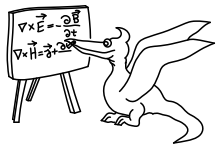

Seriál: Seriál - Zpracování dat 5. díl

V tomto díle seriálu se budeme věnovat nepřímému měření fyzikálních veličin a prokládání naměřených dat teoretickou závislostí, čili, jak by řekl matematik, regresi (v tomto díle zatím jen lineární regresi), případně, jak by řekl fyzik, fitování.

Základní problém, kterým se budeme v tomto díle (vlastně i v tom příštím) seriálu zabývat je takový problém, kdy máme naměřená dvojrozměrná data, tedy dvojice $(x_1, y_1), \ldots, (x_n, y_n)$ a chceme mezi nimi najít nějakou funkční závislost f (jak závisí y na x). Můžeme rozlišovat mezi dvěma hlavními cíli hledání funkční závislosti. Buď chceme pomocí nalezení funkční závislosti nepřímo změřit určitou fyzikální veličinu nebo chceme naměřenými daty v grafu pro názornost proložit teoretickou závislost. V obou těchto případech budeme potřebovat nějaký matematický model pro naše data.

Základní model - Lineární regrese

V našem základním modelu budeme uvažovat prokládanou funkci f pouze ve tvaru

$$f(x) = \beta_0 + \beta_1 f_1(x) + \cdots + \beta_k f_k(x)\,,$$

kde $f_1, \ldots, f_k$ jsou známé zvolené funkce a $\beta_0, \beta_1, \ldots, \beta_k$ jsou neznámé regresní koeficienty, které budeme chtít odhadovat. Na začátku statistického zpracování dat vždy musíme zvolit prokládanou funkci (pomocí volby funkcí $f_1, \ldots, f_k$), o tom, jak ji správně zvolit, bude ještě řeč.

Nyní si musíme popsat model, kterým budeme popisovat naše naměřená data. Budeme si představovat, že naše data byla vygenerována podle následujícího vztahu

$$y_i = \beta_0 + \beta_1 f_1(x_i) + \cdots + \beta_k f_k(x_i) + \varepsilon_i\,,$$

kde $f_1, \ldots, f_k$ jsou známé prokládané funkce, $\beta_0, \beta_1, \ldots, \beta_k$ jsou neznámé regresní koeficienty, které budeme chtít odhadovat, a ε_i představuje náhodné nepřesnosti měření (tedy ε_i je realizace náhodné veličiny a ostatní členy jsou deterministické, i když z části neznámé). V základním modelu budeme uvažovat, že ε_i mají rozdělení $N(0, \sigma^2)$ a že jsou pro různá měření nezávislá. Ve vztahu k teorii vyložené v prvních 4 dílech seriálu si můžeme představovat, že klasicky měříme fyzikální veličinu, jejíž hodnota je ovšem závislá na proměnné x.

Tento základní model se nazývá lineární regresní model[1]. V tomto díle seriálu budeme pracovat pouze s tímto modelem, který se také nejčastěji v praxi použije, nelineární regresní modely si popíšeme v příštím díle seriálu.

V tomto základním modelu budeme uvažovat, že hodnoty proměnné x jsou nám známé přesně, tedy bez nepřesností měření. Toto není vždy úplně oprávněný předpoklad, v praxi bývá někdy porušen, jak postupovat v tomto složitějším případě si ale popíšeme až v příštím díle seriálu.

[1] Na tomto místě musíme vyvrátit jeden rozšířený mýtus, a sice že lineární regrese znamená pouze prokládání přímky naměřenými daty. Je vidět, že prokládání přímky lze dosáhnout speciální volbou parametrů v popsaném lineárním regresním modelu (zvolit $k = 1$ a $f_1(x) = x$), ale je nutné si uvědomit, že název lineární regrese odkazuje k linearitě vzhledem k neznámým koeficientům $\beta_0, \beta_1, \ldots, \beta_k$, nikoliv k linearitě prokládané funkce (funkce f_i mohou být a často jsou nelineární, např. sin, cos, exponenciála, logaritmus, polynomy atd.).

Volba prokládané funkce

Většinou jsme v situaci, kdy máme naměřená nějaká data (tedy dvojice (x_i, y_i)) a potřebujeme jimi proložit nějakou funkci, ale nevíme jakou. Volba správné prokládané funkce je velmi důležitá a proto zde v krátkosti popíšeme, jak by se mělo správně postupovat.

Každá prokládaná funkce by měla mít jasnou fyzikální interpretaci a fyzikální opodstatnění. Měli bychom se vždy zamyslet, jaký teoretický vztah by měla naše naměřená data splňovat a takovou funkci jimi prokládat. Pokud budeme postupovat jinak, je veliké nebezpečí, že zvolíme špatnou prokládanou funkci (o tom, jak poznat, že jsme zvolili špatnou funkci, si povíme později) a budeme muset začít od znova, neboť by všechny naše výsledky byly nesprávné.

