Jak postupovat při vyhodnocení lineární regrese ve fyzikálním praktiku

Ivo Křivka

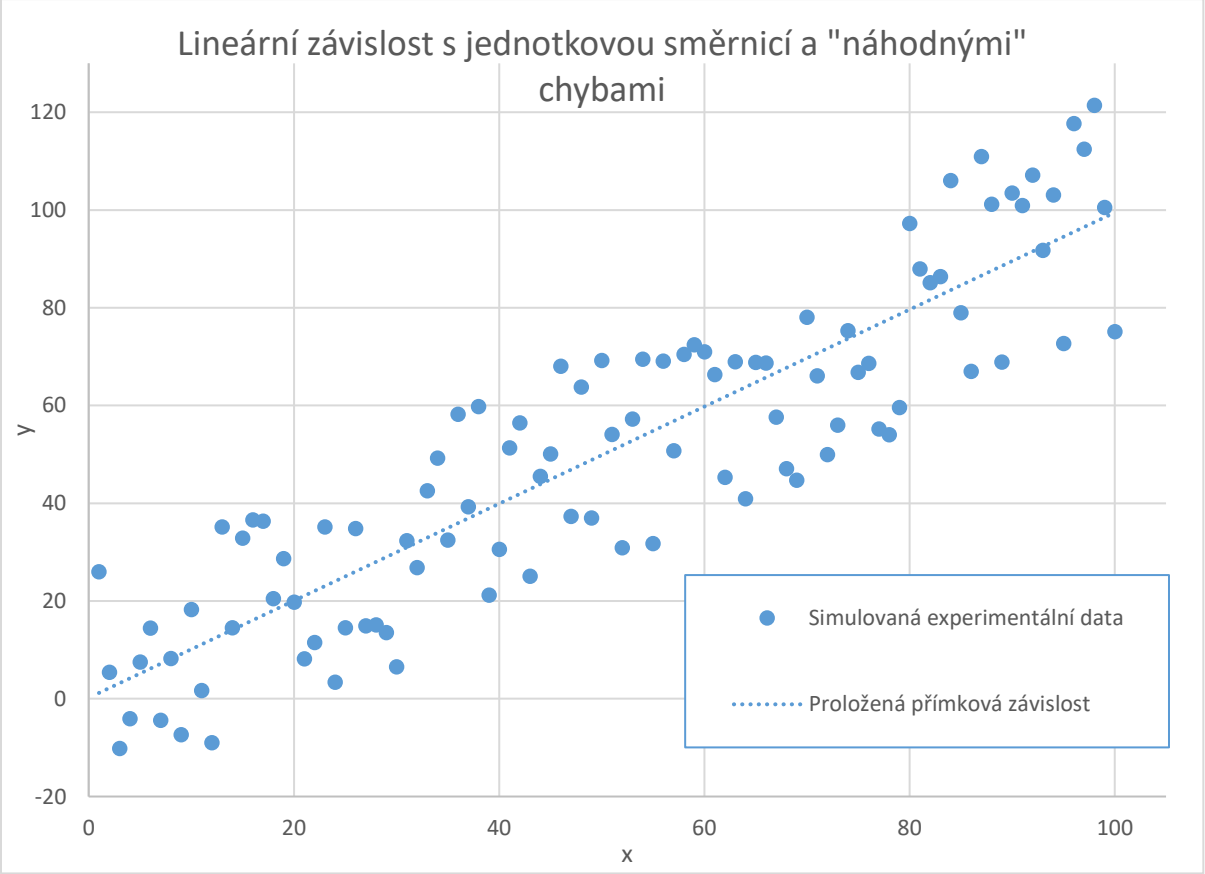
studijní text MFF UK

Namísto úvodu: Jak neprokládat přímku a proč

Uvažme jednoduchou úlohu, kdy pro ekvidistantně rozložené hodnoty x měříme hodnotu y, přičemž očekáváme lineární závislost. Naším úkolem je zjistit jen směrnici této závislosti. Popíšu jednoduchý postup, který se nabízí na první pohled a v některých případech může náhodou (to je skutečně třeba zdůraznit) poskytnout i správný výsledek. Jsem si vědom toho, že následující popis se může i přes varování v nadpisu uplatnit jako podprahový návod, jak se má takové vyhodnocení správně provádět. Proto ještě jednou zdůrazním, že ten postup popisuji jen proto, abych demonstroval, kde je jeho slabé místo.

Postup spočívá v tom, že hledanou směrnici určíme jako aritmetický průměr směrnic vypočtených pro všechny dvojice sousedních bodů v naměřené závislosti: Mějme tedy n dvojic naměřených bodů $[x_1, y_1], [x_2, y_2], \ldots [x_n, y_n]$. Podle předpokladu jsou hodnoty x rozmístěny ekvidistantně, takže $\Delta x = x_i - x_{i-1}$ pro $i = 2, \ldots, n$. První dvojice bodů vede tedy na směrnici $k_1 = (y_2 - y_1)/\Delta x$, druhá dvojice $k_2 = (y_3 - y_2)/\Delta x$ atd. Celkem tak získáme $n-1$ hodnot a z nich pak vypočteme průměrnou hodnotu. Pro dostatečně velká n tedy můžeme mít klamný pocit dobře provedeného statistického vyhodnocení.

Body v grafu (*Graf 1*) představují lineární závislost, na níž byly superponovány náhodně rozptýlené veličiny napodobující chyby měření.



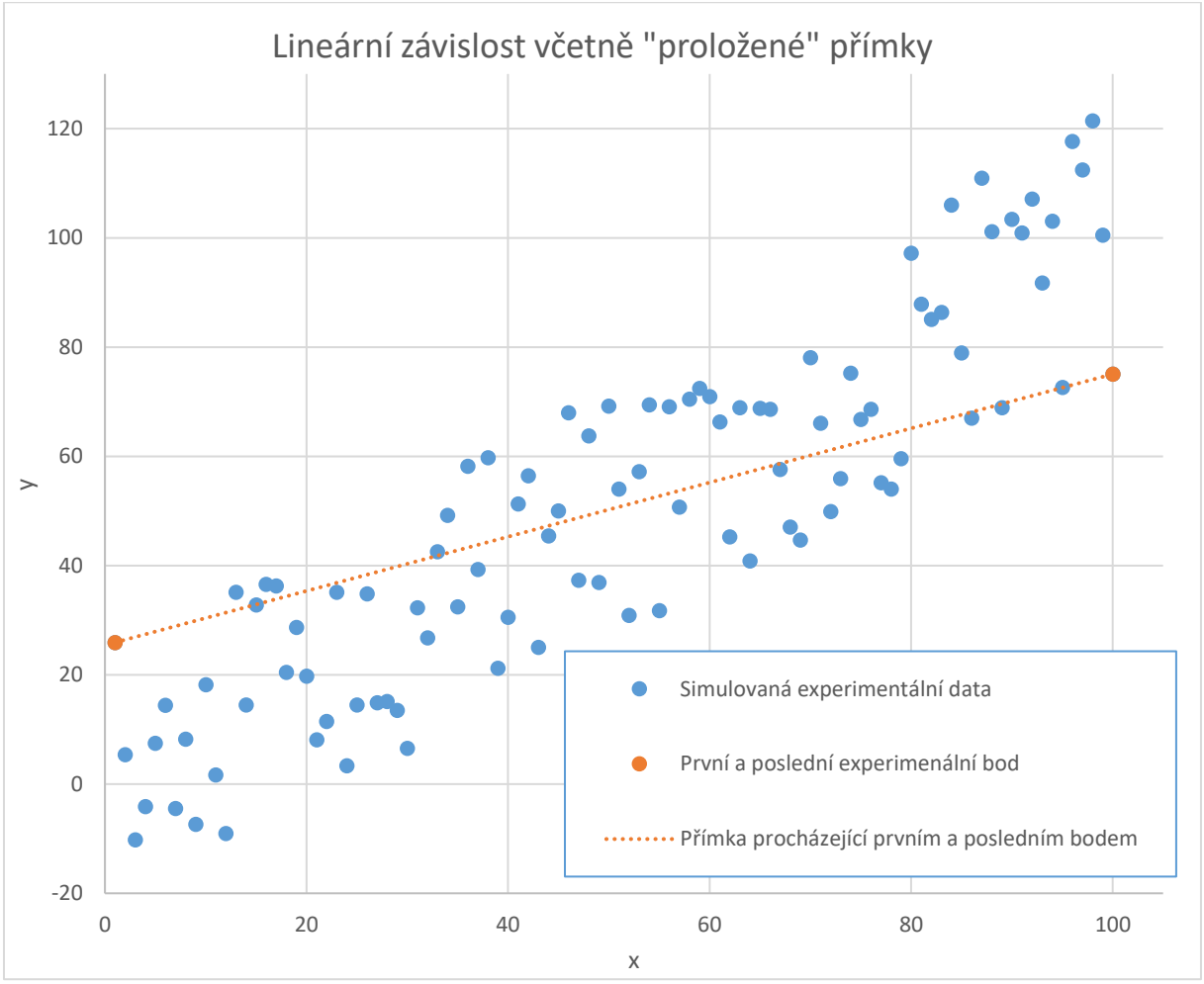
Graf 1

Při pohledu na graf je evidentní, že proložená přímka by měla mít (alespoň přibližně) jednotkovou směrnici. Výše popsané statistické vyhodnocení ale dává podstatně odlišný výsledek. Zvídavý čtenář už pravděpodobně po předchozím varování přišel na to, že ve skutečnosti vůbec nemusíme zdlouhavě průměrovat, abychom došli k výsledné hodnotě k. Když totiž dosadíme vztahy pro jednotlivé směrnice k_i do vzorce pro aritmetický průměr, dostaneme

$$k = \frac{(y_2 - y_1)/\Delta x + (y_3 - y_2)/\Delta x + \cdots + (y_n - y_{n-1})/\Delta x}{n-1}. \qquad (1)$$

Vidíme, že hodnoty y_i s výjimkou první a poslední se ve vztahu vyskytují dvakrát s různým znaménkem, takže z výpočtu vypadnou. Když navíc uvážíme, že $(n-1)$-násobek vzdálenosti Δx ve jmenovateli je právě roven rozdílu poslední a první hodnoty x_i, vidíme, že k je rovna směrnici přímky procházející prvním a posledním měřeným bodem

$$k = \frac{y_n - y_1}{(n-1)\Delta x} = \frac{y_n - y_1}{x_n - x_1}. \qquad (2)$$



Graf 2

V případě výše uvedených simulovaných dat se tedy výsledek podstatně liší od očekávání. Směrnice přímky procházející krajními body měřené série je rovna přibližně jedné polovině (viz *Graf 2*).

Obecně je nutno zdůraznit, že takovéto chování není důsledkem samotného použití průměrování, nýbrž skutečnosti, že průměrované hodnoty v tomto případě nejsou statisticky nezávislé (tj. změna jedné je vždy svázána s odpovídající změnou jiné). Postup také můžeme chápat jako extrémní případ vyhodnocení, které vstupní hodnoty započítává s různou statistickou váhou. Pokud jsou ale všechny hodnoty změřeny se stejnou přesností, pak nemáme důvod některé hodnoty upřednostňovat a měl by být naopak použit postup, kde váha všech hodnot bude stejná.

Metoda postupných měření

V předchozím příkladu bychom například mohli výpočet modifikovat tak, že bychom průměrovali pouze hodnoty k_i s lichým indexem. Přestože tak skutečně získáme $n/2$ nezávislých hodnot, ukážeme si, že ani tento způsob výpočtu není vhodný pro praktické použití.

Důležitou součástí každého experimentálního výsledku je vedle naměřené hodnoty také kvalifikovaně odhadnutá **nejistota** (z anglického **uncertainty**) této hodnoty. Dříve se pro nejistotu používal mnohoznačný pojem chyba, což mohlo vést k nedorozumění. V tomto textu budeme používat pojem chyba (ve statistickém smyslu) pro označení rozdílu mezi skutečně naměřenou hodnotou (odečtenou z přístroje) a správnou (ale neznámou) hodnotou veličiny. Nejistota naopak vymezuje šířku intervalu na obě strany od naměřené hodnoty, uvnitř něhož se správná hodnota s určitou pravděpodobností nachází.

Nejprve rozeberme šíření nejistot při průměrování směrnic spočtených z různě sestavených dvojic. Nejistoty měření budeme charakterizovat **směrodatnou odchylkou** σ (pro kterou ona výše zmíněná pravděpodobnost v případě normálního rozdělení činí zhruba 68 %). Často budeme v tomto textu pracovat přímo s jejím kvadrátem σ^2 a nazývat tuto veličinu **rozptylem**. (Podrobnosti nalezne čtenář v kapitole **Směrodatná odchylka a rozptyl** v části **Dodatek**.) Pro skládání rozptylů σ_j^2 statisticky nezávislých veličin y_j vystupujících ve funkci $f(y_1, y_2, \ldots, y_m)$ platí obecný vztah

$$\sigma_f^2 = \sum_{j=1}^{m} \left(\frac{\partial f}{\partial y_j} \right)^2 \sigma_j^2 . \qquad (3)$$

Výpočet směrnice rozdělíme do dvou kroků: Nejprve se z každé dvojice bodů vypočte směrnice proložené přímky a v druhém kroku se vypočte průměr těchto dílčích směrnic.

Krok 1: Uvažujme dvojici hodnot y_j a y_{j+l}, tedy hodnot v naměřené sadě nikoli nutně sousedních, nýbrž ležících ve vzdálenosti l. Směrnice přímky procházející těmito body bude

$$k_{j,j+l} = \frac{y_{j+l} - y_j}{l\,\Delta x} \tag{4}$$

a pro rozptyl hodnoty příslušné směrnice dostaneme vztah

$$\sigma^2_{k_{j,j+l}} = \frac{\sigma_j^2 + \sigma_{j+l}^2}{(l\,\Delta x)^2}\,. \tag{5}$$

Krok 2: Při výpočtu průměrné hodnoty směrnic k_r pro $r = 1,\ldots,m$ pomocí funkce

$$f(k_1,\ldots,k_m) \equiv \bar{k} = \frac{1}{m}\sum_{r=1}^{m} k_r \tag{6}$$

dostaneme

$$\sigma^2_{f(y_1,\ldots,y_m)} = \frac{1}{m^2}\sum_{r=1}^{m} \sigma_r{}^2\,. \tag{7}$$

Uvedené vztahy se použijí, ať už jsou indexy dvojic bodů $j,\ j+l$ „mapovány“ na pořadové indexy směrnic r jakýmkoli způsobem. Uvažujme dvě různá „mapování“:

A) Celkem $m = \frac{n}{2}$ dvojic je tvořeno sousedními body (vzdálenost $l = 1$). Indexy jsou svázány takto: $j = 2t - 1$ a $r = t$ pro $t = 1,\ldots,m$. Je zřejmé, že jde o výše popsané uspořádání.
$(1,2,\ \ 3,4,\ \ \ldots,\ \ 97,98,\ \ \underbrace{99,100}\,)$

B) Celkem $m = \frac{n}{2}$ dvojic je tvořeno body ve vzdálenosti $l = m$. Indexy jsou svázány takto: $j = t$ a $r = t$ pro $t = 1,\ldots,m$. Body ve dvojicích jsou v největší možné vzdálenosti.
$(1,51,\ \ 2,52,\ \ \ldots,\ \ 49,99,\ \ \underbrace{50,100}\,)$

Předpokládejme, že měření všech hodnot y je zatíženo stejně velkou směrodatnou odchylkou σ. Pak rozptyl střední hodnoty směrnice $\bar{k}$ ve variantě A má pro $\Delta x = 1$ velikost

$$\left(\sigma^2_{\bar{k}}\right)_A = \frac{2\sigma^2}{m}\,. \tag{8}$$

Ve variantě B je to ale jinak, protože tentokrát se ve dvojici vždy odečítají hodnoty ve vzdálenosti $m\,\Delta x$. Důsledkem je m^2-krát menší rozptyl σ_r^2 všech jednotlivých směrnic k_r a také rozptyl $\sigma^2_{\bar{k}}$ jejich střední hodnoty $\bar{k}$

$$\left(\sigma_{\bar{k}}^2\right)_B = \frac{2\sigma^2}{m^3}. \tag{9}$$

Zajímavé je také porovnat tyto dva výsledky se směrodatnou odchylkou směrnice vypočtené podle vztahu (2). Neuplatní se v něm sice průměrování, ale příznivý vliv má skutečnost, že se ve výpočtu použijí nejvzdálenější dostupné body:

$$\sigma_k^2 = \frac{2\sigma^2}{(n-1)^2}. \tag{10}$$

Pro ilustraci uveďme také číselné výsledky pro sadu měření vynesenou v grafech (*Graf 1* a *Graf 2*). Hodnoty proměnné y_i tam byly generovány předpisem $y_i = x_i + 50\left(R - \frac{1}{2}\right)$, kde R představuje náhodné číslo s rovnoměrným rozdělením v intervalu 0 až 1 (funkce NÁHČÍSLO v programu Excel). Náhodná veličina tedy pokrývá rovnoměrně interval $\langle -25, +25\rangle$. Směrodatná odchylka jednotlivého měření tedy odpovídá $\sigma = \frac{25}{\sqrt{3}} \doteq 14.4$ [viz odvození vztahu (124) v části **Dodatek**]. Po dosazení do formulí (8) a (9) dostaneme pro směrodatnou odchylku směrnice vypočtené podle varianty A

$$\left(\sigma_{\bar{k}}\right)_A = \sqrt{\frac{2(25)^2}{3\times 50}} = \frac{5}{\sqrt{3}} \doteq 2.89 \tag{11}$$

a pro směrodatnou odchylku podle varianty B

$$\left(\sigma_{\bar{k}}\right)_B = \sqrt{\frac{2(25)^2}{3\times 50^3}} = \frac{1}{10\sqrt{3}} \doteq 0.0577. \tag{12}$$

Zbývá ještě doplnit směrodatnou odchylku (10) směrnice spočtené oním nevhodným způsobem, který ve skutečnosti vede na jednoduchou formuli (2):

$$\sigma_k = \sqrt{\frac{2(25)^2}{3\times 99^2}} \doteq 0.206. \tag{13}$$

Uvážíme-li, že střední hodnota směrnice je ve všech případech rovna jedné, je zřejmé, že varianta A (navzdory průměrování) poskytuje výsledek s relativní odchylkou 290 %, zatímco jednoduchá varianta využívající jen dva krajní body nabízí relativní odchylku 21 %. Jako optimální se pak jeví varianta B s odchylkou pouze 5.7 %. Není třeba další vysvětlování.

Uspořádání výpočtu podle varianty B se využívá v **metodě postupných měření**. Proložení přímky touto metodou je výpočetně nenáročné. S výhodou lze tento postup využít v situaci, kdy příruční výpočetní technika není k dispozici. Jediným požadavkem je, aby odstupy na ose x byly ekvidistantní. To není obtížné splnit, pokud se na vodorovnou osu vynáší pořadové číslo měření. Další podrobnosti lze nalézt v literatuře [1].

Metody založené na minimalizaci odchylek

V dalších kapitolách si přiblížíme výpočetní metody, které také umožní splnit požadavek, aby všechny body byly započteny se stejnou vahou. „Proložená přímková závislost“ v prvním grafu byla vypočtena jednou z nejznámějších, a to **metodou nejmenších čtverců**. Pro porovnání uveďme, že nalezená směrnice měla hodnotu 0.99 ± 0.05. (Podrobněji to probereme v kapitole **Příklad 1 – různé způsoby práce s odchylkami.**) Metoda nejmenších čtverců tedy není výrazně přesnější ve srovnání s popsanou metodou postupných měření. Nicméně požadavky na uspořádání dat jsou daleko volnější, například nejsou nutné ekvidistantní rozestupy. Navíc popsaný princip umožní prokládat i jiné závislosti než lineární, a přitom zároveň spočítat statistické charakteristiky vstupních a výstupních hodnot. Příbuzné metody založené na minimalizaci veličiny χ^2 (v mluvené řeči často jen „metoda chí-kvadrát“) dovolují započítat také směrodatné odchylky jednotlivých vstupních hodnot, což se při zpracování fyzikálních měření často hodí. Podrobnější výklad metody nejmenších čtverců a jejích aplikací včetně spousty příkladů nalezne čtenář například ve známém Přehledu užité matematiky od profesora Rektoryse [2]. Podrobnější informace o metodách minimalizujících χ^2 zahrnující i způsob implementace pomocí klasických programovacích jazyků zase v Numerical recipes [3].

Programové nástroje na prokládání lineární závislosti

V této kapitole stručně popíšeme základní procedury, které umožní proložit lineární závislost pomocí běžně dostupných programů a programovacích prostředí. Zejména se budeme zabývat programem Excel z balíku Microsoft Office a programem Origin od firmy OriginLab. Také zmíníme volně dostupný balík LibreOffice a volně dostupný program gnuplot. Dále shrneme, co nabízejí základní knihovny programovacího prostředí Python. V popisech obsluhy nezacházíme do podrobností, které se v různých verzích programů mohou lišit. Rovnice pro výpočet regresních parametrů a statistických charakteristik jsou odvozeny v následujících kapitolách.

Funkce LINREGRESE (LINEST) v EXCELu

Jeden z možných způsobů výpočtu parametrů regresní přímky spočívá v užití funkce LINREGRESE (v anglické verzi funkce LINEST). Ta má čtyři parametry. První obsahuje odkaz na hodnoty *y*, druhý pak na hodnoty *x*. Třetí parametr určuje, zda se prokládá obecná přímka (pro hodnotu PRAVDA nebo 1 nebo když parametr není uveden) anebo se prokládá přímka procházející počátkem (pro hodnotu NEPRAVDA). Čtvrtý parametr určuje, zda funkce provede výpočet regresních statistik (pro hodnotu PRAVDA nebo 1).

Předpokládejme, že hodnoty *x* jsou na listu programu EXCEL zapsány ve sloupci A v rozsahu A1:A100 a hodnoty *y* ve sloupci B v položkách B1:B100. Do volné buňky se v tomto případě vloží „=LINREGRESE(B1:B100;A1:A100;;1)“. Regresní statistiky potom obdržíme tak, že po vložení funkce označíme oblast buněk 2 sloupce krát 5 řádků s funkcí v levém horním rohu a stiskneme postupně klávesu F2 a kombinaci Ctrl-Shift-Enter. Ve vyznačené oblasti se objeví deset hodnot, jejichž význam shrnuje následující tabulka. Odkazy na rovnice umožňují nalézt v dalším textu podrobnější popis významu většiny z nich.

- směrnice (32), resp. (91)
- směrodatná odchylka směrnice (51), resp. (92)
- koeficient determinace (58)
- statistika F
- regresní součet čtverců $S_{\mathrm{ESS}} = S_{\mathrm{TSS}} - S_{\mathrm{RSS}}$ (59), (46)
- průsečík s osou ypsilon (33)
- směrodatná odchylka průsečíku (50)
- směrodatná odchylka y_i (38), kde p=2, resp. p=1
- počet stupňů volnosti (37)
- reziduální součet čtverců S_{RSS} (29), (46)

Funkce LINEST v programu LibreOffice CALC

Naprosto stejné služby jako výše popsaná funkce LINREGRESE poskytne funkce LINEST v české i anglické verzi programu CALC z volně dostupného balíku LibreOffice.

Funkce *Analysis:::Linear Fit* programu Origin

Pokud jsou v tabulce vstupních hodnot jen dva sloupce (X a Y), provádí se výpočet podle stejných vztahů jako v Excelu:

- směrnice (*Slope*) podle (32), resp. (91) a její směrodatná odchylka podle (51), resp. (92)
- průsečík s osou ypsilon (*Intercept*) podle (33) a jeho směrodatná odchylka podle (50)
- koeficient determinace [*R-Square (COD)*] podle (58)
- počet stupňů volnosti (*Degrees of Freedom*) podle (37)
- reziduální suma čtverců (*Residual Sum of Squares*) podle (29), resp. (46)

Také je možno na záložce *Fit Control* zafixovat hodnotu průsečíku s osou ypsilon (*Fix Intercept*) nebo směrnice (*Fix Slope*) na zadané hodnotě. To se nejběžněji využije pro prokládání přímky procházející počátkem.

Tabulka vstupních hodnot může obsahovat také tři sloupce (X, Y a ErrY). Hodnoty ve třetím sloupci lze použít pro vyhodnocení regrese s vahou. Pokud třetí sloupec obsahuje směrodatnou odchylku σ_i naměřené hodnoty y_i a parametr *Errors as Weight* je nastaven na hodnotu *Instrumental*, pak se výpočty směrnice a průsečíku s osou ypsilon provedou podle (22). V důsledku toho pak regresní závislost lépe „přiléhá“ k těm naměřeným bodům, jejichž hodnota byla určena přesněji.

