4a Prova de FECD - 2019

Renato Assunção

Julho 2019

1. Numa seguradora, foi feita uma análise de 12000 apólices de seguros de automóveis emitidas para proprietários individuais. Como parte da análise, em cada apólice foram considerados a idade x (em anos) do motorista (variando de 18 a 60 anos) e o resultado Y em termos de sucesso ($Y = 1$) do motorista em conduzir o veículo por um ano sem sinistros de nenhum tipo. Caso contrário, registra-se que houve um fracasso ($Y = 0$).

   O interesse é entender como a idade está associada com a probabilidade de sucesso. Decide-se usar um modelo logístico para modelar estes dados onde $p(x) = \mathbb{P}(Y = 1|x) = \frac{1}{1+e^{-(w_0+w_1 x)}}$.

   - Esboce num gráfico qual é a relação esperada pelo modelo entre a idade x e a probabilidade $p(x)$ de sucesso.
   - Escreva a log-verossimilhança para este problema.
   - Obtenha o vetor gradiente necessário para obter o MLE.
   - Suponha que o interesse do pesquisador é estimar a idade x na qual a probabilidade dos segurados terem sucesso é maior ou igual a 0.90. Escreva essa idade como função dos parâmetros do modelo acima.

   Solução: Espera-se uma curva em forma de S com $p(x)$ decrescendo com x pois o risco de acidente diminui com a idade, fruto de maior experiência no volante e menor impulsividade. Além disso, podemos esperar nas duas idades extremas probabilidades não saturadas, longe de seus valores extremos 0 e 1. Assim, antecipamos que $p(18)$ esteja substancialmente abaixo de 1 e que $p(50)$ esteja substancialmente acima de zero.Um esboço possível da função $p(x)$ está na Figura ??.

   A log-verossmilhança do vetor de parâmetros (w_0, w_1) é:

$$\begin{aligned}
\ell(w_0, w_1) &= \log\left(\prod_{i=1} 12000 p(x_i)^{y_i}(1-p(x_i))^{1-y_i}\right) && (1)\\
&= \sum_{i=1} 12000\left(y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i))\right) && (2)\\
&= \sum_{i:y_i=1} \log(p(x_i)) + \sum_{i:y_i=0} \log(1-p(x_i)) && (3)\\
&= \sum_{i=1} 12000\left(y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i))\right) && (4)\\
&= w_0 \sum_{i=1}^{12000} y_i + w_1 \sum_{i=1}^{12000} x_i y_i - \sum_i \log(1+e^{w_0+w_1 x_i}) && (5)
\end{aligned}$$

   O vetor gradiente é o vetor das derivadas parciais com respeito aos parâmetros (w_0, w_1). Contas rotineiras levam ao resultado desejado:

$$\nabla \ell(w_0, w_1) = \begin{bmatrix} \frac{\partial \ell}{\partial w_0} \\ \frac{\partial \ell}{\partial w_1} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{12000} (y_i - p(x_i)) \\ \sum_{i=1}^{n} x_i (y_i - p(x_i)) \end{bmatrix}$$

Seja x^* a idade tal que $p(x^*) = 0.90$. Então

$$\frac{1}{1+e^{-(w_0+w_1x^*)}} = 0.90 \rightarrow -(w_0 + w_1x^*) = \log(0.1/0.9) \rightarrow x^* = \frac{1}{w_1}(log(9) - w_0)$$

Tendo estimativas de w_0 e w_1, encontramos uma estimativa da idade limite x^*

---

2. Uma operadora de planos de saúde sabe que o custo médio das internações varia muito de acordo com a idade do cliente. Aqueles com mais de 70 anos de idade acarretam a maior parte dos cursos embora eles tenham uma participação pequena no portfolio de clientes.

   A operadora decidiu investigar um pouco mais a incidência de internações entre seus clientes idosos. Para isto, escolheu uma amostra de clientes com idade acima de 70 anos e obteve o número de internações que cada um teve nos últimos dois anos. Decidiu-se adotar um modelo de Poisson para as contagens do número de internações.

   Nem todos os selecionados foram clientes por todo o período de dois anos. Aqueles que estão na operadora há pouco tempo devem apresentar, em média, menos internações do que aqueles que estão na operadora durante os últimos dois anos. Por isto, a média da Poisson deveria refletir o tempo de permanência no plano de cada cliente. Dessa forma chegou-se ao seguinte modelo estatístico.

   Sejam $Y_1, \ldots, Y_n$ a amostra de clientes. Suponha que essas sejam variáveis aleatórias independentes e que $Y_i \sim \text{Poisson}(\lambda t_i)$ onde t_i é o tempo de permanência do i-ésimo cliente na empresa (em meses). Os valores de t_i de cada cliente são conhecidos. O parâmetro $\lambda > 0$ é desconhecido e representa o número esperado de internações *por mês*. Sendo Poisson, sabe-se que $\mathbb{E}(Y_i) = \lambda t_i$.

   O interesse é estimar λ a partir dos dados que podem ser representados como na tabela abaixo:

i	t_i	y_i
1	24	4
2	12	1
3	3	0
4	24	1
...	...	...

