Regresiona analiza
Matematičke metode u fizičkoj hemiji

St. Jerosimić

5. novembar 2019

1 Uvod

Regresiona analiza je skup statističkih postupaka za procenu odnosa između zavisne promenljive i jedne ili više nezavisnih promenljivih (koji se još zovu prediktori ili kovarijati). Najznačajnija forma je linearna regresija u kojoj istraživač nalazi linearnu funkciju koja najbolje fituje podatke. Primer je dobro poznat studentima: metoda najmanjih kvadrata (engl. *Ordinary Least Squares*), koja računa jedinstvenu liniju koja minimizuje sumu kvadrata odstupanja između podataka i linije, i kojoj ćemo posvetiti poseban odeljak. Postoji i nelinearna regresija, u kojoj se suma kvadrata odstupanja mora minimizovati iterativnom procedurom. Regresiona analiza se danas koristi za potrebe na primer mašinskog učenja (engl. *machine learning*) ili standardno za otkrivanje kauzalnog odnosa između nezavisne i zavisne promenljive, što je i predmet ovog predavanja.

Istorijski, počeci regresione analize su radovi Ležanra (*Legendre*, 1805) i Gausa (*Gauss*, 1809) o metodi najmanjih kvadrata za potrebe astrofizičkih razmatranja. Pojam *regresija* je prvi put korišćen u radu iz oblasti bioloških istraživanja, pri opisivanju pojave opadanja visine kod potomaka visokih predaka, koje opadaju prema srednjoj vrednosti populacije (engl. *regress down toward the mean*). Kasnije se regresiona analiza razvijala kao značajna oblast matematičke statistike sa različitim primenama u prirodnim i društvenim naukama. Poslednjih decenija regresiona analiza nastavlja da se razvija u pravcu npr. metoda za robusnu regresiju, regresija gde su prediktori krive, slike, grafici, itd. Na ovom mestu da dodamo i to da je regresiona analiza velika oblast koja daleko premašuje dvočas iz Matematičkih metoda u fizičkoj hemiji. Stoga ćemo se na ovom predavanju usmeriti na najčešće korišćenu metodu najmanjih kvadrata.

Generalno, model regresije uključuje: nepoznate parametre β, nezavisne promenljive X_i, zavisno promenljive Y_i i članove greški e_i:

$$Y_i = f\left(X_i, \beta\right) + e_i \tag{1}$$

gde je f procenjena funkcija koja najpribližnije opisuje podatke, a β brojne vrednosti parametara. Nekada se funkcija zasniva na znanju o odnosu između X_i i Y_i, kada je samo potrebno odrediti parametre, a nekada je potrebno odrediti i formu funkcije f na osnovu empirijskih podataka.

Podsetimo se pojmova interpolacije i ekstrapolacije. Modeli za regresiju predviđaju vrednost za Y ukoliko je poznato X. Predikcija unutar opsega vrednosti podataka za model zove se interpolacija. Predviđanja van opsega podataka zove se esktrapolacija.

2 Grafičko predstavljanje funkcija

Prvi korak u primeni regresione analize jeste nacrtati i proučiti podatke. Ukoliko imamo skup empirijskih podataka veoma je korisno uraditi grafičko predstavljanje rezultata. Primer: pogledajmo u dodatku na ovo predavanje (pdf fajl **Podaci i fit**) kako se na osnovu empirijskih podataka može izvesti neka elementarna procena funkcije koja bi mogla najbolje da ih opiše uz pomoć programa Eksel. U primeru 1 su prikazani merni podaci o koncentraciji molekula ugljendioksida (u jedinicama ppm) izmerenih u Nacionalnoj okeanskoj i atmosferskoj administraciji na Havajima tokom vremena. Na osnovu podataka od januara 2015. godine do oktobra 2017. godine uradjena je analiza krive i dat je predlog rasta maksimalne koncentracije CO_2 za 2018. i 2019. godinu. Navedena ekstrapolacija se uporedila sa stvarnim mernim rezultatima iz 2019. godine i moglo se zaključiti da linearan fit koji je izgledao najslabiji (za vrlo mali broj tačaka) zapravo je najbolje predvideo maksimalnu koncentraciju za 2019. godinu. Odredjena je i periodična funkcija koja bi mogla da predstavi promene koncentracije CO_2 tokom vremena.