Na záložce *Fit Control* se nachází parametr *Scale Error with sqrt(Reduced Chi-Sqr)*, který zásadně ovlivňuje způsob výpočtu odchylek. Pokud zde nepoužíváme hodnoty σ_i jen jako nějaké relativní váhové faktory, nýbrž jde skutečně o směrodatné odchylky jednotlivých měření určené z podmínek experimentu, pak není žádoucí, aby se tyto hodnoty ještě škálovaly na základě reziduálního součtu čtverců. A právě to se děje ve výchozím nastavení, kdy je zatržena *Scale Error with sqrt(Reduced Chi-Sqr)*. **Pokud** naopak **tato volba zatržena není**, budou směrodatné odchylky směrnice a průsečíku s osou ypsilon vypočteny podle vztahů (24) a (25). To znamená, že **individuální směrodatné odchylky jednotlivých vstupních hodnot** y_i **se zahrnou do výsledných směrodatných odchylek regresních koeficientů** pomocí známého vztahu pro skládání směrodatných odchylek.

Funkce *Analysis:::Fit Linear with X Error* programu Origin

Tabulka vstupních hodnot musí obsahovat čtyři sloupce (X, Y, ErrX a ErrY). Třetí a čtvrtý sloupec jsou chápány jako směrodatné odchylky σ_{xi} a σ_{yi} naměřených hodnot x_i a y_i. Na rozdíl od předchozí funkce je takto možno do směrodatných odchylek regresních parametrů započítat navíc také směrodatné odchylky jednotlivých hodnot nezávisle proměnné. Na výběr je několik výpočetních metod. Společným rysem je nutnost prokládat vždy obecnou rovnici přímky. Výpočetní vztahy se liší podle použité metody. Ve výchozím nastavení se použije metoda Yorkova. Detaily je možno nalézt v příslušné literatuře [3,4] a v nápovědě programu.

Příkaz *stats* programu Gnuplot

Program gnuplot verze 5.4 obsahuje proložení lineární regrese jako součást příkazu ***stats***. Povinným parametrem příkazu je jméno souboru obsahujícího dva sloupce čísel – první se chápe jako hodnoty x_i a druhý jako y_i. Příkaz vyhodnotí odděleně základní statistické charakteristiky obou datových sloupců a metodou nejmenších čtverců vypočte parametry regresní přímky

- směrnici přímky (*slope*) podle (32)
- směrodatnou odchylku směrnice (*slope_err*) podle (51)
- průsečík se svislou osou (*intercept*) podle (33)
- směrodatnou odchylku průsečíku (*intercept_err*) podle (50)
- Pearsonův korelační koeficient výběru (*correlation*) podle (57)

Další podrobnosti použití příkazu ***stats*** lze nalézt v manuálu k programu gnuplot[5].

Funkce *scipy.stats.linregress* v programovacím prostředí Python

V dokumentaci ke knihovně SciPy verze 1.14.1 se uvádí volání funkce se třemi vstupními parametry [6]

***linregress*(*x*, *y*=*None*, *alternative*='*two-sided*')**

Nezbytnými vstupními parametry jsou jednorozměrná pole ***x*** a ***y*** obsahující naměřené hodnoty x_i a y_i. Ohledně správného formátování polí doporučuji řídit se dokumentací, protože funkce nemusí nutně hlásit chybu, pokud je struktura polí jiná než předepsaná. V takovém případě některé vypočtené parametry dokonce mohou i vycházet správně, zatímco jiné jsou chybné nebo jejich výpočet skončí na numerické chybě typu dělení nulou apod.

Hodnota parametru ***alternative*** je irelevantní pro výpočet regresních koeficientů, jejich směrodatných odchylek a korelačního koeficientu.

Funkce vrací šestici reálných hodnot:

- směrnice regresní přímky (***slope***) podle (32)
- průsečík s osou ypsilon (***intercept***) podle (33)
- Pearsonův korelační koeficient (***rvalue***) podle (57)
- hodnotu ***pvalue*** (pro detaily viz dokumentaci knihovny SciPy[6])
- směrodatná odchylka směrnice (***stderr***) podle (51)
- směrodatná odchylka průsečíku s osou ypsilon (***intercept_stderr***) podle (50)

Funkce ***linregress*()** využívá (přinejmenším ve verzi 1.14.1) metodu nejmenších čtverců a neumožňuje tedy započítat váhy nebo rozptyly naměřených hodnot. Také nenabízí možnost hledat regresní přímku procházející počátkem.

Funkce *numpy.polynomial.polynomial.polyfit* v programovacím prostředí Python

V knihovně NumPy verze 2.2 se funkce volá se šesti vstupními parametry [7]

***polyfit(x, y, deg, rcond**=None, **full**=False, **w**=None)*

Nezbytná jsou pole ***x*** a ***y*** obsahující naměřené hodnoty x_i a y_i. Pole ***x*** je vždy jednorozměrné, zatímco pole ***y*** může v jednotlivých sloupcích obsahovat nezávislé sady hodnot y_i pro sdílené hodnoty x_i. Jedno volání funkce tak může nezávisle proložit větší počet nezávislých polynomů. Cílem je omezit zbytečné opakování výpočtu stejných pracovních proměnných.

Parametr ***deg*** slouží k zadání stupně prokládaného polynomu. V případě prokládání obecné přímky se použije ***deg**=[0,1]* nebo jen ***deg**=1*, pro proložení přímky procházející počátkem ***deg**=[1]*.

Parametr ***rcond*** modifikuje zacházení s reálnými čísly ve výpočtu. Při běžném použití by měla postačovat výchozí hodnota. Podrobnosti lze nalézt v dokumentaci ke knihovně NumPy [7].

Parametr ***full*** zapíná a vypíná výstup podrobnějších diagnostických informací. Ve výchozím nastavení je tento výstup vypnut.

Parametr ***w*** slouží k případnému nastavení statistických vah pro jednotlivé body. Není-li použit, je všem bodům přiřazena jednotková váha. Pokud parametr ***w*** použit je, pak to musí být jednorozměrné pole stejné délky jako ***x***. Položky představují statistickou váhu w_i pro započtení rezidua $\Delta y_{\mathrm{res},i}$ dvojice x_i a y_i do reziduálního součtu podle vztahu $S_{\mathrm{RSS}} = \sum_i (w_i\, \Delta y_{\mathrm{res},i})^2$.

Pozor, statistická váha w_i se v této knihovně chápe jinak než v rovnici (161)! Zde se jí násobí ještě před provedením kvadrátu. Pro podrobnosti viz dokumentaci ke knihovně [7].

Funkce vrací

- vypočtené koeficienty polynomů uspořádané od nejnižšího řádu k nejvyššímu v poli dimenze (***deg***+1) × (počet sad naměřených bodů), v případě lineární regrese bez použití váhy jsou to tedy průsečík se svislou osou (33) a směrnice (32), resp. (91) pro každou sadu hodnot; je-li použita váha $w_i = \frac{1}{\sigma_i}$, mají průsečík a směrnice hodnoty podle (22)
- pro vstupní parametr ***full**=True* navíc vrací některé diagnostické mezivýsledky, z nichž v tomto textu je dále vysvětlen pouze ten úplně první – reziduální součet čtverců S_{RSS} (i se započtenými vahami) pro fit metodou nejmenších čtverců

Funkce ***polyfit()*** používá pro prokládání polynomů maticovou rovnici, jejímž základem je tzv. Vandermondova matice. Řešení se hledá metodou nejmenších čtverců. V případě prokládání lineární závislosti se výpočet redukuje na vztahy uvedené výše. Na rozdíl od ***linregress()*** umožňuje ***polyfit()*** prokládat polynomy vyššího stupně a také dává možnost při prokládání jednotlivým bodům přisoudit individuální statistickou váhu. Na druhou stranu nepočítá směrodatné odchylky vypočtených koeficientů polynomu.

Třída *sklearn.linear_model.LinearRegression* v programovacím prostředí Python

V knihovně scikit-learn verze 1.6.0 se tato třída inicializuje až čtyřmi parametry:

LinearRegression(*, fit_intercept**=True**, copy_X**=True**, n_jobs**=None**, positive**=False**)

Pro běžné výpočty regrese je důležitý jen ***fit_intercept***, který přepíná mezi obecnou lineární regresí (výchozí hodnota) a lineární regresí procházející počátkem. Pro ostatní parametry viz dokumentaci knihovny scikit-learn[8].

Výpočet parametrů regrese se provádí metodami ***fit()***, ***coef_***, ***intercept_*** a ***score()***.

Nejprve je nutno zavolat metodu ***fit(X, y, sample_weight**=None**)***.

Nezbytná jsou pole ***X*** a ***y*** obsahující naměřené hodnoty x_i a y_i. Pokud se omezíme se na prokládání jedné závislosti jedné proměnné, pak obě pole obsahují jeden sloupec hodnot. Pro detaily uspořádání obou polí viz dokumentaci ke knihovně.

Parametr ***sample_weight*** je určen pro případné nastavení statistických vah jednotlivých bodů. Není-li použit, je všem bodům přiřazena jednotková váha. Pokud parametr ***sample_weight*** použit je, pak to musí být jednorozměrné pole stejné délky jako ***X***. Položky představují statistickou váhu w_i pro započtení kvadrátu rezidua $\Delta y_{\mathrm{res},i}$ dvojice x_i a y_i do reziduálního součtu podle vztahu $S_{\mathrm{RSS}} = \sum_i w_i \, \Delta y_{\mathrm{res},i}^2$, což je v souladu s (161).

Po metodě ***fit()*** je možno zavolat ***coef_*** a ***intercept_***, které jako funkční hodnoty vracejí bez použití statistických vah směrnici (32), resp. (91) a průsečík se svislou osou (33), při použití vah $w_i = \frac{1}{\sigma_i^2}$ pak průsečík a směrnice mají hodnoty podle (22).

Po provedení ***fit()*** je možno pomocí ***score(X, y, sample_weight**=None**)*** se stejnými parametry získat také hodnotu koeficientu determinace (58).

Třída ***sklearn.linear_model.LinearRegression*** používá pro prokládání polynomů maticovou rovnici řešenou metodou nejmenších čtverců. V případě prokládání lineární závislosti se výpočet redukuje na vztahy uvedené výše. Na rozdíl od ***linregress()*** umožňuje jednotlivým bodům přisoudit individuální statistickou váhu, ale nepočítá směrodatné odchylky vypočtených koeficientů.

Třída *statsmodels.regression.linear_model.OLS* v programovacím prostředí Python

OLS umožní vypočítat parametry regresní přímky, jejich směrodatné odchylky, koeficient determinace a spoustu dalších statistických ukazatelů metodou nejmenších čtverců. Jde o komplikovaný nástroj pro statistické zpracování. Podrobnosti naleznete v dokumentaci [9].

Třída *statsmodels.regression.linear_model.WLS* v programovacím prostředí Python

WLS rozšiřuje možnosti výše zmiňované třídy o možnost započítat statistickou váhu při výpočtu reziduálního součtu čtverců. Důsledkem je tedy uplatnění metody vážených nejmenších čtverců. Váhy se uplatní při výpočtu parametrů regresní přímky, ale jejich směrodatné odchylky to neovlivní (využije se odhad výběrového rozptylu z reziduálního součtu, a nikoliv rozptyly použité pro výpočet vah). Procedura nenabízí žádný parametr, který by toto chování umožnil změnit [10].

Shrnutí

Všechny popisované programy nabízí nástroj, který metodou nejmenších čtverců nalezne parametry obecné lineární regrese včetně jejich směrodatných odchylek. Ne všechny už ale umožňují prokládat přímku procházející počátkem. Některé programy navíc umožní použít regresi s váhou, případně polynomiální regresi. Ve většině případů jsou směrodatné odchylky koeficientů regrese odhadnuty výhradně metodou nejmenších čtverců nebo minimalizací χ^2. Možnost stanovit je na základě zadaných směrodatných odchylek jednotlivých vstupních bodů nabízí pouze program Origin.

Všechny uvedené programy také obsahují možnost naprogramovat si potřebné výpočty. Následující kapitoly pro takový případ poskytují přehled potřebných rovnic.

Teoretické základy regresního počtu

Obecné předpoklady prokládání regresní funkce

Mějme n empiricky získaných bodů závislosti proměnné y na nezávisle proměnné x označených $[x_1, y_1], [x_2, y_2], \ldots [x_n, y_n]$. Předpokládáme, že mezi proměnnými x a y platí funkční vztah $y = f(x)$, kde $f(x)$ je funkce známého tvaru (např. lineární). Protože měření je zatíženo chybami, jsou naměřené hodnoty kolem křivky $y = f(x)$ rozptýleny, takže platí $y_i = f(x_i) + \varepsilon_i$, kde ε_i označuje chybu *i*-tého měřeného bodu.

Funkce $f(x)$ obsahuje obecně $p\,(p < n)$ parametrů, které označíme $b_1, b_2, \ldots, b_p$, takže můžeme psát $f(x) = f(x; b_1, b_2, \ldots, b_p)$. Příkladem je lineární funkce $f(x) = b_1 x + b_2$ se dvěma parametry. Měření y_i je pak možno vyjádřit rovnicí

$$y_i = f(x_i; b_1, b_2, \ldots, b_p) + \varepsilon_i \quad (i = 1, 2, \ldots, n), \tag{14}$$

kde $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ jsou chyby měření, které pokládáme za náhodné veličiny. Neklademe požadavek, aby se všechny $x_1, x_2, \ldots, x_n$ navzájem lišily. Pokud $x_i = x_k$ pro $i \neq k$, pak podle (14) platí $y_i - y_k = \varepsilon_i - \varepsilon_k$, takže dvě měření příslušná téže hodnotě x jsou obecně různá. Rovnice (14) se nazývá **určující rovnice měření** y_i.

Prokládání křivky měřenými body znamená, že se statisticky odhadují neznámé parametry $b_1, b_2, \ldots, b_p$, aby křivka k bodům co nejlépe přiléhala (v angličtině se používá sloveso „fit“). Odhad parametru b_j označíme b_j^*. Funkce $y = f(x; b_1^*, b_2^*, \ldots, b_p^*)$ se označuje jako **(empirická) regresní funkce** a její grafické znázornění jako **(empirická) regresní křivka**.

Způsob určení odhadů b_j^* závisí na zvoleném kritériu pro „přiléhavost křivky k bodům“, což se provede volbou tzv. věrohodnostní funkce (likelyhood function).

Minimalizace funkce χ^2

Ukážeme si nejprve, jak funguje jeden z běžně používaných postupů prokládání regresní funkce. Mírně předtím doplníme předpoklady týkající se statistického rozdělení náhodných veličin $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Budeme nadále předpokládat, že všechny náhodné veličiny mají nulovou střední hodnotu a nejsou korelovány. Navíc budeme uvažovat, že *i*-tá náhodná veličina ε_i se bude řídit normálním rozdělením s rozptylem velikosti σ_i^2. (Dále si ukážeme výpočetní postupy, kde hodnoty σ_i^2 budou podle potřeby použity jako vstupní nebo jako výstupní parametry výpočtu.)

Jako věrohodnostní funkci popisující konkrétní výběr měřených hodnot potom můžeme použít jeho statistickou pravděpodobnost (jako parametry použijeme hodnoty σ_i^2 a hledanou regresní funkci obsahující neznámé parametry $b_1, b_2, \ldots, b_p$). Pak budeme hledat takovou sadu parametrů $b_1^*, b_2^*, \ldots, b_p^*$, která tuto pravděpodobnost maximalizuje.

Hustota pravděpodobnosti, že pro *i*-tý bod výběru bude změřena právě hodnota y_i, bude podle předpokladu dána normálním rozdělením

$$\rho(y_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{\left[y_i - f(x_i; b_1, b_2, \ldots, b_p)\right]^2}{2\sigma_i^2}} . \tag{15}$$

Pravděpodobnost, že celý výběr bude obsahovat právě hodnoty $y_1, y_2, \ldots, y_n$, pak bude úměrná součinu

$$P(y_1, y_2, \ldots, y_n) \propto \prod_{i=1}^{n} e^{-\frac{\left[y_i - f(x_i; b_1, b_2, \ldots, b_p)\right]^2}{2\sigma_i^2}} . \tag{16}$$

Součin lze upravit do tvaru jediné exponenciely

$$P(y_1, y_2, \ldots, y_n) \propto e^{-\frac{1}{2}\sum_{i=1}^{n} \frac{\left[y_i - f(x_i; b_1, b_2, \ldots, b_p)\right]^2}{\sigma_i^2}} . \tag{17}$$

Hledání maxima pravděpodobnosti $P(y_1, y_2, \ldots, y_n)$ lze pak převést na hledání minima hodnoty součtu v exponentu (17). Pro konkrétní výběr experimentálních hodnot budou odhady neznámých parametrů $b_1^*, b_2^*, \ldots, b_p^*$ výsledkem minimalizace součtu, pro který se zavádí označení symbolem χ^2,

$$\chi^2 \equiv \sum_{i=1}^{n} \frac{\left[y_i - f(x_i; b_1^*, b_2^*, \ldots, b_p^*)\right]^2}{\sigma_i^2} . \tag{18}$$

Budeme další postup ilustrovat na případu prokládání lineární závislosti $y = a + bx$. Odhady parametrů a^* a b^* se v tomto případě určí minimalizací výrazu

$$\chi^2 = \sum_{i=1}^{n} \frac{\left[y_i - a^* - b^* x_i\right]^2}{\sigma_i^2} = \min \tag{19}$$

a vypočtou se tedy z rovnic $\frac{\partial \chi^2}{\partial a^*} = 0$ a $\frac{\partial \chi^2}{\partial b^*} = 0$. Tak dostaneme podmínky

$$\begin{aligned} -\frac{1}{2}\frac{\partial \chi^2}{\partial a^*} &= \sum_{i=1}^{n} \frac{(y_i - a^* - b^* x_i)}{\sigma_i^2} = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} - a^* \sum_{i=1}^{n} \frac{1}{\sigma_i^2} - b^* \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} = 0 \\ -\frac{1}{2}\frac{\partial \chi^2}{\partial b^*} &= \sum_{i=1}^{n} \frac{(y_i - a^* - b^* x_i) x_i}{\sigma_i^2} = \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2} - a^* \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} - b^* \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} = 0 \end{aligned} \tag{20}$$

Soustava rovnic pro výpočet a^* a b^* (označují se jako **normální rovnice**) bude mít tvar

$$\begin{aligned} a^* \underbrace{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}_{S} \quad + \quad b^* \underbrace{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}_{S_x} \quad = \quad \underbrace{\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}}_{S_y} \\ a^* \underbrace{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}_{S_x} \quad + \quad b^* \underbrace{\sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2}}_{S_{xx}} \quad = \quad \underbrace{\sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2}}_{S_{xy}} \end{aligned} \tag{21}$$

Řešením jsou odhady a^* a b^* v kompaktním tvaru zapsaném pomocí zavedených součtů

$$\begin{aligned} a^* &= \frac{S_{xx} S_y - S_x S_{xy}}{S S_{xx} - S_x^2} \\ b^* &= \frac{S S_{xy} - S_x S_y}{S S_{xx} - S_x^2} \end{aligned} \tag{22}$$

Určení přesnosti odhadů regresních parametrů

Odhady b_j^* jsou funkcemi veličin $y_1, y_2, \ldots, y_n$, které jsou náhodnými veličinami. Proto jsou odhady b_j^* také náhodnými veličinami a mají své pravděpodobnostní rozdělení. Rozptyl odhadu b_j^*, který označíme $\sigma_{b_j^*}^2$, závisí na rozptylech σ_i^2 a je tedy také neznámým parametrem. Jeho odhad označíme $s_{b_j^*}^2$. Veličina $s_{b_j^*}$ představuje odhad směrodatné odchylky odhadu b_j^*. (V tomto textu – s výjimkou příkladů – budeme označovat písmenem „s“ latinské abecedy odhady rozptylů a směrodatných odchylek vypočtené statisticky na základě výběru naměřených hodnot. Řeckým „sigma“ naopak budeme označovat rozptyly a směrodatné odchylky, které se ve výpočtech vyskytují jako vstupní parametry a mohou tedy být zadány přesně, i když ve statistických výpočtech jsou běžně i tyto hodnoty založeny na odhadech. Pokud to nepovede k nedorozumění, budeme v dalším textu namísto „odhad směrodatné odchylky“ psát jen „směrodatná odchylka“. Podobně budeme zacházet s označením „odhad rozptylu“.)

Dalším krokem je vyjádření směrodatných odchylek zmíněných odhadů pomocí skládání rozptylů na základě vztahu (3). Budeme předpokládat, že pro odhady rozptylů lze použít stejné vztahy jako pro samotné rozptyly.

Navažme na výpočet pro lineární funkci z předchozí kapitoly. Na základě vztahů (22) získáme pro derivace odhadů a^* a b^* podle *i*-té závisle proměnné vyjádření

$$\frac{\partial a^*}{\partial y_i} = \frac{S_{xx} \frac{1}{\sigma_i^2} - S_x \frac{x_i}{\sigma_i^2}}{S S_{xx} - S_x^2}$$

$$\frac{\partial b^*}{\partial y_i} = \frac{S\dfrac{x_i}{\sigma_i^2} - S_x\dfrac{1}{\sigma_i^2}}{SS_{xx} - S_x^2} \tag{23}$$

Pak pro odchylku s_{a^*} dostaneme podle (3) vyjádření

$$\begin{aligned} s_{a^*}^2 &= \sum_{i=1}^{n}\left(\frac{\partial a^*}{\partial y_i}\right)^2\sigma_i^2 = \sum_{i=1}^{n}\left(\frac{S_{xx}\dfrac{1}{\sigma_i^2} - S_x\dfrac{x_i}{\sigma_i^2}}{SS_{xx} - S_x^2}\right)^2\sigma_i^2 = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(\frac{S_{xx} - S_x x_i}{SS_{xx} - S_x^2}\right)^2 = \\ &= \sum_{i=1}^{n}\frac{1}{\sigma_i^2}\frac{S_{xx}^2 - 2S_{xx}S_x x_i + S_x^2 x_i^2}{\left(SS_{xx} - S_x^2\right)^2} = \frac{SS_{xx}^2 - 2S_{xx}S_x^2 + S_x^2 S_{xx}}{\left(SS_{xx} - S_x^2\right)^2} = \frac{SS_{xx}^2 - S_{xx}S_x^2}{\left(SS_{xx} - S_x^2\right)^2} = \frac{S_{xx}}{SS_{xx} - S_x^2} \end{aligned} \tag{24}$$

Podobně pro odchylku s_{b^*} dostaneme

$$\begin{aligned} s_{b^*}^2 &= \sum_{i=1}^{n}\left(\frac{\partial b^*}{\partial y_i}\right)^2\sigma_i^2 = \sum_{i=1}^{n}\left(\frac{S\dfrac{x_i}{\sigma_i^2} - S_x\dfrac{1}{\sigma_i^2}}{SS_{xx} - S_x^2}\right)^2\sigma_i^2 = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(\frac{Sx_i - S_x}{SS_{xx} - S_x^2}\right)^2 = \\ &= \sum_{i=1}^{n}\frac{1}{\sigma_i^2}\frac{S^2x_i^2 - 2Sx_iS_x + S_x^2}{\left(SS_{xx} - S_x^2\right)^2} = \frac{S^2S_{xx} - 2SS_x^2 + SS_x^2}{\left(SS_{xx} - S_x^2\right)^2} = \frac{S^2S_{xx} - SS_x^2}{\left(SS_{xx} - S_x^2\right)^2} = \frac{S}{SS_{xx} - S_x^2} \end{aligned} \tag{25}$$

Pro úplnost ještě uveďme bez odvození vztah pro kovarianci (její definici zmíníme později) mezi koeficienty a^* a b^* (ty totiž na rozdíl od vstupních hodnot mohou být závislé)

$$\mathrm{Cov}(a^*, b^*) = -\frac{S_x}{SS_{xx} - S_x^2} \tag{26}$$

Další postup se liší podle toho, zda hodnoty rozptylů σ_i^2 jsou předem známy anebo zda jejich určení bude součástí výpočtu.

Postup pro případ, kdy hodnoty rozptylu vstupních hodnot jsou známy

Často stojíme před zadáním, kdy z povahy experimentu známe rozptyl popisující jednotlivé měřené body. Tento rozptyl přirozeně může být pro každý z bodů odlišný. Při prokládání je pak vhodné přikládat vyšší váhu datům, která byla změřena přesněji. Zároveň je žádoucí, aby známé nejistoty jednotlivých bodů byly započteny do nejistot odhadů regresních parametrů. Tomuto zadání vyhovuje právě výše popsaná metoda založená na **minimalizaci funkce** χ^2. V případě lineární regrese vyjdou její parametry z rovnic (22) a odhady rozptylu parametrů jsou dány rovnicemi (24) a (25).

Hodnota funkce χ^2 v nalezeném minimu podléhá statistickým zákonitostem, takže je možno pro daný výběr měřených bodů usoudit, že regresní funkce je „ta správná“ (fit is good) a

hodnoty rozptylu přiřazené naměřeným bodům nejsou ani moc malé (což by znamenalo, že nejistoty někdo podcenil) ani moc velké (což by naopak znamenalo, že buď někdo nejistoty raději přecenil, anebo že jsou data upravená). Detailněji se k tomu vrátíme později.