   - Pensou-se inicialmente em estimar λ simplesmente tomando o número médio de internações e dividir pelo tempo de observação de 24 meses. Isto é, $T_1 = \overline{Y}/24$. Mostre que este estimador é viciado para estimar λ a menos que $\sum_i t_i = 24n$. Por exemplo , se todos os clientes tiverem $t_i = 24$ esta condição seria válida.
   - Tentando corrigir o vício do estimador T_1, pensou-se então em adotar

     $$T_2 = \frac{\overline{Y}}{\overline{t}} = \frac{Y_1 + \ldots + Y_n}{t_1 + \ldots + t_n}$$

     Mostre que T_2 é não-viciado para estimar λ e encontre seu risco quadrático de estimação.
   - Mais tarde, outro analista resolveu considerar o estimador

     $$T_3 = \frac{1}{n}\left(\frac{Y_1}{t_1} + \ldots + \frac{Y_n}{t_n}\right)$$

     Mostre que T_3 é não-viciado para estimar λ e encontre seu risco quadrático de estimação.
   - É possível dizer que T_2 é sempre melhor ou igual a T_3 considerando-se os riscos quadráticos dos dois. Prove isto usando a desigualdade entre a média aritmética e a média harmônica que diz que

     $$\frac{x_1 + \cdots + x_n}{n} \geq \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

     para quaisquer números reais positivos $x_1, \ldots, x_n$.

OBS: Um estimador T é não-viciado para estimar um parâmetro θ se $\mathbb{E}(T) = \theta$ para todo valor de θ.

Solução: Considerando o estimador T_1 inicialmente:

$$\begin{aligned}\mathbb{E}(T_1) &= \mathbb{E}\left(\frac{\bar{Y}}{24}\right) = \frac{1}{24}\mathbb{E}\left(Y_1 + \ldots + Y_n\right) = \frac{1}{24n}\left(\mathbb{E}(Y_1) + \ldots + \mathbb{E}(Y_n)\right) \\ &= \frac{1}{24n}\left(\lambda t_1 + \ldots + \lambda t_n\right) = \frac{\lambda}{24n}\left(t_1 + \ldots + t_n\right) = \frac{\lambda \bar{t}}{24}\,,\end{aligned}$$

que é igual a λ se, e só se, $\bar{t} = 24$.

Considerando o estimador T_2:

$$\mathbb{E}(T_2) = \mathbb{E}\left(\frac{\bar{Y}}{\bar{t}}\right) = \frac{1}{\bar{t}n}\mathbb{E}\left(Y_1 + \ldots + Y_n\right) = \frac{\lambda}{\bar{t}n}\left(t_1 + \ldots + t_n\right) = \lambda\,.$$

Portanto, T_2 é não-viciado para estimar λ. O seu risco quadrático de estimação é:

$$\begin{aligned}MSE(T_2, \lambda) &= \mathbb{E}\left[(T_2 - \lambda)^2\right] = \mathbb{V}(T_2) + \text{bias}^2(T_2, \lambda) = \mathbb{V}(T_2) + 0 = \mathbb{V}\left(\frac{\overline{Y}}{\bar{t}}\right) = \frac{\mathbb{V}(\overline{Y})}{\bar{t}^2} \\ &= \frac{1}{\bar{t}^2}\,\frac{1}{n^2}\,\mathbb{V}\left(Y_1 + \ldots + Y_n\right) = \frac{1}{\bar{t}^2}\,\frac{1}{n^2}\,\left[\mathbb{V}\left(Y_1\right) + \ldots + \mathbb{V}\left(Y_n\right)\right] \\ &= \frac{1}{\bar{t}^2}\,\frac{1}{n^2}\,\left[\lambda t_1 + \ldots + \lambda t_n\right] = \frac{\lambda}{n}\,\frac{1}{\bar{t}}\end{aligned}$$

O terceiro estimador, T_3, tem valor esperado:

$$\begin{aligned}\mathbb{E}(T_3) &= \frac{1}{n}\mathbb{E}\left(\frac{Y_1}{t_1} + \ldots + \frac{Y_n}{t_n}\right) = \frac{1}{n}\mathbb{E}\left(\frac{\mathbb{E}(Y_1)}{t_1} + \ldots + \frac{\mathbb{E}(Y_n)}{t_n}\right) = \frac{1}{n}\mathbb{E}\left(\frac{\lambda t_1}{t_1} + \ldots + \frac{\lambda t_n}{t_n}\right) \\ &= \frac{\lambda}{n}\left(\frac{t_1}{t_1} + \ldots + \frac{t_n}{t_n}\right) = \lambda\,,\end{aligned}$$

e portanto, também não-viciado para estimar λ. O seu risco quadrático de estimação é:

$$\begin{aligned}MSE(T_3, \lambda) &= \mathbb{E}\left[(T_3 - \lambda)^2\right] = \mathbb{V}(T_3) = \frac{1}{n^2}\left[\mathbb{V}\left(\frac{Y_1}{t_1}\right) + \ldots + \mathbb{V}\left(\frac{Y_n}{t_n}\right)\right] \\ &= \frac{1}{n^2}\left[\frac{\mathbb{V}(Y_1)}{t_1^2} + \ldots + \frac{\mathbb{V}(Y_n)}{t_n^2}\right] = \frac{1}{n^2}\left[\frac{\lambda t_1}{t_1^2} + \ldots + \frac{\lambda t_n}{t_n^2}\right] \\ &= \frac{\lambda}{n^2}\left[\frac{1}{t_1} + \ldots + \frac{1}{t_n}\right] = \frac{\lambda}{n}H\end{aligned}$$

onde H é a média harmônica dos tempos assegurados dos clientes:

$$H = \frac{1}{n}\left(\frac{1}{t_1} + \ldots \frac{1}{t_n}\right)$$

A comparação entre os riscos de T_2 e T_3 depende da desigualdade entre a média aritmétmica e a média harmônica dos tempos t_i. Usando a desigualdade mencionada no enunciado, temos

$$MSE(T_2, \lambda) = \frac{\lambda}{n}\,\frac{1}{\bar{t}} \leq \frac{\lambda}{n}\,H = MSE(T_3, \lambda)\,.$$

Em resumo, queremos estimar λ, o número esperado de internações mensais usando as contagens de episódios de internações de clientes expostos a diferentes tempos t_i sob o seguro. O parâmetro λ é a taxa mensal de internações por indivíduo. Temos dois estimadores não-viciados. O primeiro

deles, T_2, soma as internações de todos os clientes e divide pelo tempo total exposto ao risco de todos eles, obtendo uma estimativa intuitivamente simples. O outro, T_3, usa a taxa mensal individual ao calcular Y_n/t_n e em seguida tira sua média aritmética simples, também uma estimativa intuitivamente simples. A conclusão é que é preferível usar T_2.

---

3. Seja X uma variável aleatória discreta com distribuição dada por

$$P(X = x; \theta) = \frac{-\theta^x}{x \ \log(1-\theta)}$$

para $x = 1, 2, \ldots$ onde θ é um parâmetro desconhecido no intervalo $(0, 1)$. Suponha que $x_1 = 1$, $x_2 = 2$, $x_3 = 3$ e $x_4 = 2$ são os valores observados de uma amostra aleatória dessa distribuição. Começando com o valor inicial $\theta^{(0)} = 0.6$, encontre o valor θ_1 do processo iterativo de Newton-Raphson para obter o EMV de θ.

Solução: Como $\theta \in (0, 1)$, temos $-\log(1 - theta) > 0$. A log-verossimilhança de θ baseada em n dados $x_1, x_2, \ldots, x_n$ é igual a

$$\begin{aligned} \ell(\theta) &= \log\left(\prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i(-\log(1-\theta))}\right) = \log\left(\frac{\theta^{\sum x_i}}{(-\log(1-\theta))^n \prod x_i}\right) \\ &= \left(\sum_i x_i\right)\log(\theta) - \sum_i \log(x_i) - n\log\left(-\log(1-\theta)\right) \end{aligned}$$

A derivada da log-verossimilhança é a função escore:

$$\ell'(\theta) = \frac{\partial \ell}{\partial \theta} = \frac{\sum x_i}{\theta} + \frac{n}{(1-\theta)\log(1-\theta)}$$

A Figura ?? mostra a função log-verossimilhança $\ell(\theta)$ no lado esquerdo e a derivada (ou função escore no lado direito.

A derivada parcial de segunda ordem é:

$$\ell''(\theta) = \frac{\partial^2 \ell}{\partial \theta^2} = -\left[\frac{\sum x_i}{\theta^2} + \frac{n(1+\log(1-\theta))}{\left((1-\theta)\log(1-\theta)\right)^2}\right]$$

A equação de itereção de Newton é

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\ell'(\theta^{(t)})}{\ell''(\theta^{(t)})} = \theta^{(t)} + \frac{\frac{\sum x_i}{\theta^{(t)}} + \frac{n}{(1-\theta^{(t)})\log(1-\theta^{(t)})}}{\left[\frac{\sum x_i}{(\theta^{(t)})^2} + \frac{n(1+\log(1-\theta^{(t)}))}{\left((1-\theta^{(t)})\log(1-\theta^{(t)})\right)^2}\right]}$$

Considerando a pequena amostra de $n = 4$ observaçôes com $\sum x_i = 8$ e começando com o valor inicial $\theta^{(0)} = 0.6$, encontramos

$$\begin{aligned} \theta^{(1)} &= \theta^{(0)} - \frac{\ell'(\theta^{(0)})}{\ell''(\theta^{(0)})} = 0.6 - \frac{2.4198}{-24.7148} = 0.6979 \\ \theta^{(2)} &= \theta^{(1)} - \frac{\ell'(\theta^{(1)})}{\ell''(\theta^{(1)})} = 0.6979 - \frac{0.4012}{-10.3977} = 0.7365 \end{aligned}$$