ANEXA

Prezentare model actuarial GLM

INTRODUCERE

Societăţile de asigurare folosesc metode actuariale pentru a determina aceste variabile utilizându-se în general o modelare de tipul **generalized linear models (GLM)**. Mai jos este o prezentare generală a acestui model actuarial.

MODELUL GLM

Dauna medie şi frecventa daunelor pentru categoria autoturismelor deţinute de persoane fizice atât pentru riscul de daune materiale cat şi pentru cel de vătămări corporale s-au estimat utilizându-se o modelare de tipul Generalized Linear Models.

Modelul GLM reprezintă o extensie a modelului tradiţional liniar (care cuprinde regresiile simple sau multiple, analize ANOVA) şi asociază unei variabile pe care vrem sa o previzionam (numita variabila „răspuns" sau „variabila dependenta") mai multe variabile (numite „factori" sau „predictori" sau „variabile independente") despre care avem anumite informaţii.

Scopul modelelor liniare este redarea relaţiei dintre variabila răspuns Y şi un număr de variabile predictive, X. Atât modelul clasic linear cat şi cel generalizat privesc observaţiile Yi ca fiind realizări ale variabilei aleatoare Y.

Modelele GLM sunt compuse dintr-o gama larga de modele care includ modelele liniare ca şi caz particular. Ipotezele restrictive ale modelului liniar de Normalitate, varianta constanta şi aditivitatea efectelor sunt eliminate. în schimb, distribuţia variabilei răspuns este membra a familiei exponenţiale, iar dispersiei ii este permis sa varieze cu media distribuţiei. Membre ale familiei exponenţiale sunt: distribuţia Normala, distribuţia Poisson, distribuţia Binomiala, distribuţia Gamma, distribuţia Invers Gaussiana.

Modelul tipic pentru modelarea numărului de daune sau a frecventei daunei este Poisson multiplicativ. În afara de faptul ca este universal acceptată şi cel mai des utilizată în

modelarea numărului daunelor, distribuţia Poisson deţine şi o proprietate particulara care o face şi intuitiv adecvata şi anume aceea ca este invarianta la masurile de timp.

Similar, un model standard pentru modelarea severităţii daunelor datorita formei ei generale, este Gamma multiplicativ.

Forma GLM:

$$\bar{y} = X\bar{\beta} + \bar{\varepsilon}$$

$\bar{y}$ - vector nx1 al valorilor actuale observate, variabile aleatoare independente, normal distribuite.

X – matrice nxp a variabilelor independente (factori)

$\bar{\beta}$- vector px1 al parametrilor

$\bar{\varepsilon}$ - reprezinta termenul „eroare”, este un vector nx1 de variabile aleatoare independente, normal distribuite, cu media 0 şi dispersia σ2.

Relaţia dintre variabila dependenta şi variabilele independenta este definita prin E(y).

E($\bar{y}$)=b’(θ)=$\bar{\mu}$.

O alta componenta a modelului GLM este funcţia de legătură („link function”), diferenţiabilă, inversabila, care descrie modul în care media variabilei yi, μi relaţionează cu predictorul liniar ηi = xi’ $\bar{\beta}$:

$g(\mu_i) = x_i'\bar{\beta}$.

E($\bar{y}$)=g-1($\bar{\eta}$)= $\bar{\mu}$

Deci, o funcţie de medie şi nu media, este modelata liniar.

Sintetizând, un model GLM este structurat astfel:

$$\mu_i = E(Y_i) = g^{-1}(\sum_j X_{ij}\beta_j + \varepsilon_i)$$

$$Var(Y_i) = \frac{\emptyset V(\mu_i)}{\omega_i}$$

Unde: Yi – vectorul variabilei răspuns, g(x) – funcţia link, inversabila, care face legătura intre răspunsul aşteptat şi combinaţia liniara a factorilor observaţi, Xij – matricea factorilor, βj – vectorul parametrilor modelului (care trebuie estimaţi), εi – vectorul eroare („offset”), ∅ - parametru, V(x) – functia variatie, ωi – coeficient care atribuie o credibilitate sau o pondere fiecărei observaţii.

Tabelul de mai jos sintetizează câteva modele tipice:

$\bar{y}$	Frecventa	Nr. daune	Dauna medie	Probabilitate (de ex. de reînnoire)
Link g(x)	ln(x)	ln(x)	ln(x)	ln(x/(1-x)))
Eroare	Poisson	Poisson	Gamma	Binomial
Parametru ∅	1	1	Estimat	1
Variaţia V(x)	x	x	x2	x(1-x)
Ponderile ω	Expunere	1	Daune	1
Offset ε	0	ln(expunere)	0	0

Având definit modelul prin X, g(x), V(x), ∅, ω, ε, şi dat fiind setul de observatii y, componentele lui $\bar{\beta}$ se determina prin maximizarea functiei de verosimilitate (sau echivalent, logaritmul funcţiei de verosimilitate). în esenţă, aceasta metoda caută sa gaseasca parametrii care, odata aplicati modelului, produc datele observate cu probabilitatea cea mai mare. Verosimilitatea este definita ca fiind produsul probabilităţilor (sau a funcţiilor densitate de probabilitate în cazul distribuţiilor continue) de observare a fiecărei valori y.