Postup pro neznámý rozptyl vstupních hodnot

V tomto případě je nejjednodušší předpokládat, že pro všechny hodnoty y_i mají rozptyly σ_i^2 stejnou hodnotu $\sigma_i^2 \equiv \sigma^2$. V tom případě se minimalizovaný výraz (19) zredukuje na tvar

$$\sum_{i=1}^{n} (y_i - a^* - b^* x_i)^2 = \min ,$$

který je východiskem pro tzv. **metodu nejmenších čtverců**. Ta umožňuje proložení regresní závislosti experimentálními body za předpokladu, že všechny jsou zatíženy náhodnou chybou charakterizovanou stejným rozptylem, jehož velikost neznáme. Z toho plyne, že všechny body mají při prokládání stejnou váhu. Zároveň je implicitním předpokladem, že regresní funkce, kterou prokládáme, je „ta správná“, tj. přesně popisuje změřenou závislost a jenom kvůli náhodným chybám na ní všechny body neleží. Výsledkem proložení jsou pak nejen odhady parametrů regresní funkce, ale také odhad neznámého rozptylu a odhady nejistot stanovených parametrů. Různé aspekty metody nejmenších čtverců jsou podrobně rozebrány v dalších kapitolách.

Metoda nejmenších čtverců obecně

V metodě nejmenších čtverců se jako kritérium uvažuje součet čtverců rozdílů $y_i - f(x_i; b_1^*, b_2^*, \ldots, b_p^*)$ (označují se jako **rezidua**) a odhady $b_1^*, b_2^*, \ldots, b_p^*$ se určí jako parametry, které tento součet minimalizují. Označíme-li součet čtverců reziduí (též **reziduální součet čtverců; residual sum of squares**)

$$S_{\mathrm{RSS}} = \sum_{i=1}^{n} [y_i - f(x_i; b_1^*, b_2^*, \ldots, b_p^*)]^2 , \tag{27}$$

pak budou odhady b_j^* ($j = 1, 2, \ldots, p$) určeny z podmínky

$$S_{\mathrm{RSS}} = \min . \tag{28}$$

O křivce s rovnicí $y = f(x; b_1^*, b_2^*, \ldots, b_p^*)$, kde byly neznámé parametry určeny z podmínky (28), říkáme, že byla **proložena metodou nejmenších čtverců**.

Prokládání regresní závislosti metodou nejmenších čtverců často používáme v situacích, kdy nějakou skupinou bodů chceme proložit křivku s rovnicí známého tvaru, aniž bychom uvažovali nad tím, zda chybové veličiny ε_i mají skutečně náhodný charakter, případně jaký

mají na výsledek prokládání vliv statistické charakteristiky veličin ε_i nebo proč se použije právě kritérium nejmenších čtverců.

Základní úloha – proložení regresní přímky (aneb jak to kdysi dělaly kalkulačky)

Předpokládáme mezi proměnnými x a y platnost vztahu $y = a + bx$, kde a a b jsou neznámé parametry. Úlohou tedy bude proložit body $[x_1, y_1], [x_2, y_2], \ldots [x_n, y_n]$ metodou nejmenších čtverců přímku. Podle rovnic (27) a (28) budou odhady a^* a b^* určeny z podmínky

$$S_{\mathrm{RSS}} = \sum_{i=1}^{n} (y_i - a^* - b^* x_i)^2 = \min \tag{29}$$

a vypočtou se tedy z rovnic $\frac{\partial S_{\mathrm{RSS}}}{\partial a^*} = 0$ a $\frac{\partial S_{\mathrm{RSS}}}{\partial b^*} = 0$.

Dostaneme podmínky

$$\begin{aligned} -\frac{1}{2}\frac{\partial S_{\mathrm{RSS}}}{\partial a^*} &= \sum_{i=1}^{n} (y_i - a^* - b^* x_i) = \sum_{i=1}^{n} y_i - n a^* - (\sum_{i=1}^{n} x_i) b^* = 0 \\ -\frac{1}{2}\frac{\partial S_{\mathrm{RSS}}}{\partial b^*} &= \sum_{i=1}^{n} (y_i - a^* - b^* x_i) x_i = \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i) a^* - (\sum_{i=1}^{n} x_i^2) b^* = 0 \end{aligned} \tag{30}$$

Tak je možno sestavit soustavu rovnic (tzv. **normální rovnice**) pro výpočet a^* a b^*

$$\begin{aligned} n a^* &+ (\sum_{i=1}^{n} x_i) b^* &&= \sum_{i=1}^{n} y_i \\ (\sum_{i=1}^{n} x_i) a^* &+ (\sum_{i=1}^{n} x_i^2) b^* &&= \sum_{i=1}^{n} x_i y_i \end{aligned} \tag{31}$$

Řešením jsou odhady a^* a b^*

$$b^* = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{32}$$

$$a^* = \frac{\sum_{i=1}^{n} y_i - b^* \sum_{i=1}^{n} x_i}{n} \tag{33}$$

Pozorný čtenář jistě zaznamenal, že rovnice (31) bychom obdrželi z rovnic (21), pokud rozptyl σ_i^2 pro všechna i by byl roven stejné konstantě. V tom případě se odhady a^* a b^* podle (22) zjednoduší právě na tvar (33) resp. (32).

Pro zajímavost si rovněž povšimněme, že popsaný výpočet lineární regrese dovoluje při postupném zadávání dvojic hodnot x_i a y_i šetřit potřebnou paměť tím, že se pouze drží v paměti a průběžně aktualizuje 5 proměnných $n, \sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_i^2$ a $\sum_{i=1}^{n} x_i y_i$. Takto bylo možno prokládat regresní přímku neomezeným počtem bodů už v dobách, kdy kalkulačky měly k dispozici jen několik pamětí (přeloženo do dnešních pojmů stačila paměť 40 bajtů).

Co ještě lze zjistit metodou nejmenších čtverců

Proložením regresní křivky nejsou zdaleka využity všechny informace, které lze pomocí metody nejmenších čtverců získat. Důležitá je totiž samotná minimální hodnota reziduálního součtu čtverců. Mají na ni vliv zejména dvě skutečnosti – jak dobře model souhlasí s experimentem a jak velké jsou chyby měření.

Jednoduše řečeno, vhodnost modelu (goodness of fit) je míra toho, jak dobře statistický model vyhovuje pozorovaným datům. Předpokládejme v této kapitole, že použití určité regresní funkce je vhodné (fit is good) a všechny náhodné veličiny $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ mají nulovou střední hodnotu, nejsou korelovány a mají všechny stejný rozptyl neznámé velikosti σ^2 (nepožaduje se normální rozdělení).

Dokonce i za těchto velmi obecných předpokladů je možno pomocí metody nejmenších čtverců vedle parametrů regresní funkce současně nalézt odhad onoho neznámého společného rozptylu a také charakterizovat přesnost stanovení parametrů regrese.

Nejprve stručně shrneme postup vyhodnocení statistických parametrů pro obecný případ regresní závislosti ve tvaru polynomu stupně $p-1$, který obsahuje p neznámých konstant b_j^* $(j = 1, 2, \ldots, p)$. Nejedná se o odvození, ta by byla příliš zdlouhavá. Později se k důležitým krokům vrátíme a doplníme některá vysvětlení.

Odhad rozptylu σ^2, který označíme s^2, je dán vztahem

$$s^2 = \frac{S_{\mathrm{RSS}}}{n-p}. \tag{34}$$

Reziduální součet čtverců není nezbytné počítat pomocí jednotlivých reziduí a jejich čtverců. Lze odvodit důležitý vzorec

$$S_{\mathrm{RSS}} = \sum_{i=1}^{n} y_i^2 - \sum_{j=1}^{p} b_j^* r_j, \tag{35}$$

kde r_j jsou pravé strany normálních rovnic (viz sadu rovnic (31) sestavenou pro lineární regresi)

$$r_j = \sum_{i=1}^{n} x_i^{j-1} y_i. \tag{36}$$

Číslo

$$\nu = n - p \tag{37}$$

určuje **počet stupňů volnosti reziduálního součtu čtverců**. Použitím počtu stupňů volnosti namísto pouhého *n* ve vztahu (34) získáme tzv. **nevychýlený** odhad rozptylu. (Vrátíme se k tomuto pojmu později.) Veličina

$$s = \sqrt{\frac{S_{\mathrm{RSS}}}{n - p}} \tag{38}$$

představuje statistický odhad **směrodatné odchylky** kteréhokoli měření y_i.

Jak přesný je odhad parametrů metodou nejmenších čtverců

Zavedení veličin v tomto odstavci je prakticky totožné, jako v kapitole **Určení přesnosti odhadů regresních parametrů**. Tentokrát je ale náhodné rozložení měřených hodnot určeno společným parametrem. Odhady b_j^* jsou funkcemi veličin $y_1, y_2, \ldots, y_n$, které jsou náhodnými veličinami. Proto jsou odhady b_j^* také náhodnými veličinami a mají své pravděpodobnostní rozdělení. Rozptyl odhadu b_j^*, který označíme $\sigma^2_{b_j^*}$, závisí na neznámém rozptylu σ^2 a je tedy také neznámým parametrem. Jeho odhad označíme $s^2_{b_j^*}$. Veličina $s_{b_j^*}$ představuje odhad směrodatné odchylky odhadu b_j^*.

Obecný postup určení odhadů $s^2_{b_j^*}$ můžeme chápat jako výpočet šíření nejistot založený na vztahu (3) pro skládání rozptylů σ_j^2 statisticky nezávislých veličin y_j vystupujících ve funkci $f(y_1, y_2, \ldots, y_n)$. Pro odhady rozptylů se použijí stejné vztahy jako pro samotné rozptyly. Potřebné derivace podle všech nezávislých veličin y_j je nutno spočítat ze známého vzorce pro příslušný odhad b_j^*. Ilustrací tohoto postupu bylo nalezení vzorců (24) a (25).

Efektivnější postup pro nalezení odhadů $s^2_{b_j^*}$ využívá normálních rovnic. Ty lze v maticovém tvaru zapsat takto

$$\mathbf{A}\mathbf{b}^* = \mathbf{r}, \tag{39}$$

kde $\mathbf{b}^*$ a $\mathbf{r}$ představují vektory odhadů b_j^* a vektor pravé strany normálních rovnic r_j a matice $\mathbf{A}$ obsahuje koeficienty jejich levé strany. Označme a^{kl} $(k, l = 1, 2, \ldots, p)$ prvky matice $\mathbf{A}^{-1}$.

Pak lze odvodit, že

$$s^2_{b_j^*} = s^2 a^{jj} \tag{40}$$

resp.

$$s_{b_j^*} = s\sqrt{a^{jj}}\ . \tag{41}$$

Na rozdíl od vstupních hodnot, které jsou podle předpokladu statisticky nezávislé, můžou mezi odhady b_j^* existovat korelace. Má proto smysl uvést na tomto místě také vztah pro kovarianci odhadů b_j^* a b_k^*

$$\mathrm{Cov}(b_j^*, b_k^*) = s^2\, a^{jk}\ . \tag{42}$$

Můžeme rovněž psát

$$s_{b_j^*}^2 = \mathrm{Cov}(b_j^*, b_j^*)\,, \tag{43}$$

což je v souladu s definicí kovariance dvojice veličin x a y

$$\mathrm{Cov}(x, y) = \left\langle \left(x - \langle x\rangle\right)\left(y - \langle y\rangle\right)\right\rangle, \tag{44}$$

kde $\langle A\rangle$ označuje střední hodnotu veličiny nebo funkce A. V našem případě diskrétních veličin půjde o aritmetický nebo vážený průměr. Pokud jsou veličiny x a y statisticky nezávislé (nekorelované), je kovariance (44) rovna nule.

Odhady odchylek parametrů lineární regrese

Statistický odhad směrodatné odchylky jednotlivého měření podle (27) a (38) bude

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - a^* - b^* x_i\right)^2}{n-2}}\ . \tag{45}$$

V případě lineární regrese lze reziduální součet čtverců zapsat podle (35) a (36) ve tvaru

$$S_{\mathrm{RSS}} = \sum_{i=1}^{n} y_i^2 - a^* \sum_{i=1}^{n} y_i - b^* \sum_{i=1}^{n} x_i y_i\ . \tag{46}$$

Výpočet reziduálního součtu čtverců podle (46) je jednodušší než podle definice a opět vystačí s týmiž sumami, které jsme zmiňovali v „kalkulačkové“ úvaze o potřebné paměti, k nimž je tentokrát třeba doplnit ještě $\sum_{i=1}^{n} y_i^2$. Odhad směrodatné odchylky jednotlivého měření pak bude

$$s = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - a^* \sum_{i=1}^{n} y_i - b^* \sum_{i=1}^{n} x_i y_i}{n-2}} \tag{47}$$

Odhady odchylek parametrů regrese nejprve nalezneme na základě obecného postupu popsaného výše. V případě prokládání regresní přímky (normální rovnice ve tvaru (31)) budou prvky matice $\mathbf{A}$

$$a_{11} = n$$
$$a_{12} = a_{21} = \sum_{i=1}^{n} x_i \tag{48}$$
$$a_{22} = \sum_{i=1}^{n} x_i^2$$

Prvky inverzní matice $\mathbf{A}^{-1}$ pak budou

$$a^{11} = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$
$$a^{12} = a^{21} = -\frac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{49}$$
$$a^{22} = \frac{1}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

Tak dostaneme vyjádření pro s_{a^*} a pro s_{b^*}

$$s_{a^*} = \sqrt{a^{11}}\, s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}\; s \tag{50}$$

$$s_{b^*} = \sqrt{a^{22}}\, s = \frac{s}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}}. \tag{51}$$

Navíc lze také určit kovarianci odhadů a^* a b^*

$$\mathrm{Cov}(a^*, b^*) = a^{12}\, s^2 = -\frac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\, s^2 \tag{52}$$

Vztahy (50) a (51) můžeme také obdržet pomocí skládání rozptylů analogickým postupem jako při odvození vztahů (24) a (25). Na základě vztahu (32) získáme pro derivaci odhadu b^* podle j-té závisle proměnné vyjádření

$$\frac{\partial b^*}{\partial y_j} = \frac{x_j - \frac{\sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}. \tag{53}$$

Za všechny odhady rozptylů s_j^2 opět dosadíme s^2. Pak pro kvadrát odhadu odchylky s_{b^*} máme

$$s_{b^*}^2 = \sum_{j=1}^{n}\left(\frac{\partial b^*}{\partial y_j}\right)^2 s^2 = \sum_{j=1}^{n}\left(\frac{x_j - \frac{\sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}\right)^2 s^2 = \frac{\sum_{j=1}^{n} x_j^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)^2} s^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{54}$$

Podobně ze vztahu (33) získáme pro derivaci odhadu a^* podle *j*-té závisle proměnné vyjádření

$$\frac{\partial a^*}{\partial y_j} = \frac{1 - \frac{\partial b^*}{\partial y_j}\sum_{i=1}^{n} x_i}{n} \tag{55}$$

Potom pro kvadrát odhadu odchylky s_{a^*} máme

$$s_{a^*}^2 = \sum_{j=1}^{n}\left(\frac{\partial a^*}{\partial y_j}\right)^2 s^2 = \sum_{j=1}^{n}\left(\frac{1}{n}\left(1 - \frac{\partial b^*}{\partial y_j}\sum_{i=1}^{n} x_i\right)\right)^2 s^2 = \sum_{j=1}^{n}\frac{1}{n^2}\left(1 - \frac{x_j - \frac{\sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\sum_{i=1}^{n} x_i\right)^2 s^2 =$$

$$= \frac{1}{n^2}\left(n + \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\right) s^2 = \frac{1}{n}\left(\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\right) s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} s^2 \tag{56}$$

Čtenář si může snadno ověřit, že ke stejnému výsledku se dojde, pokud se ve vztazích (24) a (25) dosadí namísto všech σ_i^2 odhad rozptylu s^2 získaný z reziduálního součtu čtverců (34).

Korelační koeficient a koeficient determinace

I když jsme na začátku předpokládali, že použití lineárního modelu je vhodné, přece jen je občas potřeba kvantifikovat, jak dokonalá je predikce hodnot závisle proměnné. K tomu slouží jednak **korelační koeficient** a také **koeficient determinace**. V této kapitole uvedeme jen základní vztahy použitelné pro metodu nejmenších čtverců. Podrobnější rozbor jejich použití včetně obecnějšího odvození zahrnujícího statistické váhy lze nalézt v kapitole **Koeficient korelace** a následujících v části **Dodatek**.

Korelační koeficient (přesněji Pearsonův korelační koeficient výběru) $r^*(x,y)$ lze spočítat ze směrnice proložené regresní přímky b^* a směrodatných odchylek závisle a nezávisle proměnné s_y a s_x

$$r^*(x,y) = b^* \frac{s_x}{s_y}. \tag{57}$$

Korelační koeficient nabývá hodnot mezi -1 a 1. Čím vzdálenější od nulové hodnoty, tím věrnější je aproximace zkoumané závislosti přímkou. Znaménko korelačního koeficientu odpovídá směrnici přímky.

Kvadrátem korelačního koeficientu je **koeficient determinace** (v anglických textech se často nazývá **R-square**) nejčastěji definovaný vztahem

$$R^2 = 1 - \frac{S_{\mathrm{RSS}}}{S_{\mathrm{TSS}}} = \frac{S_{\mathrm{TSS}} - S_{\mathrm{RSS}}}{S_{\mathrm{TSS}}}, \tag{58}$$

kde S_{RSS} je součet čtverců reziduí (27), častěji ale vypočtený podle (46) a S_{TSS} je součet čtverců odchylek závisle proměnné y_i od její průměrné hodnoty $\bar{y}$ (**totální součet čtverců; total sum of squares**)

$$S_{\mathrm{TSS}} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2. \tag{59}$$

Je-li koeficient determinace roven 1, pak regresní přímka dokonale predikuje všechny hodnoty závisle proměnné (všechny empirické hodnoty leží přesně na přímce). Je-li naopak roven 0, pak model nepřináší žádnou užitečnou informaci.

Nejjednodušší regresní úloha – regresní proložení konstanty (alternativní pohled na výpočet průměru a s tím spojených odchylek)

Předpokládáme mezi proměnnými x a y platnost vztahu $y = c$, kde c je neznámý parametr. Úlohou tedy bude proložit body $[x_1, y_1], [x_2, y_2], \ldots [x_n, y_n]$ metodou nejmenších čtverců vodorovnou přímku. Je zřejmé, hodnota nezávisle proměnné nemá na výsledek vliv. Takže jde o zpracování pouze hodnot *y*. V této úloze je jasné, co má vyjít. Přesto bude poučné sledovat jednotlivé kroky odvození podle obecného postupu.

Podle rovnic (27) a (28) bude odhad c^* určen z podmínky

$$S_{\mathrm{RSS}} = \sum_{i=1}^{n} (y_i - c^*)^2 = \min \tag{60}$$

a vypočte se tedy z rovnice $\frac{\partial S_{\mathrm{RSS}}}{\partial c^*} = 0$.

Tak dostaneme podmínku

$$-\frac{1}{2}\frac{\partial S_{\mathrm{RSS}}}{\partial c^*} = \sum_{i=1}^{n} (y_i - c^*) = \sum_{i=1}^{n} y_i - nc^* = 0 \tag{61}$$

Sestava **normálních rovnic** zahrnuje jedinou rovnici ($p = 1$) pro výpočet c^*

$$nc^* = \sum_{i=1}^{n} y_i \,. \tag{62}$$

Řešením je odhad c^* v očekávaném tvaru (**aritmetický průměr**)

$$c^* = \frac{1}{n}\sum_{i=1}^{n} y_i \,. \tag{63}$$

S využitím vztahů (35) a (36) proto platí

$$S_{\mathrm{RSS}} = \sum_{i=1}^{n} y_i^2 - c^* \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 . \tag{64}$$

Počet stupňů volnosti reziduálního součtu čtverců

$$\nu = n - 1 \,. \tag{65}$$

Statistický odhad směrodatné odchylky kteréhokoli měření y_i je

$$s = \sqrt{\frac{S_{\mathrm{RSS}}}{n-1}} \tag{66}$$

Když vyjádříme veličinu S_{RSS} podle definice (27), dostaneme známý vztah pro **výběrovou směrodatnou odchylku** neboli **směrodatnou odchylku výběru**

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - c^*\right)^2}{n-1}} \,. \tag{67}$$

Případně lze vyjít z vyjádření S_0 pomocí (64), což umožní „kalkulačkový“ výpočet

$$s = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}{n-1}} \,. \tag{68}$$

Pro podrobnosti viz kapitolu **Směrodatná odchylka a rozptyl** v části **Dodatek**.

Směrodatná odchylka odhadu c^* označená s_{c^*} se vypočte na základě normální rovnice. V případě normální rovnice (62) bude matice $\mathbf{A}$ jednoprvková

$$a_{11} = n \tag{69}$$

Prvek inverzní matice $\mathbf{A}^{-1}$ pak bude

$$a^{11} = \frac{1}{n} \tag{70}$$

Tak dostaneme vyjádření pro s_{c^*}

$$s_{c^*} = \sqrt{a^{11}}\, s = \frac{s}{\sqrt{n}}\,. \tag{71}$$

Jak je vidět, obecný postup je úsporný a efektivní. Názornější ale opět bude alternativní výpočet založený na známé formuli pro skládání rozptylů (3). V konkrétním případě platí $f\left(y_1, y_2, \ldots, y_n\right) = c^*$, a proto na základě (63) dostaneme

$$\frac{\partial c^*}{\partial y_j} = \frac{1}{n}\,. \tag{72}$$

Za všechny rozptyly s_j^2 pak dosadíme rozptyl (jednoho každého měření) s^2, což je kvadrát výběrové směrodatné odchylky s, a získáme

$$s_{c^*}^2 = \sum_{j=1}^{n}\left(\frac{\partial c^*}{\partial y_j}\right)^2 s^2 = \sum_{j=1}^{n}\left(\frac{1}{n}\right)^2 s^2 = n\left(\frac{1}{n}\right)^2 s^2 = \frac{s^2}{n}\,, \tag{73}$$

což je ekvivalent rovnice (71). Dosazením podle (67) pak dostaneme směrodatnou odchylku odhadu parametru c^*, který představuje aritmetický průměr podle vztahu (63),

$$s_{c^*} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - c^*\right)^2}{n(n-1)}}\,, \tag{74}$$

případně „kalkulačkově“

$$s_{c^*} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}{n(n-1)}}\,. \tag{75}$$

Proto můžeme s_{c^*} označovat jako **směrodatnou odchylku aritmetického průměru.** Detaily lze nalézt kapitole **Směrodatná odchylka aritmetického průměru** v části **Dodatek**.

Jak velký výběr potřebujeme pro spolehlivý odhad směrodatné odchylky

Víme, že odhady směrodatných odchylek se s rostoucím počtem měření limitně blíží k (neznámým) přesným hodnotám těchto odchylek, které popisují statistické rozdělení měřených hodnot. Důležité je také mít představu, jak rychle se to děje. V případě s_{c^*} podle vztahu (74) nebo (75) to můžeme kvantifikovat pomocí podílu směrodatné odchylky tohoto odhadu $\sigma_{s_{c^*}}$ a limitní hodnoty σ_{c^*} přibližným vztahem (viz Annex E v [11])

$$\frac{\sigma_{s_{c^*}}}{\sigma_{c^*}} \doteq [2(n-1)]^{-1/2}. \tag{76}$$

Tato „nejistota nejistoty“ vyplývá z čistě statistické příčiny, kterou je omezený počet vzorků. Může být překvapivě velká – například pro 10 měření je to 24 procent. [Analogicky se střední hodnota malého výběru může znatelně lišit od střední hodnoty souboru. Toto chování je popsáno v kapitole **Stručně o Studentově rozdělení** v části **Dodatek**.]

Skládání individuálních směrodatných odchylek při výpočtu průměru

Průměrujeme-li hodnoty, u nichž z povahy měření známe individuální rozptyly σ_j^2 /směrodatné odchylky σ_j , můžeme tuto znalost využít a na základě vztahu (3) vypočítat rozptyl průměrné hodnoty

$$\sigma_{c^*}^2 = \sum_{j=1}^{n} \left(\frac{\partial c^*}{\partial y_j} \right)^2 \sigma_j^2 = \sum_{j=1}^{n} \left(\frac{1}{n} \right)^2 \sigma_j^2 \tag{77}$$

nebo její směrodatnou odchylku

$$\sigma_{c^*} = \frac{1}{n} \sqrt{\sum_{j=1}^{n} \sigma_j^2}. \tag{78}$$

Pozorný čtenář samozřejmě namítne, že bychom znalost individuálních rozptylů měli využít také ke zpřesnění výpočtu průměru a vstupní hodnoty zatížené menší nejistotou započítat při průměrování s větší váhou. Přesně tento postup použijeme v následující kapitole a odvodíme vztahy pro vážený průměr.

