

ALGUNS MODELOS DE REGRESSÃO EM ANÁLISE DE SOBREVIVÊNCIA.

Sílvia Akemi Fujikura, Mário Hissamitsu Tarumoto. – Probabilidade e Estatística – Estatística – Departamento de Matemática, Estatística e Computação – Faculdade de Ciências e Tecnologia – Campus de Presidente Prudente.

Nos últimos 20 anos, a área de pesquisa em análise de sobrevivência foi uma das que mais cresceu, pois a sua aplicabilidade nas diversas áreas é notável. A principal vantagem desta técnica é a possibilidade de incorporação de informação parcial do evento de interesse, conhecido como censura. A existência desta característica dos dados se explica pelo fato de que os indivíduos ou objetos são acompanhados ao longo do tempo, podendo permanecer sem experimentar o evento por muito tempo, desta forma, muitas vezes por razões econômicas é necessário interromper o estudo, gerando esta informação parcial.

Por exemplo, na área industrial, é freqüente o estudo de tempos de vida de lâmpadas elétricas, de componentes elétricos ou de vários tipos de equipamentos; ou ainda na área médica, existem estudos com o objetivo de avaliar tempos de vida de pacientes submetidos a transplante cardíaco, tempo de remissão de leucemia aguda, etc. As unidades são colocadas em observação num tempo inicial do calendário t_0 , observadas algumas condições iniciais e experimentais das unidades, de acordo com os objetivos do estudo; por exemplo, as lâmpadas devem ser aquelas nunca usadas anteriormente, de um determinado tipo ou marca.

Em experimentos clínicos, aleatorizados ou não, as condições das unidades amostrais no tempo t_0 devem ser observadas e registradas, em geral procura-se aquelas que possam influenciar no tempo de ocorrência do evento, por exemplo, idade, condição inicial da doença, entre outras. Estas condições iniciais, que são as variáveis independentes são conhecidas em geral como covariáveis ou variáveis concomitantes. Estas podem ser incorporadas ao modelo através de modelos de regressão, portanto, na área da Análise de Sobrevivência, os modelos de regressão paramétricos são utilizados, uma vez que os tempos de vida podem estar relacionados a outros fatores. Este fato implica em especificar um modelo para a distribuição de T , dado x , onde T é o tempo de vida e x é o vetor de variáveis regressoras, ou variáveis concomitantes.

O objetivo deste trabalho é o de realizar um estudo de revisão destes modelos e principalmente estudar a qualidade do ajuste destes, pois com a incorporação das censuras o ajuste pode não ser muito bom na maioria das situações. Dois modelos comumente utilizados nesta área são os modelos de regressão exponencial e Weibull, que são extensões da distribuição exponencial e Weibull, respectivamente.

O modelo exponencial é um caso especial do modelo Weibull com parâmetro $\delta = 1$ e sua função densidade de probabilidade é dada por:

$$f(t|x) = \theta_x^{-1} \exp\left(\frac{-t}{\theta_x}\right), \quad t > 0.$$

A forma mais utilizada para θ_x é:

$$\theta_x = \exp(x\beta),$$

sendo $x = (x_1, \dots, x_p)$ o vetor das variáveis regressoras e $\beta = (\beta_1, \dots, \beta_p)'$ o vetor dos parâmetros. Este modelo pode ser ajustado em casos que a função risco constante possa depender de variáveis concomitantes.

O modelo Weibull mais utilizado tem sua função densidade de probabilidade dada por:

$$f(t|x) = \frac{\delta}{\alpha(x)} \left(\frac{t}{\alpha(x)} \right)^{\delta-1} \exp \left[- \left(\frac{t}{\alpha(x)} \right)^{\delta} \right], \quad t \geq 0,$$

sendo que apenas o parâmetro de escala α depende de x .

A forma mais utilizada para $\alpha(x)$ é:

$$\alpha(x) = \exp(x\beta),$$

e assim como no caso do modelo exponencial, $x = (x_1, \dots, x_p)$ e $\beta = (\beta_1, \dots, \beta_p)'$.

Freqüentemente, é preferível trabalhar com o logaritmo do tempo de vida. Isto é possível através de uma transformação de variáveis aleatórias do tipo $Y = \log T$, que resulta num modelo denominado valor extremo, cuja função densidade de probabilidade é dada por:

$$f(y|x) = \frac{1}{\sigma} \exp \left[\frac{y - \mu(x)}{\sigma} - \exp \left(\frac{y - \mu(x)}{\sigma} \right) \right], \quad -\infty < y < \infty,$$

tal que $\sigma = 1/\delta$ e $\mu(x) = \log \alpha(x)$.

Nesta fase do trabalho, serão apresentados estudos de aplicação das principais técnicas de avaliação em modelos de regressão linear simples, que são modelos do tipo $Y = \alpha + \beta X + \varepsilon$ que descrevem o relacionamento entre uma variável resposta Y e uma variável preditora X . Uma das técnicas mais utilizadas para verificar a adequação do modelo é a Análise de Variância (ANOVA), que pode ser resumida na Tabela 1.