U drugom primeru (ne toliko fizikohemičarskom) prikazan je grafik rasta broja udžbenika u USA tokom poslednjih godina. Predložen je prikaz moguće funkcionalne zavisnosti sa R^2 vrednostima koji će biti definisani u nastavku (to je primer primene u društvenim naukama).

Kada na osnovu dobijenih podataka o zavisnosti promenljive y od x zaključimo da se funkcija može predstaviti eksponencijalno kao $y = a\,\mathrm{e}^{bx}$, tada je korisno predstaviti funkciju kao linearnu: $\ln y = \ln a + bx$, i iz nje odrediti parametre a i b.

Takodje, ukoliko je funkcija kvadratna $y = ax^2$, takodje se može predstaviti linearno ako se predstavi y u funkciji od x^2. Nadalje, racionalna funkcija

tipa $\frac{ax}{b+x}$ se može predstaviti linearnom ako se crta $\frac{1}{y}$ u funkciji od $\frac{1}{x}$:

$$\frac{1}{y} = \frac{b}{ax} + \frac{1}{a}$$

Kao što možemo videti iz navedenih primera, različite funkcije se mogu predstaviti u linearnoj formi. Drugi važan korak u primeni regresionog modela je dakle transformacija y ili x promenljivih tako da im odnos bude linearan (ukoliko je to moguće). Mi ćemo u nastavku opisati upravo linearnu regresionu analizu za statističku obradu linearnog fita.

3 Obična metoda najmanjih kvadrata

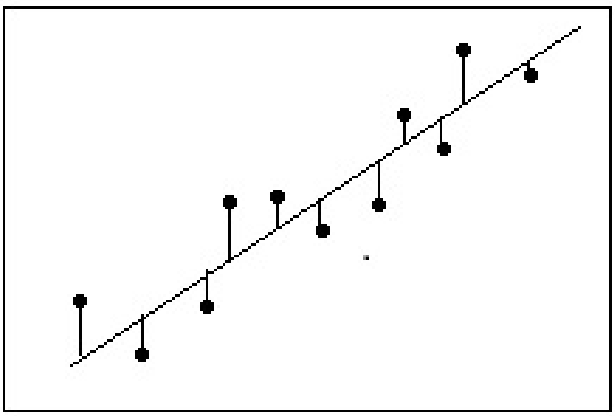
U metodi najmanjih kvadrata (koja se ne primenjuje samo u linearnoj regresiji već šire), uvodi se pojam vertikalnog rastojanja između tačke podatka i fitovane linije, koje se zove rezidual (videti sliku 1), i traži se takva kriva (tj. linija kod linearne funkcije), da suma (po svim tačkama podataka) kvadrata reziduala ima što je manju vrednost. Postoje i metode u kojima se minimizuje suma apsolutnih vrednosti reziduala, ali u metodi najmanjih kvadrata, traži se minimum sume kvadrata reziduala.

Jednostavna linearna regresija je linearna regresija za jednu varijablu (jedan prediktor). Dobijaju se dvodimenzioni grafici koji se sastoje iz jedne nezavisne i jedne zavisne promenljive. (Na ovom mestu napomenimo da u slučaju postojanja više tzv. nezavisnih promenljivih, usled toga što one ne moraju biti zaista nezavisne jer postoje u opštem slučaju korelacije među njima (međuzavisnost), bolje ih je nazivati prediktorima ili kovarijatima. U ovom slučaju postojanja samo jednog prediktora možemo bez ogradjivanja koristiti termin nezavisno promenljive). Obična linearna regresija je najčešće korišćena tehnika za odredjivanje kako na varijablu od interesa y utiče promena druge varijable x.

U linearnoj regresiji modelna funkcija je linearna $y = a+bx$, gde je b nagib, a a odsečak na ordinati. Neka je opaženo n podataka: (x_i, y_i), $i = 1, ..., n$. Ako dodamo član sa greškom e_i (jer se generalno neće sve tačke nalaziti na liniji), vrednosti y_i jednake su:

$$y_i = a + bx_i + e_i \tag{2}$$

Reziduali e_i su prikazani na slici 1.