V této kapitole se budeme zabývat otázkou, v jakém vztahu jsou směrodatná odchylka aritmetického průměru s_{c^*} podle (75) a směrodatná odchylka σ_{c^*} spočtená pro průměrnou hodnotu podle (77). [Podle zavedeného označení je s_{c^*} odhadem, zatímco σ_{c^*} nikoliv, protože je spočteno na základě známých σ_j . Jenže jak už bylo řečeno, ve skutečnosti pracujeme při statistickém zpracování většinou jen s odhady (i když slovo „odhad“ se často úplně vypouští), a to samozřejmě platí také pro hodnoty dosazované do vztahu (77) a tedy i pro jeho výsledek. Rozdílné označení zde poslouží jen k snadnějšímu rozlišení odhadů získaných různými postupy.]

Pro zjednodušení budeme dále předpokládat, že se individuální rozptyly naměřených hodnot navzájem neliší, což nás opravňuje použít běžný aritmetický průměr. Pro další úvahy totiž není důležité, zda se individuální rozptyly liší navzájem, nýbrž zda se liší od odhadu získaného metodou nejmenších čtverců.

Není složité ukázat, že jsou-li individuální směrodatné odchylky σ_j naměřených hodnot rovny výběrové směrodatné odchylce s vypočtené metodou nejmenších čtverců (68), pak s_{c^*} podle (75) a σ_{c^*} podle (77) povedou ke stejnému výsledku. Na tom není nic divného, protože jde o alternativní popisy téhož statistického chování. V reálném případě se ale s podle (68) nezanedbatelně liší od σ_j, a proto se také liší s_{c^*} od σ_{c^*}. Ani tehdy ale není žádoucí hodnoty s_{c^*} a σ_{c^*} nějak skládat nebo dokonce přímo sčítat!

Rozebereme dále několik typických situací. Vliv má jednak počet průměrovaných hodnot, jednak charakter odchylek započítaných do vztahu (77). Doporučení pro výpočty s naměřenými hodnotami rozlišují dva způsoby vyhodnocení nejistot [11]. Zjednodušeně řečeno, je to jednak statistické zpracování opakovaných měření (typ A) a pak stanovení nejistoty na základě kalibrace a dalších znalostí o měřicím zařízení (typ B), což si zhruba můžeme představit jako „chybu měřidla“. Zatímco při vyhodnocení nejistot typu A se předpokládá normální rozdělení jakožto důsledek centrální limitní věty, vyhodnocení typu B na toto spoléhat nemůže, protože „chyby měřidel“ mohou mít alespoň částečně systematický charakter. Proto se u typu B většinou předpokládá rovnoměrné rozdělení. Podrobněji porovnáme oba typy vyhodnocení nejistot v kapitolách **Nejistoty typu A a B** a **Skládání nejistot** v části **Dodatek**. A nyní slibované typické situace:

- Předpokládejme, že σ_j byly vyhodnoceny postupem typu A, přičemž jejich velikost je v souladu se skutečným náhodným rozdělením naměřených hodnot. Pak pro velký počet průměrovaných hodnot se výběrová směrodatná odchylka s (68) bude blížit k σ_j a podobně s_{c^*} (75) se bude blížit k σ_{c^*} (77). Podle vztahu (76) můžeme zjistit, že pro $n > 30$ bude odhad odchylky s_{c^*} (75) stanoven poměrně spolehlivě, protože bude zatížen relativní směrodatnou odchylkou menší než 10 procent.
- Předpokládejme, že pro σ_j platí totéž co v předchozím případě, nicméně počet průměrovaných hodnot bude malý. S klesajícím počtem průměrovaných hodnot roste rozptyl odhadů získaných metodou nejmenších čtverců. Proto se pro velmi malé počty hodnot ($n \leq 5$) mohou odhady odpovídajících veličin lišit i několikanásobně. Podle (76) bude v tomto případě relativní směrodatná odchylka odhadu s_{c^*} (75) větší než 35 procent. Známe-li tedy individuální odchylky σ_j, pak je v této situaci nanejvýš vhodné jako směrodatnou odchylku průměru použít σ_{c^*} podle (77).
- Na druhou stranu se ale s_{c^*} (75) a σ_{c^*} (77), jakožto dva odhady téže veličiny nemohou lišit přespříliš, pokud má platit předpoklad, že náhodné rozdělení naměřených hodnot odpovídá směrodatné odchylce σ_j. Vychází-li tedy v obou předchozích případech s_{c^*} podstatně větší

(viz dále) než σ_{c^*}, pak jsou naměřené hodnoty od sebe vzdáleny více, než kolik by odpovídalo směrodatné odchylce σ_j. V takové situaci je na zvážení, zda průměrování má fyzikální smysl. Naopak, σ_{c^*} podstatně větší než s_{c^*} svědčí o tom, že byly odchylky jednotlivých měření σ_j neopodstatněně nadhodnoceny. Ono nepříliš přesné vyjádření „podstatně větší“ v předchozích větách můžeme kvantifikovat na základě (76). Pro konkrétní počet průměrovaných hodnot spočteme relativní směrodatnou odchylku odhadu a její trojnásobek použijeme jako limit, jehož překročení bude známkou zřejmého nesouladu.

- Pokud jsou směrodatné odchylky σ_j vyhodnoceny postupem typu B, pak rozdíl mezi s_{c^*} (75) a σ_{c^*} (77) nemusí nutně znamenat patologickou situaci. Má-li například odchylka σ_j převážně systematický charakter, a tedy nelze její vliv eliminovat opakovaným měřením, bude vycházet σ_{c^*} podstatně větší než s_{c^*}. Opačná situace, to znamená s_{c^*} podstatně větší než σ_{c^*}, může například nastat, když měření nebylo prováděno podle doporučení výrobce měřidla a rozptyl opakovaných měření je proto příliš velký. V obou případech je třeba jako směrodatnou odchylku průměru použít ten větší z obou odhadů (protože ten menší je jím pak pokryt také).

Předchozí úvahy není složité zobecnit na případ komplikovanějšího výpočtu, do něhož vstupují různé veličiny charakterizované nejistotami získanými různým způsobem.

Odvození vztahů pro vážený průměr a jeho rozptyl

Proložíme ještě jednou naměřenými body vodorovnou přímku $y = c$, kde c je opět neznámý parametr. Tentokrát ale budeme považovat za známé kromě samotných hodnot y_i také jejich statistické rozptyly σ_i^2. Proto budeme namísto reziduálního součtu (60) minimalizovat χ^2

$$\chi^2 = \sum_{i=1}^{n} \frac{\left[y_i - c^* \right]^2}{\sigma_i^2} = \min . \tag{79}$$

Odhad neznámého parametru c^* se opět vypočte z rovnice $\frac{\partial \chi^2}{\partial c^*} = 0$. Tentokrát získáme podmínku

$$-\frac{1}{2}\frac{\partial \chi^2}{\partial c^*} = \sum_{i=1}^{n} \frac{(y_i - c^*)}{\sigma_i^2} = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} - c^* \sum_{i=1}^{n} \frac{1}{\sigma_i^2} = 0 \tag{80}$$

Normální rovnice bude mít tvar

$$c^* \sum_{i=1}^{n} \frac{1}{\sigma_i^2} = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} . \tag{81}$$

Řešením je odhad parametru c^* ve tvaru

$$c^* = \frac{\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}, \tag{82}$$

což je formule pro výpočet **váženého průměru** (weighted arithmetic mean). V tomto případě navíc jde o takovou verzi váženého průměru, kde váhy hodnot y_i jsou určeny převrácenými hodnotami jejich rozptylů $\frac{1}{\sigma_i^2}$ (inverse-variance weighting).

Ze vztahu (82) získáme pro derivaci odhadu parametru c^* podle *i*-té proměnné vyjádření

$$\frac{\partial c^*}{\partial y_i} = \frac{1}{\sigma_i^2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \tag{83}$$

Pak pro rozptyl σ_{c^*} váženého průměru dostaneme podle (3) vyjádření pomocí známých σ_i^2

$$\sigma_{c^*}^2 = \sum_{i=1}^{n} \left(\frac{\partial c^*}{\partial y_i} \right)^2 \sigma_i^2 = \sum_{i=1}^{n} \left(\frac{1}{\sigma_i^2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \right)^2 \sigma_i^2 = \frac{1}{\left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)^2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \quad . \tag{84}$$

Není složité ukázat, že po dosazení všech rozptylů σ_i^2 stejné velikosti dostaneme vztahy odvozené metodou nejmenších čtverců pro případ prostého průměrování.

Lineární regrese s pevným průsečíkem se svislou osou

Upravíme výše odvozené vztahy (19) až (25) pro prokládání lineární závislosti $y = a + bx$ na případ, kdy hodnota a průsečíku se svislou souřadnicovou osou je předem dána. Odhad parametru b^* se v tomto případě určí minimalizací výrazu

$$\chi^2 = \sum_{i=1}^{n} \frac{\left[y_i - a - b^* x_i \right]^2}{\sigma_i^2} = \min \tag{85}$$

a vypočte z rovnice $\frac{\partial \chi^2}{\partial b^*} = 0$. Tak dostaneme podmínku

$$-\frac{1}{2} \frac{\partial \chi^2}{\partial b^*} = \sum_{i=1}^{n} \frac{(y_i - a - b^* x_i) x_i}{\sigma_i^2} = \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2} - a \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} - b^* \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} = 0 . \tag{86}$$

Normální rovnice pro výpočet b^* bude mít tvar

$$a\underbrace{\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2}}_{S_x} + b^*\underbrace{\sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2}}_{S_{xx}} = \underbrace{\sum_{i=1}^{n}\frac{x_i y_i}{\sigma_i^2}}_{S_{xy}}. \tag{87}$$

Řešením je odhad b^* v kompaktním tvaru zapsaném pomocí zavedených součtů

$$b^* = \frac{S_{xy} - aS_x}{S_{xx}}. \tag{88}$$

Na základě vztahu (88) získáme pro derivaci odhadu b^* podle *i*-té závisle proměnné vyjádření

$$\frac{\partial b^*}{\partial y_i} = \frac{1}{S_{xx}}\frac{x_i}{\sigma_i^2} \quad . \tag{89}$$

Pak pro kvadrát odchylky směrnice s_{b^*} dostaneme

$$s_{b^*}^2 = \sum_{i=1}^{n}\left(\frac{\partial b^*}{\partial y_i}\right)^2\sigma_i^2 = \sum_{i=1}^{n}\left(\frac{x_i}{S_{xx}\sigma_i^2}\right)^2\sigma_i^2 = \frac{1}{S_{xx}^{\ 2}}\sum_{i=1}^{n}\frac{x_i^{\ 2}}{\sigma_i^2} = \frac{1}{S_{xx}}. \tag{90}$$

Vztah (90) využijeme, pokud známe individuální směrodatné odchylky jednotlivých měřených hodnot. Pokud tomu tak není, použijeme pro všechny body jednotný odhad rozptylu s^2 získaný z reziduálního součtu čtverců (34) a tak dostaneme pro hledanou směrnici vztah

$$b^* = \frac{\sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \tag{91}$$

a pro kvadrát odchylky směrnice s_{b^*} vztah (počet stupňů volnosti $\nu = n-1$)

$$s_{b^*}^2 = \frac{1}{S_{xx}} = \frac{1}{\sum_{i=1}^{n}\frac{x_i^2}{s^2}} = \frac{s^2}{\sum_{i=1}^{n} x_i^2} = \frac{S_{\mathrm{RSS}}}{n-1}\frac{1}{\sum_{i=1}^{n} x_i^2}. \tag{92}$$

Z porovnání vztahů (92) a (54) vidíme, že pro stejná data a stejnou proloženou přímku dává ten první vždy nižší hodnotu. To znamená, že směrnice přímky procházející pevným průsečíkem je určena s menší směrodatnou odchylkou než v případě obecné přímky. Použití tohoto postupu proto může být výhodné v případech, kdy přímka určitě pevným bodem prochází. V jednom z následujících příkladů je to Einsteinův zákon, kde prochází počátkem lineární časová závislost střední kvadratické vzdálenosti uražené difundující částicí.

Posouzení konzistentnosti individuálních rozptylů s hodnotami *x* a *y*

Ze vztahů (22), (24) a (25) je zřejmé, že zatímco hodnoty regresních koeficientů závisí jenom na vzájemných poměrech rozptylů jednotlivých hodnot, jejich směrodatné odchylky závisí na absolutních velikostech rozptylů. A podobně na jejich absolutních velikostech závisí také

hodnota funkce χ^2 v nalezeném minimu. Tato souvislost není žádným překvapením, protože v metodě nejmenších čtverců se odhad neznámého rozptylu jednotlivého měření odvozuje od reziduálního součtu čtverců vztahem (34).

Vzhledem k tomu, že tentokrát nestojíme o odhady rozptylu, jelikož ty jsme naopak použili jako vstupní parametry, můžeme statistické chování χ^2 využít naopak pro posouzení, zda statistické chování dat zadaným rozptylům odpovídá. Úpravou vztahu (34) pro individuální velikosti rozptylů u jednotlivých měření získáme vztah

$$\chi^2 = \sum_{i=1}^{n} \frac{\left[y_i - f(x_i; b_1^*, b_2^*, \ldots, b_p^*) \right]^2}{\sigma_i^2} = \nu ,$$

což v případě regresní přímky bude konkrétně

$$\chi^2 = \sum_{i=1}^{n} \frac{\left[y_i - a^* - b^* x_i \right]^2}{\sigma_i^2} = n - 2 \ .$$

Jde samozřejmě o součet pro velký počet *n*. Přesnější je proto mluvit o střední hodnotě této veličiny, takže $\left\langle \chi^2 \right\rangle = \nu$. Bez odvození uveďme, že směrodatná odchylka je v tomto případě přibližně $\sqrt{2\nu}$. Program Origin počítá veličinu *Reduced Chi-Sqr* podle vztahu

$$\chi_r^2 = \frac{\chi^2}{\nu} . \tag{93}$$

Její střední hodnota by tedy měla být rovna jedné se směrodatnou odchylkou

$$\sigma_{RCS} = \sqrt{2/\nu} . \tag{94}$$

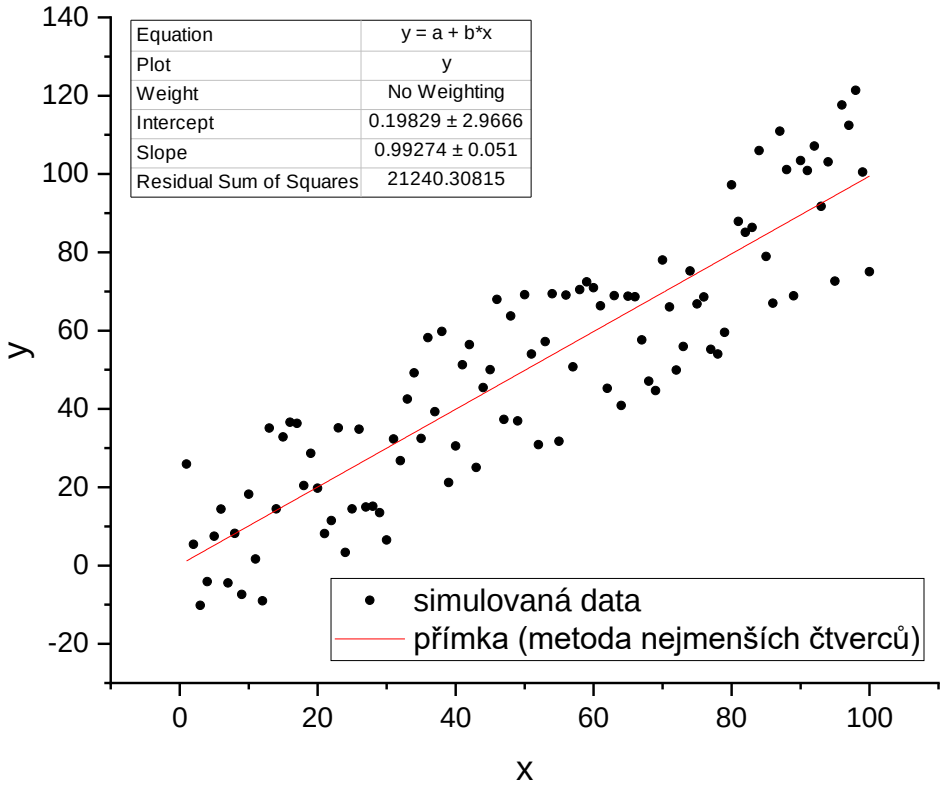
Ačkoliv rozdělení χ^2 není symetrické (pro malé počty stupňů volnosti dokonce nemá ani vrchol), s rostoucím počtem stupňů volnosti se blíží k normálnímu rozdělení (121). Uvádí se, že pro $\nu > 50$ jsou rozdíly zanedbatelné [12]. Pak můžeme využít poznatek, že 99.7 % obdržených hodnot bude ležet v intervalu $\langle 1 - 3\sigma_{RCS}, 1 + 3\sigma_{RCS} \rangle$. Pokud hodnota χ^2 pro konkrétní výběr dat tomuto kritériu nevyhovuje, pak nemusí být použitý fit optimální anebo vložené hodnoty σ_i^2 neodpovídají skutečnosti. Někdy může mít takový nesoulad i dobrý důvod, jak si později ukážeme, nicméně vždy by měl výskyt χ^2 mimo pravděpodobný interval vést k zamyšlení nad uspořádáním experimentu. Některé základní situace probereme detailněji v rámci praktických příkladů. Další podrobnosti lze nalézt například v [3].

Příklady použití lineární regrese

V následujících příkladech jsou směrodatné odchylky v souladu s běžnými zvyklostmi značeny řeckým písmenem „sigma“, i když se jedná o odhady, které jsme ve výkladové části pro názornost označovali písmenem „s“.

Příklad 1 – různé způsoby práce s odchylkami

Využijeme ještě jednou data z úvodní kapitoly a budeme si na nich ilustrovat nejdůležitější dosud zmíněné poznatky. Následující *Graf 3* zobrazuje tato data tentokrát zpracovaná programem Origin.



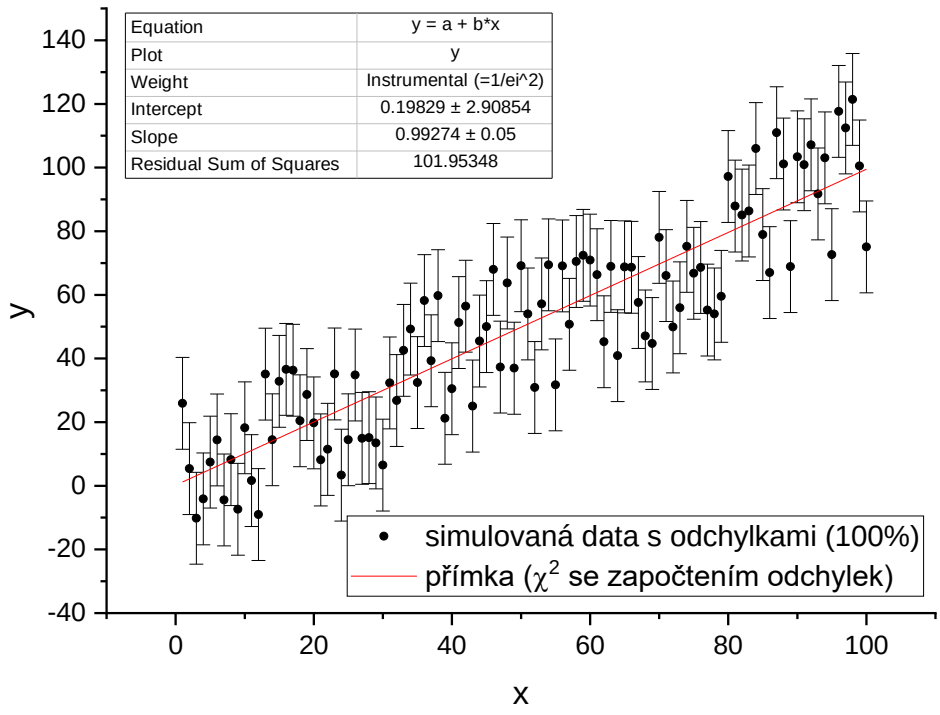
Graf 3

Lineární regrese je stovkou bodů proložena metodou nejmenších čtverců, takže do výpočtu nevstupují žádné další informace. Na základě statistického chování dat výpočet umožní mj. stanovit směrodatnou odchylku jednotlivého měření podle vztahu (38) (viz *Tabulka 1*). Tato hodnota je v dobré shodě se známou skutečnou hodnotou $\frac{25}{\sqrt{3}}$. (V tabulce vyznačeno žlutou barvou.) Pro podrobnosti viz popis přípravy simulovaných dat v kapitole **Metoda postupných měření** a odvození vztahu (124) v části **Dodatek**.

	Graf 3	*Graf 4*	*Graf 5*	*Graf 6*
směrodatná odchylka σ na ose *y* všech bodů		$\frac{25}{\sqrt{3}}$	$\frac{25}{\sqrt{3}}\frac{2}{3}$	$\frac{25}{\sqrt{3}}\frac{3}{2}$
směrodatná odchylka σ na ose *y* všech bodů číselně	---	14.4338	9.6225	21.6506
průsečík s osou *y*	0.1983	0.1983	0.1983	0.1983
směrodatná odchylka průsečíku	2.9666	2.9085	1.9390	4.3628
směrnice	0.9927	0.9927	0.9927	0.9927
směrodatná odchylka směrnice	0.0510	0.0500	0.0333	0.0750
reziduální součet čtverců (29), (46) resp. (19)	21240.3	101.95	229.40	45.313
směrodatná odchylka jednotlivého měření (38)	14.7220	---	---	---
parametr χ_r^2 (93) *(Reduced Chi-Sqr)*	---	1.0403	2.3408	0.4624

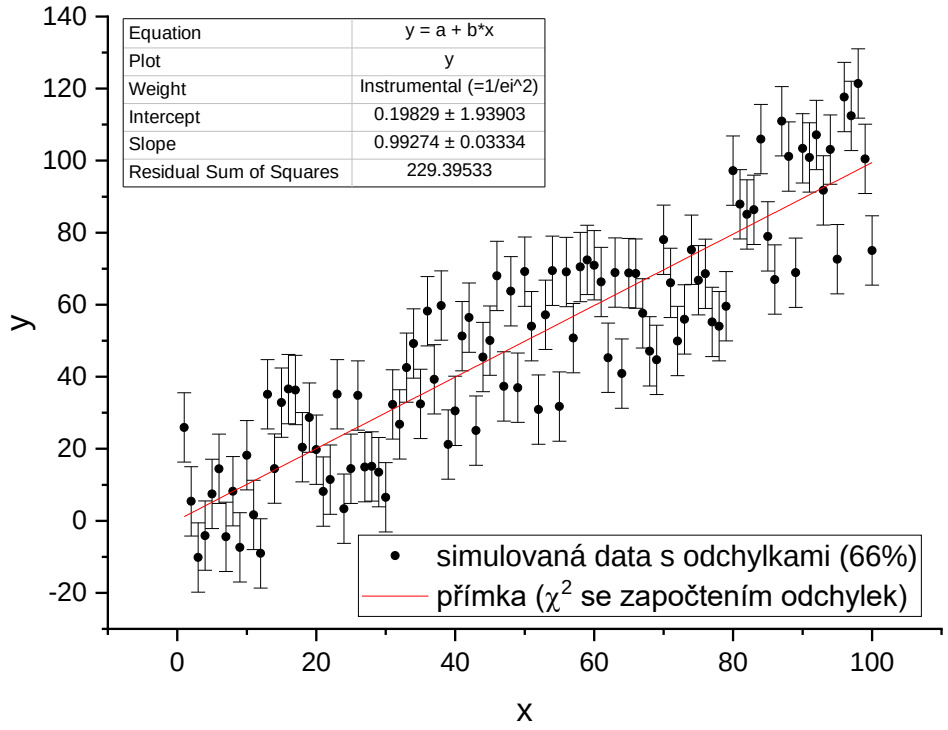
Tabulka 1

Následující tři grafy znázorňují táž data ve sloupcích *x* a *y* doplněná navíc o sloupec se směrodatnými odchylkami závisle proměnné (znázorněny chybovými úsečkami). Pro jednoduchost jsou odchylky pro všechny body v každém grafu stejné. *Graf 4* znázorňuje situaci, kdy jejich velikost je právě rovna té odchylce, která byla použita pro přípravu náhodných dat *y*, tedy mají hodnotu $\frac{25}{\sqrt{3}}$. Pro *Graf 5* byla zvolena odchylka rovná 2/3 této hodnoty. *Graf 6* naopak ukazuje situaci, kdy je odchylka jejím 1.5násobkem. V těchto třech grafech se znázorněné odchylky zahrnují do výpočtu odchylek hledaných parametrů regresní přímky. Reziduální součet čtverců χ^2 započítává podle (19) statistické váhy určené rozptylem.

