_{21}b_1 + V_{22}b_2 + \dots + V_{2g}b_g &= U_2 \\ &\vdots \\ V_{g1}b_1 + V_{g2}b_2 + \dots + V_{gg}b_g &= U_g, \end{aligned} \quad (13)$$

kde

$$V_{kj} = V_{jk} = \sum_{i=1}^n x_k(t_i) x_j(t_i) w_i; \quad U_k = \sum_{i=1}^n y_i x_k(t_i) w_i. \quad (14)$$

Soustavu g rovnic o g neznámých (b_j) pak lze standardním způsobem řešit. Nalezením všech hledaných koeficientů je pak nalezena i regresní funkce, kde $\boldsymbol{\beta} = \mathbf{b}$. Pokud nás dále nezajímá přesnost měření, hodnověrnost proložení, chyby parametrů a neurčitost předpovědi, pak jsme hotovi.

3.1. Lineární regrese užitím maticového počtu

Lineární regresi lze elegantně řešit použitím maticového počtu. Ten budeme přednostně používat i v následujícím textu.

Pozorovaný vztah mezi závisle proměnnou (nepřesně měřenou veličinou, nejčastěji hvězdnou velikostí, ale i třeba radiální rychlostí, teplotou aj.) y a nezávislou proměnnou (přesně měřenou veličinou – typicky časem) t může být proložen vhodnou **modelovou funkcí** f . Matematický model závislosti nechť je určen uspořádanou g -ticí volných parametrů β_j , ve formě sloupcového vektoru $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_g)^T$. Pokud je možné modelovou funkci f zapsat jako lineární kombinaci g různých funkcí času $x_k(t)$, tak hovoříme o tzv. lineární modelové funkci a lze psát

$$\mathbf{x} = (x_1, x_2, \dots, x_g), \quad f(\mathbf{x}, \boldsymbol{\beta}) = \sum_{k=1}^g \beta_k x_k = \mathbf{x} \boldsymbol{\beta}. \quad (15)$$

Zavedme sloupcový vektor závislé veličiny \mathbf{y} s délkou n a matici \mathbf{X} s rozměrem $n \times g$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1g} \\ x_{21} & x_{22} & \cdots & x_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{ng} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}, \quad (16)$$

kde y_i je hodnota i -tého pozorování, x_{ik} je funkční hodnota k -té funkce pro i -té pozorování,

$\mathbf{f}(t_i)$ je hodnota řádkového vektoru definovaného v (10)⁶.

$$\mathbf{f}(\mathbf{X}, \boldsymbol{\beta}) = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \boldsymbol{\beta} = \mathbf{X} \boldsymbol{\beta}; \quad \mathbf{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}. \quad (17)$$

kde \mathbf{W} je diagonální matice $n \times n$ s vahami jednotlivých měření v diagonále, $\mathbf{f}(\boldsymbol{\beta})$ je sloupcový vektor s jednotlivými hodnotami modelové funkce $f_i(\mathbf{x}_i)$ pro i -té pozorování pro zadané $\boldsymbol{\beta}$.

Jako objektivní míru úspěšnosti proložení modelovou funkcí s parametry $\boldsymbol{\beta}$ použijeme součet váhovaných čtverců odchylek pozorovaných hodnot od předpověděných $\chi^2(\boldsymbol{\beta})$

$$\chi^2(\boldsymbol{\beta}) = [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^T \mathbf{W} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})] = (\mathbf{y}^T - \boldsymbol{\beta}^T \mathbf{X}^T) \mathbf{W} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}) = \mathbf{y}^T \mathbf{W} \mathbf{y} - \boldsymbol{\beta}^T \mathbf{U} - \mathbf{U}^T \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{V} \boldsymbol{\beta} = \mathbf{y}^T \mathbf{W} \mathbf{y} - 2 \boldsymbol{\beta}^T \mathbf{U} + \boldsymbol{\beta}^T \mathbf{V} \boldsymbol{\beta}. \quad (18)$$

\mathbf{U} je řádkový vektor s délkou g , \mathbf{V} je čtvercová matice $g \times g$, jejíž inverzní matice \mathbf{H} je tzv. *kovarianční matice*:

$$\mathbf{U} = \mathbf{X}^T \mathbf{W} \mathbf{y}; \quad \mathbf{V} = \mathbf{X}^T \mathbf{W} \mathbf{X}; \quad \mathbf{H} = \mathbf{V}^{-1} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}. \quad (19)$$

Při proložení modelovou funkcí $f(t, \boldsymbol{\beta})$ metodou nejmenších čtverců se bere za optimální takové, pro něž je suma $\chi^2 = \chi^2(\boldsymbol{\beta} = \mathbf{b})$ minimální. V případě lineární modelové funkce $f(t, \boldsymbol{\beta})$ platí, že takové minimum je jen jediné. Pro řešení v podobě sady parametrů \mathbf{b} a sumu kvadrátů odchylek $\chi^2(\mathbf{b})$ platí:

$$\left. \frac{\partial \chi^2}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\mathbf{b}} = \mathbf{0} = -2 \mathbf{U} + 2 \mathbf{V} \mathbf{b} \quad \Rightarrow \quad \mathbf{b} = \mathbf{H} \mathbf{U} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \quad (20)$$

Předpověď hodnot modelové lineární funkce pro $\boldsymbol{\beta} = \mathbf{b}$, \mathbf{y}_p je dána následujícím vztahem:

$$\mathbf{y}_p = \mathbf{X} \mathbf{b} = [\mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}] \mathbf{y} = \boldsymbol{\Xi} \mathbf{y}. \quad (21)$$

Výraz v hranaté závorce – symetrická matice $\boldsymbol{\Xi}$ o rozměru $n \times n$, která zde vystupuje jako operátor, který každé hodnotě pozorování přiřadí její „vyhlazenou“ hodnotu. Toto zobrazení je tím věrnější, čím více se matice $\boldsymbol{\Xi}$ blíží jednotkové matici $\mathbf{E}(n, n)$.

