UTILIZAÇÃO DO MÉTODO *BOOTSTRAP* NA ESCOLHA ENTRE OS MODELOS DE COX E LOGÍSTICO PARA DADOS DE SOBREVIVÊNCIA COM CENSURA INTERVALAR

JEANETE ALVES MOREIRA

Bacharel em Estatística

Orientador: Prof. Dr. **JOSÉ EDUARDO CORRENTE**

Tese apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para a obtenção do título de Doutor em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

PIRACICABA
Estado de São Paulo - Brasil
Outubro - 2001

ERRATA

PÁGINA	LINHA	ONDE SE LÊ	LEIA-SE
xii	13	deviances	*deviances*
6	13	instantânea	instantâneo
14	23	it bootstrap	*bootstrap*
17	1	F	F
18	6	a a	a
19	24	desenvolveu	desenvolveram
22	3	considerando-se	considere-se
22	14	sobrevivência	sobrevivência,
24	6	$(1-\gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)})^{\Delta_{li}}(\gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)})^{1-\Delta_{li}}$.	$(1-\gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)})^{\Delta_{li}}(\gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)})^{1-\Delta_{li}}$.
29	2	$r_{Pi}=\frac{y_i-\hat{\mu}_i}{[V(\mu_i)]^{1/2}}, \quad i=1,2,\ldots,n,$	$r_{Pi}=\frac{y_i-\hat{\mu}_i}{[V(\hat{\mu}_i)]^{1/2}}, \quad i=1,2,\ldots,n,$
30	1	$\hat{\mu}$	$\hat{y}$
30	4	$y^*_{bootPi}=\hat{\mu}+V(\hat{\mu}_i)\varepsilon^*_{Pi}$	$y^*_{bootPi}=\hat{y}+V(\hat{\mu}_i)\varepsilon^*_{Pi}$
30	8	$y^*_{bootPPi}=\hat{\mu}+c_i\hat{k}V(\hat{\mu}_i)(1-h_i)]^{1/2}\varepsilon^*_{PPi}$	$y^*_{bootPPi}=\hat{y}+c_i\hat{k}V(\hat{\mu}_i)(1-h_i)]^{1/2}\varepsilon^*_{PPi}$
33	5	construiu-se	construiram-se
57	11	utilizar o esta	utilizar esta

ERRATA DAS TABELAS

- página 34 Tabela 4 - onde se lê Nula, leia-se modelo nula
- página 35 Tabela 5 - onde se lê α_2^*, leia-se α_2
- página 35 Tabela 5 - onde se lê α_3^*, leia-se α_3
- página 35 Tabela 5 - onde se lê α_4^*, leia-se α_4

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Moreira, Jeanete Alves
Utilização do método Bootstrap na escolha entre os modelos de cox e logístico para dados de sobrevivência com censura intervalar / Jeanete Alves Moreira. - - Piracicaba, 2001.
65 p.

Tese (doutorado) - Escola Superior de Agricultura Luiz de Queiroz, 2001.
Bibliografia.

1. Estatística agrícola 2. Experimento agrícola 3. Decisão estatística 4. Modelo matemático I. Título

CDD 630.2195

DEDICATÓRIA

A minha força maior,

Deus

És a minha fortaleza, meu refúgio e meu grande guia. A ti entreguei a minha vida e hoje não poderia deixar de te dedicar o meu trabalho.

Ao meu pai,

João Moreira,

pelo grande estímulo, dedicação e amor e

a minha mãezinha **Maria da Salete** (in memoriam),

os anos passaram... e, apesar de não tê-la mais por perto, sinto um enorme orgulho de ser sua filha e sempre levarei comigo seu exemplo de coragem e alegria de viver.

Aos meus filhos

Matheus e **João Augusto**,

na inocência vocês não sabem avaliar a grandiosidade deste momento, mas saibam que vocês, meus pequenos príncipes, foram um grande estímulo para que eu chegasse até aqui.

Amo vocês.

AGRADECIMENTOS

A Deus

No mundo tereis aflições, mas tendes bom ânimo, eu venci o mundo.