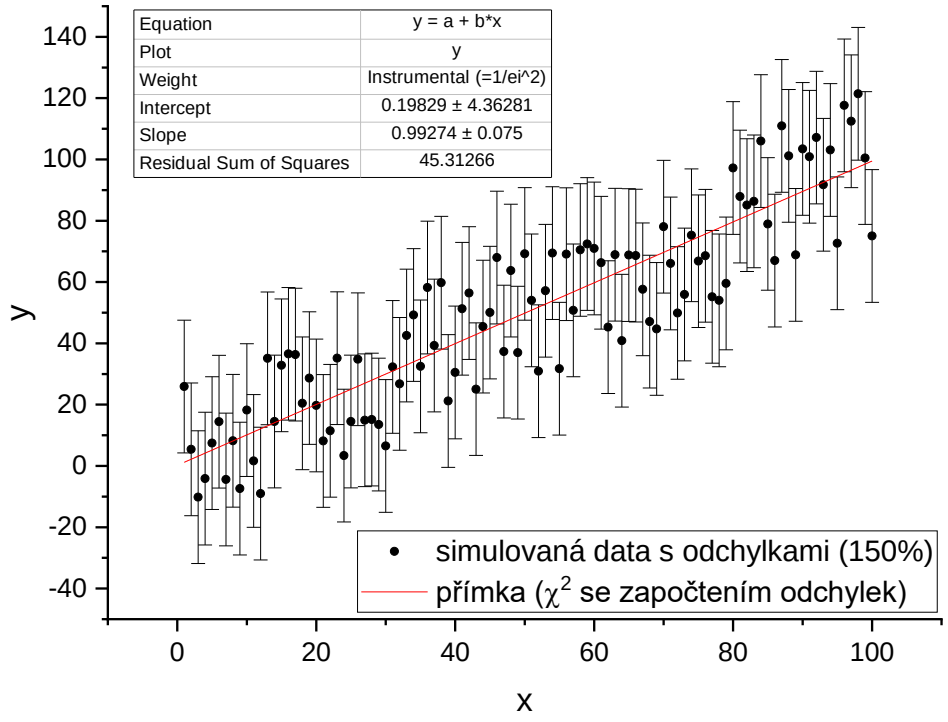


Graf 4

Nepřekvapí nás, že všechny proložené přímky mají stejné parametry. (V tabulce vyznačeno odstíny zelené.) Připomeňme, že to ale není jen důsledkem použití týchž dat, nýbrž také skutečnosti, že všechny individuální odchylky bodů jsou pro celou sadu stejné. Jen za této podmínky jsou váhové faktory $1/\sigma^2$ shodné a výpočet dává nezávisle na velikosti σ stejné parametry přímky jako metoda nejmenších čtverců. Odchylky parametrů regresní přímky ale už na σ závisí. *Graf 4* umožňuje ověřit, že pokud se popsanou metodou započtou jednotlivým bodům individuální odchylky, které odpovídají jejich skutečnému statistickému chování, pak se odchylky regresních parametrů prakticky shodují s těmi, které vypočte metoda nejmenších čtverců. (V tabulce vyznačeno odstíny modré. Drobné rozdíly jsou způsobeny omezeným počtem bodů v sadě náhodných dat.) *Graf 5* a *Graf 6* pak ukazují, jak se na odchylkách regresních parametrů odrazí změna individuálních odchylek bodů k nižší či vyšší hodnotě.



Graf 5



Graf 6

Jak bylo zmíněno v kapitole **Posouzení konzistentnosti individuálních rozptylů s hodnotami *x* a *y***, hodnota parametru χ_r^2 spočtená podle vztahu (93) umožňuje kvantifikovat, jak dobře odpovídají zadané individuální odchylky statistickému chování dat. V našem případě $\nu = 98$, takže podle (94) pak obdržíme hodnotu $\sigma_{RCS} \doteq 0.14$. Pokud tedy $\chi_r^2 > 1.42$, jsou pravděpodobně zadané odchylky podceněny. Takovou situaci zobrazuje *Graf 5*. Pokud naopak $\chi_r^2 < 0.58$, znamená to přecenění odchylek (*Graf 6*). Užitečné je, že kritérium je použitelné i ve složitějších případech, než jsou naše modelová data se všemi odchylkami stejné velikosti.

Příklad 2 – výpočet Youngova modulu pružnosti z protažení drátu

Jde o jednoduchý experiment, kde se měří prodloužení drátu způsobené známým jednoosým tahem. Detailní uspořádání experimentu je popsáno v návodu k úloze [13]. Provedeme jen stručnou rekapitulaci použitých vztahů: Tah se kvantifikuje mechanickým napětím

$$\tau = \frac{F}{S}, \tag{95}$$

tedy podílem napínací síly F a průřezu drátu S. Ten se uvažuje jako kruh o průměru d, pak

$$S = \tfrac{1}{4}\pi d^2. \tag{96}$$

Napínací síla

$$F = mg \tag{97}$$

je vytvářena závažím m. Tíhové zrychlení značíme g. Relativní prodloužení drátu

$$\varepsilon = \frac{\Delta l}{l_0} \tag{98}$$

se vyjádří jako podíl nárůstu délky Δl a výchozí délky drátu l_0. Délka drátu se snímá pomocí kladky poloměru R. Ta se při prodloužení Δl měřeného úseku drátu pootočí o malý úhel α

$$\Delta l = R\alpha. \tag{99}$$

Na ose kladky je připevněno zrcátko, přes něž se pevným dalekohledem čte centimetrová stupnice ve vzdálenosti L. Při pootočení kladky o malý úhel α se podle zákona odrazu posune hodnota v záměrném kříži o vzdálenost

$$\Delta n = 2L\alpha. \tag{100}$$

Pro malá napětí lze Youngův modul pružnosti vyjádřit jako podíl

$$E = \frac{\tau}{\varepsilon}. \tag{101}$$

Dále popíšeme některé postupy zpracování včetně numerických výsledků pro ilustrační data zahrnující hodnoty veličin d, R, l_0 a L a jejich směrodatné odchylky (*Tabulka 2*) a především výchylku n odečtenou dalekohledem na stupnici pro každé použité závaží m (*Tabulka 3*).

$$d = (6.12 \pm 0.11) \times 10^{-4}\ \mathrm{m}$$

$$R = (1.78 \pm 0.01) \times 10^{-2}\ \mathrm{m}$$

$$l_0 = (1.140 \pm 0.005)\ \mathrm{m}$$

$$L = (8.97 \pm 0.05) \times 10^{-1}\ \mathrm{m}$$

Tabulka 2

m/kg	*n*/m
0.0	0.170
0.1	0.166
0.2	0.162
0.3	0.158
0.4	0.154
0.5	0.150
0.6	0.146
0.7	0.143
0.8	0.139
0.9	0.135
1.0	0.132
1.1	0.128
1.2	0.124
1.3	0.120
1.4	0.117
1.5	0.114
1.6	0.111
1.7	0.107
1.8	0.104
1.9	0.098
2.0	0.095
2.1	0.092
2.2	0.090
2.3	0.087
2.4	0.084

Tabulka 3

Pokud napětí nepřesáhne mez úměrnosti, můžeme Youngův modul pružnosti určit ze směrnice závislosti $\varepsilon = \varepsilon(\tau)$. Ze vztahů (95), (96) a (97) se vypočte aplikované napětí

$$\tau = \frac{4mg}{\pi d^2} \tag{102}$$

a pomocí (98), (99) a (100) zase relativní prodloužení

$$\varepsilon = \frac{R\,\Delta n}{2L\,l_0}. \tag{103}$$

Těmito dvojicemi bodů se pak proloží lineární regrese. Youngův modul bude roven převrácené hodnotě směrnice. Jen je nutno uvážit, že regresní přímka nemusí procházet počátkem. Takový požadavek by znamenal, že zatímco všechny ostatní naměřené body jsou zatíženy nějakou nejistotou, počátek je znám přesně. A to není pravda. Proto se pro výpočet směrnice použije vztah (32), který slouží k proložení obecné přímky.

Protože nevyhodnocujeme skutečné absolutní hodnoty prodloužení, ale jen jeho změny, nebude nám v pracovním grafu vadit posunutí osy ε o konstantní hodnotu. Je tedy při výpočtu regrese

možno namísto Δn použít přímo hodnotu n odečtenou dalekohledem. Protože je navíc stupnice orientována tak, že při zvětšování závaží odečítané n klesá, nesmí nás překvapit, že bude směrnice záporná.

Číselně pro data z tabulek (*Tabulka 2* a *Tabulka 3*) vychází směrnice $k_1 = -9.42760 \times 10^{-12}\ \mathrm{Pa}^{-1}$ a to vede na Youngův modul v hodnotě $E_1 = 1.06072 \times 10^{11}\ \mathrm{Pa}$. Výsledky jsou záměrně uváděny s přesností přesahující možnosti experimentu, aby si čtenář mohl kontrolovat správnost svých výpočtů.

Je-li cílem získat vedle hodnoty Youngova modulu také příslušnou směrodatnou odchylku, je třeba postupovat obezřetně. Nelze jednoduše použít směrodatnou odchylku směrnice podle (51), protože ta pouze charakterizuje, jak hodně se body odchylují od proložené přímky. Nejistoty vstupních veličin do (51) vůbec nevstupují. Pro naše data takto vychází směrodatná odchylka směrnice $\sigma_{k_1} = 7.631 \times 10^{-14}\ \mathrm{Pa}^{-1}$, takže směrodatná odchylka Youngova modulu je $\sigma_{E_1} = 8.58 \times 10^{8}\ \mathrm{Pa}$, což představuje zhruba 0.8 %.

Jako nejjednodušší se proto může zdát prostě doplnit k výše vypočteným dvojicím bodů ε a τ ještě směrodatné odchylky σ_ε a σ_τ spočtené na základě principu šíření nejistot (3) pomocí směrodatných odchylek hodnot d, R, l_0 a L a pak použít lineární regresi, která umí takové odchylky zahrnout, například některou z metod nabízených programem Origin. Ukažme si, proč takový postup, i když se může jevit jako elegantní, není rozumné používat.

Ze vztahu (103) lze vypočítat směrodatnou odchylku relativního prodloužení

$$\sigma_\varepsilon = \varepsilon \sqrt{\left(\frac{\sigma_{l_0}}{l_0}\right)^2 + \left(\frac{\sigma_L}{L}\right)^2 + \left(\frac{\sigma_R}{R}\right)^2 + \left(\frac{\sigma_n}{n}\right)^2} \tag{104}$$

a pomocí (102) zase směrodatnou odchylku aplikovaného napětí

$$\sigma_\tau = \tau \sqrt{\left(\frac{\sigma_m}{m}\right)^2 + \left(\frac{2\sigma_d}{d}\right)^2},$$

což ještě upravíme, aby bylo možno dosazovat i v případě nulové hmotnosti m

$$\sigma_\tau = \frac{4g}{\pi d^2} \sqrt{{\sigma_m}^2 + \left(\frac{2m\sigma_d}{d}\right)^2}. \tag{105}$$

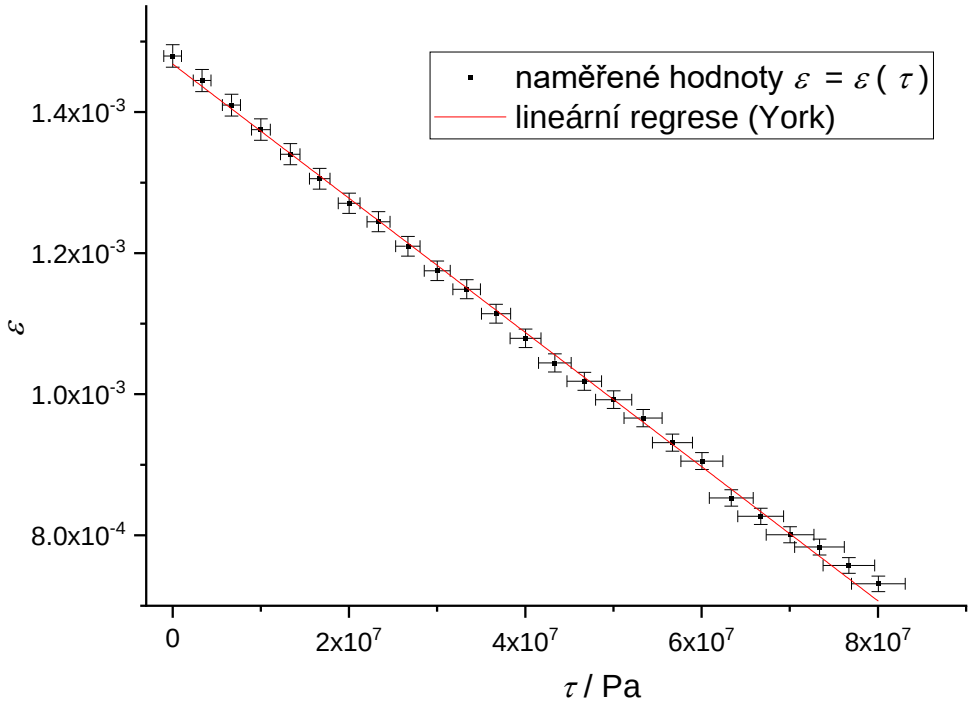
Vidíme, že kromě směrodatných odchylek veličin d, R, l_0 a L (*Tabulka 2*) potřebujeme také odchylky veličin m a n (*Tabulka 3*). Ty nejsou dány jen přesností závaží či odečítání v dalekohledu, ale také nepravidelnostmi v chodu aparatury (tření v ose kladky, nerovnoměrnosti v ohybu drátu, způsob kladení závaží na misku). Použijme hodnoty, které obsahuje *Tabulka 4*. Později si vysvětlíme, jak je lze odhadnout z parametrů regresní přímky

proložené metodou nejmenších čtverců. Nejprve ale vyzkoušíme, co s daty provede Yorkova metoda. (viz **Funkce *Analysis:::Fit Linear with X Error* programu Origin**)

$$\sigma_n = 1 \text{ mm}$$

$$\sigma_m = 30 \text{ g}$$

Tabulka 4



Graf 7

Na pracovním grafu (*Graf 7*) vidíme naměřené hodnoty τ a ε spolu s odchylkami σ_τ a σ_ε a také regresní přímku proloženou pomocí Yorkovy metody. Posun osy ε a nezvyklý sklon závislosti jsme již zmiňovali. *Tabulka 5* ukazuje parametry regrese.

Parameters

		Value	Standard Error
epsilon	Intercept	0.00147	7.60408E-6
	Slope	-9.51537E-12	1.92792E-13

Statistics

	epsilon
Number of Points	25
Degrees of Freedom	23
Reduced Chi-Sqr	0.14377
Residual Sum of Squares	3.30682

Tabulka 5

V tomto případě směrnice $k_2 = -9.51537\times10^{-12}\,\mathrm{Pa}^{-1}$ a směrodatná odchylka $\sigma_{k_2} = 1.9279\times10^{-13}\,\mathrm{Pa}^{-1}$. Směrnice k_2 získaná Yorkovou metodou se liší od hodnoty k_1 spočtené podle (51), protože v Yorkově výpočtu se měřené body započítávají s různou vahou. Youngův modul pak vyjde $E_2 = 1.05093\times10^{11}\,\mathrm{Pa}$ se směrodatnou odchylkou $\sigma_{E_2} = 2.129\times10^{9}\,\mathrm{Pa}$, což představuje 2 %. Přestože jsou nejistoty vstupních hodnot tentokrát započteny ve vyšší míře, signalizuje nízká hodnota χ_r^2 (93), tedy parametru *Reduced Chi-Sqr* (viz *Tabulka 5*), že zadané odchylky jsou nadhodnoceny a neodpovídají skutečnému chování dat. Připomeňme, že střední hodnota tohoto statistického ukazatele je rovna jedné a směrodatná odchylka (94) pro $\nu = 23$ vyjde 0.295.

Základní potíž spočívá v tom, že metody, které započítávají individuální statistické odchylky, jsou určeny pro prokládání skutečně naměřenými body, které se opravdu chovají podle specifikovaných individuálních rozdělení. V případě výše popsaného započítání odchylek dalších vstupních veličin to ale neplatí, protože tyto veličiny ovlivňují všechny naměřené body obdobně. V našem případě d, R, l_0 a L mění především náklon celé závislosti, ale mají zanedbatelný vliv na její tvar. Důsledkem tohoto korelovaného účinku není jen nízká hodnota *Reduced Chi-Sqr* (ta nám jen oznamuje, co už víme), ale také nedostatečné započtení individuálních odchylek do výsledků regrese.

Jako korektní lze doporučit následující postup: Spojením vztahů (95) až (101) můžeme získat vztah pro Youngův modul

$$E = \frac{8gl_0L}{\pi d^2 R}\frac{m}{\Delta n} = \frac{8gl_0L}{\pi d^2 R k_{n(m)}}, \tag{106}$$

kde veličina $k_{n(m)} = \dfrac{\Delta n}{m}$ představuje posunutí stupnice v dalekohledu vyvolané jednotkovým závažím, neboli je to směrnice závislosti $n = n(m)$. Nejprve se tedy pomocí vztahů (32) a (51) vypočte směrnice k závislosti $n = n(m)$ včetně směrodatné odchylky, Youngův modul se spočte podle (106) a pomocí principu šíření nejistot (3) se pak určí také jeho směrodatná odchylka

$$\sigma_E = E\sqrt{\left(\frac{\sigma_{l_0}}{l_0}\right)^2 + \left(\frac{\sigma_L}{L}\right)^2 + 4\left(\frac{\sigma_d}{d}\right)^2 + \left(\frac{\sigma_R}{R}\right)^2 + \left(\frac{\sigma_k}{k}\right)^2}\,. \tag{107}$$

Na výpočty stačí nástroje, které nabízí Microsoft EXCEL. Přeneseme sloupce (*Tabulka 2)* na list programu EXCEL. Předpokládejme, že hodnoty m/kg jsou ve sloupci A v rozsahu A2:A26 a hodnoty n/m ve sloupci B v položkách B2:B26. Do volné buňky tedy vložíme „=LINREGRESE(B2:B26;A2:A26;;1)“ a vyznačíme oblast pro vypočtené parametry (viz **Funkce LINREGRESE (LINEST) v EXCELu**). Zobrazí se následující výsledky (*Tabulka 6*).

-0.036123077	0.168387692
0.000292396	0.000409354
0.998495312	0.001054248
15262.56524	23
0.016963397	2.55631E-05

Tabulka 6

To znamená, že směrnice má hodnotu $k_{n(m)} = -3.6123077 \times 10^{-2}\,\mathrm{m\,kg^{-1}}$ a její směrodatná odchylka je $\sigma_{k_{n(m)}} = 2.92396 \times 10^{-4}\,\mathrm{m\,kg^{-1}}$, což představuje opět oněch 0.8 %. Po dosazení do (106) dostaneme $E_3 = 1.06072 \times 10^{11}\,\mathrm{Pa}$, což se podle očekávání shoduje s hodnotou E_1, protože regrese byla počítána stejnou metodou. Nicméně směrodatná odchylka podle (107) zahrnuje nejen neurčitost bodů závislosti $n = n(m)$, ale také neurčitosti všech ostatních měřených veličin. Proto nepřekvapí, že v tentokrát vyjde $\sigma_{E_3} = 4.024 \times 10^{9}\,\mathrm{Pa}$, což je 3.8 %.

Ještě se vraťme k směrodatným odchylkám veličin m a n (viz *Tabulka 4*). Pomocí dalších parametrů regrese, konkrétně reziduálního součtu čtverců a počtu stupňů volnosti (*Tabulka 6*) můžeme snadno vypočítat směrodatnou odchylku měřených hodnot na vertikální ose. Dosazením do vztahu (38) dostaneme $\sigma_n = 1.054248 \times 10^{-3}\,\mathrm{m}$. Stejnými daty můžeme ovšem proložit také závislost $m = m(n)$. Stačí jen upravit předpis funkce do tvaru „=LINREGRESE(A2:A26;B2:B26;;1)“. Z regresních parametrů pro inverzní závislost (viz *Tabulka 7*) pak analogicky vyjde $\sigma_m = 2.91629 \times 10^{-2}\,\mathrm{kg}$. Můžeme si také povšimnout, že platí $\sigma_n \approx k_{n(m)} \sigma_m$.

-27.64148011	4.656290673
0.223742006	0.028578222
0.998495312	0.029162915
15262.56524	23
12.98043906	0.019560938

Tabulka 7

Příklad 3 – výpočet aktivity částice konající Brownův pohyb

Úloha reprodukuje Perrinův pokus, kde se z pozorování Brownova pohybu určuje Avogadrova konstanta. Podrobnosti lze nalézt v návodu k úloze [14]. Zde zmíníme jen základní informace.

Hlavní součástí aparatury je mikroskop vybavený kamerou, která umožňuje přenést obraz na počítačový monitor. Při měření se pak v programu BROWN pomocí myši v pravidelných intervalech zaznamenávají polohy částice. Výsledkem je datový soubor, který jednak obsahuje kalibrační konstanty, které umožní přepočítat vzdálenosti na obrazovce na skutečné posuny částice, a dále obsahuje sérii záznamů zahrnujících pro každé kliknutí na tlačítko myši souřadnice jejího kurzoru na obrazovce v pixelech a časové razítko.

Program BROWN slouží také ke zpracování zaznamenaných dat. Pro zajímavost stručně popíšeme postup zpracování, které program průběžně provádí po každém přidání bodu. Jde o několik oddělených výpočtů:

1) Vyhodnocení středního intervalu záznamu *t*. Jak je vysvětleno v úvodní kapitole, není možné jednoduše průměrovat časové prodlevy mezi záznamy. Proto je použita lineární regrese: Na osu *x* se vynáší pořadové číslo značky a osu *y* pak časové razítko vyjádřené jako počet sekund od pevného časového okamžiku. Střední doba trvání intervalu, resp. její směrodatná odchylka je pak přímo rovna směrnici proložené přímky, resp. její směrodatné odchylce. Ukazuje se, že takto získané hodnoty intervalu se velmi dobře shodují s nastavením časovače dokonce i v případech, kdy odchylka jednotlivého měření převyšuje jednu sekundu (například když experimentátor zaznamenává dlouhou sekvenci a z nějakého důvodu dočasně ztratí koncentraci). Pro určení směrodatné odchylky jednotlivého měření je použit reziduální součet čtverců a počet stupňů volnosti podle vztahu (38). Poznamenejme, že kromě kontroly pravidelnosti se při dalším zpracování odchylka jednotlivého měření času nijak nevyužívá, protože důsledky nepravidelných kliknutí se případně projeví ve větším rozptylu střední kvadratické vzdálenosti.

2) Vyhodnocení střední hodnoty kvadrátu vzdálenosti uražené za dobu *t*, 2*t*, 3*t* a 4*t* včetně směrodatných odchylek. Výpočet pro *n* zaznamenaných poloh se provede podle vztahu

$$\left\langle r_k^2 \right\rangle = \frac{1}{n-k} \sum_{i=1}^{n-k} \left[(x_{i+k} - x_i)^2 + (y_{i+k} - y_i)^2 \right] \tag{108}$$

pro k=1,2,3 a 4. Dále se pro tyto hodnoty vypočítá směrodatná odchylka aritmetického průměru

$$\sigma_{\left\langle r_k^2 \right\rangle} = \sqrt{\frac{1}{(n-k)(n-k-1)} \sum_{i=1}^{n-k} \left[(x_{i+k} - x_i)^2 + (y_{i+k} - y_i)^2 - \left\langle r_k^2 \right\rangle \right]^2} \,. \tag{109}$$

3) Vyhodnocení poměrů středních kvadratických posunutí Q_2, Q_3 a Q_4 podle vztahu

$$Q_k = \frac{\left\langle r_k^2 \right\rangle}{\left\langle r_1^2 \right\rangle} \tag{110}$$

včetně směrodatných odchylek

$$\sigma_{Q_k} = Q_k \sqrt{\left(\frac{\sigma_{\langle r_1^2 \rangle}}{\langle r_1^2 \rangle}\right)^2 + \left(\frac{\sigma_{\langle r_k^2 \rangle}}{\langle r_k^2 \rangle}\right)^2} \,. \tag{111}$$

Pokud pro danou částici poměry středních kvadratických posunutí odpovídají poměrům časových intervalů, neboli $1{:}Q_2{:}Q_3{:}Q_4$ se blíží k 1:2:3:4, pak se předpokládá, že chování částice není narušeno vnějšími vlivy a ta se pohybuje podle Einsteinova zákona. Jinými slovy to znamená, že střední kvadratické posunutí $\langle r^2 \rangle$ je lineárně závislé na délce časového intervalu t a hledaná aktivita částice A je úměrná směrnici této závislosti (faktor 2 souvisí s uspořádáním experimentu)

$$\langle r^2 \rangle = 2At \,. \tag{112}$$

Střední kvadratická posunutí jsou zpravidla zatížena nezanedbatelnou směrodatnou odchylkou $\sigma_{\langle r^2 \rangle}$. Ta je principiálním důsledkem použitého způsobu měření – každý statistický výpočet se zde zabývá jedinou částicí namísto souboru částic a ta je navíc pozorována jen po velmi omezenou dobu. Směrodatné odchylky středních kvadratických posunutí je nezbytné započítat do odchylky výsledné aktivity.

Ukážeme několik způsobů výpočtu aktivity a zaměříme se především na postup při vyhodnocení odchylek. Porovnáme číselné výsledky získané různými postupy pro tři datové sady (výstupy programu BROWN pro tři zaznamenané trajektorie) obsahující vyhodnocené hodnoty středního kvadratického posunutí $\langle r^2 \rangle$ a jejich směrodatné odchylky $\sigma_{\langle r^2 \rangle}$ pro čtyři násobky základního měřicího intervalu t a jeho směrodatné odchylky σ_t (*Tabulka 8*, *Tabulka 9* a *Tabulka 10*).