Minimální sumu kvadrátů odchylek χ^2 lze pro lineární regresi zapsat různě

$$\chi^2 = (\mathbf{y} - \mathbf{X} \mathbf{b})^T \mathbf{W} (\mathbf{y} - \mathbf{X} \mathbf{b}) = \mathbf{y}^T \mathbf{W} \mathbf{y} - \mathbf{b}^T \mathbf{U} = \mathbf{y}^T \mathbf{W} \mathbf{y} - \mathbf{y}_p^T \mathbf{W} \mathbf{y}_p. \quad (22)$$

V posledních dvou variantách vystupuje i váhovaná suma čtverců funkčních hodnot, což je veličina vstupní, vyplývající z pozorování, tudíž zcela nezávislá na modelování. Metodu

⁶Standardně používanými modely lineárních regresních funkcí jsou běžné nebo trigonometrické polynomy vhodných stupňů. Jako příklad lze zvolit parabolický model, jenž je nejjednodušším modelem části světelné křivky s extrémem. Parabolický model lze předpokládat ve formě: $f(t) = \beta_1 t^2 + \beta_2 t + \beta_3$, $\mathbf{f}(t) = [t^2, t, 1]$, $\mathbf{X} = [\{t_i^2\} \{t_i\} \{1\}]$.

nejmenších čtverců tak lze alternativně chápat i jako metodu největších čtverců modelových předpovědí. Tento pohled lze s výhodou využít např. při hledání nejlepších period, tedy při tvorbě LSM periodogramů.

Sumu čtverců odchylek $\chi^2(\boldsymbol{\beta})$ pro lineární regresní model lze po určitých úpravách zapsat v následujícím instruktivním tvaru:

$$\chi^2(\boldsymbol{\beta}) = \chi^2 + \sum_{k=1}^g (\beta_k - b_k)^2 \sum_{i=1}^n \frac{x_{ki}^2}{\sigma_i^2}. \quad (23)$$

Ze zápisu je okamžitě patrné, že funkce $\chi^2(\boldsymbol{\beta})$ má tvar paraboloidu s minimem v bodu $\boldsymbol{\beta} = \mathbf{b}$. Má tedy jediné a tudíž absolutní minimum.

3.2. Nejistoty parametrů modelu a předpovědí

V rámci řešení úlohy lineární regrese lze též odhadnout střední rozptyl měření⁷ s^2 , dále odhad nejistoty předpovědi jednotlivých vstupních dat $\delta \mathbf{y}_p$ a odhad nejistot parametrů modelu $\delta \mathbf{b}$

$$s^2 = \frac{\chi_\mu^2}{w}; \quad \delta \mathbf{y}_p = \sqrt{\chi_\mu^2 \text{diag}(\mathbf{X} \mathbf{H} \mathbf{X}^T)}; \quad \delta \mathbf{b} = \sqrt{\chi_\mu^2 \text{diag}(\mathbf{H})}, \quad \text{kde } \chi_\mu^2 = \frac{\chi^2}{n-g}. \quad (24)$$

χ_μ^2 je pomocná bezrozměrná funkce, jejíž velikost závisí na adekvátnosti volby regresního modelu a správnosti odhadu nejistot použitých dat. Operátor „diag“, aplikovaný na čtvercovou matici, vytvoří sloupcový vektor sestavený z prvků nacházejících se na její diagonále; operátor může fungovat i v opačném směru, aplikací na sloupcový vektor obdržíme čtvercovou matici, jejíž diagonálu tvoří prvky vektoru v odpovídajícím pořadí. Je-li vše v pořádku, pak platí $\chi_\mu^2 \approx 1 \pm \sqrt{2/(n-g)}$.

Složky sloupcového vektoru $\delta \mathbf{b}$ se často uvádějí jako rigorózní odhad nejistot jednotlivých parametrů modelu. Bohužel, tento význam mají jen výjimečně, nicméně na nich občas trvají recenzenti odborných článků a oponenti diplomových prací. Naproti tomu velmi cenný je následující odhad předpovědi modelu $\delta f(t, \mathbf{b})$

$$\delta f(t, \mathbf{b}) = \sqrt{\chi_\mu^2 \mathbf{x} \mathbf{H} \mathbf{x}^T} = \sqrt{w s^2 \mathbf{x} \mathbf{H} \mathbf{x}^T} = \sqrt{\chi_\mu^2 \vec{\nabla} f \mathbf{H} (\vec{\nabla} f)^T}. \quad (25)$$