Slika 1: Prikaz reziduala, vertikalno rastojanje y_i od linije

3.1 Odredjivanje parametara a i b

Cilj je naći takve vrednosti parametara a i b koji će dati najbolji linearan fit za podatke. Najbolji fit prema pristupu najmanjih kvadrata daje linija koja minimizuje sumu kvadrata reziduala (ostataka) $e_i = y_i - a - bx_i$. Označimo sumu kvadrata reziduala sa $Q(a, b)$:

$$Q(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{3}$$

tada minimizacija funkcije $Q(a, b)$ znači da se odrede izvodi funkcije po parametrima a i b i izjednače sa nulom.

Odredimo izvod Q po parametru a:

$$Q = \sum_{i=1}^{n} \left[(y_i - a)^2 - 2(y_i - a)bx_i + b^2 x_i^2 \right] =$$

$$\sum_{i=1}^{n} \left(y_i^2 - 2ay_i + a^2 - 2bx_i y_i + 2abx_i + b^2 x_i^2 \right)$$

$$\frac{\partial Q}{\partial a} = \sum_{i=1}^{n} (-2y_i + 2a + 2bx_i) = 0$$

$$\sum_{i=1}^{n} (y_i - a - bx_i) = (y_1 - a - bx_1) + (y_2 - a - bx_2) + ... = 0$$

$$\sum_{i=1}^{n} y_i - na - b \sum_{i=1}^{n} x_i = 0$$

Ukoliko poslednju jednakost podelimo sa n (ukupan broj tačaka), dobija se:

$$\bar{y} = a + b\bar{x} \tag{4}$$

gde smo koristili oznaku $\bar{y}$ za srednju vrednost n izmerenih vrednosti y_i, a $\bar{x}$ za srednju vrednost x_i. To znači da je konstanta a (y-odsečak) takva da linija mora da prođe kroz srednje vrednosti $\bar{x}$ i $\bar{y}$. Srednje vrednosti su centar "oblaka"tačaka na grafiku. Vrednost a je sada optimizovana vrednosti parametara.

Odredimo izvod $Q(a, b)$ po promenljivoj b:

$$Q = \sum_{i=1}^{n} \left(y_i^2 - 2ay_i + a^2 - 2bx_i y_i + 2abx_i + b^2 x_i^2 \right)$$

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} \left(-2x_i y_i + 2ax_i + 2bx_i^2 \right) = 0$$

$$\sum_{i=1}^{n} \left(x_i y_i - ax_i - bx_i^2 \right) = \sum_{i=1}^{n} \left(x_i y_i - \bar{y}x_i + b\bar{x}x_i - bx_i^2 \right) = 0$$

Odavde direktno sledi:

$$b = \frac{\sum_{i=1}^{n} (x_i y_i - x_i \bar{y})}{\sum_{i=1}^{n} \left(x_i^2 - \bar{x}x_i \right)} \tag{5}$$

i budući da važe relacije (iz definicija srednjih vrednosti):

$$\sum_{i=1}^{n}(x_i-\bar{x})=0 \quad \sum_{i=1}^{n}\bar{x}\,(x_i-\bar{x})=0$$
$$\sum_{i=1}^{n}(y_i-\bar{y})=0 \quad \sum_{i=1}^{n}\bar{x}\,(y_i-\bar{y})=0$$

može se b predstaviti kao:

$$b=\frac{\sum_{i=1}^{n}\left(x_iy_i-x_i\bar{y}+\bar{x}\bar{y}-\bar{x}y_i\right)}{\sum_{i=1}^{n}\left(x_i^2-\bar{x}x_i-\bar{x}x_i+\bar{x}^2\right)}$$

pa dobijamo krajnji izraz za optimizovanu vrednost b u obliku:

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{6}$$

Parametar b se na osnovu prethodne jednakosti može izjednačiti sa sledećim veličinama:

$$b=\frac{S_{xy}}{S_x^2}=r_{xy}\frac{S_y}{S_x} \tag{7}$$

gde je S_{xy} kovarijansa između skupa podataka x i y, S_x^2 varijansa od x, S_y^2 varijansa od y, S_x i S_y odgovarajuće nekorigovane standardne devijacije (koren iz varijansi), i r_{xy} veličina koja se zove **korelacioni koeficijent** između x i y za razmatrani uzorak:

$$S_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \tag{8}$$

$$S_x^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \tag{9}$$

$$S_y^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} \tag{10}$$

$$S_x=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{11}$$

$$S_y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} \tag{12}$$

$$r_{xy}=\frac{S_{xy}}{S_xS_y} \tag{13}$$

vidimo da u imeniocu stoji $n-1$ usled toga što se radi o uzorku (Beselova korekcija), a ne kompletnoj populaciji.

