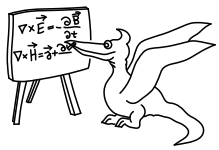

Seriál: Seriál - Zpracování dat 6. díl

V tomto posledním díle seriálu se budeme věnovat pokročilejším partiím regresní analýzy. Nejprve si popíšeme, jak vyřešit případy, kdy není zřejmé, jakou máme zvolit prokládanou funkci, což vyřešíme lokální aproximací polynomy. Dále si popíšeme případy, ve kterých potřebujeme naměřenými daty proložit funkční závislost, která není lineární v neznámých regresních parametrech. Poznáme, že z hlediska zpracování dat a práce s matematickým softwarem je nelineární regrese prakticky stejná jako regrese lineární (liší se jen v matematických modelech v pozadí, jejichž znalost ale není pro praxi nutná). Dále se naučíme zahrnovat do naší analýzy i nejistoty měření našich hodnot (zejména nejistoty typu A způsobované použitým postupem měření). Nakonec stručně zmíníme, jak postupovat při testech hypotéz o hodnotách regresních koeficientů.

Nelineární regrese

Až do teď jsme se zabývali pouze případem, kdy prokládaná funkce byla lineární v neznámých regresních koeficientech, prokládaná funkce tedy měla vždy podobu

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

Toto byla tzv. lineární regrese. Nyní se budeme zabývat případem, kdy potřebujeme naměřenými daty proložit funkci, která není lineární v neznámých regresních koeficientech. Budeme tedy prokládat funkci tvaru

$$f(x, \beta_0, \beta_1, \ldots, \beta_k)\,,$$

kde $\beta_0, \beta_1, \ldots, \beta_k$ jsou neznámé regresní koeficienty, v nichž není tato funkce lineární. Takovémuto typu regrese se říká nelineární regrese. Naším cílem nadále bude stejně jako v případě regrese lineární odhadnout z naměřených dat hodnoty neznámých regresních koeficientů a vykreslit do grafu proloženou funkci.

Nelineární regrese se od lineární regrese liší jen v několika detailech. Na odhadování neznámých parametrů v nelineární regresi použijeme stejně jako v případě regrese lineární metodu maximální věrohodnosti, která za předpokladu normálního rozdělení chyb měření bude ekvivalentní metodě nejmenších čtverců. Výpočetní aspekt odhadu parametrů v nelineární regresi je značně komplikovanější a musí proto používat numerické algoritmy (na rozdíl od lineární regrese, kde existoval explicitní vzorec). Vlastnosti odhadů budou v případě nelineární regrese velice podobné regresi lineární. Všechny tyto odlišnosti si nyní popíšeme podrobněji.

Metoda maximální věrohodnosti

O principu fungování metody maximální věrohodnosti už jsme psali v minulém dílu seriálu, ale pro jistotu ji zde ještě stručně zopakujeme. Při metodě maximální věrohodnosti předpokládáme, že naměřená data (tedy dvojice (x_i, y_i)) byla vygenerována pomocí následujícího schématu. Pro každou hodnotu nezávisle proměnné x_i jsme naměřili určitou hodnotu y_i podle vztahu

$$y_i = f(x_i, \beta_0, \beta_1, \ldots, \beta_k) + \varepsilon_i\,,$$

kde ε_i představuje náhodnou nepřesnost měření, o které předpokládáme, že má rozdělení $N(0,\sigma^2)$, pro nějakou neznámou hodnotu σ^2. Věrohodnost pro určitou volbu hodnot regresních koeficientů $\beta_0,\ldots,\beta_k$ je definována jako pravděpodobnost, že při takovéto volbě regresních koeficientů naměříme právě taková data, která jsme naměřili. Metoda maximální věrohodnosti říká, že za odhady regresních koeficientů vezmeme takové hodnoty $\widehat{\beta_0},\ldots,\widehat{\beta_k}$, které maximalizují věrohodnost pro naše naměřená data. Věrohodnost pro naše naměřená data je tvaru

$$L(\beta_0,\ldots,\beta_k,x_1,\ldots,x_n,y_1,\ldots,y_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{1}{2}\frac{(y_i-f(x_i,\beta_0,\ldots,\beta_k))^2}{\sigma^2}}\,.$$

Tento výraz se dá pomocí algebraických úprav přepsat na výraz

$$\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-f(x_i,\beta_0,\ldots,\beta_k))^2}\,.$$

Podobně jako v minulém díle seriálu můžeme odvodit, že tento výraz je maximalizován, právě když je minimalizována hodnota výrazu

$$\sum_{i=1}^{n}\left(y_i-f(x_i,\beta_0,\ldots,\beta_k)\right)^2\,.$$

Chceme tedy za odhady regresních koeficientů zvolit takové hodnoty $\widehat{\beta_0},\ldots,\widehat{\beta_k}$, aby byl minimalizován součet čtverců vzdáleností naměřených hodnot od proložené funkce. Používáme tedy metodu nejmenších čtverců.