Odhady nejistoty jednotlivých parametrů obsažených ve vektoru řešení \mathbf{b} , $\delta \mathbf{b}$ se zdají být důležité, neboť přece pomocí nich lze odhadnout i nejistotu libovolného výrazu $Q(\boldsymbol{\beta}, t)$, a to podle notorického zákona o šíření chyb

$$\delta Q(\boldsymbol{\beta}, t) = \sqrt{\sum_{k=1}^g \left(\frac{\partial Q}{\partial \beta_k} \delta b_k \right)^2}, \quad (26)$$

který lze přepsat do elegantnějšího tvaru zahrnujícího i výpočet vektoru chyb $\delta \mathbf{b}$

$$\delta Q(\boldsymbol{\beta}, t) = \sqrt{\chi_\mu^2 \vec{\nabla} Q \text{diag}(\mathbf{H}) (\vec{\nabla} Q)^T}, \quad \text{kde } \vec{\nabla} Q(\boldsymbol{\beta}) = \left(\frac{\partial Q}{\partial \beta_1}, \frac{\partial Q}{\partial \beta_2}, \dots, \frac{\partial Q}{\partial \beta_g} \right), \quad (27)$$

⁷Tato veličina má ovšem fyzikální význam pouze tehdy, zpracováváme-li měření stejného druhu (se stejnou fyzikální jednotkou - mag, km/s apod.). V opačném případě je význam veličiny s^2 čistě formální.

kde $\vec{\nabla}Q(\boldsymbol{\beta})$ je řádkový vektor gradientu funkce Q podle jednotlivých parametrů.

Jenže výrazy (26,27) platí pouze tehdy, je-li kovarianční matice \mathbf{H} diagonální, jinými slovy – jednotlivé parametry v daném výrazu nejsou korelované. V obecném případě takto dostaneme jen horní hranici nejistoty. Chcete-li postupovat korektně, měli byste použít následující, jistě ještě elegantnější vztah

$$\delta Q = \sqrt{\chi_\mu^2 \vec{\nabla}Q \mathbf{H} (\vec{\nabla}Q)^T}. \quad (28)$$

Funkcí Q může být i první nebo druhá derivace modelové funkce podle času \dot{f} , \ddot{f} , což jsou veličiny nezbytné např. k výpočtu nejistoty určení okamžiku extrému světelné křivky:

$$\delta \dot{f}(t, \mathbf{b}) = \sqrt{\chi_\mu^2 \vec{\nabla} \dot{f} \mathbf{H} (\vec{\nabla} \dot{f})^T} = \sqrt{\chi_\mu^2 \dot{\mathbf{x}} \mathbf{H} \dot{\mathbf{x}}^T}; \quad (29)$$

$$\delta \ddot{f}(t, \mathbf{b}) = \sqrt{\chi_\mu^2 \vec{\nabla} \ddot{f} \mathbf{H} (\vec{\nabla} \ddot{f})^T} = \sqrt{\chi_\mu^2 \ddot{\mathbf{x}} \mathbf{H} \ddot{\mathbf{x}}^T}, \quad (30)$$

kde $\dot{\mathbf{x}}(t) = (\dot{x}_1(t), \dot{x}_2(t), \dots, \dot{x}_g(t))$ a $\ddot{\mathbf{x}}(t) = (\ddot{x}_1(t), \ddot{x}_2(t), \dots, \ddot{x}_g(t))$.

3.3. Základní regresní modely - aplikace lineární regrese

Následuje několik praktických příkladů aplikace lineární regrese metody nejmenších čtverců, které mají ilustrovat způsob, jak lze metodu lineární regrese v maticové podobě používat. Pokud tyto příklady někomu připadnou jako triviální, pak se nemýlí, neboť jde o záměr. Pokud ovšem zvládnete toto, můžete si troufnout na složitější modely.

V řadě příkladů budou s výhodou použity některé střední veličiny, nezávislých i závislých veličin t a y :

$$\overline{t^m y^l} = \sum_{i=1}^n t_i^m y_i^l w_i / \sum_{i=1}^n w_i, \quad (31)$$

$$u_{tt} = \overline{t^2} - \bar{t}^2, \quad s_t = \sqrt{u_{tt}}, \quad u_{yy} = \overline{y^2} - \bar{y}^2, \quad s_y = \sqrt{u_{yy}}, \quad u_{ty} = \overline{ty} - \bar{t}\bar{y}, \quad (32)$$

$$r = \frac{\overline{ty} - \bar{t}\bar{y}}{s_t s_y} = \sqrt{\frac{u_{ty}^2}{u_{tt} u_{yy}}} = \frac{u_{ty}}{s_t s_y} \quad (33)$$

Korelační koeficient r je bezrozměrná veličina nabývající hodnotu mezi -1 a 1, přičemž 0 je roven tehdy, kdy mezi veličinami t a y neexistuje žádná lineární korelace, ± 1 je roven tehdy, kdy jsou všechny hodnoty $\{t_i, y_i\}$ vyskládány na jediné přímce. Individuální váha souvisí s nejistotou takto: $w_i = \sigma_i^{-2}$.

3.4. Průměrná hodnota

V případě, že mezi n dvojicemi t a y datového souboru $\{t_i, y_i, \sigma_i\}$ neexistuje žádná závislost (korelační koeficient je blízký nule), bude hodnota $y(t)$ v mezích chyb nejspíš konstantní. Regresní model pak můžeme sestavit takto: $y_i = \beta + e_i, f(\beta) = \beta$. Optimální hodnotu β , při níž je vážená suma čtverců modifikovaných odchylek $\tilde{e}_i = e_i/\sigma_i$ minimální, b , nazveme váženou střední

hodnotou. Můžeme ji najít přímo minimalizací výrazu $\chi^2(\beta)$:

$$\chi^2(\beta) = \sum_{i=1}^n \tilde{\epsilon}_i^2 = \sum_{i=1}^n \left(\frac{y_i - \beta}{\sigma_i} \right)^2 = \sum_{i=1}^n \frac{y_i^2}{\sigma_i^2} - 2\beta \sum_{i=1}^n \frac{y_i}{\sigma_i^2} + \beta^2 \sum_{i=1}^n \frac{1}{\sigma_i^2}, \quad (34)$$

$$\frac{\partial \chi^2(b)}{\partial \beta} = -2 \sum_{i=1}^n \frac{y_i}{\sigma_i^2} + 2b \sum_{i=1}^n \frac{1}{\sigma_i^2} = 0; \quad \Rightarrow \quad b = \frac{\sum y_i \sigma_i^{-2}}{\sum \sigma_i^{-2}} = \frac{\sum y_i w_i}{\sum w_i} = \bar{y}; \quad (35)$$

$$\chi^2(\bar{y}) = \sum_{i=1}^n \frac{y_i^2 - \bar{y}^2}{\sigma_i^2}; \quad \chi^2(\beta) = \chi^2(\bar{y}) + (\beta - \bar{y})^2 \sum_{i=1}^n \sigma_i^{-2}. \quad (36)$$

Grafem funkce $\chi^2(\beta)$ je parabola s minimem v $\beta = \bar{y}$ a funkční hodnotou $\chi^2(\bar{y})$ (viz (36)).

I když minimalizací funkce $\chi^2(\beta)$ lze střední hodnotu vypočítat přímo, zkusme si nyní ze cvičných důvodů všechny potřebné vztahy odvodit pomocí maticových vztahů.

$$\mathbf{X} = [1, 1, \dots, 1]^T, \quad \mathbf{Y} = [y_1, y_2, \dots, y_n]^T, \quad \mathbf{W} = \text{diag}[\sigma_1^{-2}, \sigma_2^{-2}, \dots, \sigma_n^{-2}]; \quad (37)$$

$$\mathbf{V} = \mathbf{X}^T \mathbf{W} \mathbf{X} = \sum \sigma_i^{-2}; \quad \mathbf{H} = \mathbf{V}^{-1} = \frac{1}{\sum \sigma_i^{-2}}, \quad (38)$$

$$\mathbf{U} = \sum y_i \sigma_i^{-2}, \quad b = \mathbf{H} \mathbf{U} = \frac{\sum y_i \sigma_i^{-2}}{\sum \sigma_i^{-2}} = \bar{y}, \quad (39)$$

$$\chi^2(\bar{y}) = \mathbf{Y}^T \mathbf{W} \mathbf{Y} - \mathbf{b}^T \mathbf{U} = \sum (y_i^2 - \bar{y}^2) \sigma_i^{-2}; \quad s^2 = \frac{\chi^2(\bar{y})}{\sigma^{-2}(n-1)} = s_y^2 \frac{n}{n-1}, \quad (40)$$

$$\chi_\mu^2 = \frac{\chi^2}{n-1}, \quad \delta b = \sqrt{\chi_\mu^2 \text{diag}(\mathbf{H})} = \frac{s}{\sqrt{n}}, \quad \delta y_p = s \sqrt{\chi_\mu^2 \text{diag}(\mathbf{x} \mathbf{H} \mathbf{x}^T)} = s. \quad (41)$$

Za povšimnutí jistě stojí, že vztahy pro b , σ , δb a δy_p jsou formálně stejné jako v případě bez vah. Rozdíl ovšem je v tom, jak jsou definovány střední veličiny, z nichž se při výpočtu vychází.

3.5. Přímka jdoucí počátkem

Občas se můžeme setkat se situací, kdy je jeden nebo více bodů závislosti pevně fixováno. Z této skutečnosti musíme při volbě regresního modelu vycházet. Nejjednodušším příkladem toho druhu je naše očekávání, že n bodů o souřadnicích $[t_i, y_i]$ se stejnými váhami lze proložit přímkou jdoucí bodem o souřadnicích $[0, 0]$, neboli počátkem. Regresní model je pak: $y_i = \beta t_i + e_i$, $f(\beta, t) = \beta t$. Optimální hodnotu $\beta = b$, při níž je vážená suma kvadrátů odchylek e_i minimální, nazveme tentokrát koeficientem úměrnosti.

I zde budeme předpokládat, že každému z bodů měření bude přisouzena určitá individuální

váha $w_i = 1/\sigma_i^2$.

$$\mathbf{X} = [t_1, t_2, \dots, t_n]^T, \quad \mathbf{y} = [y_1, y_2, \dots, y_n]^T, \quad \mathbf{W} = \text{diag}[w_1, w_2, \dots, w_n], \quad (42)$$

$$\mathbf{V} = \mathbf{X}^T \mathbf{W} \mathbf{X} = n \bar{w} \bar{t}^2, \quad \mathbf{H} = \mathbf{V}^{-1} = \frac{1}{n \bar{w} \bar{t}^2}, \quad \mathbf{U} = \mathbf{X}^T \mathbf{W} \mathbf{y} = \sum_{i=1}^n y_i t_i = n \bar{w} \bar{t} \bar{y}, \quad (43)$$

$$\mathbf{b} = \mathbf{H} \mathbf{U} = \frac{\sum_{i=1}^n t_i y_i w_i}{\sum_{i=1}^n t_i^2 w_i} = \frac{\bar{t} \bar{y}}{\bar{t}^2}, \quad (44)$$