João 16:33.

Ao conquistarmos um objetivo sempre temos muito a agradecer e durante esse período conheci pessoas que me estenderam as mãos e me ajudaram bastante a conviver com a distância e a saudade. Portanto, registro aqui meus agradecimentos a todos aqueles que de algum modo colaboraram para a realização deste trabalho e consequentemente para minha maturidade.

Ao Prof. Dr. José Eduardo Corrente, meu orientador e meu amigo, pela orientação, dedicação e confiança que muito contribuiram para que eu chegasse até aqui. Você é muito especial para mim.

Às prof^{as} Dr.^{as} Clarice Garcia Borges Demétrio e Maria Cristina Stolf Nogueira, pelos ensinamentos, orientações e atenção.

À prof^{a} Dr.^{a} Liciana Vaz de Arruda S. Chalita, pelos esclarecimentos, troca de idéias e principalmente pelo seu carinho.

Aos professores e funcionários do Departamento de Ciências Exatas da ESALQ - USP, pelo carinho, respeito e amizade.

Aos professores e funcionários do Departamento de Estatística da UFRN, pelo apoio, estímulo e credibilidade.

Ao programa PICDT/CAPES, pela bolsa de estudos concedida.

Aos professores Francisco Canindé de Oliveira e Robson Santana Pacheco do Departamento de Matemática da UFRN pela contribuição e amizade.

A minha família pelo apoio e palavras de conforto, feliz de quem tem uma família e nela pode buscar o mais nobre dos sentimentos, o amor.

A minha irmã Josilete Alves M. de Azevêdo, pela grande ajuda nas correções do Português e seu incentivo constante.

A minha querida sobrinha Ângela Patrícia pela sua dedicação e amor.

Ao Augusto que soube suportar com paciência a saudade dos nossos filhos e por ter me apoiado nesta decisão.

À Maria Helena, pela sua presença amiga/irmã em todas as horas, pelo seu estímulo e carinho sempre presentes em minha vida.

Às amigas Denise e Sílvia, vocês foram mais que amigas, vocês foram irmãs.

Aos amigos da Pós-graduação, pela convivência agradável, pelo carinho e amizade que tornaram momentos difíceis em momentos de muita alegria, em especial ao Afrânio, Adilson, Cristián, Wilson, Sandra, Cecília, Lêda, Suely, Silvano, André, Maria Cristina, Jomar, Heyder e a turma do mestrado/2000.

A Gilmara, pelo seu abraço amigo e grande incentivo nesta etapa final.

Ao Idemauro, em um mundo de muita maldade e ambições encontrar pessoas com a sua dignidade nos faz acreditar que ainda vale a pena investir no ser humano. Obrigada pelo seu companheirismo e amizade.

Aos meus filhinhos Matheus e João Augusto, pelo amor infinito que nos une.

SUMÁRIO

LISTA DE FIGURAS

Página

LISTA DE TABELAS

Página

UTILIZAÇÃO DO MÉTODO *B*OOTSTRAP NA ESCOLHA ENTRE OS MODELOS DE COX E LOGÍSTICO PARA DADOS DE SOBREVIVÊNCIA COM CENSURA INTERVALAR

Autora: JEANETE ALVES MOREIRA
Orientador: Prof. Dr. JOSÉ EDUARDO CORRENTE

RESUMO

Em muitos trabalhos científicos, o interesse dos pesquisadores está em identificar variáveis ou fatores que influenciem no tempo de ocorrência de um determinado evento, o qual geralmente é conhecido como tempo de falha. Nem sempre, porém, é possível observar o tempo exato de ocorrência da falha, conhecendo-se somente o intervalo em que a mesma ocorreu. Dados deste tipo são conhecidos como agrupados ou de sobrevivência com censura intervalar e apresentam observações empatadas. Quando o número de empates é pequeno, a análise desses dados pode ser feita através do ajuste do modelo de Cox (Cox, 1972) considerando-se aproximações para a verossimilhança parcial. No caso de ocorrerem muitos empates, deve-se considerar o tempo como discreto e modelar a probabilidade de ocorrência do evento num determinado intervalo, dado que ele não ocorreu no intervalo imediatamente anterior.