Odhadování parametrů

Když jsme si popsali základní model, který budeme používat, můžeme začít odvozovat, jak bude odhad neznámých parametrů $\beta_0, \beta_1, \ldots, \beta_k$ vypadat. K odhadu parametrů použijeme metodu nejmenších čtverců.

Metoda nejmenších čtverců

Metoda nejmenších čtverců určí odhad parametrů $\beta_0, \beta_1, \ldots, \beta_k$ (odhady značíme tak, že přidáme stříšku např. $\widehat{\beta_i}$) tak, že za odhady těchto parametrů vezme taková čísla, aby byl součet čtverců odchylek naměřených hodnot od odhadnuté hodnoty prokládané funkce co možná nejmenší. Matematicky zapsáno, metoda nejmenších čtverců se snaží odhadnout regresní parametry tak, aby výraz

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} f_1(x_i) - \cdots - \widehat{\beta_k} f_k(x_i) \right)^2$$

byl co možná nejmenší. Intuitivní vysvětlení, proč postupujeme právě takto je, že se snažíme, aby proložená funkce procházela co nejblíže naměřeným hodnotám (vzdálenost v tomto případě měříme jako druhou mocninu rozdílu naměřené hodnoty a proložené funkce). Výraz uvnitř závorky se označuje jako reziduum a značí se U_i. Residuum je rozdíl mezi naměřenou hodnotou a proloženou závislostí.

Není úplně zřejmé, proč chceme minimalizovat právě druhou mocninu residuí, mohli bychom minimalizovat třeba absolutní hodnotu residuí nebo absolutní hodnotu ze třetí mocniny nebo zvolit nějakou jinou váhovou funkci. Vysvětlení není úplně triviální, trochu zjednodušené vysvětlení může být takové, že druhá mocnina zvýrazní velmi odlehlá měření, ale zároveň příliš neupozadí neodlehlá měření (tj. chceme odhadnout regresní koeficienty tak, abychom neměli žádné měření příliš odlehlé od proložené závislosti). Toto vysvětlení je trochu nepřesné, existuje i lepší vysvětlení, které zde uvedeme.

Metoda maximální věrohodnosti[2]

Odhady parametrů metodou maximální věrohodnosti na základě naměřených dat fungují na tom principu, že za odhad parametrů vždy vezmeme takové hodnoty parametrů, které maximalizují pravděpodobnost naměření takových dat, které jsme zrovna v našem případě naměřili.

[2] Toto už je opravdu pokročilé téma, pokud chcete (nebo pokud se vám tento odstavec nepodaří pochopit) můžete tento odstavec bez obav přeskočit, na pochopení dalších částí to nebude mít vliv.

Když budeme uvažovat všechny výše popsané předpoklady (tedy navzájem nezávislé nepřesnosti měření ε_i s rozdělením $N(0,\sigma^2)$), můžeme si napsat, jak vypadá hustota pravděpodobnosti v závislosti na neznámých parametrech. Jelikož jsou jednotlivá měření na sobě nezávislá, výsledná sdružená ("vícerozměrná") hustota pravděpodobnosti bude součin jednotlivých hustot pravděpodobnosti. Matematicky zapsáno sdružená hustota pravděpodobnosti L (L z anglického likelihood, česky věrohodnost) vypadá následovně

$$\begin{aligned} &L(x_1,\ldots,x_n,y_1,\ldots,y_n,\sigma^2,\beta_0,\beta_1,\ldots,\beta_k) = \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\frac{(\beta_0+\beta_1 f_1(x_i)+\cdots+\beta_k f_k(x_i)-y_i)^2}{\sigma^2}}, \end{aligned}$$

neboť náhodná veličina Y_i představující výsledek jednoho měření příslušný hodnotě nezávisle proměnné x_i má rozdělení $N(\beta_0+\beta_1 f_1(x_i)+\cdots+\beta_k f_k(x_i),\sigma^2)$. Chceme najít takové hodnoty parametrů $\beta_0,\beta_1,\ldots,\beta_k$, které tuto věrohodnost maximalizují. Abychom lépe viděli, jak tyto parametry zvolit, je výhodné si tento výraz upravit[3], čímž dostaneme

$$\begin{aligned} &L(x_1,\ldots,x_n,y_1,\ldots,y_n\sigma^2,\beta_0,\beta_1,\ldots,\beta_k) = \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \mathrm{e}^{-\frac{1}{2}\frac{(\beta_0+\beta_1 f_1(x_i)+\cdots+\beta_k f_k(x_i)-y_i)^2}{\sigma^2}}, \end{aligned}$$

Nyní se musíme pokusit odvodit, jak zvolit parametry $\beta_0,\beta_1,\ldots,\beta_k$, aby byla hodnota věrohodnosti L co možná největší. Na začátek si můžeme všimnout, že tyto parametry vystupují pouze v exponentu, nikde jinde, tudíž nám stačí zabývat se jen členem s exponentem. Chceme tedy maximalizovat hodnotu exponentu (protože exponenciála je rostoucí funkce, tedy čím větší exponent tím větší hodnota exponenciály), tedy chceme mít co největší hodnotu členu

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\beta_0+\beta_1 f_1(x_i)+\cdots+\beta_k f_k(x_i)-y_i\right)^2.$$