Veličine koje se dobijaju u regresionoj analizi se mogu napisati na mnogo različitih načina koji su matematički ekvivalentni. Ukoliko bi se formula (5) raspisala, dobio bi se izraz:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{14}$$

a pomoću relacije (4), za vrednost a bi se dobilo (možete izvesti):

$$a = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{15}$$

Formule (6) i (14) za parametar b se mogu prevesti jedna u drugu matematičkim operacijama (radi se o drugačijem zapisu). Iz poslednjih formula za a i b se korišćenjem npr. programa Eksel na jednostavan način mogu odrediti nagib b i odsečak na ordinati a.

3.2 Kriterijum r^2

U dodatku (**Podaci i fit**) su predstavljene linearne, kao i eksponencijalne funkcije, zatim polinomi sa greškama fita R^2 dobijenih iz Eksela. Sada ćemo navesti šta predstavlja R^2 u metodi najmanjih kvadrata za linearnu regresiju. Radi se zapravo o kvadratu korelacionog koeficijenta definisanog jednačinom (13), a koji se eksplicitno može napisati kao:

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{16}$$

Osim parametara a i b treći parametar koji je od fundamentalnog značaja jeste korelacioni koeficijent r_{xy} ili koeficijent odredjenja r^2 (engl. *coefficient of determinantion*). r je odnos između varijanse y predstavljene preko regresione formule i ukupne varijanse y, drugim rečima:

$$r_{xy}^2 = \frac{b^2 S_x^2}{S_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} \tag{17}$$

r^2 je statistička mera za kvalitet odredjenog fitovanja krive ili linije. Radi se o kvadratu odnosa varijanse zavisno promenljive koja se predstavlja linearnim fitom i stvarne varijanse podataka y_i. r^2 je uvek između 0 i 1 (ili 0% i 100%). 0% pokazuje da model ne objašnjava ništa od varijabilnosti zavisne

promenljive oko njene srednje vrednosti. 100% pokazuje da model objašnjava svu varijabilnost podataka zavisne promenljive oko njene srednje vrednosti. Generalno, što je veća vrednost r^2 (što je bliži 1) to model bolje fituje podatke. Treba, međutim imati u vidu da uopšteno kod regresione analize visok korelacioni koeficijent ne znači odmah da se radi o dobroj regresiji (ima primera kod nelinearne regresije da se nekada podaci bolje fituju polinomom višeg reda, jer se povećanjem reda polinoma r^2 povećava, ali to ne znači da se poboljšava i cela regresiona analiza).

3.3 Pitanje nesigurnosti

Reziduali e_i su odstupanja svake vrednosti y_i od njene vrednosti procenjene formulom, tj od optimizovane linije, i možemo ih smatrati greškom za svaku pojedinu tačku. Već smo rekli da je suština metode da se minimizuje suma kvadrata pojedinačnih odstupanja (grešaka), tj. veličina $Q(a, b)$

$$Q(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Međutim, nakon što se odredi optimalna linija koja će predstaviti zavisnost y od x, može se govoriti o veličini koju zovemo "varijansa greške", a koja je srednja vrednost kvadrata grešaka (u oznaci MSE, engl. *Mean Square Error*):

$$\text{MSE} = \frac{Q(a,b)}{n-2} = \frac{\sum_{i=1}^{n} (y_i - a - bx_i)^2}{n-2} \tag{18}$$

MSE se računa podelom sa $n-2$ (a ne sa $n-1$ kako se koristi kod varijanse promenljive nekog uzorka), zato što linearna regresija uklanja dva stepena slobode od podataka, procenom dva parametra a i b. Dakle, kao što smo ranije definisali varijansu promenljive y da je jednaka srednjoj vrednosti kvadrata odstupanja y_i od srednje vrednosti y,

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

tako se sada definiše varijansa greške (MSE) koja umesto srednje vrednosti $\bar{y}$ sadrži vrednost $a + bx_i$, dakle pojedinačnu vrednost određenu linearnim fitom, tj. regresijom.