Výpočetní aspekty odhadů parametrů

Jak už jsme naznačili, počítání odhadů regresních koeficientů v nelineární regresním modelu je značně komplikované. V praxi se pro tento výpočet musejí používat numerické metody. Nejčastěji používanou metodou je tzv. gradient descent metoda, kterou si zde stručně popíšeme.

Tato metoda se snaží najít extrém funkce tím způsobem, že začne v nějakém námi zadaném bodě (později bude upřesněno, jak zvolit počáteční bod), následně určí směr, ve kterém minimalizovaná funkce co nejvíce klesá[1] a posune se v tomto směru o předem zvolený krok. Tento postup se pořád opakuje, čímž se postupně posouváme k hledanému minimu funkce. Pokud se dostaneme k hledanému minimu funkce, začne se tato metoda většinou cyklit na místě (už se nikam systematicky neposouvá). Když se toto stane, je to signál, že můžeme skončit a prohlásit poslední bod za aproximaci hledaného minima.

Tato metoda požaduje, aby uživatel na začátku zadal hodnotu počátečního bodu. Ve většině případů na volbě počátečního bodu nezáleží (tj. můžeme zadat libovolný počáteční bod), ale dobrá volba počátečního bodu dokáže zejména u složitějších úloh značně urychlit výpočet. Jako počáteční bod by se měl volit bod, o kterém si myslíme, že je blízko hledanému řešení (aby nebyla cesta k hledanému minimu příliš dlouhá). Jak už bylo řečeno dříve, špatná volba počátečního bodu obvykle nemá fatální důsledky (kromě delší doby výpočtu), ale může se stát, že kvůli špatné volbě počátečního bodu metoda nebude vracet požadované výsledky.

[1] Toto se provede pomocí gradientu funkce (odtud název gradient descent), což je něco jako zobecněná derivace pro funkce více proměnných.

Toto je nejčastěji způsobeno tím, že metoda nenajde globální minimum, ale pouze nějaké lokální minimum. Je proto dobré věnovat pozornost vhodné volbě startovacího bodu a potom také zkontrolovat, zda výstup programu odpovídá intuitivní představě o správném výsledku. V opačném případě bychom měli zkusit jinou volbu počátečního bodu. V přiloženém vzorovém skriptu je opět uvedena ukázka toho, jak správně volit hodnoty počátečních bodů, abychom se vyhnuli nesprávným závěrům.

Pokud budeme chtít v programu R odhadovat regresní koeficienty v nelineární regresi, po zavolání příslušného příkazu na pozadí proběhne podobný algoritmus, jaký byl popsán výše. Z tohoto důvodu je tedy potřeba společně s ostatními vstupními parametry modelu zadat i počáteční bod pro metodu gradient descent. Nyní máme alespoň základní představu, proč se po nás něco takového chce.

Vlastnosti odhadů

V případě lineární regrese jsme odvodili poměrně dost vlastností, které mají odhady regresních koeficientů, a naučili jsme se konstruovat intervalové odhady pro hodnoty regresních koeficientů i prokládané funkce (viz minulý díl seriálu). Nyní bychom chtěli něco podobného odvodit i pro případ nelineární regrese.

Dobrá zpráva je, že v případě nelineární regrese mají naše odhady úplně analogické vlastnosti jako odhady v případě lineární regrese a dají se konstruovat také intervalové odhady pro regresní koeficienty i pro prokládanou funkci. Špatná zpráva je, že matematická teorie, která všechno toto dovoluje, je v případě nelineární regrese o dost složitější. Na tomto místě tedy tuto teorii nebudeme uvádět a jen napíšeme, že se dají zkonstruovat asymptotické intervalové odhady, které mají pro dostatečně velký počet měření (bude upřesněno dále) vlastnosti

$$P\left(\beta_i \in \left(\widehat{\beta_i} \pm u_{1-\frac{\alpha}{2}}\, s_n^{K_i}\right)\right) \doteq 1-\alpha\,,$$
$$P\left(f(x,\beta_0,\ldots,\beta_k) \in \left(f(x,\widehat{\beta_0},\ldots,\widehat{\beta_k}) \pm u_{1-\frac{\alpha}{2}}\, s_n^f(x)\right)\right) \doteq 1-\alpha\,,$$

kde $s_n^{K_i}$ je nejistota měření regresního koeficientu a $s_n^f(x)$ je nejistota měření funkční hodnoty prokládané funkce v bodě x. Obě tyto hodnoty poskytuje matematický software jako standardní výstup. Jak už bylo zmíněno dříve, na výpočet těchto nejistot sice existují explicitní vzorce, my je zde ale nebudeme uvádět.

Regresní diagnostika

I v případě nelineární regrese je potřeba, aby byly splněny všechny předpoklady modelu, jinak obdržené výsledky nebudou správné. Předpoklady jsou v podstatě stejné jako v případě lineární regrese, pro jistotu je zde ale zopakujeme. Předpoklady nelineární regrese jsou

- Správná volba prokládané funkce.
- Stejný rozptyl chyb měření.
- Nezávislost jednotlivých měření.
- Normalita chyb měření.