In practica, când exista mulţi factori cum ar fi vârsta, capacitatea cilindrica, judeţul şi durata poliţei, fiecare conţinând mai multe niveluri, este mai util sa se parametrizeze GLM considerând, pe lângă factorii observaţi, un termen „intercept" (constanta GLM), care reprezintă un parametru ce se aplica tuturor observaţiilor.

O analiză de tip GLM a daunelor necesită un anumit volum de date.

Cu toate ca GLM sunt metode multivariate, este utila şi benefica efectuarea unor analize univariate sau bivariate ale datelor înaintea modelarii GLM.

In primul rând, analiza univariata a expunerilor şi a daunelor pentru toţi factorii indica daca o variabila conţine suficiente informaţii pentru a fi inclusa în model (de exemplu, daca 99% din expunerea unei variabile se afla într-un singur nivel, atunci aceasta nu este potrivita pentru modelare).

In al doilea rând, presupunând ca exista distribuţii viabile ale factorilor pe niveluri, atenţie suplimentara trebuie acordata nivelurilor individuale care conţin expunere şi număr de daune reduse. Daca aceste niveluri nu sunt în cele din urma combinate cu alte niveluri, algoritmul de verosimilitate maxima a GLM s-ar putea sa nu conveargă (daca un nivel al unui

factor conţine zero daune şi si se modelează printr-un model multiplicativ, coeficientul teoretic corect pentru acel nivel va fi apropiat de zero, şi parametrul estimat corespunzător logaritmului acelui coeficient s-ar putea sa fie atât de mare şi negativ încât algoritmul numeric corespunzător verosimilităţii maxime nu va converge).

Pe lângă investigarea expunerilor şi daunelor, analiza unifactoriala a frecventei şi daunei medii furnizează indicaţii preliminare a efectului produs de fiecare factor în parte.

Deoarece în analiza se utilizează factori categoriali, este necesara o atenţie sporita cu privire la modul în care aceşti factori sunt categorizati astfel încât GLM sa beneficieze de factorii care afectează sistematic experienţa şi sa se excludă factorii care nu au un efect sistematic. Pentru a distinge intre factorii cu efect sistematic şi cei cu efect aleator (si deci neprobabil sa se repete în viitor) exista mai multe criterii, printre care: erorile standard ale estimaţiilor parametrilor, teste de devianta (teste de tipul III), consistenta în timp şi nu în ultimul rând intuiţia şi simţul realităţii.

Pe lângă faptul ca furnizează estimaţiile de verosimilitate maxima a parametrilor, GLM furnizează şi informaţii suplimentare ce indica certitudinea estimărilor. Un astfel de diagnostic util este reprezentat de erorile standard ale estimaţiilor, acestea fiind definite ca fiind rădăcina pătrata a elementului diagonal al lui -H-1 unde H (Hessian) este matricea derivatelor de ordin 2 a log verosimilităţii.

In general, se presupune ca estimaţiile parametrilor sunt distribuite asimptotic Normal. În consecinţă, se poate efectua un simplu test statistic asupra estimaţiilor parametrilor pentru a vedea daca efectul fiecărui nivel al factorului este semnificativ diferit fata de nivelul de baza al acelui factor. Pentru aceasta se utilizează de obicei testul $\chi 2$ prin care se compara rădăcina pătrata a estimaţiei parametrului împărţită la dispersia ei cu distribuţia $\chi 2$

In practica, deseori interpretarea grafica a estimărilor şi a erorilor standard sunt mai utile în testarea adecvării unui model.

Pentru compararea masurilor de devianta a doua modele se utilizează teste de “tipul III” ($\chi 2$ or F-tests) pentru a determina semnificaţia teoretica a factorilor individuali. Devianta măsoară cu cat diferă valorile previzionate de cele observate.

Pentru validarea modelului, după ce s-a stabilit cat de semnificativi sunt factorii modelaţi, trebuie sa se investigheze şi alte elemente care include: reziduurile (testarea adecvării erorilor), levierul (identifica observaţiile care au o influenta excesiva asupra modelului), transformarea Box-Cox (testarea adecvării funcţiei legătură).

După ce au fost definiţi factorii categoriali, este utila o stabilire a corelaţiilor dintre factori (de exemplu prin utilizarea statisticii de corelaţie Cramer’s V).

Chiar daca nu sunt utilizate direct în procesul GLM, o analiza a corelaţiilor din portofoliu este utila în interpretarea rezultatelor produse de către GLM. în particular, acestea pot explica de ce rezultatele multivariate pentru un anumit factor diferă fata de rezultatele univariate şi pot indica ce factori pot fi afectaţi de eliminarea sau includerea altor factori în GLM.