Nyní si musíme všimnout, že člen $-\frac{1}{2\sigma^2}$ je nezávislý na našich koeficientech, stačí nám tedy zabývat se pouze sumou. Nakonec si musíme všimnout, že celý tento člen bude mít vždy záporné znaménko, takže bude maximální právě, když bude suma minimální. Chceme tedy, aby výraz

$$\sum_{i=1}^{n}\left(\beta_0+\beta_1 f_1(x_i)+\cdots+\beta_k f_k(x_i)-y_i\right)^2$$

byl co možná nejmenší. Toto nás ale vede na minimalizaci součtu druhých mocnin residuí, tedy na metodu nejmenších čtverců (pokud vnitřek závorky vynásobíme -1 nic to nezmění díky druhé mocnině).

Ukázali jsme tedy, že odhad metodou maximální věrohodnosti, který má jasnou intuitivní interpretaci, se v tomto případě shoduje s odhadem metodou nejmenších čtverců, čímž jsme získali hlubší pochopení toho, proč odhadovat parametry metodou nejmenších čtverců.

[3] Používáme známého vzorce $\mathrm{e}^a\mathrm{e}^b = \mathrm{e}^{a+b}$.

Výpočetní aspekty odhadu parametrů

Nyní jsme si popsali, jakým způsobem budeme chtít parametry $\beta_0, \beta_1, \ldots, \beta_k$ odhadovat, ale slušelo by se i stručně zmínit postup výpočtu, který k tomu povede. V praxi budeme všechny tyto odhady počítat za použití počítače, nicméně nikdy neuškodí vědět, jak přesně to počítač počítá. Jak později (a v zadání úloh) poznáme, v praxi se tato znalost občas také hodí. Existuje mnoho způsobů, jak se lze dobrat hodnot koeficientů $\beta_0, \beta_1, \ldots, \beta_k$, my zde uvedeme dva nejběžnější. Je zcela evidentní, ale mnohdy se na to zapomíná a je důležité si to uvědomovat, že odhady regresních parametrů závisí na zformulovaném modelu (tedy na volbě prokládané funkce) a na naměřených datech (v určitých chvílích bude výhodné chápat odhad regresních parametrů jako transformaci naměřených dat).

Prvním možným postupem je použití diferenciálního počtu[4], kdy parciálně zderivujeme sumu čtverců podle všech proměnných $\beta_0, \beta_1, \ldots, \beta_k$, tyto parciální derivace položíme rovny nule a snažíme se vyřešit vzniklou soustavu rovnic (obecný postup hledání extrémů funkcí), která má tvar

$$\begin{aligned}
\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}\left(\beta_0+\beta_1 f_1(x_i)+\cdots+\beta_k f_k(x_i)-y_i\right)^2 &= 0\,,\\
&\vdots\\
\frac{\partial}{\partial\beta_k}\sum_{i=1}^{n}\left(\beta_0+\beta_1 f_1(x_i)+\cdots+\beta_k f_k(x_i)-y_i\right)^2 &= 0\,.
\end{aligned}$$

Tato soustava rovnic se v obecném případě neřeší úplně lehce, proto je lepší použít druhou metodu výpočtu odhadů koeficientů.

Druhá metoda, která je asi výpočetně schůdnější, je založena na lineární algebře[5]. Pokud si sestavíme následující matici (někdy se jí říká matice modelu) a vektor naměřených dat

$$\mathbb{X}=\begin{pmatrix} 1 & f_1(x_1) & \ldots & f_k(x_1)\\ \vdots & & \ddots & \vdots\\ 1 & f_1(x_n) & \ldots & f_k(x_n)\end{pmatrix},\quad \mathbf{Y}=\begin{pmatrix} y_1\\ \vdots\\ y_n\end{pmatrix},$$

potom je odhad koeficientů $\widehat{\beta}=(\widehat{\beta_0},\widehat{\beta_1},\ldots,\widehat{\beta_k})^T$ určen následující identitou[6]

$$\widehat{\beta}=(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}\,,$$

kde T značí transpozici matice a $^{-1}$ značí inverzní matici.

Vlastnosti odhadů

Výše jsme si odvodili, jak odhadovat regresní prametry $\beta_0, \beta_1, \ldots, \beta_k$. Použitím metody nejmenších čtverců za pomoci počítače jsme schopni najít odhady těchto parametrů, které budeme dále značit $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}$, nyní si povíme něco o jejich vlastnostech.

[4] Pokud neznáte diferenciální počet nezoufejte, pokud tuto pasáž přeskočíte, nepřijdete o nic podstatného.

[5] Pokud nejste v lineární algebře (transponování, násobení a invertování matic) příliš zběhlí, nevadí, můžete tuto část textu přeskočit, pro další pochopení to nebude vadit.

[6] V extrémním případě by se mohlo stát, že inverzní matice nebude existovat (pokud by matice $\mathbb{X}^T\mathbb{X}$ nebyla regulární, což se ale v praxi nestává), potom bychom tuto metodu použít nemohli.

