Atividade 1

Guilherme Navarro NUSP: 8943160

22 de junho de 2020

Atividade 1

Algumas vezes, dados de sobrevivência são reduzidos a respostas binárias. Nestes casos, a informação utilizada é se o tempo de falha T é maior ou não do que t_0 fixado. Em geral, utiliza-se um instante t_0 com relevância clínica para aquele problema particular.

(a) Considere um modelo de regressão Weibull (utilizando a parametrização de riscos proporcionais) e defina $p_0(x) = \mathbb{P}(T > t_0|x)$, em que x representa o vetor de covariáveis. Obtenha uma expressão para $p_0(x)$ em termos dos parâmetros do modelo Weibull utilizado. Defina um modelo de regressão adequado para a variável resposta binária, especificando a função de ligação apropriada para este caso.

Dica: Calcule $log(-log(p_0(x)))$.

Resolução

O modelo de riscos proporcionais é dado por:

$$\alpha(t|x) = \alpha_0(t)g(x)$$

Usualmente $g(x) = e^{x'\beta}$

Para o modelo Weibull, temos:

$$\alpha(t|x) = \rho t^{\rho-1}e^{x'\beta}$$

E assim temos também que:

$$f(x|t) = \rho t^{\rho-1}e^{x'\beta}e^{-(e^{x'\beta})t^{\rho}}$$

E como $S(t|x) = \frac{f(x|t)}{\alpha(x|t)}$, logo:

$$S(t|x) = \mathbb{P}(T > t) = \frac{\rho t^{\rho-1}e^{x'\beta}e^{-(e^{x'\beta})t^{\rho}}}{\rho t^{\rho-1}e^{x'\beta}} = e^{-(e^{x'\beta})t^{\rho}}$$

Assim:

$$p_0(x) = \mathbb{P}(T > t_0|x) = S(t_0|x) = e^{-(e^{x'\beta})t_0^{\rho}}$$

No modelo de regressão binária, modelamos a probabilidade p_0 de um evento acontecer, assim definindo a variável resposta:

$$Y = \begin{cases} 1 \; se \; \mathbb{P}(T > t_0|x) \\ 0 \; c.c \end{cases}$$

Logo chegamos a um modelo de regressão logística com função de probabilidade definida:

$$f(y|p_0) = p_0^y(1-p_0)^{1-y}$$

Com $0 < p_0 < 1$

para chegarmos a uma expressão para p_0, pode se calcular $ln(-ln(p_0(x)))$:

$$ln(-ln(p_0(x))) = ln(-ln(e^{-(e^{x'\beta})t_0^\rho})) = ln(e^{x'\beta}t_0^\rho) = x'\beta + \rho ln(t_0)$$

O que implica que a função de ligação log-glog

$$ln(-ln(p_0(x)))$$

.

(b) Repita o item (a) considerando um modelo de regressão log-logístico.

Resolução

Pela parametrização de locação escala o modelo log-logístico, pode ser escrito:

$$ln(T) = x'\gamma + \sigma w$$

Sendo x a matriz com os dados, γ o vetor de parâmentros e σ o parâmetro de escala e com $w \sim$ Logística padrão.

Análogamente ao exercício anterior $p_0(x) = \mathbb{P}(T > t_0|x) = S(t_0|x)$, então basta calcular $S(t|x)$ para encontrar uma expressão para p_0, assim:

$$S(t|x) = \mathbb{P}(T > t|x) = \mathbb{P}(lnT > ln(t|x)) = \mathbb{P}(x'\gamma + \sigma w > ln(t|x)) = \mathbb{P}\Big(w > \frac{ln(t) - x'\gamma}{\sigma}\Big)$$

Como $w \sim$ Logística padrão, então:

$$S(t|x) = \mathbb{P}(T > t|x) = \frac{1}{1 + e^{(ln(t) - x'\gamma)/\sigma}} = \frac{1}{1 + t^{1/\sigma}e^{-x'\gamma/\sigma}}$$

Logo:

$$p_0(x) = \mathbb{P}(T > t_0|x) = S(t_0|x) = \frac{1}{1 + t_0^{1/\sigma}e^{-x'\gamma/\sigma}}$$

No modelo de regressão binária, modelamos a probabilidade p_0 de um evento acontecer, assim definindo a variável resposta:

$$Y = \begin{cases} 1 \ se \ \mathbb{P}(T > t_0|x) \\ 0 \ c.c \end{cases}$$

Logo chegamos a um modelo de regressão logística com função de probabilidade definida:

$$f(y|p_0) = p_0^y(1 - p_0)^{1-y}$$

Com $0 < p_0 < 1$

$$p_0(x) = \frac{1}{1 + t_0^{1/\sigma}e^{-x'\gamma/\sigma}} \Rightarrow p_0(x)(1 + t_0^{1/\sigma}e^{-x'\gamma/\sigma}) = 1 \Rightarrow \frac{1}{p_0(x)} = 1 + t_0^{1/\sigma}e^{-x'\gamma/\sigma}$$
$$\Rightarrow \frac{1}{p_0(x)} - 1 = t_0^{1/\sigma}e^{-x'\gamma/\sigma} \Rightarrow \frac{1 - p_0(x)}{p_0(x)} = t_0^{1/\sigma}e^{-x'\gamma/\sigma}$$

Agora para ficar uma expressão linear, basta aplicar a função $ln(.)$ em $\frac{1-p_0(x)}{p_0(x)}$, resultando em:

$$ln\Big(\frac{1-p_0(x)}{p_0(x)}\Big) = \frac{1}{\sigma}\Big[ln(t_0) - x'\gamma\Big]$$

O que implica que a função de ligação é logito

$$ln\Big(\frac{1-p_0(x)}{p_0(x)}\Big)$$

.

(c) Assuma que os dados estão sujeitos a censura à direita tipo I, com o mesmo tempo de acompanhamento para todas as observações no estudo. Discuta em que situações é possível utilizar um modelo binário e em que situações o modelo binário não é adequado.

Resolução

Como os dados estão sujeitos a censura à direita tipo I, a definição censura à direita do tipo I é que o estudo será terminado após um período pré-estabelecido de tempo. As observações cujo evento de interesse não foram observadas até este tempo são ditas censuradas. Isso quer dizer que se difinirmos um modelo binário com o tempo de falha T sendo maior ou não do que t_0 fixado, logo se escolhermos um t_0 fora do tempo de acompanhamento teremos um problema, pois não saberiamos ao certo se a observação foi uma falha ou censura, outro problema seria se com a escolha de um t_0 tal que a proporção de eventos fique desbalanceada e a regressão logística não se da bem com problemas de desbalancemanto, já um caso onde é seria mais adequado é quando o t_0 é fixado na metade do período de estudo ou próximo, para garantir que todos os indivíduos serão capturados no modelo.