Stejně jako v lineární regresi platí, že čtvrtý předpoklad není nijak zásadní, pokud pracujeme s dostatečným počtem měření. Splnění ostatních předpokladů je ale poměrně zásadní. Je proto nutné pokaždé provést alespoň základní regresní diagnostiku, abychom ověřili, že byly všechny předpoklady splněny (nebo že alespoň nebyly porušeny zásadním způsobem).

V zásadě platí, že na ověření platnosti předpokladů můžeme použít stejné metody, které používáme na ověření předpokladů v lineární regresi. Nemělo by smysl všechny tyto metody zde znovu opakovat, proto jen odkážeme na kapitolu Regresní diagnostika v minulém dílu seriálu.

Statistické testy o hodnotách regresních koeficientů

Metody popsané v tomto odstavci jsou aplikovatelné na lineární i nelineární regresi. Poslední věc, kterou bychom chtěli v souvislosti s regresní analýzou pokrýt, je testování hypotéz o hodnotách regresních parametrů. Je to tedy situace, kdy chceme pomocí naměřených dat otestovat hypotézu proti alternativě tvaru

$$\begin{aligned} H &: \beta_j = \vartheta\,, \\ A &: \beta_j \neq \vartheta\,, \end{aligned}$$

kde ϑ je nějaká předem zvolená konstanta.

Jak bylo popsáno ve čtvrtém dílu seriálu, k odvození testu potřebujeme odvodit podobu testové statistiky a kritického oboru. Jako testovou statistiku v tomto případě zvolíme

$$T = \frac{\widehat{\beta}_j - \vartheta}{s_n^{K_j}}\,,$$

kde $s_n^{K_j}$ je nejistota měření regresního koeficientu. Z předchozího dílu seriálu víme, že v případě lineární regrese je tato nejistota rovna prvku na pozici (j,j) v matici $\widehat{\sigma}^2(\mathbb{X}^T\mathbb{X})^{-1}$, kde $\mathbb{X}$ je matice modelu. V případě nelineární regrese jsme si explicitní vzorec neuvedli a spoléháme se v tomto na matematický software. Takto zvolená testová statistika bude za platnosti hypotézy konvergovat v distribuci k rozdělení $N(0,1)$. Kritický obor tedy zvolíme jako

$$C = (-\infty, u_{\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}, \infty)\,.$$

Zdůvodnění je stejné jako ve čtvrtém dílu seriálu u klasického t-testu.

Jednoduchou modifikací by se dal takovýto test upravit, aby testoval mírně pozměněnou hypotézu proti alternativě

$$\begin{aligned} H &: \beta_j < \vartheta\,, \\ A &: \beta_j \geq \vartheta\,. \end{aligned}$$

Testová statistika by se zvolila stejně, jen volba kritického oboru by se lišila. Kritický obor testu by v tomto případě byl

$$C = (u_{1-\alpha}, \infty)\,.$$

Zdůvodnění je opět zcela analogické jako v případě t-testu. V případě opačných znamének nerovností v definici testované hypotézy a alternativy bychom dostali stejným způsobem kritický obor tvaru

$$C = (-\infty, u_\alpha)\,.$$

Několik poznámek k nelineární regresi

- Stejně jako v případě lineární regrese platí, že v každém případě musíme mít alespoň takový počet měření, jako máme neznámých regresních koeficientů (jinak nemůžeme nelineární regresi ani použít). Dále platí, že aby byly intervalové odhady pro regresní koeficienty a hodnoty prokládané funkce, které jsme popsali v předchozích kapitolách, přesné, potřebujeme mít alespoň 4-5 krát více měření než regresních koeficientů.

- Stejně jako v případě lineární regrese platí, že čím více naměřených hodnot použijeme pro odhad regresních koeficientů, tím přesnější výsledky dostaneme (dostaneme menší nejistotu měření regresních parametrů).

- Stejně jako v případě lineární regrese platí, že bychom měli naměřenými daty prokládat jen takové funkce, které mají určité fyzikální opodstatnění. Z opačném případě se vystavujeme velkému riziku, že zvolíme prokládanou funkci špatně, což by mělo na závěry naší analýzy fatální důsledky.

- Další věc, na kterou si musíme dát v případě nelineární regrese pozor, je identifikovatelnost regresních parametrů v našem modelu (v lineární regresi jsme toto řešit nemuseli). Identifikovatelnost regresních parametrů znamená, že se nesmí stát, že bychom pro různé hodnoty regresních koeficientů dostali stejnou prokládanou funkci.

  Pro lepší pochopení uvedeme příklad. Pokud bychom zvolili prokládanou funkci ve tvaru

  $$f(x) = a + \frac{b}{cx + d}\,,$$

  stalo by se, že by regresní parametry nebyly identifikovatelné. Například pro volby parametrů

  $$\begin{aligned}(a, b, c, d) &= (1, 1, 1, 1)\,,\\ (a', b', c', d') &= (1, 2, 2, 2)\end{aligned}$$

  bychom dostali naprosto stejnou prokládanou funkci (po krácení ve zlomku). Tento problém se dá vždy odstranit vhodnější parametrizací (s méně regresními koeficienty). V tomto případě bychom si museli prokládanou funkci přepsat do tvaru

  $$f(x) = a + \frac{1}{\frac{c}{b}x + \frac{d}{b}} = \beta_0 + \frac{1}{\beta_1 x + \beta_2}\,,$$

  kde máme jen 3 regresní koeficienty $\beta_0, \beta_1, \beta_2$.