Dopředu upozorňujeme, že všechna odvození v této kapitole budou pouze naznačená a že není nutné snažit se je detailně pochopit (i když pilnosti se meze nekladou). Pro fyzika je spíše důležité a plně postačující mít obecné povědomí o teoretických odvozeních a současně vědět, jak se tyto metody aplikují na konkrétní řešené problémy.

Intervalové odhady pro jednotlivé parametry

Podobně jako jsme konstruovali intervalové odhady v případě, kdy jsme měřili jen jednu fyzikální veličinu (viz 2. a 3. díl seriálu), můžeme i nyní zkonstruovat intervalové odhady pro jednotlivé regresní koeficienty. Vyjdeme z tvrzení[7], že pro následující vhodnou transformaci naměřených dat platí

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{S^2 v_{j,j}}} \xrightarrow{D} N(0,1)\,,$$

kde $\widehat{\beta}_j$ je odhad regresního koeficientu metodou nejmenších čtverců, β_j je skutečná (pro nás v praxi neznámá) hodnota regresního koeficientu, $v_{j,j}$ je prvek na místě (j,j) v matici $(\mathbb{X}^T\mathbb{X})^{-1}$ a člen S^2 je určen vztahem

$$S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} U_i^2\,,$$

kde U_i jsou residua v našem lineárním regresním modelu. Tento vztah nám velmi připomíná vztah, ze kterého jsme odvozovali intervalové odhady v případě měření jedné fyzikální veličiny (odlišné jsou jen členy ve jmenovateli zlomku), můžeme tedy naprosto stejným způsobem odvodit intervalový odhad i pro skutečnou hodnotu regresních koeficientů. Vyjdeme ze vztahu, že asymptoticky (tj. pro velký počet měření, později bude upřesněno, co to znamená velký počet měření) platí pro libovolnou hladinu spolehlivosti $\alpha \in (0,1)$

$$P\left(u_{\frac{\alpha}{2}} < \frac{\widehat{\beta}_j - \beta_j}{\sqrt{S^2 v_{j,j}}} < u_{1-\frac{\alpha}{2}}\right) \doteq 1-\alpha\,.$$

Z tohoto vztahu algebraickými úpravami (analogicky jako v případě intervalových odhadů pro jednu měřenou fyzikální veličinu) dostaneme, že platí

$$P\left(\beta_j \in \left(\widehat{\beta}_j \pm u_{1-\frac{\alpha}{2}}\sqrt{S^2 v_{j,j}}\right)\right) \doteq 1-\alpha\,.$$

Toto je intervalový odhad pro skutečnou hodnotu regresního koeficientu β_j, tento intervalový odhad se zkráceně zapisuje jako

$$\widehat{\beta}_j \pm \sqrt{S^2 v_{j,j}}$$

a člen $\sqrt{S^2 v_{j,j}}$ se nazývá nejistota měření regresního koeficientu[8] (neplést s chybou měření regresního koeficientu[9]).

Na konec tohoto odstavce jen poznamenejme, že člen S^2 je odhadem rozptylu našich měření σ^2 (tj. pro velký počet měření bude hodnota členu S^2 s největší pravděpodobností velice

[7] Odvození tohoto tvrzení je dosti náročné, proto ho zde nebudeme uvádět.
[8] Anglicky standard error (zkráceně S.E.).
[9] Význam stejný jako u měření jedné fyzikální veličiny (viz 2. díl seriálu).

blízká skutečné hodnotě rozptylu σ^2) a je označován jako střední čtvercová chyba[10]. Dále je dobré poznamenat, že čím více měření provedeme, tím menší bude hodnota členu $v_{j,j}$ a tím pádem budeme mít užší intervalový odhad, tedy budeme mít regresní koeficient určen přesněji (odvození tohoto tvrzení už je ale nad možnosti tohoto seriálu).

Bodový odhad prokládané funkce

Nyní jsme si odvodili, jak vypadají intervalové odhady pro jednotlivé parametry β_j. Na začátku jsme popsali, že někdy je naším cílem také proložit naměřenými daty odhadnutou závislost, čemuž se budeme teď věnovat. Jako bodový odhad funkční závislosti (což lze chápat jako nejpravděpodobnější tvar prokládané funkce) vezmeme jednoduše funkci $\widehat{f}$ ve tvaru

$$\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 f_1(x) + \cdots + \widehat{\beta}_k f_k(x),$$

kde $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ představují odhady regresních koeficientů získané metodou nejmenších čtverců. Tento odhad má opět tu vlastnost, že pro velký počet měření bude tato funkce s největší pravděpodobností velmi podobná skutečné funkční závislosti měřených dat (ve smyslu, že proložená křivka a skutečná teoretická křivka budou téměř identické).

Tento odhad se obvykle zakresluje do grafu jako proložená křivka. Je sice pěkné, že víme, jak bodově odhadovat prokládanou funkci, ale musíme si uvědomit, že to je poměrně málo. Musíme si ještě ukázat, jak vypadá intervalový odhad pro funkční hodnoty v jednotlivých bodech, abychom měli představu s jakou nejistotou jsme prokládanou funkci určili.