Tabela 1 - Tabela de Análise de Variância (ANOVA).				
Fonte de variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F ₀
Regressão	1	$SQ_{Reg} = \sum_{i=1}^k n_i (\hat{y}_i - \bar{y})^2$	$QM_{Reg} = SQ_{Reg}/1$	QM_{Reg}/QM_{Res}
Resíduo	n-2	$SQ_{Res} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$	$QM_{Res} = SQ_{Res}/(n-2)$	
Falta de Ajuste	k-2	$SQ_{Fal} = \sum_{i=1}^k n_i (\bar{y}_i - \hat{y}_i)^2$	$QM_{Fal} = SQ_{Fal}/(k-2)$	QM_{Fal}/QM_{EP}
Dentro (Erro Puro)	n-k	$SQ_{EP} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$QM_{EP} = SQ_{EP}/(n-k)$	
Total	n-1	$SQ_{Tot} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$		

Fonte: Bussab (1986).

Além de todas essas informações, é possível calcular o coeficiente de determinação:

$$R^2 = \frac{SQ\text{ Reg}}{SQ\text{ Tot}},$$

que fornece a porcentagem da variação total de Y explicada pela variável X . Nesta tabela, também estão apresentados o erro puro e a falta de ajuste, uma vez que a verificação da qualidade do ajuste é tão importante quanto ajustar um modelo. Entretanto, para a aplicação do teste da falta de ajuste é necessária a existência de medidas repetidas de Y para um ou mais valores de X . A estatística do teste para a falta de ajuste é:

$$F_0 = \frac{QM\text{ Fal}}{QMEP}.$$

Se a regressão linear em X for verdadeira, a estatística F_0 seguirá distribuição F com $k-2$ e $n-k$ graus de liberdade. Desta forma, se o p -valor associado a F_0 for suficientemente pequeno, chega-se a conclusão que a função de regressão é não-linear em X .

Um exemplo de aplicação do modelo de regressão linear simples e da construção da Tabela de ANOVA com análise da falta de ajuste pode ser encontrado em Bussab (1986). Trata-se do caso de um psicólogo que está investigando a relação entre a acuidade visual de 20 indivíduos e suas respectivas idades.

Tabela 2 – Idade e acuidade visual dos 20 indivíduos.

Idade	Acuidade Visual			
20	90	100	80	90
25	100	90	80	90
30	70	90	90	80
35	90	80	70	90
40	90	90	60	80

Fonte: Bussab (1986).

Utilizando o *software* estatístico SAS, é possível construir um programa que faça o ajuste do modelo de regressão e o teste da falta de ajuste

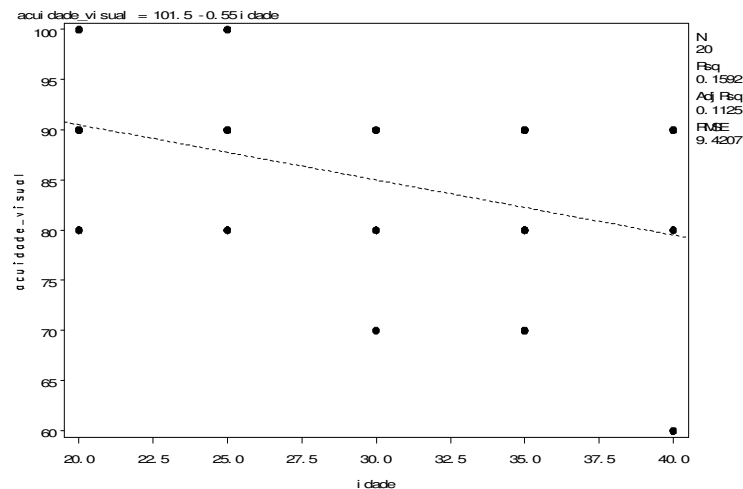


Figura 1 – Ajuste do modelo de regressão linear simples.

Graficamente, o modelo de regressão linear simples ajusta-se bem aos dados. Entretanto, no SAS, é preciso utilizar dois procedimentos: proc reg e proc glm, e combinar seus resultados para encontrar a estatística F_0 . Para estes dados, no procedimento de regressão é obtido o erro total, cuja soma de quadrados é dada por: $SQ_{\text{erro total}}=1597,50$, com 18 graus de liberdade. Pelo procedimento de modelo de regressão linear generalizado é obtido a soma de quadrados do erro puro, cuja soma de quadrados é dada por $SQ_{\text{erro puro}}=1550$, com 15 graus de liberdade. Desta forma, a soma de quadrados da falta de ajuste é dado por: $SQ_{\text{falta de ajuste}}=SQ_{\text{erro total}} - SQ_{\text{erro puro}}=47,5$ com 3 graus de liberdade, indicando que o p_valor calculado para esta estatística é dada por $p=0,926$. Assim para este ajuste pode concluir que não há evidências de que haja falta de ajuste do modelo.

É possível fazer o teste da falta de ajuste utilizando outros pacotes estatísticos, como o R e o Minitab. O *software* R apresenta os resultados da mesma forma que o SAS. Já o Minitab fornece a Tabela de ANOVA, incluindo a análise da falta de ajuste. Para estes dados, utilizando o Minitab, obteve-se as mesmas conclusões.

Referências Bibliográficas

BUSSAB, W.O. *Análise de variância e de regressão*. São Paulo: Atual, 1986. 147 p.

CHATTERJEE, S.; HADI, A.S.; PRICE, B. *Regression analysis by example*. 3. ed. New York: John Wiley & Sons, 2000. 359 p.

LAWLESS, J.F. *Statistical models and methods for lifetime data*. New York: John Wiley & Sons, 1982. 580 p.

Bolsa: FAPESP