- Pokud používáme pro zpracování měření určitého experimentu nelineární regresi, měli bychom do závěrečného protokolu uvést dostatečně mnoho informací, aby čtenář byl schopen přesně zrekonstruovat náš postup, poskytnout i grafické výstupy a dbát přitom na přehlednost. Stejně jako v případě aplikace lineární regrese platí, že bychom jako výstup měli uvést minimálně

  – 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. 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.

- Pozorného čtenáře jistě napadne, že nelineární regrese je univerzálnější než lineární regrese a že bychom mohli používat pouze nelineární regresi (tedy aplikovat ji i na problémy řešitelné lineární regresí). Toto je do určité míry pravda, ale stále existuje mnoho důvodů, proč je dobré používat lineární regresi v případech, kdy to jde. Mezi hlavní patří, že lineární bude poskytovat přesnější výsledky (díky jednodušeji formulovaným modelům a jednodušší matematické teorii). Jako druhý hlavní důvod můžeme uvést snadnější výpočet koeficientů. V problémech, které v tomto seriálu řešíme, a při uvážení výpočetní síly dnešních počítačů se toto nezdá být jako nevýhoda nelineární regrese, ale ocenili bychom to v případě zpracovávání opravdu velkých objemů měřených dat.

Pokročilé partie regresní analýzy

V této poslední kapitole se budeme věnovat několika problémům pokročilé regresní analýzy. Uvádíme je zde spíše pro zajímavost, neboť na ně v praxi moc často nenarazíme a dá se říci, že je v určitém přiblížení můžeme řešit už známými metodami nebo tyto problémy ignorovat a nedopustili bychom se velkých nepřesností (alespoň tedy ve velké většině případů).

Omezení na regresní koeficienty

Někdy v praxi potkáme případ, kdy předem víme, že regresní koeficienty musí nabývat jen určitých hodnot. Například mohou nabývat jen kladných hodnot nebo jen "malých hodnot" (řekněme menší než 100). V takovémto případě je možno numerickému algoritmu tuto informaci poskytnout a ten bude hledat regresní koeficienty tak, aby tyto naše podmínky splňovaly. V přiloženém vzorovém skriptu najdete příklad, kdy je toto potřeba udělat, a také, jak to správně udělat.

Toto se nejčastěji stane, když prokládáme naměřenými daty periodické funkce (jako např. funkce $\sin(x)$, $\cos(x)$ atd.). Představme si, že chceme naměřenými daty proložit funkci

$$f(x) = a + b\sin(cx + d)\,.$$

Pokud bychom si nestanovili žádnou dodatečnou podmínkou na hodnoty regresních koeficientů, velmi pravděpodobně by nám software na výstupu poskytl velice vysokou hodnotu odhadu regresního koeficientu c. Pokud bychom si následně nakreslili graf proložené funkce, dostali bychom sinusoidu s velikou frekvencí, která by sice skoro dokonale prošla naměřenými daty (tedy součet čtverců residuí by byl velice malý), ale velice špatně by aproximovala pozorovanou závislost. Řešením by v tomto případě bylo omezit možné hodnoty regresního koeficientu c. Příklad správného použití omezujících podmínek najdete ve vzorovém skriptu.

Problematická volba prokládané funkce - spliny

V praxi se může někdy stát, že závislost, kterou chceme naměřenými daty proložit, je příliš komplikovaná na to, abychom ji mohli vyjádřit pomocí jedné prokládané funkce s regresními koeficienty. Pokud bychom chtěli i v této situaci vykreslit graf, ve kterém bude kromě naměřených dat i proložená teoretická křivka, můžeme k tomu použít lokální aproximaci polynomy.

Lokální aproximace funkce polynomy zjednodušeně znamená, že definiční obor funkce rozdělíme na menší intervaly a na každém takovémto intervalu budeme chtít najít polynom, který bude co nejvíce podobný aproximované funkci (podobný ve smyslu, že vzdálenost mezi aproximovanou funkcí a polynomem bude co nejmenší). Jistě někteří z vás už slyšeli o Taylorových polynomech. Taylorovy polynomy jsou jedním ze způsobů, jak lokálně aproximovat funkci pomocí polynomů[2].

Idea, jak budeme chtít postupovat, je taková, že si naměřená data rozdělíme do několika intervalů podle hodnot nezávisle proměnné (dělící body budeme v tomto kontextu nazývat uzly) a na každém z těchto intervalů budeme chtít naměřenými daty prokládat polynomiální funkci určitého stupně p. Jednotlivým polynomům se v tomto kontextu říká regresní spliny[3]. V praxi stačí zvolit p rovno dvěma nebo třem, větší stupně prokládaných polynomů už obvykle nevedou k výrazně lepším výsledkům.

