Stockholms universitet, Matematiska institutionen

Tentamen: Prissättning inom sakförsäkring, MT7028
2021-02-03 9–15

Examinator: Filip Lindskog, lindskog@math.su.se
Återlämning: Meddelas via kursforum.
Tillåtna hjälpmedel: Alla hjälpmedel som inte innebär kommunikation med någon annan.
Allmänt: Resonemang skall vara klara och tydliga att följa.

Uppgift 0

[Obligatorisk] Försäkrar du att du under skrivtiden inte kommunicerat med någon annan och att du kommit fram till dina lösningar helt på egen hand?

Uppgift 1

Betrakta data (x_{ij}), $i = 1,\dots,m_1$, $j = 1,\dots,m_2$, från en Poissonmodell för antalet skadehändelser: för varje indexpar (i,j), antag att x_{ij} är en observation av $X_{ij} \sim \text{Pois}(w_{ij}\gamma_0\gamma_{1i}\gamma_{2j})$ där w_{ij} är en känd konstant svarande mot duration (försäkringsår). Antag att X_{ij} och $X_{i'j'}$ är oberoende då $(i,j) \neq (i',j')$.

(a) Visa att fördelningen för $Y_{ij} := X_{ij}/w_{ij}$ är en EDM och bestäm loglikelihood-funktionen för ML-skattning av $\gamma_0, \gamma_{1i}, \gamma_{2j}$, $i = 1,\dots,m_1$, $j = 1,\dots,m_2$, givet data (y_{ij}). (5 poäng)

(b) Visa från loglikelihood-functionen i (a) att ML-skattningarna av $\gamma_0, \gamma_{1i}, \gamma_{2j}$, $i = 1,\dots,m_1$, $j = 1,\dots,m_2$, löser ett ekvationssystem på formen

$$\begin{aligned}
\gamma_0 &= A,\\
\gamma_{1i} &= B_i, \quad i = 1,\dots,m_1,\\
\gamma_{2j} &= C_j, \quad j = 1,\dots,m_2
\end{aligned}$$

och bestäm uttrycken för A, B_i, C_j ovan. (5 poäng)

Uppgift 2

Antag att Tabell 1 visar historisk skadedata från år 2020, enligt en uppdelning i tariffceller, för en viss försäkringsprodukt. Antag att skadefrekvens modelleras med en multiplikativ Tweedie(1)-modell. Antag att durationerna för de olika tariffcellerna år 2021 bedöms vara dubbelt så stora som de motsvarande historiska durationerna. Beskriv hur relationstalen skattas med GLM. Baserat på en GLM-analys, vad blir prediktionen för antalet skador under 2021 för försäkringsprodukten? Hur ändras prediktionen om modellen reduceras genom att ett av premieargumenten utgår? (10 poäng)

Uppgift 3

Antag att Tabell 1 visar historisk skadedata från år 2020, enligt en uppdelning i tariffceller, för en viss försäkringsprodukt. Antag att riskpremie (pure premium) modelleras med en multiplikativ Tweedie(1.5)-modell. Antag att den historiska durationen för de olika tariffcellerna svarar exakt mot durationen för år 2021. Antag att relationstalen för Zon är kända: relationstalet 1 för Zon 1 och relationstalet 2 för Zon 2. Antag att riskpremie anses ej bero på övriga kovariater i termer av premieargumenten Klass och Ålder med tillhörande klasser. Baserat på en GLM-analys, vad blir prediktionen för skadekostnaden under 2021 för försäkringsprodukten? Hur fördelas den predikterade skadekostnaden över Zon 1 och Zon 2? (10 poäng)

Uppgift 4

Antag att Tabell 1 visar sammanställning av historisk skadedata från år 2015-2020, enligt en uppdelning i tariffceller, för fordon som används för transporttjänster av 37 företag. Antag att en GLM-analys för skattning av förväntad skadekostnad per försäkringsår mynnat ut i en anpassad multiplikativ modell $\mu_{ijk} = \gamma_0\gamma_{1i}\gamma_{2j}\gamma_{3k}$, $i = 1, 2$, $j = 1, 2$, $k = 1, 2$. GLM-analysen inkluderade inte företag som premieargument. Antag att skadedata finns på formen

$$(w_{ijklt}, x_{ijklt}), \quad l = 1, \ldots, 37, \quad t = 2015, \ldots, 2020,$$

där w_{ijklt} och x_{ijklt} betecknar duration och skadekostnad för företag l år t, och där (i, j, k) betecknar klasser för premieargumenten som analyserats i GLM-analysen. Föreslå en förbättrad prissättningsmodell som även betraktar eventuella skillnader mellan företag. Beskriv hur förväntad skadekostnad för en 1-årig försäkring skattas i den förfinade modellen utifrån tillgänglig data och anpassad multiplikativ modell. (10 poäng)