Intervalový odhad prokládané funkce

Podobně jako můžeme konstruovat intervalové odhady pro měření jedné fyzikální veličiny a pro regresní koeficienty, můžeme také konstruovat intervalové odhady pro hodnotu prokládané funkce. Při konstrukci intervalového odhadu pro hodnotu prokládané funkce v bodě x vyjdeme z toho, že platí

$$\frac{\widehat{f}(x) - f(x)}{\sqrt{S_n^2 \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}}} \xrightarrow{D} N(0,1)\,,$$

kde $f(x)$ je skutečná hodnota teoretické funkční závislosti v bodě x, vektor $\mathbf{x}$ je definován jako $\mathbf{x} = (1, f_1(x), \ldots, f_k(x))^T$ a ostatní členy jsou stejné jako v předchozím textu. Analogickým způsobem jako v předchozích případech můžeme potom dojít k výsledku

$$P\left(f(x) \in \left(\widehat{f(x)} \pm u_{1-\frac{\alpha}{2}} \sqrt{S^2 \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}}\right)\right) \doteq 1 - \alpha\,.$$

Toto je intervalový odhad pro hodnotu prokládané funkce v bodě x o spolehlivosti $1-\alpha$, který se obvykle zkráceně zapisuje jako

$$\widehat{f}(x) \pm \sqrt{S^2 \mathbf{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}}\,.$$

Tento intervalový odhad můžeme zkonstruovat pro libovolný bod x, tedy i pro takový bod x, který leží mimo všechna naše měření (tomu se potom říká extrapolace[11]). Jak uvidíte při procházení vzorového skriptu, šířka tohoto intervalu spolehlivosti také závisí na tom, jak daleko

[10] Anglicky mean square error (zkráceně MSE).

[11] Musíme si ale uvědomit, že při extrapolování mlčky předpokládáme, že proložená závislost platí i mimo námi naměřená data (tedy, že prokládaná funkce lze „protáhnout“), což nemusí být vždy správný předpoklad

je bod x od námi naměřených hodnot. Je vcelku opodstatněné předpokládat, že v místě, kde máme změřeno hodně hodnot bude interval spolehlivosti užší než v místě, kde máme hodnot změřeno méně. Tento efekt je způsoben vlastnostmi vektorového a maticového součinu ve výrazu $\mathbf{x}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}$, jeho podrobné vysvětlení je ale nad rámec tohoto seriálu.

Nakonec poznamenejme, že vůbec není nutné si tento výraz pamatovat ani ho umět vyčíslit, udělá to za nás matematický software (v našem případě R) po použití jednoduchého příkazu.

Regresní diagnostika

V tomto odstavci se budeme věnovat ověřování, zda jsou splněny všechny předpoklady pro použití lineárního regresního modelu, které jsme uvedli na začátku tohoto dílu seriálu. Pokud bychom aplikovali lineární regresní model a nebyly by splněny předpoklady pro jeho použití, obdržené výsledky by nebyly správné. Je proto vždy důležité ověřit, zda jsou tyto předpoklady splněny. Pro jistotu zde všechny tyto předpoklady ještě zopakujeme, jedná se o (uvedeno od nejdůležitějšího předpokladu po nejméně důležitý)

- Správná volba prokládané funkce.
- Stejný rozptyl pro všechna měření.
- Nezávislost jednotlivých měření.
- Normální rozdělení našich měření.

Na začátek uveďme, že předpoklad o normálním rozdělení našich měření se v praxi velice těžko ověřuje [12] a jelikož byly všechny uvedené tvrzení formulovány tak, že platí i bez tohoto předpokladu (se splněním tohoto předpokladu platí pro libovolný počet měření, bez splnění platí asymptoticky pro velký počet měření) nebudeme se jím dále zabývat. Tento předpoklad se téměř nikdy v praxi neověřuje.

Ostatní předpoklady už jsou ovšem velice důležité a pokud by nebyly splněné a my bychom přesto aplikovali lineární regresní model, naše výsledky by byly nesprávné. Naštěstí existují poměrně spolehlivé způsoby, jak ověřit, zda jsou tyto předpoklady splněny. Všechny tyto postupy jsou založeny na myšlence, že zkusíme aplikovat lineární regresní model a až následně (většinou na základě residuí našeho modelu) zkoumáme, zda byly všechny předpoklady splněny. Pokud zjistíme, že splněny nebyly, nemůžeme takovýto model používat k vyhodnocování našeho experimentu a musíme přijít s lepším modelem (většinou změníme prokládanou funkci) nebo se smíříme s tím, že obdržené výsledky nejsou tolik přesné.

Nyní už k jednotlivým metodám ověřování předpokladů.

Grafická kontrola správnosti prokládané funkce

Tato metoda je velice jednoduchá a nevyžaduje pochopení žádné matematické teorie, její nevýhoda ovšem je, že není tolik přesná. Jejím základem je, že na vykresleném grafu zkontrolujeme, zda proložená funkce odpovídá naměřeným datům. Pokud bychom v grafu našli, že např.