$\langle r^2 \rangle$ / μm²	$\sigma_{\langle r^2 \rangle}$ / μm²	t / s	σ_t / s
8.6	0.7	5.00	0.01
18.0	1.4	10.00	0.02
26.4	2.1	15.00	0.03
33.0	2.7	20.00	0.04

Tabulka 8

$\langle r^2 \rangle$ / μm²	$\sigma_{\langle r^2 \rangle}$ / μm²	t / s	σ_t / s
13.30	0.80	4.80	0.01
25.20	1.60	9.60	0.02
41.40	2.50	14.40	0.03
55.70	3.50	19.20	0.04

Tabulka 9

$\langle r^2 \rangle$ / μm²	$\sigma_{\langle r^2 \rangle}$ / μm²	t / s	σ_t / s
7.10	1.10	4.80	0.01
14.20	2.00	9.60	0.02
20.90	3.40	14.40	0.03
28.20	5.60	19.20	0.04

Tabulka 10

Vyhodnotíme poměry středních kvadratických posunutí Q_k podle (110) včetně směrodatných odchylek σ_{Q_k} (111). Pro všechny čtyři délky intervalu spočteme aktivitu A z Einsteinova vztahu (112)

$$A = \frac{\langle r^2 \rangle}{2t} \tag{113}$$

a také příslušné směrodatné odchylky

$$\sigma_A = A\sqrt{\left(\frac{\sigma_{\langle r^2 \rangle}}{\langle r^2 \rangle}\right)^2 + \left(\frac{\sigma_t}{t}\right)^2}\,. \tag{114}$$

Všechny uvedené veličiny nalezneme v horní části tabulek (*Tabulka 11, Tabulka 12 a Tabulka 13*).

Průměrovat nezávisle vypočtené hodnoty aktivity má podle teorie smysl, protože průměrováním 4 hodnot se nejistota výsledku zmenší dvakrát. Spočteme tedy na základě měřených hodnot aritmetický průměr aktivity A, dále výběrovou směrodatnou odchylku podle vztahu (67) a také směrodatnou odchylku aritmetického průměru (74). A vedle toho pro srovnání vypočteme složenou směrodatnou odchylku průměrné hodnoty také podle (78) pomocí známých odchylek měřených hodnot. Pokud ovšem známe odchylky měřených hodnot, jeví se jako korektnější použití nikoli aritmetického průměru, nýbrž průměru váženého podle (82) s výpočtem složené směrodatné odchylky podle (84).

Všechny tyto hodnoty jsou k nalezení ve střední části tabulek (*Tabulka 11, Tabulka 12 a Tabulka 13*). Přívlastek „složená“ ve spojení s pojmem směrodatná odchylka nemá žádný ustálený význam; v tomto textu je toto sousloví použito jen pro odlišení od běžně užívané směrodatné odchylky aritmetického průměru.

Jako alternativu k průměrování aktivit lze využít proložení lineární regrese na základě Einsteinova vztahu (112). Vedlejším výsledkem tohoto postupu zpracování je grafické ověření linearity. Nicméně i zde můžeme postupovat několika způsoby. Ukážeme si, jak to ovlivní výsledky. Pro porovnání tedy proložíme závislostí středního kvadratického posunutí na délce časového intervalu lineární regresi několika způsoby:

- obecná přímka metodou nejmenších čtverců (Excel)
- obecná přímka se započtením odchylek Y Error (Origin)

- obecná přímka se započtením odchylek X Error a Y Error (Origin, York)
- přímka procházející počátkem metodou nejmenších čtverců (Excel)
- přímka procházející počátkem se započtením odchylek Y Error (Origin)

Nalezené odhady směrnic proložených přímek včetně jejich směrodatných odchylek [přepočteno na aktivitu podle vztahu (112), tj. vyděleno dvěma] jsou uvedeny v dolní části tabulek (*Tabulka 11, Tabulka 12 a Tabulka 13*).

$\langle r^2 \rangle$ / μm²	$\sigma_{\langle r^2 \rangle}$ / μm²	t / s	σ_t / s	Q_k	σ_{Q_k}	A / μm²/s	σ_A / μm²/s
8.60	0.70	5.00	0.01	1.00		0.860	0.070
18.00	1.40	10.00	0.02	2.09	0.24	0.900	0.070
26.40	2.10	15.00	0.03	3.07	0.35	0.880	0.070
33.00	2.70	20.00	0.04	3.84	0.44	0.825	0.068

aritmetický průměr změřených hodnot A/(μm²/s)	0.866
výběrová směrodatná odchylka A/(μm²/s)	0.032
směrodatná odchylka aritmetického průměru A/(μm²/s)	0.016
složená směrodatná odchylka aritmetického průměru A/(μm²/s)	0.035
vážený průměr změřených hodnot A/(μm²/s)	0.865
složená směrodatná odchylka váženého průměru A/(μm²/s)	0.035

	A / μm²/s	σ_A / μm²/s
obecná přímka (Excel)	0.816	0.045
obecná přímka se započtením Y Error (Origin)	0.857	0.071
obecná přímka se započtením X Error a Y Error (Origin; York)	0.857	0.071
přímka procházející počátkem (Excel)	0.853	0.018
přímka procházející počátkem se započtením Y Error (Origin)	0.865	0.035

Tabulka 11

$\langle r^2 \rangle$ / μm²	$\sigma_{\langle r^2 \rangle}$ / μm²	t / s	σ_t / s	Q_k	σ_{Q_k}	A / μm²/s	σ_A / μm²/s
13.30	0.80	4.80	0.01	1.00		1.385	0.083
25.20	1.60	9.60	0.02	1.89	0.17	1.313	0.083
41.40	2.50	14.40	0.03	3.11	0.27	1.438	0.087
55.70	3.50	19.20	0.04	4.19	0.36	1.451	0.091

aritmetický průměr změřených hodnot A/(μm²/s)	1.396
výběrová směrodatná odchylka A/(μm²/s)	0.063
směrodatná odchylka aritmetického průměru A/(μm²/s)	0.031
složená směrodatná odchylka průměrné hodnoty A/(μm²/s)	0.043
vážený průměr změřených hodnot A/(μm²/s)	1.393
složená směrodatná odchylka váženého průměru A/(μm²/s)	0.043

	A / μm²/s	σ_A / μm²/s
obecná přímka (Excel)	1.494	0.060
obecná přímka se započtením Y Error (Origin)	1.431	0.090
obecná přímka se započtením X Error a Y Error (Origin; York)	1.431	0.090
přímka procházející počátkem (Excel)	1.426	0.027
přímka procházející počátkem se započtením Y Error (Origin)	1.393	0.043

Tabulka 12

$\langle r^2 \rangle$ / μm²	$\sigma_{\langle r^2 \rangle}$ / μm²	t / s	σ_t / s	Q_k	σ_{Q_k}	A / μm²/s	σ_A / μm²/s
7.10	1.10	4.80	0.01	1.00		0.740	0.115
14.20	2.00	9.60	0.02	2.00	0.42	0.740	0.104
20.90	3.40	14.40	0.03	2.94	0.66	0.726	0.118
28.20	5.60	19.20	0.04	3.97	1.00	0.734	0.146

aritmetický průměr změřených hodnot A/(μm²/s)	0.735
výběrová směrodatná odchylka A/(μm²/s)	0.007
směrodatná odchylka aritmetického průměru A/(μm²/s)	0.003
složená směrodatná odchylka průměrné hodnoty A/(μm²/s)	0.061
vážený průměr změřených hodnot A/(μm²/s)	0.735
složená směrodatná odchylka váženého průměru A/(μm²/s)	0.059

	A / μm²/s	σ_A / μm²/s
obecná přímka (Excel)	0.729	0.008
obecná přímka se započtením Y Error (Origin)	0.728	0.128
obecná přímka se započtením X Error a Y Error (Origin; York)	0.728	0.128
přímka procházející počátkem (Excel)	0.733	0.003
přímka procházející počátkem se započtením Y Error (Origin)	0.735	0.059

Tabulka 13

Z horní části všech tří tabulek je zřejmé, že poměry středních kvadratických posunutí se shodují s poměry délek trvání odpovídajících časových intervalů daleko přesněji, než jak běžně chápeme slovní spojení „v rámci směrodatných odchylek“. Datové sady, které nevyhovují kritériu „1:2:3:4“, se totiž v tomto experimentu zpravidla dále nezpracovávají. Cílem takového předvýběru je vyloučit případy se špatně připraveným vzorkem a nevhodně nastavenou aparaturou. Nevyhnutelným dalším důsledkem je ale také změna statistického charakteru dat – všechny započtené hodnoty aktivity jsou daleko blíže aritmetickému průměru, než dovoluje jejich skutečná směrodatná odchylka.

Popsané chování můžeme sledovat v horní části tabulek pro všechny tři datové sady (*Tabulka 11, Tabulka 12 a Tabulka 13*). I když samozřejmě neznáme skutečnou směrodatnou odchylku, máme k dispozici její odhad, který vychází z hodnot vypočtených programem BROWN na základě statistického vyhodnocení trajektorie částice. Vidíme, že směrodatné odchylky aktivity pro všechny čtyři délky časového intervalu vycházejí podobně.

Jak jsme si už vysvětlili, výpočtem jen z naměřených hodnot podle (67) nebo (74) získáme v důsledku předvýběru podhodnocené směrodatné odchylky. U nich navíc musíme očekávat nízkou spolehlivost, protože výběr obsahuje pouze čtyři body. [Podle (76) je relativní nejistota vypočtené odchylky zhruba 42 procent.] Zmíněné chování můžeme sledovat tentokrát ve střední části tabulek: Výběrová směrodatná odchylka podle (67) je ve všech případech podstatně menší než odhady vypočtené z dat programu BROWN (v horní části tabulek). Nejmarkantnější je tento rozdíl v případě třetí datové sady (*Tabulka 13*), což je dáno velmi přesným splněním podmínky předvýběru. Zároveň můžeme porovnat směrodatnou odchylku aritmetického průměru podle (74) se směrodatnou odchylkou, která byla získána podle (78) složením odhadů z programu BROWN. Zase vidíme systematické podcenění nejistoty průměrné hodnoty při použití (74). A opět je největší v případě třetí sady.

Poučné je také porovnání s výsledky získanými váženým průměrováním (82) a skládáním rozptylů (84), kterým by měla být obecně dána přednost v případě, kdy jsou známy rozptyly průměrovaných hodnot. Vidíme, že v našem případě, kdy individuální rozptyly se navzájem odlišují nejvýše o dvacet procent, přináší toto zdokonalení změnu jen zanedbatelnou v porovnání s velikostí směrodatné odchylky. [Porovnávané hodnoty jsou ve střední části tabulek zvýrazněny zelenou (vážený průměr a jeho odchylka) a žlutou barvou (aritmetický průměr a jeho odchylka).]

Závěr pro vyhodnocení aktivity je zřejmý: Zmenšit nejistotu výsledné aktivity je možno pomocí váženého průměrování (82). Při výpočtu směrodatné odchylky průměrné hodnoty spočívá jediný korektní způsob ve složení odchylek průměrovaných hodnot pomocí (84). Nicméně použitím aritmetického průměru (74) a zjednodušeného skládání individuálních odchylek (78) se (v této úloze) dopustíme jen zanedbatelné nepřesnosti ve srovnání se směrodatnou odchylkou výsledku.

Nakonec ještě shrneme poznatky o použití lineární regrese:

- Započtení odchylek na obou osách nevede (v tomto měření) na žádnou změnu oproti pouhému započtení odchylek na ose svislé. To nepřekvapuje, protože relativní odchylky času jsou zde zanedbatelné oproti relativním odchylkám středního kvadratického posunutí. Podobně (při použitém zaokrouhlení) nedojde zanedbáním nejistoty času ke změně výsledku formule (114).

- Všechny testované typy lineární regrese poskytují směrnice blízké [po přepočtu podle rovnice (112)] průměru změřených hodnot aktivity. Jsou tedy použitelné pro výpočet nějaké „středované“ hodnoty. (V každém případě je ale nutné specifikovat konkrétní použitý postup.) Výsledek nejbližší aktivitě počítané průměrováním dávají přímky procházející počátkem.

- Přímky prokládané metodou nejmenších čtverců (Excel) podceňují odchylku směrnice tím více, čím lépe data splňují podmínku předvýběru. Nejmarkantnější je to opět u třetí sady dat. Je to tedy situace obdobná jako při průměrování.

- Regresní přímka procházející počátkem se započtením odchylek (jen na svislé ose) poskytuje ve všech testovaných případech výsledek (směrnici a její odchylku) prakticky shodný s váženým průměrem (82) a odchylkou počítanou skládáním individuálních odchylek (84). [Hodnoty jsou v tabulkách zvýrazněny zelenou barvou.]

- Obecná regresní přímka se od té procházející počátkem mírně liší směrnicí, ale hlavně dává zhruba dvojnásobnou odchylku [vysvětleno při odvození vztahu (92)]. To platí se započítáním individuálních odchylek i bez něj. V případě započtení odchylek čtyř vzorků je tak odchylka výsledku na úrovni jednotlivého měření. Pokud tedy lineární závislost určitě prochází počátkem, může být použití obecné přímky pokládáno za nevhodné.

Závěr pro použití lineární regrese při vyhodnocení aktivity Brownova pohybu: Vyhodnocení aktivity částice pomocí lineární regrese je přípustné, ale ne nezbytné. Vzhledem k tomu, že statistický soubor obsahuje v tomto případě pouze čtyři hodnoty, je vhodnější prokládat přímku procházející počátkem. Neumožňuje-li nástroj pro výpočet lineární regrese započítat odchylky vstupních hodnot, pak korektním řešením je namísto regrese použít vážený (případně aritmetický) průměr aktivit a odhadnout jeho směrodatnou odchylku pomocí příslušného skládání odchylek.

Dodatek. Několik poznámek k základním statistickým výpočtům

V tomto studijním textu se používá řada pojmů a vztahů z oblasti matematické statistiky a zpracování experimentálních měření. Některé jsou běžně známé, jiné byly v teoretické části stručně vysvětleny. Nicméně další důležité nebo alespoň zajímavé informace byly odsunuty do tohoto dodatku, aby hlavní linie výkladu týkající se regresního počtu nebyla rušena zbytečnými odbočkami.

Směrodatná odchylka a rozptyl

Jak jsme výše předpokládali, vstupní hodnoty (typicky výsledky opakovaného měření) se liší od skutečné hodnoty a hledané funkce o náhodné veličiny, které mají nulovou střední hodnotu, nejsou korelovány a mají všechny stejný **rozptyl** (anglicky **variance**) neznámé velikosti σ^2. (Naopak jsme zatím nepředpokládali konkrétní rozdělení.)

Kvadrát směrodatné odchylky výběru (67) je odhadem právě tohoto neznámého rozptylu. Kdybychom měli k dispozici celou populaci hodnot (tj. zhruba řečeno všechny analogické měřené hodnoty, které mohly být za stejných podmínek získány a jejichž počet N se blíží nekonečnu), pak by skutečná hodnota a funkce byla rovna střední hodnotě populace

$$a = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{115}$$

a rozptyl σ^2 (pro jednoznačnost bychom ho měli nazývat **rozptylem populace** neboli anglicky **population variance**) by byl určen vztahem

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a\right)^2 . \tag{116}$$

Protože však výpočty provádíme pouze pro výběr z této populace ($n << N$), můžeme spočítat pouze odhad a^* podle (63) a s^2 (**výběrový rozptyl** neboli **sample variance**) jakožto kvadrát výběrové směrodatné odchylky (67)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - a^*\right)^2 . \tag{117}$$

To znamená, že **výběrovou směrodatnou odchylku** s můžeme zároveň chápat jako odhad **směrodatné odchylky celé populace** σ.

Zaznamenejme, že ve jmenovateli (117) vystupuje počet měření zmenšený o jedna oproti vztahu (116). Tato korekce je nezbytná, abychom získali **nevychýlený** odhad (unbiased estimator), tzn. takový odhad, který se při středování přes všechny možné zkoumané výběry blíží skutečné hodnotě odhadovaného parametru. Důležitá je totiž skutečnost, že v (116) se v závorce odečítá parametr a, zatímco v (117) jeho odhad a^*. Ukážeme si později, jak z této zdánlivé maličkosti vyplyne právě ona jedničková korekce.

V programu Excel se pro výpočet **výběrové směrodatné odchylky** podle (117) použije funkce SMODCH.VÝBĚR.S nebo STDEV. Naopak funkce SMODCH.P nebo STDEVP pracují podle vzorce (116); proto je na konci písmeno P jako populace. Jméno funkce je v anglické variantě odvozeno od termínu **standard deviation**, případně pro jednoznačnost **sample standard deviation.** Do češtiny se pak vrací v překladu **standardní odchylka**.

Jako příklad použití výběrové směrodatné odchylky může posloužit vyhodnocení velikosti testovacích částic pro výpočet Avogadrovy konstanty na základě záznamu Brownova pohybu. V této úloze se z mikrofotografie pro další výpočet vyhodnocuje průměrný rozměr částic. Pro stanovení nejistoty výsledné hodnoty je také třeba statisticky určit, k jaké odchylce povede náhodná volba některé z existujících částic pro záznam trajektorie. A právě toto vyjadřuje výběrová směrodatná odchylka. (Pro experiment se nepoužije hypotetická průměrná částice, takže není vhodné použití směrodatné odchylky průměru, která říká, jak přesně umíme tento průměr určit.)

Směrodatná odchylka aritmetického průměru

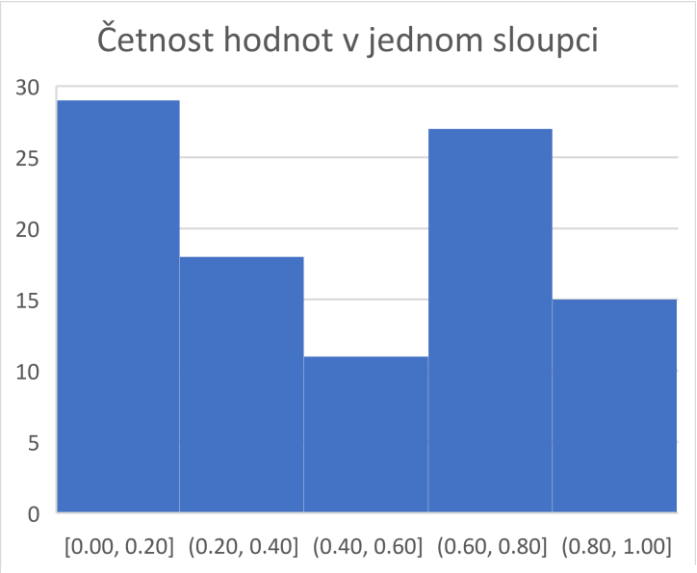
Jak bylo již několikrát řečeno, zatím jsme nepředpokládali konkrétní rozdělení měřených náhodných hodnot kolem skutečné hodnoty neznámé funkce. V případě diskrétních rozdělení (házení kostkou) se toto popisuje pravděpodobností jednotlivých generovaných hodnot. V případě spojitých rozdělení se pro popis použije takzvaná hustota pravděpodobnosti $\rho(x)$, která po vynásobení šířkou intervalu hodnot dx udává pravděpodobnost výskytu hodnoty uvnitř intervalu dx v okolí hodnoty x. Podobně jako v případě diskrétního rozdělení musí součet pravděpodobností všech hodnot být roven jedné, bude pro diskrétní rozdělení platit

$$\int_{-\infty}^{\infty} \rho(x)dx = 1. \tag{118}$$

Pojďme pomocí Excelu prozkoumat chování součtu velkého počtu náhodných hodnot, které budou simulovat výsledky měření. Záměrně použijeme pro vstupní hodnoty velmi jednoduché rozdělení. Vstupní hodnoty budou rovnoměrně rozprostřeny pouze v intervalu 0 až 1, neboli měřené hodnoty budou součtem konstanty $\frac{1}{2}$ a náhodné hodnoty s rozdělením symetrickým kolem počátku souřadnic (pak střední hodnota je nula) určeným hustotou pravděpodobnosti

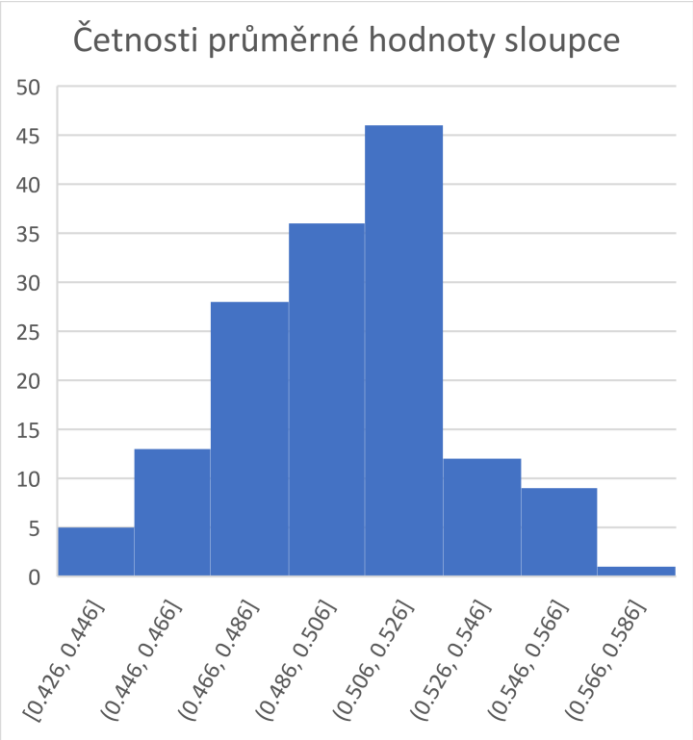
$$\rho(x) = \begin{cases} 1 & \text{pro } x \in \left\langle -\frac{1}{2}, \frac{1}{2} \right) \\ 0 & \text{pro } x \notin \left\langle -\frac{1}{2}, \frac{1}{2} \right) \end{cases} \tag{119}$$

vyhovující podmínce (118). V Excelu právě takové hodnoty generuje funkce NÁHČÍSLO nebo RAND. Vytvořme si kopírováním v tabulce Excelu sloupec A zahrnující 100 buněk obsahujících „=NÁHČÍSLO()“ a pak na jeho konec doplňme buňku obsahující „=PRŮMĚR(A1:A100)“ a další buňku obsahující „=SMODCH.VÝBĚR(A1:A100)“. Tento sloupec opět kopírujme tak, aby tabulka obsahovala 150 stejných sloupců. Pak můžeme pomocí grafu typu HISTOGRAM zkoumat rozdělení četností hodnot v tabulce. Jednotlivé sloupce tabulky budou nezávislými výběry (n=100), souhrn všech sloupců bude populace (N=15000).



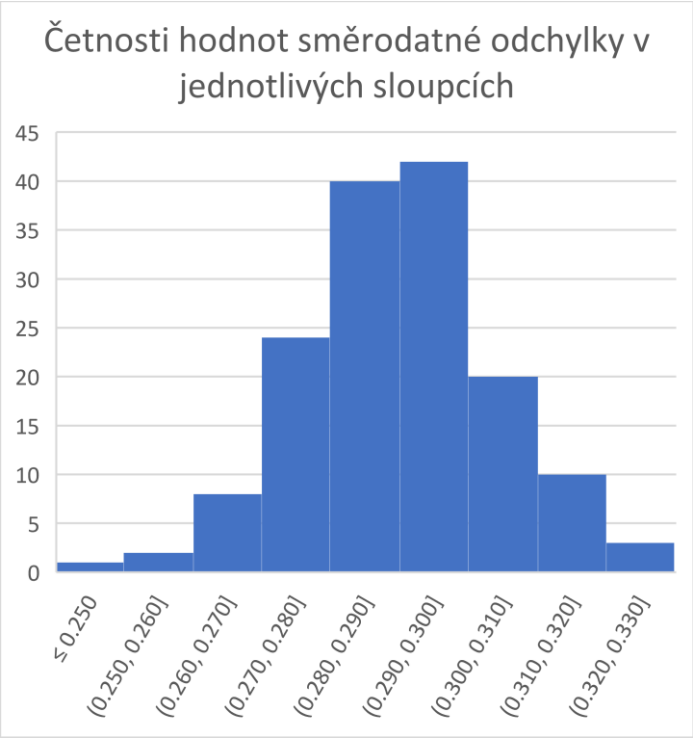
Histogram 1

Nejprve zobrazme *Histogram 1* obsahující četnosti hodnot v prvním sloupci – stačí do histogramu vložit jako zdrojová data oblast A1:A100. Po každém přepočítání stránky klávesou F9 se zobrazí jiná náhodná čísla a histogramy se změní. Společným znakem všech těchto histogramů bude, že v celém intervalu 0 až 1 budou četnosti srovnatelné. Sto hodnot je ale příliš málo na ověření, že použitá funkce skutečně generuje náhodná čísla s rovnoměrným rozdělením.



