ANALIZA DE REGRESIE ORIENTATĂ-OBIECT

ing. Marius Nițu

Institutul de Cercetări în Informatică.

Rezumat: Lucrarea reprezintă o încercare de a ordona și ierarhiza diferite tipuri de regresii (simple și multiple, lineare sau nelineare) în scopul unei analize a implicațiilor posibile într-o abordare bazată pe programarea orientată-obiect. Modelul matematic al fiecărui tip de regresie este discutat sumar, (pentru detalii, cititorii interesați pot consulta [1],[2],[3] și sînt reliefate unele probleme clasice legate de dificultăți de calcul numeric. Se propune o taxonomie a analizei de regresie, iar în final sînt discutate avantajele unei abordări OOP.

Cuvinte cheie: modelare matematică, analiza de regresie, programare orientată-obiect (OOP).

Analiza de regresie reprezintă un instrument statistic intens folosit într-o varietate de domenii între care se regăsesc prelucrarea datelor experimentale, analiza de marketing, etc. În special în acest din urmă domeniu analiza de regresie a fost aplicată într-o multitudine de probleme de "predicție" cum ar fi: prezicerea consumului săptămînal de bere al unei persoane în funcție de vîrsta, sexul, venitul și nivelul de educație al acesteia; estimarea atractivității exercitate de diferite jurnale medicale asupra medicilor în funcție de stilul de scriere, calitatea ilustrațiilor, utilitatea reclamelor prezentate și calitatea științifică a informațiilor, iar lista poate continua fiind aproape fără sfîrșit. Din acest motiv aparatul matematic utilizat a fost puternic dezvoltat și există referințe vaste asupra algoritmilor ce pot fi utilizați. Ceea ce ne propunem în acest articol este ca, după o scurtă prezentare a modelelor matematice implicate de regresia simplă și multiplă, lineară sau neliniară să investigăm posibilitățile și avantajele oferite de abordarea problemei în spiritul conceptelor programării OOP.

1. Regresia simplă. Modelul matematic.

Să presupunem că plecînd de la o serie de valori tabelate obținute experimental pentru două variabile aleatoare X și Y, ne propunem să găsim o relație generală între X și Y. În cazul cel mai simplu, căutăm o relație liniară de forma:

$$Y^* = a + b\,X \tag{1}$$

unde:

Y^* - desemnează valorile criteriu prezise de modelul linear;

a - desemnează termenul liber sau valoarea lui Y^* cînd X este zero;

b - desemnează panta dreptei sau variația în Y^* pe unitatea de variație în X.

Metoda folosită, numită **metoda celor mai mici pătrate** (sau **principiul Gauss** [1]), constă în esența din minimizarea expresiei:

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{2}$$

unde x_i, y_i reprezintă datele empirice.

În plus, așa numitul model "clasic" al regresiei, operează în următoarele ipoteze [1], [2]:

a. pentru fiecare valoare fixată a lui X, presupunem o repartiție normală a lui Y;

b. media tuturor acestor repartiții normale ale lui Y - condiționate de X - se află pe o dreaptă cu panta τ;

c. toate repartițiile normale ale lui Y au aceeași dispersie.

Exprimat algebric modelul este:

$$Y = \alpha + \tau X + \varepsilon \tag{3}$$

unde: α - media populației Y cînd X = 0;

τ - variația în media populației Y pe unitatea de variație în X;

ε - un termen de eroare independent, dedus dintr-un univers normal distribuit, și care are media 0.

În acest caz pentru calculul parametrilor (a și b) modelului de regresie lineară se folosesc în mod uzual următoarele relații:

$$b = \frac{\sum_{i=1}^{n} YX - n\overline{YX}}{\sum_{i=1}^{n} X^2 - n\overline{X^2}} \tag{4}$$

$$a = \overline{Y} - b\overline{X}$$

unde n desemnează mărimea eșantionului, iar $\overline{Y}$ și $\overline{X}$ desemnează mediile aritmetice pentru Y și X. Deducerea relațiilor (4) se face anulînd derivatele parțiale ale expresiei (2) în raport cu a și b, transformîndu-le apoi, prin intermediul unor identități, pentru a minimiza erorile legate de calculul numeric și rezolvînd sistemul astfel rezultat [1], [3].

Fără a intra în detalii menționăm că, de obicei, calculul impune evaluarea unor parametrii statistici ce măsoară tăria corelației, dintre care menționăm coeficientul determinării R^2 și coeficientul corelației R_{yx}(vezi [2], [3] pentru detalii).