K proložení polynomů na jednotlivých intervalech můžeme použít lineární regresi, která byla popsána v minulém dílu seriálu. Regresní koeficienty budeme tedy odhadovat pomocí metody nejmenších čtverců. Pokud bychom vše provedli jen tak, jak jsme dosud popsali, velice pravděpodobně bychom dostali nespojitou proloženou funkci, což není úplně to, co by odpovídalo intuitivním požadavkům na prokládanou funkci. Proto si navíc přidáme podmínku na to, aby bylo napojení proložených polynomů na krajích našich intervalů spojité. Takovouto podmínku přidáme jednoduše tím, že při počítání odhadů regresních koeficientů budeme uvažovat jen taková řešení, která zároveň splňují podmínku na spojitost na krajích intervalů. Pokud si nejste jistí, jak by se takováto soustava řešila, nevadí, v praxi toto vždy bude obstarávat matematický software. Tento odstavec slouží jen na získání základní představy, co se v počítači po spuštění příkazu děje.

Poslední problém, který musíme vyřešit, je volba dělení na intervaly, tedy jaké množství a jaké rozmístění uzlů zvolit. Obecně se dá říci, že je dobré zvolit uzly tak, aby na každém vzniklém intervalu stále byl dostatečný počet měření (ideálně alespoň 4 až 5 krát více, než je stupeň prokládaného polynomu). Uzly bychom měli potom volit tak, aby hranice intervalů odpovídaly bodům, ve kterých se mění chování měřených dat. Obvykle se v praxi postupuje metodou pokus–omyl, tedy zkusíme podle výše popsaných pravidel uzly nějak zvolit, vykreslíme si proloženou funkci a pokud nejsme spokojeni, pokoušíme se jinou volbou uzlů výsledek zlepšit. Zejména bychom volbou uzlů měli zajistit, aby proložená funkce nevykazovala příliš rozkolísané a neočekávané chování. Tato metoda vyžaduje cvik a zkušenosti, proto ve vzorovém skriptu najdete příklady toho, jak by správná volba uzlů měla a neměla vypadat.

Pomocí tohoto postupu dostaneme pěkně vypadající graf, ale je nutné si uvědomit několik omezení, která tento postup má:

- Je potřeba poměrně hodně měření k tomu, aby byla tato metoda spolehlivá (což je logické vzhledem k tomu, že naměřená data rozdělíme na několik skupin a na nich zvlášť prokládáme jednotlivé funkce).

[2] Pokud jste o Taylorových polynomech ještě neslyšeli, nevadí, uvádíme to zde jen pro zajímavost.
[3] Čti [splajny].

- Obdržené odhady regresních koeficientů na rozdíl od klasické lineární regrese nemají naprosto žádnou fyzikální interpretaci. Jediné, čeho tímto postupem docílíme, je názorně proložená křivka v grafu naměřených hodnot.

- Vyvstává problém, jaké dělení na jednotlivé intervaly bychom měli správně zvolit, neboť obdržený výsledek závisí na zvoleném dělení.

Aplikace regrese v případě heteroskedasticity

V tomto dílu a v minulém dílu seriálu jsme opakovaně uváděli, že jedním z předpokladů lineárních i nelineárních regresních modelů je shodnost rozptylu měřených dat (homoskedasticita). Uvedli jsme, že pokud není tento předpoklad porušen výrazně, můžeme toto porušení ignorovat (případně slovně upozornit na možné nepřesnosti v obdržených výsledcích). Nyní představíme metodu, jak můžeme heteroskedasticitu vzít v potaz při odhadování modelu tak, abychom obdrželi správné výsledky.

Pro začátek je nutné si uvědomit, že odhady regresních koeficientů budou správné i při porušení předpokladu homoskedasticity, jediné, co bude tímto předpokladem pokaženo, jsou nejistoty odhadů regresních koeficientů. Tyto nejistoty odhadů budou ve většině případů podhodnoceny, budeme tedy mít falešný pocit, že jsme koeficienty změřili přesně, i když opak je ve skutečnosti pravdou. Naštěstí existuje metoda, jak upravit nejistoty měření regresních koeficientů tak, aby reflektovaly nesplněný předpoklad homoskedasticity. Této metodě se říká Whiteův odhad kovarianční matice (někdy také sendvičový odhad). Je nad možnosti tohoto seriálu podrobně popsat tento odhad, proto to na tomto místě ani nebudeme dělat. Jediné, co je potřeba si pamatovat, je, že pokud pomocí regresní diagnostiky odhalíme nesplnění předpokladu homoskedasticity, je potřeba použít jiný výpočet nejistot měření, než je ten základní. Ve vzorovém skriptu můžete nalézt příklad, jak takovýto odhad v praxi konstruovat.

Typickým důsledkem použití Whiteova odhadu kovarianční matice namísto základního odhadu kovarianční matice je zvětšení nejistot odhadů regresních parametrů. Toto je nepříjemné z hlediska snížení přesnosti výsledků, ale je nutné si uvědomit, že je to jediná možnost, neboť předchozí nejistoty měření regresních koeficientů byly sice menší, ale nebyly správné. Přítomnost heteroskedasticity v našich datech snižuje přesnost, se kterou jsme schopni určit regresní koeficienty.