Koren iz MSE je **standardna devijacija regresije**, odnosno standardna devijacija reziduala, u oznaci $S_{y\bullet x}$:

$$S_{y\bullet x} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - a - bx_i)^2}{n-2}} \tag{19}$$

Može se pokazati da je standardna greška nagiba regresije b jednaka:

$$S_b = \frac{S_{y\bullet x}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{20}$$

koja se može dovesti u vezu i sa parametrima b i r_{xy}:

$$S_b = \frac{b}{\sqrt{n-2}} \frac{\sqrt{1-r^2}}{r} \tag{21}$$

Ukoliko su reziduali normalno distribuisani (prema normalnoj raspodeli), tada će i nagib podleći t -raspodeli. Da bi se odredilo da li je nagib regresione linije statistički značajan, može se izračunati t vrednost:

$$t = \frac{b}{S_b} = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}} \tag{22}$$

i da se zatim koristi t-tabela da se proceni vrednost α za tu vrednost t (i za $n-2$ stepena slobode). Interval pouzdanosti je tada:

$$b \pm t_{\alpha(2),n-2} S_b \tag{23}$$

Bez izvodjenja navodimo i standardnu devijaciju za korelacioni koeficijent:

$$S_r = \sqrt{\frac{1-r^2}{r-2}} \tag{24}$$

a da bi odredili da li je korelacija između x i z statistički značajna, poredimo r sa njenom standardnom greškom S_b:

$$t = \frac{r}{S_r} = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}} \tag{25}$$

i nalazeći vrednost u t tabeli (ista je vrednost kao t za nagib), pa je statistički značaj korelacionog koeficijenta ekvivalentan statističkom značaju nagiba b.

Da bi se linearna regresija sprovela kako treba, potrebno je analizirati da li su dobijeni nagib i odsečak fizički prihvatljivi, zatim proučiti reziduale, nacrtati ih u funkciji od x i od y, jer ako se povećavaju ili smanjuju sa x možda nije izabrana dobra funkcija (nije u osnovi linearna). Ukoliko postoje autlajeri (engl. *outlier*), to su tačke koje odskaču i nalaze se van očekivanog opsega vrednosti, i mogu značajno izmeniti srednje vrednosti $\tilde{x}$ i $\bar{y}$ ili procenjenu vrednost a i b parametara, treba proveriti njihovu validnost ili koristiti robusne metode regresije, koje će biti "otporne"na autlajere. Itd.

3.4 Pretpostavke iza linearne regresije

1. Uzorak je reprezentativan. Podaci koji se koriste za fitovanje su reprezenti cele populacije.

2. Odnos između x i y je u osnovi linearan.

3. Prve dve tačke su dovoljne za predvidjanje y od xa. Medjutim, ukoliko želimo odrediti standardnu grešku predvidjanja svake pojedine tačke y_i nužno je da su varijanse reziduala konstantne.

4. Da bi se dobro procenila prava y vrednost na osnovu regresije, treba pretpostaviti i da su reziduali nezavisni.

5. Da bi se dale izjave o verovatnoćama, npr. izvodili statistički testovi za b ili r, određivali intervali pouzdanosti, osim do sada navedenih tačaka mora se pretpostaviti i da reziduali podležu normalnoj raspodeli. Suprotno široko rasprostranjenom mišljenju, linearna regresija ne pretpostavlja ništa u vezi sa raspodelama x ili y varijabli, ona samo uključuje pretpostavke u vezi sa distribucijom reziduala! Kao i za mnoge druge statističke tehnike, nije nužno da su podaci normalno distribuisani, samo je važno da su greške normalno distribuisane. A ovo poslednje je važno samo u slučaju da se žele sprovesti validni statistički testovi.

4 Robusno fitovanje

Mi zapravo nemamo prostora za analizu ove vrste regresije koja je značajna da bi se izbegao negativan uticaj autlajera na rezultat regresione analize (primer pokazan na tabli na predavanju). Robusne metode fitovanja mogu npr uvesti pojam medijane (središnja vrednost niza, a ako niz ima paran broj onda je jednaka aritmetičkoj sredini dve središnje vrednosti), pa traži da se medijana niza reziduala

$$Med\left\{(y_1 - a - bx_1)^2, (y_2 - a - bx_2)^2, ..., \right\}$$

minimizuje, a ne suma kvadrata reziduala. Zainteresovane studente za temu upućujem na niz dobrih tekstova i referenci na wikipediji, u slučaju da u svom budućem eksperimentalnom radu naiđu na autlajere, koje bi svi najviše voleli jednostavno da izbace. Na pominjem da ne treba učiti napamet komplikovane relacije za greške.