Este evident că regresia lineară simplă poate fi adeseori nepotrivită. Dacă desenăm graficul datelor, acesta poate să sugereze uneori o hiperbolă, o curbă exponențială, o curbă geometrică sau chiar o curbă trigonometrică. Modele matematice pentru aceste tipuri de curbe pot fi, respectiv:

$$y = \frac{1}{a_0 + a_1 x}$$

$$y = a(b^x)$$

$$y = a \cdot x^b$$

$$y = a_0 + a_1 \cos(\omega \cdot x)$$

De remarcat că, unele lucrări trec în revistă chiar mai multe tipuri de curbe, posibil de utilizat [3]. Să menționăm că regresia în acest caz nu mai este liniară, dar că **metoda celor mai mici pătrate nelineare** poate fi utilizată în continuare. Algoritmul de rezolvare în acest caz se bazează pe linearizarea expresiei funcției utilizate prin intermediul unei schimbări de variabilă adecvată [1], [3], iar apoi se ajunge la cazul clasic. Se mai poate menționa că nu întotdeauna funcțiile propuse pot fi transformate ușor în forme lineare și că exceptînd funcția trigonometrică, în celelate cazuri baza statistică discutată anterior nu mai este valabilă (nu se mai minimizează suma pătratelor abaterilor din y, ci mai degrabă se minimizează suma pătratelor abaterilor din schimbarea de variabilă utilizată [1]).

2. O propunere de taxonomie pentru regresia simplă.

Se poate încerca acum, în baza celor prezentate anterior, să se definească mai întîi o clasă de obiecte pentru regresia lineară simplă, folosind un pseudocod bazat pe limbajul Turbo-Pascal:

```
type RL2 = object
     {datele obiectului RL2}
     X1,X2,Xest:TNVector;
     n          :integer;
     a,b        : float;
   direct:boolean;Stat_Param: TNVector; {vector de
                                  parametri statistici}
     {procedurile obiectului RL2}
     constructor Init(...);
     procedure Compute_Reg; virtual;
     procedure Plot_Data; virtual;
     destructor Done;
   end;
```

Se poate remarca, față de cele prezentate anterior, notația X1, X2 pentru Y respectiv X, prezența unei variabile boolene (direct) prin intermediul căreia se poate comuta sensul regresiei (regresie în Y sau regresie în X) că și prezența unei proceduri de reprezentare grafică a datelor inițiale și a celor estimate (procedura Plot_Data). De asemenea, nu sînt precizați parametrii procedurii de inițializare, lucru nerelevant în contextul actual.

Acum se poate încerca să se deriveze următoarea clasă de obiecte pentru regresia neliniară. Această clasă generică, este considerată ca o moștenire a clasei definite anterior (RL2). Se moștenesc atît datele clasei anterioare, cît și procedurile definite, dar se mai poate introduce o procedură generică numită "Change_Variable" cu rolul de a efectua schimbarea de variabilă adecvată:

```
type RN2 = object(RL2)
                Z:TNVector;
     constructor Init(...);
     procedure Change_Variable;
     procedure Compute_Reg; virtual;
     procedure Plot_Data; virtual;
     destructor Done;
   end;
```

Este evident acum că, scriind în mod adecv procedura de schimbare de variabilă, practic prelucrîn în mod corespunzător vectorul de date inițiale, se poat particulariza această clasă pentru diferite tipuri d funcții nelineare, dorite a fi utilizate ca mod matematic. Se poate, în acest moment, să se defineasc efectiv obiectele de tip regresie hiperbolic exponențială, geometrică etc.

În cele ce urmează se discută foarte pe scurt regresia polinomială bivariată.

3. Regresia polinomială. O nouă clasă.

Motivul abordării regresiei polinomiale este datora aceluiași fapt care a dus la prezentarea regresie nelineare și anume nepotrivirea dintre forma sugerat de graficul valorilor și modelul linear. Abordare separată a acestui tip de regresie (evident tot nelinear este justificată în continuare.

Mai întîi se formulează însă, modelul matematic:

$$Y^* = a + b_1 x + b_2 x^2 + \ldots + b_k x^k \qquad (5)$$

Se dorește minimizarea expresiei:

$$\sum_{i=1}^{n} (y_i - b_k x_i^k - b_{k-1} x_i^{k-1} - \ldots - b_1 x_i - a)^2 \qquad (6)$$

Este evident că, urmînd aceeași cale cu ce menționată anterior, se ajunge la un sistem de k+1 ecuații cu k+1 necunoscute, ce poate fi rezolvat prir eliminare gaussiană [1], [3], [4], [5].