Uppgift 5

Ett aktuarie vill anpassa en kurva med målet att beskriva hur förväntad skadekostnad (tSEK) per skada beror av bilens motoreffekt (hästkrafter). Som utgångspunkt finns historisk data i form av tripletter

i	1	2	$\ldots$	n
(e_i, w_i, x_i)	$(150, 46, 620)$	$(210, 12, 241)$	$\ldots$	$(170, 62, 744)$

där (e, w, x) betyder att historisk data innehåller w rapporterade skadehändelser med total skadekostnad x för bilar som har e hästkrafter. Kurvanpassningen ska ges som lösningen $\widehat{f_\lambda}$ till

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y_i - f(e_i))^2 + \lambda \int_{100}^{350} (f''(t))^2 dt,$$

där funktionsrummet $\mathcal{F}$ har basfunktionerna $e \mapsto 1$, $e \mapsto e$, $e \mapsto e^2$, $e \mapsto e^3$. Bortse från effekter av andra kovariater än motoreffekt.

(a) Aktuarien funderar på att, i uttrycket ovan, ersätta $w_i(y_i - f(e_i))^2$ med ett deviansuttryck $w_i d(y_i, f(e_i))$. Är det rimligt och i så fall hur bör det väljas? Vad är y_i? (4 poäng)

(b) Aktuarien funderar på att, i uttrycket ovan, ersätta $\int_{100}^{350} (f''(t))^2 dt$ med $\int_{100}^{350} (f''(t))^m dt$ där m är ett icke-negativt heltal. Är det rimligt och i så fall hur bör det väljas? (3 poäng)

(c) Aktuarien funderar på att, i uttrycket ovan, ersätta $\mathcal{F}$ med en annan functionsklass $\mathcal{G}$. Är det rimligt och i så fall hur bör det väljas? (3 poäng)

Klass	Ålder	Zon	Duration	Antal skador	Skadekostnad
1	1	1	62.9	17	310 352
1	1	2	112.9	7	95 424
1	2	1	352.1	52	428 064
1	2	2	840.1	69	511 842
2	1	1	191.6	43	333 422
2	1	2	237.3	34	235 722
2	2	1	844.8	94	444 432
2	2	2	1296.0	99	420 948

Table 1: Skadedata, duration i år, skadekostnad i kronor

Uppgift 1

$$f_{X_{ij}}(x) = P(X_{ij} = x) = \exp(-w_{ij}\gamma_0\gamma_{1i}\gamma_{2j})\frac{(w_{ij}\gamma_0\gamma_{1i}\gamma_{2j})^x}{x!}$$

vilket ger, för $w_{ij}y$ heltal,

$$\begin{aligned} f_{Y_{ij}}(y) = P(X_{ij} = w_{ij}y) &= \exp(-w_{ij}\gamma_0\gamma_{1i}\gamma_{2j})\frac{(w_{ij}\gamma_0\gamma_{1i}\gamma_{2j})^{w_{ij}y}}{(w_{ij}y)!} \\ &= \exp\Big(w_{ij}\big(y\log(\gamma_0\gamma_{1i}\gamma_{2j}) - \gamma_0\gamma_{1i}\gamma_{2j}\big) + c_{ij}\Big) \end{aligned}$$

där

$$c_{ij} = w_{ij}y\log w_{ij} - \log((w_{ij}y)!)$$

inte beror på $\gamma_0, \gamma_{1i}, \gamma_{2j}$. Alltså,

$$\begin{aligned} l((\gamma_0, \gamma_{1i}, \gamma_{2j})_{i,j}; (y_{ij})_{i,j}) &= \log\prod_{i=1}^{m_1}\prod_{j=1}^{m_2} f_{Y_{ij}}(y_{ij}) \\ &= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\log f_{Y_{ij}}(y_{ij}) \\ &= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}\big(y_{ij}\log(\gamma_0\gamma_{1i}\gamma_{2j}) - \gamma_0\gamma_{1i}\gamma_{2j}\big) \\ &\quad + \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} c_{ij} \end{aligned}$$