Pozorného čtenáře určitě napadne, proč bychom nemohli Whiteův odhad kovarianční matice používat vždy (je to přece obecnější odhad). Toto by bylo možné a obdržené výsledky by byly správné, ale všechny naše výpočty by se výrazně zkomplikovaly. Navíc je nutné zmínit, že mírné porušení předpokladu homoskedasticity příliš nevadí (nijak výrazně to neovlivní správnost získaných výsledků) a můžeme ho ignorovat (jak bylo zmiňováno dříve). Whiteův odhad je nutné používat až v případech opravdu vážného porušení předpokladu homoskedasticity. Ve vzorovém skriptu opět můžete najít ukázky použití Whiteova odhadu včetně ukázky toho, kdy už je nutné tento odhad používat a kdy to ještě není nutné.

Regrese na základě průměrů z měření

Pokud se provádí nějaké rozsáhlejší experimentální měření, většinou se postupuje tak, že se pro každou hodnotu nezávisle proměnné změří více hodnot závisle proměnné[4]. Matematická teorie

[4] Potom se (mimo jiné výhody) dá provádět test vhodnosti prokládané funkce popsaný v minulém díle seriálu.

regresní analýzy potom umožňuje dělat regresi jen na základě průměru naměřených hodnot pro jednotlivé volby nezávisle proměnné.

Pro lepší představu si uvedeme příklad. Představme si, že naměříme následující data

$$\begin{aligned}&(x_1, y_{1,1}), \dots, (x_1, y_{1,n_1}),\\ &\vdots\\ &(x_m, y_{m,1}), \dots, (x_m, y_{m,n_m}).\end{aligned}$$

Takováto situace není nic neobvyklého, na základě dosud vyložené teorie si s ní umíme poradit (tj. umíme těmito daty proložit libovolnou funkční závislost). Co chce tento odstavec říci, je, že pokud by nám někdo dal jen částečnou informaci o těchto měřeních, a to sice průměry hodnot závisle proměnné v jednotlivých skupinách odpovídajících hodnotám nezávisle proměnné, příslušnou výběrovou směrodatnou odchylku průměru a počty měření v jednotlivých skupinách, tedy

$$(x_1, \overline{y_{1,\bullet}}, s_{n_1}, n_1), \dots, (x_m, \overline{y_{m,\bullet}}, s_{n_m}, n_m),$$

stále bychom byli schopni takovýmito daty proložit libovolnou funkční závislost. Někdy se místo surových dat poskytují právě jen průměry hodnot nezávisle proměnné, příslušné výběrové směrodatné odchylky průměrů a počty měření v jednotlivých skupinách z důvodu úspory času a paměťové náročnosti.

Toto tvrzení si nyní také odvodíme. Budeme postupovat tak, že si napíšeme součet čtverců, který chceme při odhadování neznámých regresních koeficientů minimalizovat, na základě všech měření (tedy zatím nebudeme zavádět zjednodušení měřených dat). Následně pomocí algebraických úprav ukážeme, že jsme schopni tento součet čtverců přepsat jen pomocí symbolů $(x_j, \overline{y_{j,\bullet}}, s_{n_j}, n_j)$, čímž bude odvození dokončeno, neboť při praktickém výpočtu můžeme používat jen tento alternativní zápis. Nyní už k samotnému odvození, součet čtverců, který chceme minimalizovat, má tvar

$$S(x_1, \dots, x_j, y_{1,1}, \dots, y_{j,n_j}, \beta_0, \dots, \beta_k) = \sum_{i=1}^{N} (y_i - f(x_j, \beta_0, \dots, \beta_k))^2 ,$$

kde N představuje celkový počet měření. Tento výraz budeme nyní algebraicky upravovat, nejprve rozdělíme sumu přes všechna měření na dvě sumy, nejprve podle hodnot nezávisle proměnné x a poté podle hodnot měření příslušných měření závisle proměnné, čímž dostaneme

$$\sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_i - f(x_j, \beta_0, \dots, \beta_k))^2 .$$

Nyní provedeme důležitý trik, a sice ten, že k vnitřku závorky přičteme a odečteme výběrový průměr příslušné skupiny měření, tedy člen

$$\overline{y_{j,\bullet}} = \frac{1}{n_j} \sum_{i=1}^{n_i} y_{j,i} .$$

Tímto nezměníme hodnotu našeho výrazu, ale následně si tím výrazně pomůžeme, dostáváme

$$\sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{j,i} - \overline{y_{j,\bullet}} + \overline{y_{j,\bullet}} - f(x_j, \beta_0, \dots, \beta_k))^2 .$$

Nyní roznásobíme závorku, čímž dostaneme

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left[(y_{j,i}-\overline{y_{j,\bullet}})^2+2\,(y_{j,i}-\overline{y_{j,\bullet}})\,(\overline{y_{j,\bullet}}-f(x_j,\beta_0,\ldots,\beta_k))+(\overline{y_{j,\bullet}}-f(x_j,\beta_0,\ldots,\beta_k))^2\right].$$