$$y_p = b t, \quad R = \mathbf{y}^T \mathbf{W} \mathbf{y} - \mathbf{b}^T \mathbf{U} = n \bar{w} \left(\bar{y}^2 - b \bar{t} \bar{y} \right) = n \bar{w} \left[\bar{y}^2 - \frac{(\bar{t} \bar{y})^2}{\bar{t}^2} \right], \quad (45)$$

$$s^2 = \frac{\chi^2}{\bar{w}(n-1)} = \frac{n \left[\bar{t}^2 \bar{y}^2 - (\bar{t} \bar{y})^2 \right]}{(n-1) \bar{t}^2}, \quad \delta b = s \sqrt{\bar{w} H} = \frac{s}{\sqrt{n \bar{t}^2}}, \quad (46)$$

$$\mathbf{x} = \frac{\partial f}{\partial \beta} = t; \quad \delta y_p = s \sqrt{\bar{w} \mathbf{x}(t) \mathbf{H} \mathbf{x}(t)^T} = s \sqrt{\frac{t^2}{n \bar{t}^2}}. \quad (47)$$

3.6. Proložení obecnou přímkou

Při zpracování časově proměnných pozorovacích dat se můžeme často setkat s úlohou nalezení parametrů časové trendu, přičemž se v prvním přiblížení nejčastěji předpokládá, že mezi závislou veličinou y a nezávislou veličinou t (standardně časem měření) existuje lineární závislost. Jinými slovy body v grafu lze proložit přímkou. Regresní model pro takovou situaci je zřejmý: $y_i = \beta_1 + \beta_2 t_i + e_i$.

Přímka nechť je prokládána n body o souřadnicích $[t_i, y_i]$, přičemž každému z bodů je přiřazena jeho individuální váha w_i . Řešením úlohy je nalezení vektoru \mathbf{b} se složkami b_1, b_2 , pro něž je suma $\chi^2(\beta_1, \beta_2)$ minimální:

$$\chi^2(\beta_1, \beta_2) = \sum_{i=1}^n w_i (y_i - \beta_1 - \beta_2 t_i)^2, \quad (48)$$

$$\frac{\partial \chi^2}{\partial \beta_1} = -2 \sum_{i=1}^n w_i (y_i - b_1 - b_2 t_i) = 0, \quad \frac{\partial \chi^2}{\partial \beta_2} = -2 \sum_{i=1}^n w_i (y_i - b_1 - b_2 t_i) t_i = 0. \quad (49)$$

Soustavu dvou rovnic o dvou neznámých (49) řešíme prostředky maticového počtu:

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}; \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}; \quad (50)$$

$$\mathbf{V} = \mathbf{X}^T \mathbf{W} \mathbf{X} = n \bar{w} \begin{bmatrix} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{bmatrix}; \quad \mathbf{U} = \mathbf{X}^T \mathbf{W} \mathbf{y} = n \bar{w} \begin{bmatrix} \bar{y} \\ \bar{t} \bar{y} \end{bmatrix}; \quad (51)$$

$$\mathbf{H} = \mathbf{V}^{-1} = \frac{1}{n \bar{w} u_{tt}} \begin{bmatrix} \bar{t}^2 & -\bar{t} \\ -\bar{t} & 1 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \mathbf{H} \mathbf{U} = \frac{1}{u_{tt}} \begin{bmatrix} \bar{t}^2 \bar{y} - \bar{t} \bar{t} \bar{y} \\ -\bar{t} \bar{y} + \bar{t} \bar{y} \end{bmatrix}. \quad (52)$$

Přesvědčte se, že platí: $y_p = \bar{y}$, tedy, že regresní přímka prochází těžištěm.

$$\chi^2 = \mathbf{y}^T \mathbf{W} \mathbf{y} - \mathbf{b}^T \mathbf{U} = n \bar{w} \left(\bar{y}^2 - b_1 \bar{y} - b_2 \bar{t} \bar{y} \right), \quad \chi_\mu^2 = \frac{\chi^2}{n-2}, \quad (53)$$

$$s^2 = \frac{\chi_\mu^2}{\bar{w}}, \quad \mathbf{x} = [1, t]; \quad y_p = \mathbf{x} \mathbf{b}, \quad \delta y_p = \sqrt{\chi_\mu^2 \mathbf{x} \mathbf{H} \mathbf{x}^T} = \frac{s}{\sqrt{n}} \sqrt{1 + \frac{(t - \bar{t})^2}{s_t^2}}, \quad (54)$$

$$\delta b_2 = \sqrt{\chi_\mu^2 H_{22}} = \frac{s}{s_t \sqrt{n}}, \quad \delta b_1 = \sqrt{\chi_\mu^2 H_{11}} = \frac{s}{s_t} \sqrt{\frac{t^2}{n}} = \delta b_2 \sqrt{t^2}. \quad (55)$$

Nejistota směrnice přímky δb_2 tedy nezávisí na umístění počátku, zatímco chyba absolutního členu δb_1 ano. Minimální je tato chyba v případě, kdy počátek souřadnic ztotožníme s těžištěm. Nejistota pak bude $\delta b_1 = s/\sqrt{n}$. Absolutní člen b_1 lze geometricky interpretovat jako úsek na ose y , který na ní vytíná regresní přímka. Neurčitost polohy tohoto průsečíku udává chyba předpovědi $\delta y_p(t=0)$ v bodě 0. Číselně je tato chyba rovna chybě absolutního členu δb_1 , tak jak je uvedeno v (55).