Se poate demonstra [1] însă că, dacă k este rezonabi de mare ($k > 7$), matricea coeficienților sistemului se apropie de așa-numita matrice Hilbert, matrice care este cunoscută ca fiind rău condiționată. Oricum, chiar și neglijind dificultățile legate de calculul numeric creșterea nejustificată a gradului polinomului (k aproximativ de același ordin de mărime cu n) duce la c expresie ce descrie exact toate datele, inclusiv erorile aleatorii de măsurare.

Aceste considerații fac adeseori ca pentru problemele de regresie de tip polinomial să fie preferată utilizarea funcțiilor ortogonale. Cel mai adesea sînt preferate polinoame ortogonale de tip Cebîșev, datorită relației de recurență a acestora ca și datorită faptului că valorile parametrilor b_j nu depind de gradul polinomului [1], [3].

Revenind însă la ierarhia în dezvoltare pe care o propunem, se poate adăuga acum o nouă clasă de obiecte, exploatînd și de această dată posibilitățile

oferite de proprietatea de moştenire:

```
type RNP2 = object(RL2)
      B:TNVector;
        k:integer;
      constructor Init(...);
      procedure Compute_Reg; virtual;
      procedure Plot_Data; virtual;
      destructor Done;
    end;
```

Este pusă în evidenţă în acest moment şi mai clar proprietatea de polimorfism ce caracterizează metoda de programare orientată-obiect. La transmiterea de către programul principal a mesajului de calcul de regresie, procedura Compute_Reg se va comporta adecvat tipului de obiect creat.

Se atrage atenţia că se poate pune în evidenţă şi o soluţie elegantă legată de deficienţele eliminării gaussiene, ca şi de alegerea judicioasă a gradului k al polinomului utilizat. Testînd raportul k/n şi calculînd criteriul de concordanţă al lui Gauss [3] se poate comuta între metoda rezolvării sistemului de ecuaţii pentru calculul coeficienţilor regresiei şi metoda bazată pe polinoame Cebîşev, comutare ce poate să fie transparentă utilizatorului.

4. Regresia multiplă. Definitivarea taxonomiei.

Modelul matematic al regresiei lineare multiple poate fi definit restrictiv prin introducerea unei a doua variabile predictor în modelul bivariat linear:

$$Y = \alpha + \tau_1 X_1 + \tau_2 X_2 + \varepsilon \qquad (7)$$

cu parametrii estimaţi de:

$$Y^* = a + b_1 X_1 + b_2 X_2$$

în acest caz se caută un plan particular - şi anume acel plan a cărui sumă a pătratelor abaterilor din y_i este minimă. Metoda generală de rezolvare constă în minimizarea expresiei:

$$\sum_{i=1}^{n} (y_i - a - b_1 \cdot x_{i,1} - b_2 \cdot x_{i,2})^2 \qquad (8)$$

După derivare rezultă şi în acest caz un sistem de ecuaţii ce poate fi rezolvat [1], [3].

Dar, revenind la formularea restrictivă a problemei (7), se pot observa unele lucruri interesante. Studiind interpretarea ce se poate da coeficientului b_1 [2] putem să imaginăm mai întîi o regresie a lui X_1 în X_2. Calculînd apoi reziduurile $x_{i,1}$-$x_{i,2}^*$ din acestă regresie şi regresînd apoi valorile lui Y pe mulţimea reziduurilor lui X_1, coeficientul regresiei "simple" va fi egal cu coeficientul regresiei parţiale b_1. În mod analog se poate trata şi coeficientul b_2. Ca o consecinţă practică, înseamnă că putem rezolva problema formulată de relaţia (7), efectuînd trei regresii simple în loc să se rezolve un sistem de 3 ecuaţii în cele 3 necunoscute a, b_1, b_2.