Nyní tento výraz rozdělíme na jednotlivé sumy a uvědomíme si, že některé vzniklé výrazy můžeme z definice přepsat pomocí výrazů $s^2_{n_i}$ a n_i, dostáváme

$$\begin{aligned}&\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{j,i}-\overline{y_{j,\bullet}})^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\overline{y_{j,\bullet}}-f(x_j,\beta_0,\ldots,\beta_k))^2+\\&\qquad\sum_{j=1}^{m}\sum_{i=1}^{n_j}2\,(y_{j,i}-\overline{y_{j,\bullet}})\,(\overline{y_{j,\bullet}}-f(x_j,\beta_0,\ldots,\beta_k))=\\&=\sum_{j=1}^{m}(n_j-1)s^2_{n_j}+\sum_{j=1}^{m}n_j\,(\overline{y_{j,\bullet}}-f(x_j,\beta_0,\ldots,\beta_k))^2+\\&+\sum_{j=1}^{m}\sum_{i=1}^{n_j}2\,(y_{j,i}-\overline{y_{j,\bullet}})\,(\overline{y_{j,\bullet}}-f(x_j,\beta_0,\ldots,\beta_k))\,.\end{aligned}$$

Nyní vidíme, že první dva členy našeho součtu už jsou vyjádřeny jen pomocí symbolů $(x_i,\overline{y_{j,\bullet}},s_{n_i},n_i)$, zbývá nějak se vypořádat se třetím členem. Na tomto místě si musíme uvědomit, že třetí člen je roven nule, což plyne z jednoduchých algebraických úprav, které zde nyní podrobně provedeme.

$$\begin{aligned}&\sum_{j=1}^{m}\sum_{i=1}^{n_j}2\,(y_{j,i}-\overline{y_{j,\bullet}})\,(\overline{y_{j,\bullet}}-f(x_j,\beta_0,\ldots,\beta_k))=\\&\quad=2\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left[y_i\overline{y_{j,\bullet}}-y_{j,i}f(x_j,\beta_0,\ldots,\beta_k)-(\overline{y_{j,\bullet}})^2+\overline{y_{j,\bullet}}f(x_j,\beta_0,\ldots,\beta_k)\right]=\\&\quad=2\sum_{j=1}^{m}\left[n_j(\overline{y_{j,\bullet}})^2-n_j\overline{y_{j,\bullet}}f(x_j,\beta_0,\ldots,\beta_k)-n_j(\overline{y_{j,\bullet}})^2+n_j\overline{y_{j,\bullet}}f(x_j,\beta_0,\ldots,\beta_k)\right]=\\&\quad=0\,.\end{aligned}$$

Shrneme si, co jsme celkově dokázali. Zapsali jsme součet čtverců, který chceme při odhadování neznámých regresních koeficientů minimalizovat, pouze pomocí symbolů $(x_i,\overline{y_{i,\bullet}},s_{n_i},n_i)$ jako

$$S(x_1,\ldots,x_j,y_{1,1},\ldots,y_{j,n_j},\beta_0,\ldots,\beta_k)=\sum_{j=1}^{m}(n_j-1)s^2_{n_j}+\sum_{j=1}^{m}n_j\,(\overline{y_{j,\bullet}}-f(x_j,\beta_0,\ldots,\beta_k))^2\,.$$

Je tedy vidět, že k určení odhadů regresních koeficientů nepotřebujeme znát hodnoty všech měření, ale stačí nám znát hodnoty $(x_j,\overline{y_{j,\bullet}},s_{n_j},n_j)$. V praxi se potom na hledání odhadů regresních koeficientů použijí stejné metody, které byly popsány dříve, a bude to za nás vždy provádět počítač.

Když se do takovéto situace dostaneme a chceme vykreslit graf naměřených hodnot a proložené funkce, vetšinou do grafu nekreslíme samotné naměřené hodnoty, ale pouze průměry měření odpovídajících hodnotám nezávisle proměnné $\overline{y_{j,\bullet}}$ a nejistoty s_{n_j}, které vyjádříme pomocí tzv. error barů (česky chybová úsečka). Chybová úsečka je vertikální úsečka o velikosti $2s_{n_j}$ se středem v bodě $y_{j,\bullet}$, pomocí které v některých případech lépe vyjádříme rozptýlenost měřených hodnot (zejména v případech, kdy máme hodně naměřených hodnot a klasický graf by byl kvůli tomu nepřehledný). Delší chybová úsečka znamená, že měřené hodnoty byly více rozptýleny. Chybová úsečka se dá také interpretovat tak, že je pravděpodobnost přibližně 68 %, že skutečná hodnota prokládané funkce leží uvnitř chybové úsečky (pokud máme dostatek měření, abychom mohli použít centrální limitní větu). V přiloženém vzorovém skriptu také naleznete příklad na použití této metody a kreslení chybových úseček.

Regrese při uvažování nepřesností měření nezávisle proměnné

Až dosud jsme uvažovali pouze nepřesnosti měřené závisle proměnné y, hodnoty nezávisle proměnné x byly považovány za přesně známé. Nyní si rozebereme obecnější případ, kdy uvažujeme i hodnoty nezávisle proměnné x za zatížené nepřesnostmi měření. Naším cílem bude opět prokládat naměřenými daty teoretickou závislost, k tomu budeme chtít odhadovat neznámé regresní koeficienty z naměřených dat.