- V některé části grafu je proložená funkce výrazně pod nebo nad naměřenými daty.

[12] Pro správné ověření tohoto předpokladu je potřeba velké množství měření, ale pokud máme velké množství měření, můžeme se už spolehnout na asymptotické vlastnosti všech našich odhadů a ani nepotřebujeme normální rozdělení našich měření.

- Tvar proložené funkce neodpovídá naměřeným datům.

potom bychom museli prohlásit, že jsme prokládali špatnou funkci a pokusit se najít vhodnější funkci.

Zejména v případě, že naměřené hodnoty jsou velice blízko proložené křivce, může být tato metoda značně komplikovaná (nejde jednoznačně odlišit data pod a nad proloženou křivkou). V tomto případě je výhodnější místo na graf naměřených hodnot s proloženou závislostí koukat na graf residuí (graf kde jsou vynesena residua a příslušné hodnoty nezávisle proměnné). Na tomto grafu bychom měli vidět náhodně rozeseté body kolem osy x, pokud tam vidíme cokoliv jiného (např. že v nějaké části grafu jsou residua výrazně nad nebo pod osou x), znamená to, že jsme pravděpodobně zvolili špatně prokládanou funkci a měli bychom se pokusit najít vhodnější funkci.

Obě tyto metody potřebují trochu cviku a příklady, co je ještě akceptovatelné a co už nikoliv. Několik takovýchto příkladů proto najdete v přiloženém vzorovém skriptu.

Statistický test správnosti prokládané funkce (lack of fit test)

Ve speciálním případě můžeme provést také statistický test správnosti prokládané funkce, což je přesnější metoda než výše popsaná grafická metoda. Tím speciálním případem se myslí případ, kdy pro jednu hodnotu nezávisle proměnné x máme naměřeno vždy více měření závisle proměnné (většinou se uvažuje alespoň 5 měření pro každou hodnotu nezávisle proměnné).

Jak jsme si popsali v minulém díle seriálu, k určení statistického testu potřebujeme formulovat hypotézu a alternativu, testovou statistiku a kritický obor testu. Nebudeme zde vše podrobně odvozovat, protože je to nad rámec tohoto dílu seriálu. Tomuto testu se také někdy říká χ^2 test kvality fitu a testuje následující dvojici hypotéza a alternativa

$$H : \text{Prokládaná funkce } f \text{ je správně zvolená.}$$
$$A : \text{Prokládaná funkce } f \text{ není správně zvolená.}$$

Testová statistika má následující tvar

$$CH = \frac{\dfrac{\sum_{i=1}^{n} n_i \left(\overline{y_{i\bullet}} - \widehat{f}(x_i)\right)^2}{n-p}}{\dfrac{\sum_{i=1}^{n}\sum_{j=1}^{n_i} \left(y_{i,j} - \overline{y_{i,\bullet}}\right)^2}{N-n}},$$

kde x_i je hodnota nezávisle proměnné, $y_{i,j}$ je hodnota j-tého měření příslušícího i-té nezávislé proměnné, $\overline{y_{i,\bullet}}$ je průměr naměřených hodnot odpovídající hodnotě nezávisle proměnné x_i, N je celkový počet měření, n je počet různých hodnot nezávislé proměnné a p je počet regresních parametrů. Za platnosti nulové hypotézy má tato statistika Fisherovo F- rozdělení o $(n-p)$ a $(N-n)$ stupních volnosti. Když si zkusíme uvědomit, co testová statistika vyjadřuje, zjistíme, že čitatel je vážený průměr druhých mocnin vzdáleností průměrů naměřených hodnot od proložené křivky a jmenovatel je průměrná druhá mocnina vzdálenosti naměřených hodnot od průměrů naměřených hodnot. Uvědomme si, že za platnosti hypotézy by měla testová statistika nabývat malých hodnot (neboť $\overline{y_{i,\bullet}}$ by měly být velmi podobné $\widehat{f}(x_i)$), proto kritický obor zvolíme na základě Fisherova F-rozdělení následovně

$$C = \left(F_{n-p,N-n}(1-\alpha), \infty\right),$$

kde $F_{n-p,N-n}(1-\alpha)$ je příslušný kvantil Fisherova rozdělení a α je hladina testu.

V praktickém případě budeme tento test provádět pomocí jednoduchého příkazu v matematickém softwaru, není tedy nutné si pamatovat všechny tyto odvozené vzorce. Matematický software jako výstup nabídne také numericky spočítanou p-hodnotu testu.

Poznamenejme, že pokud bychom naměřenými daty prokládali špatnou funkci, bude to mít na správnost našich výsledků velice negativní vliv (prakticky se dá říct, že budou všechny závěry špatné). Je proto naprosto nutné vždy provést alespoň grafickou kontrolu správnosti prokládané funkce.

Grafická kontrola homoscedasticity

Homoscedasticita je označení pro konstantní hodnotu rozptylu. V praxi se občas stane, že rozptyl našich měření není konstantní, ale závisí buď na hodnotě nezávisle proměnné nebo na hodnotě měřené veličiny. Proto bychom měli pokaždé vykreslit grafy, ve kterých budou vynesena

- Residua oproti nezávisle proměnné.
- Residua oproti hodnotě proložené funkce.