Histogram 2

Zobrazme analogicky *Histogram 2*, zobrazující četnosti průměrné hodnoty vypočtené pro jednotlivé sloupce (A101:ET101). Očekáváme, že střední hodnota bude přibližně 0.5, ale podoba histogramu může leckoho překvapit. Tentokrát totiž přísluší nenulové četnosti pouze hodnotám v okolí 0.5 a navíc četnost prudce klesá se vzdáleností od této střední hodnoty. Právě jsme si ilustrovali důsledek **centrální limitní věty**, která říká, že se rozdělení výběrového průměru vždy blíží k normálnímu rozdělení, a to bez ohledu na to, jaké je rozdělení průměrované náhodné veličiny. Zobrazený histogram bude tedy pro rostoucí počet měření ve výběru konvergovat k tzv. „gaussovce“.



Histogram 3

Dále zobrazme *Histogram 3*, kde jsou vyneseny četnosti hodnot výběrové směrodatné odchylky spočtené pro jednotlivé sloupce (A102:ET102). Opět má histogram podobný tvar, i když výpočet funkce tentokrát neobsahuje jenom sčítání.

Využijeme-li známé parametry náhodné veličiny, můžeme pro ni vypočítat hodnotu rozptylu populace

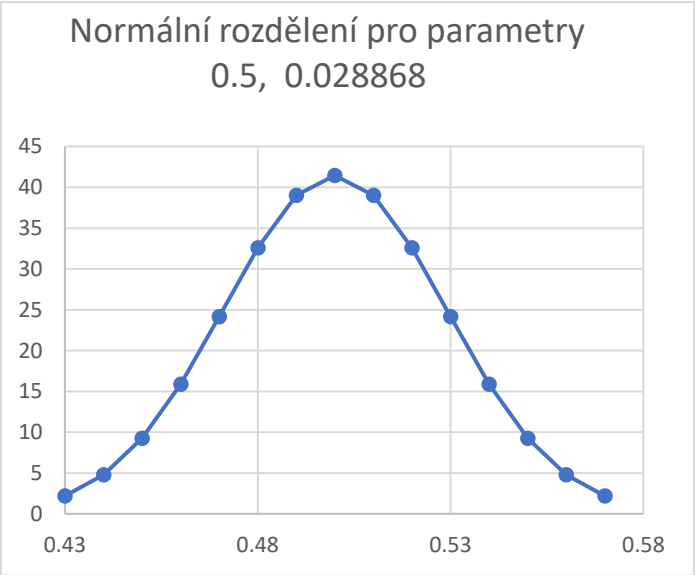
$$\sigma^2 = \int_{-\infty}^{\infty} x^2 \rho(x) dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x^2 dx = \frac{1}{12} \tag{120}$$

Z histogramu je zřejmé, že kvadrát střední hodnoty výběrové směrodatné odchylky (přibližně 0.29) skutečně je dobrým odhadem rozptylu σ^2 (platí $\sqrt{\frac{1}{12}} \doteq 0.288$), který charakterizuje rozdělení náhodných hodnot vystupujících v určující rovnici.

Vraťme se teď k rozdělení četnosti hodnot aritmetického průměru (*Histogram 2*). Onou zmiňovanou „gaussovkou“ se v běžné mluvě rozumí funkce, jejíž tvar odpovídá průběhu hustoty pravděpodobnosti v normálním (též Gaussově) rozdělení

$$\rho(x) = \frac{1}{\sigma_N \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma_N^2}}, \tag{121}$$

kde parametr μ představuje střední hodnotu a σ_N^2 představuje rozptyl. Často se pro normální rozdělení s těmito parametry používá označení $N(\mu, \sigma_N^2)$. V programu Excel je k dispozici funkce NORM.DIST s číselnými parametry x, μ a σ_N. Pokud čtvrtý parametr funkce NORM.DIST má hodnotu NEPRAVDA, vrací funkce hustotu pravděpodobnosti normálního rozdělení.



Graf 8

Zajímá nás, jaké hodnoty parametrů má normální rozdělení, kterému odpovídá *Histogram 2* četnosti průměrných hodnot. Kvůli porovnání s histogramem se do grafu (*Graf 8*) budou vynášet četnosti, proto hustotu pravděpodobnosti násobíme šířkou intervalu (0.02) a celkovým počtem výběrů (zde 150). Jako střední hodnotu samozřejmě použijeme $\mu = \frac{1}{2}$. Proporce vrcholu se nastaví jediným zbývajícím parametrem σ_N. Ten sice nemůžeme z histogramu spočítat, ale můžeme vizuálně ověřit, že průběh četnosti v histogramu (*Histogram 2*) je dobře vystižen, pokud pro σ_N platí

$$\sigma_N^2 = \frac{\sigma^2}{n}, \tag{122}$$

kde σ^2 je rozptyl populace (náhodných veličin $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$; zde výše zmíněná hodnota $\frac{1}{12}$) a n je počet průměrovaných vzorků výběru (zde 100). Za povšimnutí stojí skutečnost, že na levé straně jde o parametr normálního rozdělení, zatímco parametr na pravé straně jsme modelovali pomocí rovnoměrného rozdělení.

Vztah mezi rozptyly (122) nebudeme odvozovat. Postup by byl obdobný jako v případě vztahu (73) pro odhady rozptylů, tj. pro veličiny, které v tomto textu označujeme symbolem s^2. Mezi odhady patří i výběrový rozptyl. Analogický vztah (71) platí i mezi odmocninami rozptylů (tj. mezi směrodatnými odchylkami), ať jsou vztaženy k celé populaci nebo jen k výběru.

Zabývejme se ještě výběrovou směrodatnou odchylkou (67) a směrodatnou odchylkou aritmetického průměru (74). Kromě toho, že teď už víme, proč mezi nimi platí (71), je také jasnější význam analogických veličin v případě obecné lineární regrese: Výběrová směrodatná odchylka charakterizuje rozdělení náhodných veličin $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ (její kvadrát představuje jejich výběrový rozptyl), přičemž typ rozdělení nemusí být specifikován. V případě odhadů parametrů regrese jde vždy o normální rozdělení, jehož střední hodnota je rovna odhadu příslušného parametru a rozptyl (přesněji jeho odhad) je pak roven kvadrátu (odhadu) směrodatné odchylky tohoto parametru. Toto normální rozdělení říká, jaké by při opakovaném měření s různými výběry vzorků z téže populace byly možné hodnoty odhadu příslušného parametru a jejich pravděpodobnosti.

V programu Excel neexistuje pro výpočet **směrodatné odchylky aritmetického průměru** podle (71) samostatná funkce. Do příslušné buňky v tabulce je nutno vložit vzorec =SMODCH.VÝBĚR.S/ODMOCNINA(n) nebo =STDEV/SQRT(n). V anglických textech se používá spojení **standard error**, což pak vede na český překlad **standardní chyba**.

Nejistota není chyba

Text této kapitoly vychází především z doporučení pro zpracování naměřených hodnot vydaného mezinárodním výborem pro míry a váhy [11].

Každé měření je určitým způsobem nedokonalé, důsledkem čehož vznikají chyby měřených hodnot. Jako chyba se označuje rozdíl mezi skutečně naměřenou hodnotou (odečtenou z přístroje) a správnou (ale neznámou) hodnotou veličiny. Podle tradiční představy má chyba měření dvě složky – *náhodnou* a *systematickou*.

Náhodná složka se chápe jako důsledek nepředvídatelných změn veličin ovlivňujících měření. Tyto náhodné efekty pak způsobují rozdíly mezi výsledky opakovaných měření. Není sice možno náhodné chyby kompenzovat, ale s rostoucím počtem opakování lze očekávat, že střední hodnota chyby se bude blížit nule.

Systematická složka také nemůže být eliminována, ale často ji lze zredukovat. Pokud je důsledkem známého jevu závislého na ovlivňujících veličinách (proto systematický efekt), pak je (alespoň principiálně) možno doplnit korekční postup pro kompenzaci tohoto efektu. V důsledku by se pak střední hodnota chyby dané systematickým efektem měla blížit nule.

Ale pozor, v moderních textech se směrodatná odchylka aritmetického průměru série měření striktně odlišuje od náhodné složky chyby průměru. Směrodatná odchylka se chápe jako míra nejistoty stanovení průměru, která je důsledkem působení náhodných efektů. Skutečnou velikost chyby plynoucí z náhodných efektů naopak zjistit nemůžeme, zrovna jako neznáme správnou hodnotu měřené veličiny. Podobně je nutno odlišit (principiálně neznámou) systematickou složku chyby, jež je důsledkem nedokonalé kompenzace, od nejistoty korekce měření nutné pro kompenzaci systematického efektu, která je důsledkem neúplné znalosti potřebných parametrů korekce.

Nejistota se v novějším pojetí chápe nejen abstraktně jako nedostatek určitosti, nýbrž také jako nová statistická veličina (**nejistota měření** nebo často jen zkráceně **nejistota**), která

charakterizuje neurčitost procesu měření. Přesněji řečeno, charakterizuje rozptyl hodnot, jichž může nabýt naměřená hodnota. Jednou z možností, jak nejistotu měření parametrizovat, je použití směrodatné odchylky (anglicky standard deviation). Taková nejistota se pak označuje jako **standardní nejistota**. Chyba a nejistota (odchylka) se už nepoužívají jako synonyma, což ve starší literatuře bývalo časté. Tak může být bezrozporně popsána situace, kdy měřená hodnota (včetně korekce) se přiblíží ke skutečné (ale neznámé) hodnotě měřené veličiny (a má tedy zanedbatelnou chybu), přestože nejistota měření (vyjádřená na základě statistického vyhodnocení) zůstává podstatně větší.

Nejistoty typu A a B

Doporučení pro výpočty s naměřenými hodnotami rozlišují dva způsoby vyhodnocení nejistot [11]. Jako typ A se označuje vyhodnocení statistickou analýzou série opakovaných měření. Vyhodnocení typu B stanovuje nejistotu na základě předpokládaného rozložení pravděpodobnosti, které vychází ze znalosti zkoumaných procesů.

Charakter nejistot vyhodnocených podle A a B přibližně odpovídá zmiňovaným náhodným a systematickým chybám, ale tato podobnost má své meze. Nejde jen o to, že nejistota není totéž co chyba, ale hlavně se atribut nevztahuje přímo k nejistotě určité veličiny, nýbrž ke způsobu jejího získání. Například nejistotu korekce systematického efektu lze někdy vyhodnotit postupem A, ale jindy je nutný postup B.

Standardní nejistota typu A měřené hodnoty je rovna směrodatné odchylce této hodnoty, přičemž podstatný je samozřejmě také výpočetní postup. Pokud například výsledná měřená hodnota je průměrem série naměřených hodnot, použije se směrodatná odchylka aritmetického průměru. Je-li ovšem použit vážený průměr, pak nejistotu určuje směrodatná odchylka váženého průměru. Při použití regrese je měřená hodnota v závislosti na parametrech dána rovnicí regresní závislosti, jejíž každý koeficient je vyhodnocen včetně příslušné směrodatné odchylky.

Standardní nejistota typu B měřené hodnoty také představuje její směrodatnou odchylku (takže je možno ji skládat s ostatními standardními nejistotami), nicméně je vyhodnocena na základě kalibrace a dalších znalostí o měřicím zařízení.

Při zpracování série naměřených dat (vyhodnocení nejistot typu A) se zpravidla uplatní centrální limitní věta, takže je přirozené předpokládat normální rozdělení průměrných hodnot nebo jiných výsledných parametrů (např. koeficientů regrese). V případě zpracování typu B je situace odlišná, protože tehdy mohou být zahrnuty také systematické efekty. Jako příklad můžeme uvést nejistotu délkových měřidel. Ta jsou jistě při výrobě zatížena chybou danou složením náhodných faktorů a při měření se k tomu přidají další náhodné nepřesnosti, takže bychom mohli očekávat, že délka změřená náhodně vybraným měřidlem se bude od té správné lišit o hodnotu charakterizovanou normálním rozdělením. Výrobce ale také musí zohlednit například teplotní roztažnost materiálu. (U délkových měřidel se projevuje jako systematický efekt, který se navíc nesnadno kompenzuje. Třímetrové ocelové pásmo s koeficientem roztažnosti cca 1.5×10^{-5} K^{-1} se při zvýšení teploty o 25 °C protáhne o více než jeden milimetr.) Nejjednodušší způsob spočívá v tom, že se případná změna délky započte jako součást nejistoty měřidla typu B.

U jednoduchých měřidel se často předpokládá, že hodnota měřené veličiny leží při každém měření v intervalu $-\Delta X_{\max}$ až $\Delta X_{\max}$ s pravděpodobností prakticky rovnou jedné a naopak pravděpodobnost, že hodnota měřené veličiny leží vně tohoto intervalu, je v podstatě nulová. Pokud není k dispozici specifická znalost rozložení hodnot uvnitř intervalu, je přirozené předpokládat rovnoměrné rozdělení. Takto zadaná nejistota $\Delta X_{\max}$ se označuje jako **mezní**. Analogicky k výpočtu rozptylu pro jednotkový interval (120) dostaneme pro interval $-\Delta X_{\max}$ až $\Delta X_{\max}$ rozptyl

$$\sigma^2 = \frac{\Delta X_{\max}^2}{3}, \tag{123}$$

takže výběrová směrodatná odchylka takového měření se z mezní odchylky spočte takto

$$\sigma = \frac{\Delta X_{\max}}{\sqrt{3}}. \tag{124}$$

Povšimněme si, že v případě rovnoměrného rozdělení leží veškeré možné hodnoty v intervalu $\pm\sqrt{3}\,\sigma$ od středu.

Skládání nejistot

Kdyby bylo možno nezávisle měnit všechny vstupní parametry, na nichž závisí výsledek měření, pak by jeho nejistotu bylo možno získat statistickými nástroji. (Tomu by samozřejmě muselo být přizpůsobeno uspořádání experimentu.) Protože takový přístup je vzhledem k omezením času a zdrojů zřídka možný, stanoví se namísto toho (standardní) nejistoty σ_j vstupních hodnot y_j a pak se předpokládá platnost matematické závislosti [modelu vyjádřeného funkcí $f(y_1, y_2, \ldots, y_m)$] a zákona šíření nejistot, kdy se složená (kombinovaná) standardní nejistota σ_f výsledku spočte jako odmocnina ze složeného rozptylu σ_f^2 podle formule (3) pro skládání standardních nejistot, ať byly vyhodnoceny postupem typu A nebo B

$$\sigma_f^2 = \sum_{j=1}^{m} \left(\frac{\partial f}{\partial y_j} \right)^2 \sigma_j^2 .$$

Stojí nicméně za připomenutí, že pokud ve funkci $f(y_1, y_2, \ldots, y_m)$ jako argumenty vystupují korelované veličiny, musí být při skládání nejistot započteny také kovariance. V případech, kdy funkce $f(y_1, y_2, \ldots, y_m)$ je silně nelineární, může zase být nezbytné doplnit do rovnice (3) další členy Taylorova rozvoje.

Dva popsané postupy vyhodnocení nejistoty výsledku lze aplikovat nejen na celý experiment, ale také na jeho libovolnou část. To znamená, že pokud je součástí zpracování např. výpočet průměrné hodnoty nebo lineární regrese, je možno nejistotu jeho výsledku vypočítat jednak statisticky (směrodatná odchylka aritmetického průměru nebo směrodatné odchylky regresních koeficientů), jednak je možno použít skládání nejistot vstupních hodnot. Pro nejistoty typu A

by vzhledem k tomu, že popisují stejné statistické chování, měly oba způsoby dávat zhruba stejné hodnoty, jejichž rozdíl by měl klesat s počtem opakovaných měření.

Při kombinaci s nejistotami typu B je situace složitější: Nejistota typu B může mít systematický charakter, takže statistické vyhodnocení opakovaných měření provedených jen za určitých podmínek vede na menší nejistotu (příkladem může být výše zmíněné délkové měřítko, jehož kalibrace se mění s teplotou); naopak větší nejistota statisticky vyhodnocená než nejistota typu B uvedená v kalibrační dokumentaci zařízení může svědčit o nesprávném způsobu měření.

Nejistoty vyhodnocené z opakovaných měření a ty spočtené pomocí šíření nejistot se rozhodně neskládají (to by vedlo na zdvojené započtení nejistoty téže veličiny, před kterým zmíněná doporučení přímo varují [11]), nicméně pokud jedna z nich značně převyšuje druhou, měla by se pro charakterizaci výsledků použít ta větší, která už tu menší v sobě zahrnuje.

Toto samozřejmě platí i pro měření jednotlivých vstupních veličin: Předpokládejme, že byl měřidlem opakovaně změřen určitý rozměr, vypočten průměr a jeho standardní nejistota (typu A). Předpokládejme dále, že vedle toho je k dispozici dokumentace k měřidlu, kde se uvádí mezní chyba v závislosti na měřené délce, a na jejím základě byla podle vztahu (124) stanovena standardní nejistota typu B. Pak se jako změřená hodnota použije vypočtený průměr a jeho nejistota bude ta větší z obou výše stanovených.

Pozor na překlady z angličtiny

Již byla zmíněna nejednotnost v zacházení s pojmy chyba a odchylka. V českých textech (zejména těch starších) se často například **směrodatná odchylka**, **směrodatná chyba** a **střední kvadratická chyba** chápou jako synonyma. Zdrojem nedorozumění při překladu z cizích jazyků (vedle doby vzniku dokumentu) mohou být také jiné zvyklosti v označování veličin. Například v českém textu se upřesňuje, o jaký typ směrodatné odchylky jde, pomocí dalších atributů (výběrová směrodatná odchylka nebo směrodatná odchylka aritmetického průměru). V anglickém textu označuje **standard deviation** zpravidla tu první, zatímco **standard error** tu druhou. Ani v případě překladu ze starších anglických pramenů tedy **standardní odchylka** neznamená totéž co **standardní chyba**. Podobně **mean squared error** nelze překládat jako **střední kvadratickou chybu**.

Stručně o Studentově rozdělení

Pivovarský sládek W. S. Gosset prováděl testy kvality produkovaného moku na základě odběru malého počtu vzorků. Pro pravděpodobnostní hodnocení výsledků vypracoval novou teorii využívající tzv. *t*-rozdělení. Podle zjištění historiků měl zaměstnavatelem kvůli utajení povoleno publikovat výsledky výhradně pod pseudonymem a nesměl použít slovo „pivo“ ani jméno pivovaru. Protože si prý zapisoval poznámky do školního sešitu, zvolil si pseudonym Student. (To jen na vysvětlenou, proč se píše Studentovo rozdělení s velkým „S“.) Možná ještě stojí za zmínku, že než strávil zbytek kariéry ve světoznámém pivovaru Guinness, stačil Gosset vystudovat matematiku v Oxfordu.

Předpokládejme normální rozdělení výsledků experimentu. (Dosud jsme to nepožadovali.) Pak opakovaným měřením můžeme získat nejen střední hodnotu μ a rozptyl σ^2, ale také budeme znát pravděpodobnostní rozložení měřených hodnot. (Můžeme pak například tvrdit, že v intervalu šíře σ okolo střední hodnoty se nachází 68 % hodnot apod.) Potíže nastanou, pokud je změřených vzorků malý počet. Pak totiž musíme zohlednit skutečnost, že skutečná střední hodnota rozdělení a rozptyl se liší od statistických odhadů. V předchozí kapitole jsme si ukázali vliv takového rozdílu. Řešení spočívá v použití Studentova rozdělení.

Chápejme dále hodnoty $x_1, x_2, \ldots, x_n$ jako výběr n nezávislých vzorků normálního rozdělení $N(\mu, \sigma^2)$. Skutečné hodnoty μ ani σ^2 neznáme (pro výpočet by bylo nutno znát celou populaci), takže nemůžeme vyčíslit hustotu pravděpodobnosti. Pro velká n, kdy odhady statistických veličin se od jejich skutečných hodnot příliš neliší, můžeme nahradit μ průměrnou hodnotou výběru $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \tag{125}$$

rozptyl σ^2 můžeme nahradit kvadrátem výběrové směrodatné odchylky

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{126}$$

a pro výpočty pravděpodobnosti použít rozdělení $N(\bar{x}, s^2)$. Pro malá $n < 30$ už je ale nezanedbatelný rozdíl mezi asymptotickým normálním rozdělením, jehož parametry neznáme, a rozdělením odvozeným od statistických odhadů.

Není složité ilustrovat zavedení Studentova rozdělení. Vycházíme z toho, že průměrná hodnota výběru $\bar{x}$ se řídí normálním rozdělením $N(\mu, \frac{\sigma^2}{n})$. Ze vztahu (121) je pak zřejmé, že náhodná veličina

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \tag{127}$$

má standardní normální rozdělení $N(0,1)$. Dá se ukázat, že pokud nahradíme σ odhadem s, pak veličina

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \tag{128}$$

bude mít Studentovo *t*-rozdělení pro $n-1$ stupňů volnosti. Pro daný výběr jsou hodnoty $\bar{x}$, s a také n známy. Jediný neznámý parametr μ můžeme chápat jako proměnnou a spočítat pro něj průběh pravděpodobnosti. Detaily se zabývat nebudeme. Stačí si jen zapamatovat, že průběh Studentova *t*-rozdělení se s rostoucím počtem stupňů volnosti blíží k normálnímu rozdělení.

Pro malá n má ale tato „gaussovka“ výrazně širší základnu, takže se rozšiřují také intervaly okolo střední hodnoty, kde se skutečná hodnota vyskytuje s požadovanou pravděpodobností.

Pro praktické použití při měření a zpracování dat se mohou hodit dvě skutečnosti týkající se Studentova rozdělení:

1) Pro výběry obsahující aspoň 10 vzorků odpovídá hladině pravděpodobnosti 68 % interval spolehlivosti zhruba $\langle -s, +s \rangle$ (zanedbáním korekčního faktoru se dopustíme chyby srovnatelné se zaokrouhlením na dvě platné cifry). Takže pokud se při zpracování dat chceme vyhnout nutnosti použít koeficienty Studentova rozdělení, změřme vždy aspoň 10 vzorků.
2) Intervaly spolehlivosti pro hladiny pravděpodobnosti nad 90 % zůstávají ve srovnání s normálním rozdělením významně rozšířené ještě pro 50 vzorků. Proto musíme být obezřetní při posuzování hrubých chyb. Při deseti vzorcích totiž bude hladině pravděpodobnosti 99.7 % odpovídat interval spolehlivosti $\langle -4s, +4s \rangle$ a nikoli $\langle -3s, +3s \rangle$ jako v případě normálního rozdělení.

Proč je ve jmenovateli vzorce pro výběrovou směrodatnou odchylku $n-1$ a ne jenom n

Budeme zkoumat, jak se chová střední hodnota odhadu (117) přes všechny možné výběry.

Nejprve upravíme (117) s využitím normální rovnice

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - a^*\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^2 - 2y_i a^* + a^{*2}\right) =$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - 2a^*\underbrace{\sum_{i=1}^{n} y_i}_{na^*} + a^{*2}\underbrace{\sum_{i=1}^{n} 1}_{n}\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - na^{*2}\right) \tag{129}$$

Pak upravíme reziduální součet, aby obsahoval neznámou správnou hodnotu parametru a a provedeme obdobné úpravy jako v předchozím případě.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[(y_i - a) - (a^* - a)\right]^2 =$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left[(y_i - a)^2 - 2(y_i - a)(a^* - a) + (a^* - a)^2\right] =$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(y_i - a)^2 - 2(a^* - a)\underbrace{\sum_{i=1}^{n}(y_i - a)}_{\sum_{i=1}^{n} y_i - na} + (a^* - a)^2 \sum_{i=1}^{n} 1\right] =$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(y_i - a)^2 - n(a^* - a)^2\right] = \tag{130}$$

$$= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - a)^2 - (a^* - a)^2\right]$$

Když budeme právě obdržený výraz (130) pro s^2 středovat přes všechny možné výběry, poznáváme v závorce na pravé straně dva členy se známým významem

$$\left\langle s^2 \right\rangle = \frac{n}{n-1}\left[\left\langle \frac{1}{n}\sum_{i=1}^{n}(y_i - a)^2 \right\rangle - \left\langle (a^* - a)^2 \right\rangle\right]. \tag{131}$$

První člen (131) představuje rozptyl σ^2, jehož hodnotu neznáme a snažíme se ji odhadnout. Druhý člen představuje rozptyl odhadu parametru a^* okolo neznámé skutečné hodnoty parametru a. Protože pro každý výběr velikosti n je odhad parametru a^* roven aritmetickému průměru hodnot y_i přes celý výběr a parametr a je roven střední hodnotě y_i přes celou populaci, můžeme druhý člen (131) chápat jako rozptyl aritmetického průměru, pro nějž platí σ^2/n. Pak můžeme zapsat

$$\left\langle s^2 \right\rangle = \frac{n}{n-1}\left[\sigma^2 - \frac{\sigma^2}{n}\right] = \sigma^2. \tag{132}$$

Vidíme, že bez korekce ve jmenovateli (117) bychom dostali vychýlený odhad (biased estimator), který by vedl na podcenění rozptylu zejména při malých výběrech.