ML-skattningarna löser

$$\frac{\partial}{\partial\gamma_0}l(\dots) = 0, \quad \frac{\partial}{\partial\gamma_{1i}}l(\dots) = 0,\ i = 1,\dots,m_1 \quad \frac{\partial}{\partial\gamma_{2j}}l(\dots) = 0,\ j = 1,\dots,m_2$$

där vi får

$$\frac{\partial}{\partial\gamma_0}l(\dots) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}\left(y_{ij}\frac{1}{\gamma_0} - \gamma_{1i}\gamma_{2j}\right) = \frac{1}{\gamma_0}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}y_{ij} - \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}\gamma_{1i}\gamma_{2j}$$

och

$$\frac{\partial}{\partial\gamma_{1i}}l(\dots) = \sum_{j=1}^{m_2} w_{ij}\left(y_{ij}\frac{1}{\gamma_{1i}} - \gamma_0\gamma_{2j}\right) = \frac{1}{\gamma_{1i}}\sum_{j=1}^{m_2} w_{ij}y_{ij} - \gamma_0\sum_{j=1}^{m_2} w_{ij}\gamma_{2j}$$

och

$$\frac{\partial}{\partial\gamma_{2j}}l(\dots) = \sum_{i=1}^{m_1} w_{ij}\left(y_{ij}\frac{1}{\gamma_{2j}} - \gamma_0\gamma_{1i}\right) = \frac{1}{\gamma_{2j}}\sum_{i=1}^{m_1} w_{ij}y_{ij} - \gamma_0\sum_{i=1}^{m_1} w_{ij}\gamma_{1i}$$

Uppgift 2

För Tweedie(1) ges ML-ekvationerna för anpassning till historisk data av

$$0 = \frac{1}{\phi}\sum_{k=1}^{n} d_k \frac{y_k - \mu_k}{\mu_k^{1-1}} x_{kj}, \quad \mu_k = \exp\Big\{\sum_j x_{kj}\beta_j\Big\}, \quad j = 1, \dots, r,$$

där d_k är historiska durationer och $n_k = d_k y_k$ är historiska skadeantal. ML-skattningarna $\widehat{\mu}_k$ uppfyller därför (tag $j = 1$ och notera att $x_{k1} = 1$ för alla k)

$$\sum_{k=1}^{n} n_k = \sum_{k=1}^{n} d_k \widehat{\mu}_k.$$

Prediktionen för 2021 ges därför av

$$\sum_{k=1}^{n} 2 \cdot d_k \widehat{\mu}_k = 2 \cdot \sum_{k=1}^{n} n_k$$

och notera att denna prediktion inte beror på hur indelning i premieargument och klasser är gjord.

Uppgift 3

För Tweedie(p) ges ML-ekvationerna för anpassning till historisk data, givet offsets, av

$$0 = \frac{1}{\phi}\sum_{k=1}^{n} w_k \frac{y_k - u_k\mu_k}{(u_k\mu_k)^{p-1}} x_{kj}, \quad \mu_k = \exp\Big\{\sum_j x_{kj}\beta_j\Big\}, \quad j = 1, \dots, r,$$

där w_k är historiska durationer och y_k är historiska riskpremier. Vi kan skriva om ekvationen som

$$0 = \frac{1}{\phi}\sum_{k=1}^{n} w_k u_k^{2-p} \frac{y_k/u_k - \mu_k}{\mu_k^{p-1}} x_{kj}, \quad \mu_k = \exp\Big\{\sum_j x_{kj}\beta_j\Big\}, \quad j = 1, \dots, r.$$

Notera att här är $p = 3/2$, $n = 8$, $r = 1$ och $X = (1, \dots, 1)^{\mathrm{T}}$. Eftersom μ_k inte beror på k sätter vi $\mu := \mu_k$. Vidare har vi $u_1 = 1, u_2 = 2, u_3 = 1, \dots, u_8 = 2$. Alltså blir ML-ekvationen

$$0 = \frac{1}{\phi}\sum_{k=1}^{8} w_k u_k^{1/2} \frac{y_k/u_k - \mu}{\mu^{1/2}}$$