Oba tyto grafy by měly vypadat jako náhodně rozeseté body kolem osy x, zejména by se nemělo stávat, že bude v určitých místech větší variability residuí (tj. že budou residua více rozesety do prostoru).

Opět platí, že obě tyto grafické metody vyžadují trochu cviku. Ve vzorovém skriptu naleznete několik příkladů, jak by residua měla a neměla správně vypadat.

Poznamenejme, že porušení předpokladu homoscedasticity nevadí, pokud není velké, což se prakticky nestává. V případě malého porušení předpokladu homoscedasticity se tento fakt zmíní v diskuzi a přidá se varování, že výsledky odvozené z takového modelu mohou být mírně nepřesné. Co dělat, když budeme pracovat s daty, kde rozptyl měření silně závisí na hodnotách nezávislé proměnné nebo měřené veličiny si povíme v příštím díle seriálu.

Jen pro doplnění uvedeme, že existuje i statistický test, který testuje homoscedasticitu residuí. Nicméně podrobné odvození tohoto testu je nad rámec tohoto seriálu, proto zde jen uvedeme, že jeho název je Breusch-Paganův test[13].

Grafická kontrola nezávislosti měření

Jestli jsou naše měření na sobě nezávislá se kontroluje velmi těžko, splnění tohoto předpokladu si musí pohlídat experimentátor při měření. Existuje jedna metoda, jak se dá odhalit, zda jsou naše měření na sobě nezávislá a sice vykreslení grafu residuí oproti posunutým residuím. Zjednodušeně řečeno, pokud jsou naše měření na sobě nezávislá, neměla by hodnota jednoho residua ovlivňovat hodnotu ostatních residuí. Naopak typickým typem závislosti měřených dat je, že jsou na sobě residua sériově závislá (jsou tzv. autokorelovaná), tedy, že hodnota jednoho residua ovlivňuje hodnotu toho následujícího. Toto si lze představovat tak, že pokud jsme v jednom měření dostali hodnotu vyšší než skutečnou, potom v dalším měření pravděpodobně také dostaneme hodnotu vyšší než skutečnou a naopak. Toto nám pomůže odhalit graf residuí oproti posunutým residuím, což je jen graf kde jsou vykresleny body

$$(U_1, U_2), (U_2, U_3), \ldots, (U_{n-1}, U_n).$$

[13] Více informací o tomto testu například zde: `https://en.wikipedia.org/wiki/Breusch-Pagan_test`

V případě, že jsou naše měření na sobě nezávislá, měl by takovýto graf vypadat jako náhodně rozeseté body kolem počátku soustavy souřadné. Pokud jsou na sobě residua závislá, budou typicky body koncentrovány hlavně v 1. a 3. kvadrantu.

Porušení předpokladu nezávislosti měřených dat je poměrně závažné a neexistuje jednoduchý způsob, jak takovýto problém spravit. Je potřeba na toto myslet už při měření experimentálních dat a dát si na pozor. Pokud při zpracování dat zjistím, že jsou naměřená data na sobě silně závislá, je potřeba buď data změřit znova nebo v diskuzi uvést, že obdržené výsledky mohou být kvůli závislým datům značně nepřesné.

Opět platí, že tato grafická metoda vyžaduje trochu cviku, proto je ve vzorovém skriptu uvedeno několik příkladů použití této metody.

Jen pro doplnění uvedeme, že také existuje statistický test, který testuje nezávislost měřených dat. Podrobné odvození tohoto testu je nad rámec tohoto seriálu a proto jen uvedeme, že jeho název je Durbin- Watsonův test[14].

Koeficient determinace

Pokud chceme zhodnotit, jak dobře proložená funkce vysvětluje opravdovou závislost měřené hodnoty y na hodnotě vysvětlující proměnné x, zadefinujeme si k tomu tzv. koeficient determinace, který budeme značit R^2, pomocí vztahu

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y_n})^2 - \sum_{i=1}^{n} \left(y_i - \widehat{f}(x_i)\right)^2}{\sum_{i=1}^{n} (y_i - \overline{y_n})^2},$$

kde $\overline{y_n}$ představuje výběrový průměr všech naměřených hodnot (bez prokládání nějaké funkce, nezávisle na hodnotách vysvětlující proměnné x).

Koeficient determinace[15] vyjadřuje, jak velký podíl celkové variability (součtu druhých mocnin vzdáleností měření od výběrového průměru) naměřených dat se nám povedlo vysvětlit tím, že jsme daty proložili funkci f. Je to vlastně podíl variability okolo proložené funkce ku variabilitě okolo výběrového průměru. Koeficient determinace slouží jako pomocný nástroj pro hodnocení toho, jak dobrý máme pro naše naměřená data model. Čím je R^2 vyšší, tím více je chování naměřených dat y vysvětleno vysvětlujícími proměnnými x (poznamenejme, že R^2 vždy nabývá hodnot z intervalu $(0, 1)$).