Korelační koeficient r je dobrou mírou toho, jak dobře právě přímka vystihuje pozorovanou časovou závislost

$$r = \frac{\bar{t} \bar{y} - \bar{t} \bar{t}}{s_t s_y} = \frac{u_{ty}}{s_t s_y}. \quad (56)$$

3.7. Proložení časových řad polynomem

Při zpracování delších časových řad často aproximujeme vývoj pozorované veličiny y polynomem řádu $g-1$. Lineární regresní model předpokládáme ve tvaru: $y_i = \beta_1 + \beta_2 t_i + \dots + \beta_g t_i^{g-1} + e_i$.

Polynomiální závislost nechť je prokládána n body o souřadnicích $[t_i, y_i]$, přičemž každému z bodů je přisouzena jeho individuální váha w_i . Řešením úlohy je nalezení sloupcového vektoru \mathbf{b} s g složkami b_1, b_2, \dots, b_g , pro něž je suma váhovaných čtverců odchylek $\chi^2(\beta_1, \beta_2, \dots, \beta_g) = \chi^2(\boldsymbol{\beta})$ minimální. Řešíme pomocí maticového počtu. Definice matic \mathbf{W} a \mathbf{y} je též jako v (50), jediný rozdíl je v matici \mathbf{X} :

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{g-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{g-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^{g-1} \end{bmatrix}, \quad (57)$$

nazývané též matice Vandermondova.

3.8. Proložení časových řad harmonickým polynomem

Řada astrofyzikálních dějů probíhá více či méně periodicky. Známe-li z dřívějšíka parametry periodicity, lze si zavést tzv. *fázovou funkci* ϑ , kterou dostanete jako součet běžné fáze φ a epochy E . Pokud je perioda P konstantní, lze si fázovou funkci vypočítat jednoduchým vztahem:

$$\vartheta = \frac{t - M_0}{P}, \quad (58)$$

kde t je juliánské datum pozorování, M_0 je juliánské datum počátku počítání fázové funkce, P je fixní perioda ve dnech.

Pozorované periodicky se měnící veličiny y (jasnosti, radiální rychlosti, intenzity spektrálních čar, indukce magnetického pole aj.) vytvářejí *fázovou křivku*, kterou nejčastěji znázorňujeme jako

závislost proměnné veličiny na fázi $\varphi = \text{frac}(\vartheta)$. Fázové křivky zpravidla prokládáme harmonickým polynomem stupně $q = (g-1)/2$, kde g je počet stupňů volnosti. Matematický model s harmonickým polynomem stupně q lze zapsat: $y_i = \beta_1 + \sum_{k=1}^q \beta_{2k} \cos(2k\pi\vartheta_i) + \beta_{2k+1} \sin(2k\pi\vartheta_i) + e_i$.⁸ Odpovídající matice \mathbf{X} :

$$\mathbf{X} = \begin{bmatrix} 1 & \cos(2\pi\vartheta_1) & \sin(2\pi\vartheta_1) & \cos(4\pi\vartheta_1) & \sin(4\pi\vartheta_1) & \cdots & \cos(2q\pi\vartheta_1) & \sin(2q\pi\vartheta_1) \\ 1 & \cos(2\pi\vartheta_2) & \sin(2\pi\vartheta_2) & \cos(4\pi\vartheta_2) & \sin(4\pi\vartheta_2) & \cdots & \cos(2q\pi\vartheta_2) & \sin(2q\pi\vartheta_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \cos(2\pi\vartheta_n) & \sin(2\pi\vartheta_n) & \cos(4\pi\vartheta_n) & \sin(4\pi\vartheta_n) & \cdots & \cos(2q\pi\vartheta_n) & \sin(2q\pi\vartheta_n) \end{bmatrix}. \quad (59)$$

3.9. Zobecnění lineární regrese II - více nezávisle proměnných

Až doposud jsme jako jedinou nezávislou proměnnou brali čas a vše jsme nahlíželi z hlediska časové proměnnosti. Složky vektoru $\mathbf{x} = (x_1, x_2, \dots, x_g)$ pak byly funkcemi času. To však metoda nejmenších čtverců vůbec nevyžaduje. Jednotlivé položky mohou být třeba funkcemi prostorových souřadnic, rychlosti nebo to mohou být jen indikace popisující povahu měření (zda šlo třeba o fotometrické měření či měření radiálních rychlostí nebo intenzity spektrálních čar). Vše to jsou nezávislé, nenáhodné veličiny charakterizující konkrétní měření v rámci zvoleného komplexního modelu. Proto má smysl dívat se na celý soubor veličin obsažených ve vektoru $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ig})$ přímo jako na soubor g nezávislých veličin, které mohou nabývat různých hodnot. Pro určitý typ měření mohou být některé z nezávislých proměnných rovny 0, pro jiný typ měření mohou být nulové jiné nezávislé proměnné. Ve vektoru $\mathbf{y}_i = (y_1, y_2, \dots, y_n)^T$ s naměřenými veličinami jsou pak jednotlivé položky řazeny často v pořadí, v jakém byly naměřeny.