Här kan vi multiplicera med $\mu^{1/2}$ och därefter explicit lösa ut μ som ML-skattaren $\widehat{\mu}$:

$$\widehat{\mu} = \frac{\sum_{k=1}^{8} w_k u_k^{-1/2} y_k}{\sum_{k=1}^{8} w_k u_k^{1/2}} = \frac{2410008}{4967.559} = 485.1493$$

Eftersom durationen 2021 antas vara identisk med den historiska blir prediktorn $\widehat{S}_z$ för skadekostnaden för Zon $z \in \{1, 2\}$

$$\widehat{S}_1 = \widehat{\mu}(w_1 + w_3 + w_5 + w_7) = 704145.7,$$
$$\widehat{S}_1 = 2\widehat{\mu}(w_2 + w_4 + w_6 + w_8) = 2412453$$

Man kan notera att de givna relationstalen för Zon inte har stöd i historisk data.

Uppgift 4

På listform har vi en modell

$$\mu_i = \mu\gamma_i, \quad \mu_i = \exp\left\{\sum_{j=1}^{4} x_{ij}\beta_j\right\}, \quad \gamma_i = \exp\left\{\sum_{j=2}^{4} x_{ij}\beta_j\right\},$$

där designmatrisen är

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

och, med $(i, j, k) = (1, 1, 1)$ som bascell, $\beta_1 = \log\gamma_0$, $\beta_2 = \log\gamma_{12}$, $\beta_3 = \log\gamma_{22}$, $\beta_4 = \log\gamma_{32}$. Vi kan kombinera vanliga premieargument med flernivå-faktorer (MLF) genom kredibilitetsteori. Låt modellen vara

$$\mathrm{E}[Y_{ijt} \mid U_j] = \mu_i U_j, \quad i = 1, \ldots, 8, \quad j = 1, \ldots, 37.$$

Notera att $i = 1, \ldots, 8$ svarar mot ordningen

$$(1, 1, 1), (1, 1, 2), (1, 2, 1), (1, 2, 2), (2, 1, 1), (2, 1, 2), (2, 2, 1), (2, 2, 2)$$

Tex svarar observationen $y_{1,15,t}$ mot $x_{1,1,1,15,t}/w_{1,1,1,15,t}$, $y_{3,8,t}$ svarar mot $x_{1,2,1,8,t}/w_{1,2,1,8,t}$, etc. Sätt

$$\widetilde{y}_{ijt} = y_{ijt}/\gamma_i, \quad \widetilde{w}_{ijt} = w_{ijt}\gamma_i^{2-p}$$

och

$$\widehat{U}_j = \widetilde{z}_j \frac{\overline{\widetilde{y}}_{\cdot j\cdot}}{\mu} + (1 - \widetilde{z}_j), \quad \widetilde{z}_j = \frac{\widetilde{w}_{\cdot j\cdot}}{\widetilde{w}_{\cdot j\cdot} + \sigma^2/\tau^2}.$$

Slutligen fås en förbättrad tariff genom att ersätta μ_i med $\mu_i\widehat{U}_j$ där i svarar mot lämplig kombination av vanliga premieargument och j svarar mot företaget.

Uppgift 5

(a) Om nyckeltalsdatan (y-värden) avviker från normalfördelning är det lämpligt att istället välja en deviansfunktion som bättre svarar mot data. Detaljer finns i kursboken.

(b) Udda positiva heltal är mycket olämpliga eftersom sådana val inte straffar fladdriga (ej släta) kurvor - vilket är syftet med strafftermen. Jämna positiva heltal är möjliga val, ju större desto hårdare straffas fladdriga kurvor. Dock är det rimligare att välja $m = 2$ och istället variera λ för att få en lämplig balans mellan anpassning och släthet.

(c) Genom att införa knutpunkter och välja en funktionsklass enligt kubiska splines med dessa knutpunkter fås en mer flexibel funktionsklass som bättre kan hantera ett större datamaterial. Valet av kubiska splines kan motiveras med goda teoretiska egenskaper, se kursboken.