Je nutné si uvědomit, čím je způsobena variabilita naměřených dat kolem proložené funkce. Tato variabilita může být způsobena zaprvé nepřesností měření nebo špatně zvolenou prokládanou funkcí (prokládáme naměřenými daty jinou funkci, než jaká je opravdová závislost naměřených dat). Je nutné si uvědomit, že tyto dva zdroje od sebe nedokážeme použitím pouze R^2 odlišit, proto není dobré používat R^2 jako jediný ukazatel toho, jak je náš model dobrý.

I naprosto správný model může mít malé R^2 (pokud máme velkou nepřesnost měření) a na druhou stranu i špatný model (špatně zvolená prokládaná funkce) může mít vysoký R^2. Proto je nutné R^2 vždy používat současně s metodami regresní diagnostiky, zejména těmi metodami, které zkoumají správnost proložené funkce.

[14] Více informací o tomto testu například zde: `https://en.wikipedia.org/wiki/Durbin-Watson_statistic`

[15] Někdy se zavádí ještě tvz. upravený koeficient determinace vztahem $R^2_{\text{adj}} = R^2 - (1 - R^2)\frac{n-1}{n-k}$, kde k vyjadřuje počet neznámých regresních parametrů. Význam R^2_{adj} je stejný jako význam R^2, jediný rozdíl je, že se snaží zohlednit počet neznámých regresních parametrů.

Několik poznámek na závěr

Nakonec uvedeme opět několik důležitých poznámek.

- Několikrát jsme se odvolávali na nutnost použít dostatečný počet měření a nikde jsme přesně nespecifikovali, kolik měření už je dostatečně. Nyní to napravíme. Obecně se dá říci, že pokud máme alespoň 10 krát více měření než odhadovaných regresních parametrů, všechny popsané metody už budou spolehlivě fungovat. Pokud budeme mít alespoň 5 krát více měření než regresních parametrů, popsané metody budou pořád poměrně spolehlivé, ale méně měření už bychom mít neměli. Pokud jsme v situaci, že je velmi těžké nebo nemožné získat dostatek měření, můžeme použít metody lineární regrese i pro menší počet měření, ale musíme si být vědomi, že použité metody nemusí být úplně spolehlivé (je dobré to na závěr zmínit v diskuzi). Vždy je ale naprosto nutné mít alespoň o 1 měření více, než kolik máme regresních parametrů, jinak dojde k tomu, že prokládaná funkce projde přesně naměřenými body a nebude nám nic říkat o obecné závislosti[16].

- Rozhodně není nutné znát zpaměti umět ani dopodrobna rozumět konstrukci bodových ani intervalových odhadů, které jsme výše uvedli. Důležitá věc je znát a chápat rozdíl mezi bodovým a intervalovým a mít detailně rozmyšlené, co přesně vyjadřují intervalové odhady. V praxi za nás bude všechny výpočty provádět matematický software. Jediné důležité je umět tyto výsledky správně interpretovat.

- Opravdu silně doporučujeme projít si (třeba i několikrát) přiložený vzorový skript a rozmyslet si všechny uvedené příklady. Na praktických příkladech se toho člověk nejvíce naučí.

- Když v praxi používáme lineární regresi ke zpracování naměřených dat, je nutné do protokolu uvést alespoň tolik informací, aby čtenář zjistil, co přesně jsme dělali a mohl tento postup sám reprodukovat (když chce například hledat chyby v použitém postupu a výpočtech). Jako minimální výčet věcí, které by měly být vždy uvedeny, můžeme považovat

  - Tvar prokládané funkce (tedy vzorec $f(x) = \dots$)
  - Bodový odhad a nejistota měření všech regresních koeficientů.
  - Alespoň stručný komentář, zda jsou splněny všechny předpoklady použití regresního modelu (případně upozornění na možné nepřesnosti způsobené nesplněním předpokladů). Není nutné přikládat všechny popsané grafy.
  - Pokud se rozhodnete vykreslit graf s naměřenými daty a proloženou funkcí, měl by být v legendě uveden tvar prokládané funkce a odhadnuté hodnoty regresních koeficientů včetně nejistoty měření. Je také dobré (i když ne úplně nutné) do grafu vykreslit interval spolehlivosti pro prokládanou funkci. Ve vzorovém skriptu najdete podrobný návod a ukázku, jak by toto mělo správně vypadat.

- Ačkoliv se v praxi lineární regrese používá nejčastěji, je nutné si uvědomit, že není aplikovatelná na všechny případy. Existují i případy, kdy je potřeba daty proložit funkci nelineární v regresních parametrech. V těchto případech je nutné využít nelineární regresi, které se budeme věnovat v příštím díle seriálu.

[16] Takovému problému se říká přefitování a vzniká vždy, když pracujeme s málo měřeními v porovnání s počtem regresních koeficientů.

Fyzikální korespondenční seminář je organizován studenty MFF UK. Je zastřešen Oddělením pro vnější vztahy a propagaci MFF UK a podporován Ústavem teoretické fyziky MFF UK, jeho zaměstnanci a Jednotou českých matematiků a fyziků.