Příklad: Takovým lineárním modelem může být funkce se dvěma stupni volnosti popisující měření šířky a délky nějakého obdélníku. V případě, že v i -tém měření měříme šířku, je $\mathbf{x}_i = (0, 1)$, jde-li naopak o měření délky, pak je $\mathbf{x}_i = (1, 0)$, y_i je ona naměřená veličina. Modelová funkce pro i -té měření pro $f_i = \beta_1 x_{i1} + \beta_2 x_{i2} = \mathbf{x}_i \boldsymbol{\beta}$, β_1 je délka, β_2 je šířka. Cílem zpracování je najít střední velikost těchto parametrů \mathbf{b} na základě n měření. Při výpočtu budeme předpokládat, že váhy všech měření jsou jednotkové - tedy že je měříme se stejnou chybou.

$$\begin{array}{l|l} 3.16 & \text{\textit{s}} \\ 2.15 & \text{\textit{d}} \\ 2.18 & \text{\textit{d}} \\ 3.13 & \text{\textit{s}} \\ 2.15 & \text{\textit{d}} \\ 2.19 & \text{\textit{d}} \\ 3.13 & \text{\textit{s}} \end{array} ; \mathbf{y} = \begin{pmatrix} 3.16 \\ 2.15 \\ 2.18 \\ 3.13 \\ 2.15 \\ 2.19 \\ 3.13 \end{pmatrix} ; \mathbf{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} ; \mathbf{H} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} ; \mathbf{b} = \begin{matrix} 2.168 \pm 0.009 \\ 3.140 \pm 0.010 \end{matrix}. \quad (60)$$

Výhodou tohoto přístupu je, že můžeme solidně odhadnout směrodatnou odchylku a tedy i nejistotu určení hledané délky a šířky. Vzhledem k tomuto zobecnění se takto mohou pod sebe dostat i velmi odlišné typy měření s velmi odlišným rozsahem měřených veličin. Proto je důležité, aby byly jednotlivé typy měření správně oceněny svou vahou w_i nepřímo úměrnou své disperzi.

⁸Zde je třeba mít na paměti skutečnost, že fázová funkce je funkcí periody, která se může v průběhu času měnit. Úlohu, kde bychom kromě tvaru světelné křivky řešili i časový vývoj periody, lze zvládnout až prostředky nelineární regrese.

Nalezení okamžiku minima ze dvou sad pozorování - domácí úloha

Cílem této domácí úlohy je aplikace zobecněné lineární regrese na problém, který simuluje situaci, do níž se pozorovatelé proměnných hvězd často dostávají.

Představme si, že dva pozorovatelé v odlišných časových pásmech spolupracovali při pozorování minima jasnosti určité dlouhodobě periodické proměnné hvězdy, přičemž spolupracujícímu Číňanovi ($q = 1$) se podařilo provést celkem 15 pozorování, vesměs na sestupné větvi. Český pozorovatel ($q = 2$) zachytil až výstup světelné křivky z minima v 30 pozorováních ovšem s poněkud horší kvalitou. Samotné minimum žádný z pozorovatelů nezachytil.

V obou případech se pozorování vedla ve filtru V , hvězdné velikosti se vztahovaly k vybrané srovnávací hvězdě, pozorovatelé se však neshodli na její volbě, takže světelné křivky na sebe nenavazovaly. Světelné křivky byly simulovány parabolou

$$\Delta m(t) = a_1 (t - t_{\min})^2 + a_5 \delta_{i1} + a_6 \delta_{i2} = a_1 t^2 + a_2 t + a_3 \delta_{i1} + a_4 \delta_{i2}, \quad t_{\min} = -\frac{a_2}{2a_1}, \quad (61)$$

kde a_1 je koeficient parabolického členu (pro simulaci zvoleno $a_1 = 1$), t_{\min} je okamžik minima (zvoleno $t_{\min} = 0,350$), a_5 , a_6 jsou rozdíly hvězdné velikosti v minimu jasnosti pro čínského a českého pozorovatele (zvoleno $a_5 = 0,000$, $a_6 = 0,400$). Funkce $\delta_{i1} = 1$, pokud jde o pozorování Číňana, jinak $\delta_{i1} = 0$, naproti tomu $\delta_{i2} = 1$, pokud jde o pozorování Čecha, jinak $\delta_{i2} = 0$. a_2 je lineární člen, a_3 , a_4 jsou hodnoty $\Delta m(t = 0)$ pro jednotlivé pozorovatele. Okamžiky pozorování jsou udávány ve dnech od začátku určitého juliánského dne. Jednotlivé okamžiky t_i byly voleny náhodně v intervalu 0 až 0,3 ($q = 1$) a 0,4 až 0,8 ($q = 2$). K simulovaným hodnotám rozdílu hvězdné velikosti $\Delta m(t_i)$ určeným vztahem (61) pro dané hodnoty časů t_i byl přičten náhodný gaussovský šum o standardních odchylkách postupně: $s_1 = 0.005$ mag a $s_2 = 0.007$ mag. Tabulka s takto nasimulovanými časy t_i a hodnotami $\Delta m(t_i)$ včetně příznaku q následuje.

t_i	Δm_i	q	t_i	Δm_i	q	t_i	Δm_i	q
0,013	0,117	1	0,428	-0,037	2	0,596	0,014	2
0,039	0,093	1	0,455	-0,035	2	0,609	0,015	2
0,053	0,086	1	0,473	-0,042	2	0,623	0,026	2
0,100	0,058	1	0,486	-0,036	2	0,623	0,002	2
0,112	0,054	1	0,488	-0,031	2	0,634	0,033	2
0,114	0,055	1	0,489	-0,024	2	0,672	0,049	2
0,120	0,056	1	0,502	-0,035	2	0,672	0,056	2
0,131	0,041	1	0,502	-0,032	2	0,681	0,063	2
0,132	0,051	1	0,543	-0,017	2	0,697	0,086	2
0,206	0,014	1	0,549	-0,005	2	0,739	0,102	2
0,220	0,020	1	0,561	0,005	2	0,740	0,095	2
0,248	0,019	1	0,568	-0,005	2	0,743	0,097	2
0,252	0,006	1	0,572	0,006	2	0,743	0,101	2
0,264	0,005	1	0,573	0,005	2	0,761	0,123	2
0,294	-0,006	1	0,587	0,007	2	0,772	0,133	2

Vaším úkolem bude:

- Nakreslit graf pozorovaných světelných křivek.
- Pomocí lineární regrese se stejnými vahami jednotlivých měření vypočítat zvlášť pro 1. a 2. sadu pozorování hodnotu koeficientů a_1 , a_2 , a_3 , případně a_4 , včetně odhadu jejich nejistot, hodnotu standardní odchylky. Výsledné hodnoty mezi sebou porovnejte a srovnajte je se zadanými parametry simulace.