Je třeba zdůraznit, že ačkoliv jsme takto získali nevychýlený odhad rozptylu, neznamená to, že odhad výběrové směrodatné odchylky je také nevychýlený. Příčinou je konkávní charakter odmocniny. Obecná formule pro nevychýlený odhad výběrové směrodatné odchylky neexistuje.

Nevychýlený odhad rozptylu při lineární regresi

Pro získání nevychýleného odhadu rozptylu v případě regresní závislosti obsahující p konstant musí ve jmenovateli vždy vystupovat počet stupňů volnosti reziduálního součtu čtverců (37) a nikoli přímo počet bodů. Jako zdůvodnění se zpravidla uvádí, že pouze $n-p$ naměřených hodnot je nezávislých, protože jsou svázány pomocí p konstant.

Obdobným postupem jako pro výběrovou směrodatnou odchylku můžeme v případě lineární regrese ověřit, že odhad rozptylu bude nevychýlený, pokud ve jmenovateli bude $n-2$. Kromě rozptylu obou odhadů bude tentokrát třeba započítat také jejich kovarianci.

Budeme zkoumat, jak se chová kvadrát střední hodnoty (45) přes všechny možné výběry. Nejprve upravíme kvadrát (45) s využitím normálních rovnic

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}[y_i-(a^*+b^*x_i)]^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[y_i^2-2y_i(a^*+b^*x_i)+(a^*+b^*x_i)^2\right] = \\ &= \frac{1}{n-2}\left[\sum_{i=1}^{n}y_i^2-2a^*\underbrace{\sum_{i=1}^{n}y_i}_{na^*+b^*\sum_{i=1}^{n}x_i}-2b^*\underbrace{\sum_{i=1}^{n}x_iy_i}_{a^*\sum_{i=1}^{n}x_i+b^*\sum_{i=1}^{n}x_i^2}+\underbrace{\sum_{i=1}^{n}(a^*+b^*x_i)^2}_{na^{*2}+2a^*b^*\sum_{i=1}^{n}x_i+b^{*2}\sum_{i=1}^{n}x_i^2}\right] = \\ &= \frac{1}{n-2}\left[\sum_{i=1}^{n}y_i^2-na^{*2}-2a^*b^*\sum_{i=1}^{n}x_i-b^{*2}\sum_{i=1}^{n}x_i\right] \end{aligned} \tag{133}$$

V dalším kroku upravíme reziduální součet, aby obsahoval neznámé správné hodnoty parametrů a a b. Pak použijeme normální rovnice obdobně jako v předchozím případě.

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\left[\left(y_i-(a+bx_i)\right)-\left((a^*-a)+(b^*-b)x_i\right)\right]^2 = \\ &= \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i-(a+bx_i))^2-\sum_{i=1}^{n}\left((a^*-a)+(b^*-b)x_i\right)^2\right] = \\ &= \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i-(a+bx_i))^2-\sum_{i=1}^{n}\left((a^*-a)^2+2(a^*-a)(b^*-b)x_i+(b^*-b)^2x_i^2\right)\right] = \\ &= \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i-(a+bx_i))^2-n(a^*-a)^2-2(a^*-a)(b^*-b)\sum_{i=1}^{n}x_i-(b^*-b)^2\sum_{i=1}^{n}x_i^2\right] \end{aligned} \tag{134}$$

Takto obdržený výraz (134) pro s^2 budeme středovat přes všechny možné výběry vedoucí na různé kombinace odhadů a^* a b^*. Vzhledem k tomu, že n a součty mocnin x_i jsou pro všechny výběry společné, je možno s nimi zacházet jako s konstantami.

$$\begin{aligned} \left\langle s^2\right\rangle = \frac{n}{n-2}\Bigg[&\underbrace{\left\langle\frac{1}{n}\sum_{i=1}^{n}(y_i-(a+bx_i))^2\right\rangle}_{\sigma^2}-\underbrace{\left\langle(a^*-a)^2\right\rangle}_{\sigma_{a^*}^2}- \\ &-\frac{2}{n}\sum_{i=1}^{n}x_i\underbrace{\left\langle(a^*-a)(b^*-b)\right\rangle}_{\mathrm{Cov}(a^*,b^*)}-\frac{1}{n}\sum_{i=1}^{n}x_i^2\underbrace{\left\langle(b^*-b)^2\right\rangle}_{\sigma_{b^*}^2}\Bigg] \end{aligned} \tag{135}$$

První člen (135) představuje rozptyl σ^2, jehož hodnotu neznáme a snažíme se ji odhadnout. Druhý člen představuje rozptyl $\sigma_{a^*}^2$ odhadu parametru a^* okolo neznámé skutečné hodnoty parametru a. Vztah mezi $\sigma_{a^*}^2$ a σ^2 je určen rovnicí (56). Ve třetím členu vystupuje kovariance odhadů a^* a b^*, která je dána vztahem (52). Ve čtvrtém členu pak vystupuje rozptyl $\sigma_{b^*}^2$ odhadu

parametru b^* okolo neznámé skutečné hodnoty parametru b. Vztah mezi $\sigma_{b^*}^2$ a σ^2 je určen rovnicí (54). Pak můžeme zapsat

$$\left\langle s^2 \right\rangle = \frac{n}{n-2}\left[\sigma^2 - \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sigma^2 - \frac{-\frac{2}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sigma^2 - \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sigma^2 \right] =$$
$$= \sigma^2 \frac{n}{n-2}\left[1 - \frac{2}{n}\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \right] = \sigma^2 \frac{n}{n-2}\left[1 - \frac{2}{n}\right] = \sigma^2 \tag{136}$$

Vidíme, že bez korekce ve jmenovateli (45) bychom dostali vychýlený odhad (biased estimator), který by vedl na podcenění rozptylu zejména při malých výběrech. V případě polynomiálních regresí vyššího stupně bude analogický výpočet obsahovat více členů, ale princip zůstává stejný.

Koeficient korelace

Při fitování bývá užitečné kvantifikovat, jak dokonalou predikci hodnot závisle proměnné nalezená regresní funkce poskytuje. **Pearsonův korelační koeficient** dvou veličin je definován obecně jako poměr jejich vzájemné kovariance (44) a součinu odmocnin kovariancí každé z nich se sebou

$$r(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Cov}(x,x)}\sqrt{\mathrm{Cov}(y,y)}} = \frac{\left\langle \left(x - \langle x\rangle\right)\left(y - \langle y\rangle\right)\right\rangle}{\sqrt{\left\langle \left(x - \langle x\rangle\right)^2\right\rangle}\sqrt{\left\langle \left(y - \langle y\rangle\right)^2\right\rangle}} \tag{137}$$

V dalším odvození použijeme pro určení střední hodnoty populace vážený průměr

$$\langle y\rangle = \frac{\sum_i w_i y_i}{\sum_i w_i} \tag{138}$$

čímž umožníme zahrnutí individuálních statistických vah do výpočtů. V případě výběru bude vážený průměr jen odhadem střední hodnoty — podobně jako aritmetický průměr (63) v běžné metodě nejmenších čtverců. Nevychýlený odhad směrodatné odchylky výběru pak bude analogicky k (67) dán vztahem

$$s = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(y_i - \langle y\rangle\right)^2}{n-1}}, \tag{139}$$

což můžeme podobným postupem jako v případě (129) upravit do tvaru

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} w_i y_i^2 - \frac{\left(\sum_{i=1}^{n} w_i y_i \right)^2}{\sum_{i=1}^{n} w_i} \right)} . \tag{140}$$

Vážený průměr použijeme k vyjádření korelačního koeficientu populace

$$r(x,y) = \frac{\frac{\sum_i w_i \left(x_i - \langle x \rangle \right)\left(y_i - \langle y \rangle \right)}{\sum_i w_i}}{\sqrt{\frac{\sum_i w_i \left(x_i - \langle x \rangle \right)^2}{\sum_i w_i}} \sqrt{\frac{\sum_i w_i \left(y_i - \langle y \rangle \right)^2}{\sum_i w_i}}} = \frac{\sum_i w_i \left(x_i - \langle x \rangle \right)\left(y_i - \langle y \rangle \right)}{\sqrt{\sum_i w_i \left(x_i - \langle x \rangle \right)^2} \sqrt{\sum_i w_i \left(y_i - \langle y \rangle \right)^2}} , \tag{141}$$

což lze dále upravit do tvaru

$$r(x,y) = \frac{\sum_i w_i \sum_i w_i x_i y_i - \sum_i w_i x_i \sum_i w_i y_i}{\sqrt{\sum_i w_i \sum_i w_i x_i^2 - \left(\sum_i w_i x_i \right)^2} \sqrt{\sum_i w_i \sum_i w_i y_i^2 - \left(\sum_i w_i y_i \right)^2}} . \tag{142}$$

Při výpočtu regrese se ovšem nepočítá korelační koeficient celé populace. Namísto toho se jako jeho odhad použije Pearsonův korelační koeficient výběru $r^*(x,y)$ vypočítaný pomocí výše uvedeného vztahu (57). Ukážeme si nyní, jak ho lze odvodit.

Pearsonův korelační koeficient výběru $r^*(x,y)$ samozřejmě můžeme spočítat podle (142). Započtení všech bodů výběru je prakticky snadno proveditelné (na rozdíl od celé populace). Pro přehlednost ve výpočtu $r^*(x,y)$ použijeme statistické váhy dané rozptylem $w_i = \sigma_i^{-2}$ (inverse-variance weighting), což spolu s využitím značení součtů podle (21) vede na kompaktnější zápis

$$r^*(x,y) = \frac{SS_{xy} - S_x S_y}{\sqrt{SS_{xx} - S_x^2} \sqrt{SS_{yy} - S_y^2}} = \frac{SS_{xy} - S_x S_y}{SS_{xx} - S_x^2} \frac{\sqrt{SS_{xx} - S_x^2}}{\sqrt{SS_{yy} - S_y^2}} . \tag{143}$$

Ten lze dále upravit použitím (21), (22) a (140) na tvar (57)

$$r^*(x,y) = b^* \frac{\sqrt{\frac{1}{n-1}\left(S_{xx} - \frac{S_x^2}{S} \right)}}{\sqrt{\frac{1}{n-1}\left(S_{yy} - \frac{S_y^2}{S} \right)}} = b^* \frac{s_x}{s_y} . \tag{144}$$

Tento vztah mezi korelačním koeficientem výběru a odhadem směrnice stojí za povšimnutí. Není těžké ověřit, že nezávisí na typu použitých statistických vah. Odhady směrnice (22) a (32) lze rovněž zobecnit na tvar s obecnou statistickou váhou

$$b^* = \frac{\sum_i w_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i w_i (x_i - \bar{x})^2} \,, \tag{145}$$

kde pruhem označíme střední hodnotu výběru – zde vážený průměr výběru. Pak po dosazení do (144) dostaneme pro korelační koeficient výběru

$$r^*(x,y) = \frac{\sum_i w_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i w_i (x_i - \bar{x})^2} \frac{s_x}{s_y} = \frac{\sum_i w_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\underbrace{\frac{1}{n-1}\sum_i w_i (x_i - \bar{x})^2}_{s_x^2}} \frac{s_x}{s_y} \,, \tag{146}$$

což skutečně vede na vztah, podle něhož se (nevychýlený) Pearsonův korelační koeficient výběru také vypočítá

$$r^*(x,y) = \frac{\sum_i w_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x s_y} \,. \tag{147}$$

Matematické vlastnosti koeficientu korelace

Z definice korelačního koeficientu je zřejmé, že jde o symetrickou vlastnost $r(x,y) = r(y,x)$.

Klíčová matematická vlastnost Pearsonova korelačního koeficientu (populace i výběru) spočívá v jeho invariantnosti při nezávislé lineární transformaci obou proměnných.

Pokud zavedeme takovou transformaci

$$\begin{aligned} x_i' &= A + Bx_i \\ y_i' &= C + Dy_i \end{aligned} \,, \tag{148}$$

kde A, B, C, D jsou libovolné konstanty splňující $B > 0, D > 0$, pak dosazením do (142) snadno ukážeme, že

$$r(x', y') = r(x, y) \,. \tag{149}$$

Jak koreluje velikost opravy s velikostí odhadu

Vysvětleme podrobněji: Ke každé změřené hodnotě závisle proměnné y_i jsme schopni po nalezení regresní závislosti dopočítat odhad $\tilde{y}_i$. Rozdíl $\tilde{y}_i - y_i$ nazvěme velikostí opravy. Zajímá nás, zda existuje korelace mezi touto hodnotou a velikostí odhadu $\tilde{y}_i$. Ukážeme, že pokud regresní závislost byla získána metodou nejmenších čtverců, pak zmiňovaná korelace je nulová. To je ekvivalentní platnosti vztahu

$$\sum_i w_i \left(\tilde{y}_i - y_i\right) \tilde{y}_i = 0 \tag{150}$$

Pro jednoduchost ukážeme platnost na případu lineární regrese, ale postup je možno jednoduše rozšířit na polynomiální regresi libovolného stupně. Regresní přímku hledáme minimalizací reziduálního součtu čtverců (odhad $\tilde{y}_i$ vyjádříme pomocí rovnice regresní přímky)

$$S_{\mathrm{RSS}} = \sum_i w_i \underbrace{(a^* + b^* x_i}_{\tilde{y}_i} - y_i)^2 \tag{151}$$

Podmínky pro výpočet odhadů a^* a b^* se získají pomocí podmínek

$$\frac{\partial S_{\mathrm{RSS}}}{\partial a^*} = 2\sum_i w_i \underbrace{(a^* + b^* x_i}_{\tilde{y}_i} - y_i) = 0 \tag{152}$$

$$\frac{\partial S_{\mathrm{RSS}}}{\partial b^*} = 2\sum_i w_i \underbrace{(a^* + b^* x_i}_{\tilde{y}_i} - y_i) x_i = 0 \tag{153}$$

Do vztahu (150) můžeme dosadit a dostaneme

$$\begin{aligned} &\sum_i w_i \left(\tilde{y}_i - y_i\right)\tilde{y}_i = \sum_i w_i \left(a^* + b^* x_i - y_i\right)\left(a^* + b^* x_i\right) = \\ &= a^* \underbrace{\sum_i w_i \left(a^* + b^* x_i - y_i\right)}_{0} + b^* \underbrace{\sum_i w_i \left(a^* + b^* x_i - y_i\right) x_i}_{0} = 0 \end{aligned} \tag{154}$$

Korelace mezi naměřenými hodnotami a jejich odhady

Když hovoříme o korelačním koeficientu v případě lineární regrese, můžeme ho chápat buď jako $r^*(x, y)$, který charakterizuje linearitu sledované závislosti $y(x)$, anebo jako $r^*(y, \tilde{y})$, který udává korelaci mezi pozorovanými hodnotami y_i a hodnotami $\tilde{y}_i = a^* + b^* x_i$ fitovanými lineární regresí. Jak bylo výše vysvětleno, při lineární transformaci proměnných x a y se korelační koeficient zachovává. Zároveň můžeme snadno ověřit, že pro střední hodnoty platí

$$\overline{\tilde{y}} = \overline{a^* + b^* x} = a^* + b^* \overline{x} = \overline{y} \,. \tag{155}$$

Tyto skutečnosti využijeme a ve vztahu pro $r^*(x, y)$ ve tvaru podle (141) provedeme transformaci $x_i \to a^* + b^* x_i$

$$r^*(x, y) = \frac{\sum_i w_i \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_i w_i \left(x_i - \overline{x}\right)^2}\sqrt{\sum_i w_i \left(y_i - \overline{y}\right)^2}} =$$

$$= \frac{\sum_i w_i \left(\underbrace{a^* + b^* x_i}_{\tilde{y}_i} - \underbrace{(a^* + b^* \bar{x})}_{\bar{\tilde{y}}} \right) (y_i - \bar{y})}{\sqrt{\sum_i w_i \left(\underbrace{a^* + b^* x_i}_{\tilde{y}_i} - \underbrace{(a^* + b^* \bar{x})}_{\bar{\tilde{y}}} \right)^2} \sqrt{\sum_i w_i (y_i - \bar{y})^2}} = \tag{156}$$

$$= \frac{\sum_i w_i (y_i - \bar{y})(\tilde{y}_i - \bar{y})}{\sqrt{\sum_i w_i (y_i - \bar{y})^2} \sqrt{\sum_i w_i (\tilde{y}_i - \bar{y})^2}} = r^*(y, \tilde{y})$$

Ukázali jsme, že korelační koeficienty pro obě pojetí se rovnají.

Pojďme si doplnit některé souvislosti, které vyplynou z druhého možného přístupu. Sumu v čitateli předchozího zlomku rozdělíme na dvě části

$$r^*(y, \tilde{y}) = \frac{\sum_i w_i (y_i - \tilde{y}_i + \tilde{y}_i - \bar{y})(\tilde{y}_i - \bar{y})}{\sqrt{\sum_i w_i (y_i - \bar{y})^2} \sqrt{\sum_i w_i (\tilde{y}_i - \bar{y})^2}} = \frac{\sum_i w_i \left[(y_i - \tilde{y}_i)(\tilde{y}_i - \bar{y}) + (\tilde{y}_i - \bar{y})^2 \right]}{\sqrt{\sum_i w_i (y_i - \bar{y})^2} \sqrt{\sum_i w_i (\tilde{y}_i - \bar{y})^2}}. \tag{157}$$

První z nich je podle (150) a (155) rovna nule

$$\sum_i w_i (y_i - \tilde{y}_i)(\tilde{y}_i - \bar{y}) = \underbrace{\sum_i w_i (y_i - \tilde{y}_i) \tilde{y}_i}_{0} - \bar{y} \sum_i w_i (y_i - \tilde{y}_i) = -\bar{y} \underbrace{(\bar{y} - \bar{\tilde{y}})}_{0} \sum_i w_i = 0 \tag{158}$$

Tak se vztah pro korelační koeficient zjednoduší na tvar

$$r^*(y, \tilde{y}) = \frac{\sum_i w_i (\tilde{y}_i - \bar{y})^2}{\sqrt{\sum_i w_i (y_i - \bar{y})^2} \sqrt{\sum_i w_i (\tilde{y}_i - \bar{y})^2}} = \sqrt{\frac{\sum_i w_i (\tilde{y}_i - \bar{y})^2}{\sum_i w_i (y_i - \bar{y})^2}} \tag{159}$$

Suma ve jmenovateli představuje totální součet čtverců S_{TSS} v zobecněném tvaru, kde jsou na rozdíl od (59) započteny jednotlivé čtverce se statistickou váhou w_i

$$S_{\text{TSS}} = \sum_i w_i (y_i - \bar{y})^2 . \tag{160}$$

Dále navíc ukážeme, že sumu v čitateli lze zapsat jako rozdíl S_{TSS} a reziduálního součtu čtverců S_{RSS}, kde jsou ovšem oproti (27) použity také váhy w_i

$$S_{\text{RSS}} = \sum_i w_i (\tilde{y}_i - y_i)^2 . \tag{161}$$

Vztahy mezi součty čtverců rozdílů různých veličin

Ze vztahu (158) plyne rovnost

$$\sum_i w_i \tilde{y}_i \bar{y} = \sum_i w_i \left(\tilde{y}_i^2 + y_i \bar{y} - y_i \tilde{y}_i \right). \tag{162}$$

Pro sumu v čitateli vztahu (159) pak dostaneme

$$\begin{aligned}
&\sum_i w_i \left(\tilde{y}_i - \bar{y} \right)^2 = \sum_i w_i \left(\tilde{y}_i^2 - 2 \underbrace{\tilde{y}_i \bar{y}}_{\tilde{y}_i^2 + y_i \bar{y} - y_i \tilde{y}_i} + \bar{y}^2 \right) = \sum_i w_i \left(\tilde{y}_i^2 - 2(\tilde{y}_i^2 + y_i \bar{y} - y_i \tilde{y}_i) + \bar{y}^2 \right) = \\
&= \sum_i w_i \left(-\tilde{y}_i^2 + 2 y_i \tilde{y}_i - 2 y_i \bar{y} + \bar{y}^2 \right) = \sum_i w_i \left(-\tilde{y}_i^2 + 2 y_i \tilde{y}_i - y_i^2 + y_i^2 - 2 y_i \bar{y} + \bar{y}^2 \right) = \\
&= \sum_i w_i \left(y_i^2 - 2 y_i \bar{y} + \bar{y}^2 \right) - \sum_i w_i \left(\tilde{y}_i^2 - 2 y_i \tilde{y}_i + y_i^2 \right) = \underbrace{\sum_i w_i \left(y_i - \bar{y} \right)^2}_{S_{\mathrm{TSS}}} - \underbrace{\sum_i w_i \left(\tilde{y}_i - y_i \right)^2}_{S_{\mathrm{RSS}}}
\end{aligned} \tag{163}$$

Můžeme tedy na základě získaného vztahu mezi součty čtverců různých rozdílů psát vztah mezi rozptyly

$$\underbrace{\frac{\sum_i w_i \left(y_i - \bar{y} \right)^2}{\sum_i w_i}}_{\sigma_{\mathrm{tot}}^2} = \underbrace{\frac{\sum_i w_i \left(\tilde{y}_i - \bar{y} \right)^2}{\sum_i w_i}}_{\sigma_{\mathrm{expl}}^2} + \underbrace{\frac{\sum_i w_i \left(\tilde{y}_i - y_i \right)^2}{\sum_i w_i}}_{\sigma_{\mathrm{res}}^2}, \tag{164}$$

kde σ_{tot}^2 je celkový rozptyl naměřených hodnot y kolem střední hodnoty, σ_{expl}^2 pak představuje tu jeho část, kterou lze vysvětlit závislostí na x podle proložené regresní závislosti a σ_{res}^2 je naopak ta část, kterou tak zdůvodnit nelze. Na základě této interpretace se pak suma

$$S_{\mathrm{ESS}} = \sum_i w_i \left(\tilde{y}_i - \bar{y} \right)^2 \tag{165}$$

označuje v anglické literatuře jako **explained sum of squares.**

Tato nově zavedená suma umožní výše odvozené vztahy zapsat v kompaktnějším tvaru. Jednak platí

$$S_{\mathrm{TSS}} = S_{\mathrm{ESS}} + S_{\mathrm{RSS}}. \tag{166}$$

Zároveň můžeme zjednodušit vztah (159) pro korelační koeficient

$$r(y, \tilde{y}) = \sqrt{\frac{S_{\mathrm{ESS}}}{S_{\mathrm{TSS}}}}. \tag{167}$$

Koeficient determinace (58) můžeme na základě (166), (164) a (167) také spočítat ze vztahů

$$R^2 = \frac{S_{\mathrm{ESS}}}{S_{\mathrm{TSS}}} = \frac{\sigma_{\mathrm{expl}}^2}{\sigma_{\mathrm{tot}}^2} = r^2(y, \tilde{y}). \tag{168}$$

Literatura

[1] Brož J. a kol.: Základy fyzikálních měření I, SPN Praha 1967

[2] Rektorys K. a kol.: Přehled užité matematiky, Státní nakladatelství technické literatury Praha 1981

[3] Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P.: Numerical Recipes in Fortran 77: The Art of Scientific Computing, Cambridge University Press 1992

[4] York D.: Unified equations for the slope, intercept, and standard error of the best straight line, American Journal of Physics, vol. 72 (2004), pp. 367-375.

[5] gnuplot 5.4 An Interactive Plotting Program, online manuál:

http://www.gnuplot.info/docs_5.4/Gnuplot_5_4.pdf (citováno 31. 12. 2024)

[6] SciPy API reference > scipy.stats.linregress:

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html (citováno 31. 12. 2024)

[7] NumPy API reference > numpy.polynomial.polynomial.polyfit:

https://numpy.org/doc/stable/reference/generated/numpy.polynomial.polynomial.polyfit.html (citováno 31. 12. 2024)

[8] scikit-learn API reference > LinearRegression:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (citováno 31. 12. 2024)

[9] statsmodels.regression.linear_model.OLS:

https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html (citováno 31. 12. 2024)

[10] statsmodels.regression.linear_model.WLS:

https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.WLS.html (citováno 31. 12. 2024)

[11] JCGM 100:2008 Evaluation of measurement data — Guide to the expression of uncertainty in measurement, vydáno JCGM/WG1 (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP a OIML), Bureau Indernational des Poids et Mesures (BIPM), Sèvres Cedex France, 2008; tyto dokumenty jsou dostupné na adrese https://www.bipm.org/documents/20126/2071204/JCGM-pack.zip (citováno dne 28.1.2025).

[12] Box G.E., Hunter W.G., Hunter J.S.: Statistics for experimenters. Wiley New York 1978

[13] Návod fyzikálního praktika I k úloze IX. Měření modulu pružnosti v tahu - Měření modulu E z protažení drátu: https://physics.mff.cuni.cz/vyuka/zfp/_media/zadani/texty/txt_109.pdf (citováno 16. 1. 2024)

[14] Návod fyzikálního praktika I k úloze XVI. Studium Brownova pohybu: https://physics.mff.cuni.cz/vyuka/zfp/_media/zadani/texty/txt_116.pdf (citováno 18. 1. 2024)

Obsah