Se pot adăuga acum două noi ramuri taxonomiei precedente:

```
type RL3 = object(RL2)
           ob1,ob2:^RL2; {^ - pointer la}
           constructor Init(...);
           procedure Compute_Reg; virtual;
           procedure Plot_Data; virtual;
           destructor Done;
         end;
type RN3 = object(RN2)
           ob1,ob2:^RN2; {^ - pointer la}
           constructor Init(...);
           procedure Compute_Reg; virtual;
           procedure Plot_Data; virtual;
           destructor Done;
         end;
```

Obiectul RN3 poate fi considerat ca o extrapolare a regresiei nelineare bivariate şi care este, bineînţeles, o clasă generică. Particularizarea funcţiilor de schimbare de variabilă conduce la forma dorită a clasei pentru obiecte de acest tip. Prin similitudine, se poate adăuga o nouă clasă pentru regresia polinomială multiplă:

```
type RNP3 = object (RNP2)
      {date şi proceduri proprii}
      .....
    end;
```

În sfîrşit, revenind la cazul general al regresiei multiple lineare în "v" variabile, se poate alege şi în acest caz o variabilă dependentă (afectată de erori statistice) şi se regresează în cele "v-1" variabile rămase. Modelul matematic în acest caz devine:

$$Y = \alpha + \tau_1 X_1 + \tau_2 X_2 + ... + \tau_{v-1} X_{v-1} + \varepsilon \qquad (9)$$

care prin minimizarea expresiei:

$$\sum_{i=1}^{n} (y_i - a - b_1 \cdot x_{1,i} - b_2 \cdot x_{2.i} - ... - b_{v-1} \cdot x_{v-1,i})^2$$

conduce la un sistem de "v" ecuaţii cu "v" necunoscute ce poate fi rezolvat [1], [3], [4], [5].

Se poate completa taxonomia construind această nouă clasă ca moştenire a clasei RL3 şi rescriind în mod adecvat procedurile de calcul:

```
type RLV = object (RL2)
      {date şi proceduri proprii}
      .....
    end;
```

5. Concluzii.

În această lucrare s-a făcut o sumară trecere în revistă a problemelor legate de regresia simplă şi multiplă lineară sau neliniară şi s-a încercat să se construiască o taxonomie în spiritul conceptelor

programării orientate-obiect. Fără a intra în detaliile legate de algoritmii propriu-zişi, s-a încercat relevarea particularităţilor legate de dificultăţile de calcul numeric şi exploatarea cunoştinţelor acumulate în domeniu. Ierarhizarea propusă a avut ca principal criteriu de clasificare unitatea algoritmilor de rezolvare.

Această abordare are în mod clar un mare avantaj evident. Utilizarea proprietăţii de moştenire şi a celei de încapsulare a datelor permit simplificarea la maxim a procedurilor de calcul prin preluarea automată a codului identic din obiectul părinte. Un alt avantaj constă în faptul că este posibilă o afişare în paralel a rezultatelor calculelor diferitelor tipuri de regresii (chiar şi sub formă grafică) pentru unul şi acelaşi set de date, şi,, chiar mai mult decît atît, printr-o îmbinare judicioasă a secvenţelor de creare a obiectelor şi a celor de calcul în programul principal, este posibilă implementarea cu succes a metodei de "regresie în trepte" (Stepwise Regression), intens folosită în cercetarea de marketing.

De asemenea, abordînd în acelaşi spirit metodele de analiză de varianţă şi covarianţă, analiza factorială şi discriminatorie, se poate anticipa că pot fi obţinute rezultate interesante din punct de vedere al efortului de programare.

NOTĂ:

Acest articol reprezintă versiunea în limba română revizuită a materialului "An Object Oriented approach for Regression Analysis" care a fost prezentat la TWELTH IASTED INTERNATIONAL CONFERENCE, "Modelling, Identification and Control" 15-17 februarie 1993, Innsbruck, Austria.

Bibliografie

1. **DORN, W., MCCRACKEN, D.**: Metode numerice cu programe în Fortran IV, Editura Tehnică, Bucureşti 1976.
2. **GREEN, P., TULL, D., Albaum, G**: Research for Marketing Decisions. **Prentice-Hall**, Englewood Cliffs, New Jersey, 1988.
3. **CONSTANTINESCU,I., ş.a.**:Prelucrarea datelor experimentale cu calculatoare numerice, Editura Tehnică, Bucureşti, 1980.
4. **SALVADORI, M., BARON, L.**: Metode numerice în tehnica, Editura Tehnică, Bucureşti, 1972.
5. **SIMIONESCU I., ş.a.**: Metode Numerice, IPB, Bucureşti, 1992.
6. **DEMIDOVICH, B.,MARON I.**: Computational Mathematics, MIR Publisher, Moscow, 1981.
7. **MEYER, B.**: Object Oriented Software Construction, Prentice Hall, Englewood Cliffs, New Jersey, 1986.