Stejně jako v předchozích případech se na odhad hodnot regresních koeficientů použije metoda maximální věrohodnosti. Nyní ale bude analytické vyjádření řešeného problému o něco složitější. Obecně budeme předpokládat, že pro měřená data platí teoretická závislost

$$y = f(x, \beta_0, \ldots, \beta_k)\,,$$

ale my nemůžeme přímo měřit hodnoty x a y. Tyto veličiny budeme vždy měřit nepřesně, výsledné hodnoty měření $\tilde{x_i}, \tilde{y_i}$ tedy budou rovny

$$\begin{aligned} \tilde{y_i} &= f(x_i, \beta_0, \ldots, \beta_k) + \varepsilon_i\,, \\ \tilde{x_i} &= x_i + \delta_i\,, \end{aligned}$$

kde ε_i je náhodná veličina s rozdělením $N(0, \sigma_y^2)$ a δ_i je náhodná veličina s rozdělením $N(0, \sigma_x^2)$.

Nebudeme na tomto místě podrobně odvozovat, jak budou odhady v takovémto případě vypadat, neboť je to příliš komplikované a v praxi je za nás bude vždy dělat počítač. Jediné, co zde uvedeme, je, jak si můžeme takovouto metodu představit ve srovnání s klasickou regresí při uvažování měření hodnot nezávisle proměnné bez nepřesností. V případě klasické regrese jsme chtěli minimalizovat součet druhých mocnin vertikálních vzdáleností naměřených hodnot od prokládané funkce. Vertikální vzdálenost naměřené hodnoty od prokládané funkce byla nazývána residuum, matematicky zapsáno

$$U_i = y_i - f(x_i, \beta_0, \ldots, \beta_k)\,.$$

V případě, kdy uvažujeme i nepřesnosti měření nezávisle proměnné, ale nebudeme uvažovat vertikální vzdálenost, ale přímo vzdálenost naměřené hodnoty od prokládané funkce. Abychom mohli tuto vzdálenost správně upravit, potřebujeme znát poměr mezi rozptyly náhodných nepřesností měření příslušných nezávisle a závisle proměnné, tedy poměr

$$\frac{\sigma_x^2}{\sigma_y^2}\,.$$

Na závěr poznamenejme, že v praxi nastane téměř vždy případ, kdy měříme i hodnoty nezávisle proměnné s nepřesnostmi (minimálně nejistota typu B způsobená použitím nedokonalého měřidla). Téměř nikdy se nestane, že bychom hodnoty nezávisle proměnné znali přesně. Pokud ale je rozptyl hodnot měření nezávisle proměnné σ_x^2 výrazně menší, než rozptyl hodnot měření závisle proměnné, tato metoda dá téměř shodné výsledky jako klasická regrese (bez uvažování nepřesností nezávisle proměnné). Je to způsobeno tím, že úprava v chápání vzdálenosti nebude příliš odlišná od klasické vertikální vzdálenosti. V takovýchto případech je potom naprosto postačující použít metody klasické regrese, není nutné věci příliš komplikovat (výsledky budou jen zanedbatelně odlišné). Tato popsaná metoda by se měla aplikovat jen v případech, kdy jsou hodnoty σ_x^2 a σ_y^2 srovnatelné, což se v praxi často nestává.

Toto je způsobeno tím, že hodnoty nezávisle proměnné jsou typicky zatíženy jen nejistotou typu B (nepřesnosti měřidla), zatímco u hodnot závisle proměnné se musí počítat s nejistotami typu A (náhodnost měřených dat) i B (nepřesnosti měřidla). Toto způsobí, že hodnoty závisle proměnné mají řádově vetší rozptyl než hodnoty nezávisle proměnné. Je ale nutné mít na paměti, že v případech, kdy jsou nepřesnosti měření závisle a nezávisle proměnné srovnatelné (což se také občas stává), metody klasické regrese jsou nedostačující a musíme proto použít výše popsaný zobecněný postup.

Jen úplně na závěr poznamenejme, že i v takovémto obecnějším případě lze konstruovat intervalové odhady pro hodnoty regresních koeficientů i pro hodnotu prokládané funkce. Teoretické odvození je však příliš složité na to, abychom ho zde uváděli. Ve vzorovém skriptu můžete najít příklady použití této metody na konkrétních datech.

Závěr

Toto je ze seriálu 30. ročníku FYKOSu všechno. Pokud jste dočetli až sem, musím Vám pogratulovat a zároveň poděkovat za trpělivost. Doufám, že znalosti, které jste v tomto seriálu načerpali, v budoucnosti budete moci zužitkovat, ať už v rámci různých fyzikálních soutěží nebo během studia na střední (nebo i vysoké) škole.

Na závěr bych chtěl ještě kromě čtenářů seriálu a řešitelů seriálových úloh poděkovat i korektorům, kteří se výrazným způsobem podíleli na finální podobě jednotlivých dílů tohoto seriálu.

Michal Nožička

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.