Vypočítejte dále okamžiky t_{\min} , včetně nejistoty jejich určení, přičemž využijete vztah uvedený v (61) a vztah pro výpočet odhadu chyby funkce koeficientů (28) a funkční hodnotu v minimu proložené paraboly a_5 a a_6 , včetně nejistoty. Výsledné hodnoty mezi sebou porovnejte a srovnajte je se zadanými parametry simulace.

- Spojte obě pozorování dohromady a předpokládejte, že absolutní členy lineární regrese jsou různé. Předpokládejte nejprve, že váhy všech pozorování jsou identické, rovné 1. Vypočtete koeficienty a_1 , a_2 , a_3 , a_4 , včetně odhadu jejich nejistot, hodnotu standardní odchylky. Výsledné hodnoty mezi sebou porovnejte a srovnajte je se zadanými parametry simulace.
- Vypočítejte standardní odchylky vzhledem k předpovědi vůči tomuto modelu zvlášť pro čínské a české pozorování. Pomocí nich vypočtete normalizovanou váhu jednotlivých čínských a českých pozorování. S těmito vahami pak opakujte výpočet parametrů a_1 , a_2 , a_3 , a_4 , včetně odhadu jejich nejistot, hodnotu standardní odchylky. Výsledné hodnoty mezi sebou porovnejte a srovnajte je se zadanými parametry simulace.
- Vypočítejte okamžik t_{\min} , včetně nejistoty jeho určení, a funkční hodnotu v minimu proložené paraboly a_5 a a_6 , včetně nejistoty. Výsledné hodnoty mezi sebou porovnejte a srovnajte je se zadanými parametry simulace.
- Pro spojené sady pozorování předpovězte funkční hodnoty a jejich nejistoty pro obě sady pozorování. Diskutujte, vynesete do grafu.

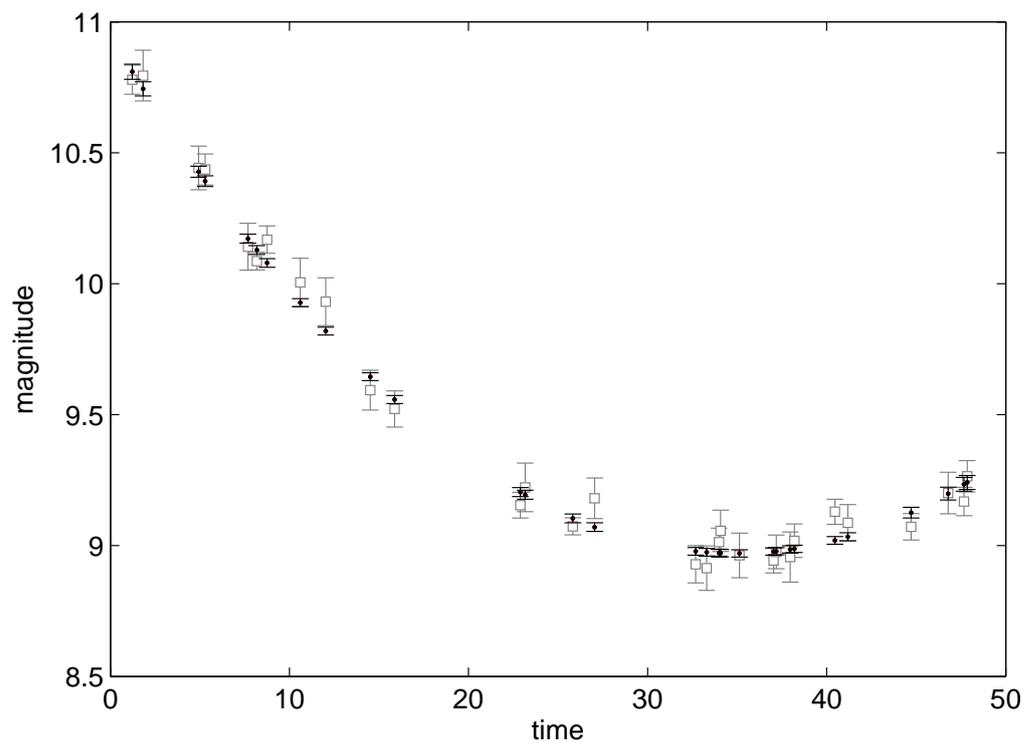


Fig. 2. Na obrázku jsou kolečky znázorněna simulovaná pozorování proměnné hvězdy v okolí jejího minima jasnosti. Vnitřní přesnost jednotlivých měření je znázorněna šedými chybovými úsečkami. Proložená parabola je naznačena černými tečkami s chybovými úsečkami odpovídajícími nejistotě předpovědi pomocí zvoleného parabolického lineárního modelu.