Isto pode ser feito considerando-se a variável resposta como uma variável indicadora de falha, que é binária, e ajustando-se os modelos de Cox para dados agrupados ou o modelo logístico, através dos modelos lineares generalizados com funções de ligação complemento *log-log* e *logit*, respectivamente. Após o ajuste dos modelos, existe o interesse em selecionar qual deles melhor se adequa aos dados experimentais. Para isto, algumas técnicas de seleção de modelos estão disponíveis na literatura, como o Critério de Informação de Akaike (Akaike, 1973), os Testes Escores propostos por Colosimo et al. (2000), dentre outras. Este trabalho propõe uma nova metodologia para a seleção entre os modelos de Cox e logístico para dados de sobrevivência com censura intervalar, utilizando o método de reamostragem *bootstrap* dos resíduos. Isto é feito extraindo-se os resíduos dos ajustes dos modelos propostos e calculando-se a diferença de deviances entre os modelos ajustados. Após isto, as novas observações são recompostas a partir dos resíduos *bootstrap* e novas diferenças de deviances são calculadas. Um histograma para esses valores é, então, obtido e um nível de significância empírico é proposto comparando-se os valores das diferenças de deviances das amostras *bootstrap* com a diferença de deviances inicial. Os resíduos utilizados são os resíduos simples, de Pearson e de Pearson padronizados. A aplicação da técnica é realizada através de dois conjuntos de dados de experimentos agronômicos, em que, no primeiro, a variável resposta foi o tempo até a murcha de um cultivar de linho susceptível ao patógeno *Fusarium oxysporum* e, no segundo, a variável resposta é o tempo até o inseto *Podysus nigrispinus* atingir a fase adulta. O método *bootstrap* indicou evidência de que o modelo logístico é o mais adequado para ajustar os dados do experimento com linho. Já no caso do experimento com insetos, o método indicou o modelo de Cox como o mais adequado aos dados e o resíduo de Pearson e Pearson padronizado, em ambos os casos, foram os que forneceram uma maior indicação para a escolha dos modelos em questão.

THE UTILIZATION OF THE BOOTSTRAP METHOD TO CHOOSE BETWEEN THE COX AND THE LOGISTIC MODELS FOR INTERVAL-CENSORED DATA

Author: JEANETE ALVES MOREIRA
Adviser: Prof. Dr. JOSÉ EDUARDO CORRENTE

SUMMARY

In many scientific works, researchers are interested in identifying variables or factors that influence the time of occurrence of certain events, which is generally known as failure time. However, it's not always possible to observe the exact failure time knowing only the interval in which failure occurred. Such data are known as grouped data or interval-censored data and present tied observations. When the number of ties is small, analysis of such data can be made by fitting the Cox Model (Cox, 1972) considering approximations for partial likelihood. In case there are many ties, the time must be considered as discreet and the probability of occurrence of the event must be modeled in a certain interval, since it did not take place in the immediately preceding interval. This can be done by considering the response variable as a binary variable that indicates failure and by fitting the Cox

model for grouped data or the logistic model through generalized linear models with complementary *log-log* and *logit* link functions, respectively. After fitting the models, it is necessary to decide which of them is more adequate for the experimental data. To that end, some model selection techniques are available in the literature, such as the Akaike Information Criterion (Akaike, 1973) and the Score Tests proposed by Colosimo et al. (2000), among others. This work proposes a new methodology to select between the Cox and logistic models for interval-censored data using the bootstrap method for the residues. This is done by extracting the residues from the fit of the proposed models and by calculating the deviance difference between the fitted models. After that, new observations are recomposed from the bootstrap residues and new deviance differences are calculated. A histogram for such values is then obtained and an empirical significance level is proposed by comparing the differences deviance values in the bootstrap samples with the deviance difference in the initial fit. The residues used are the simple residues, Pearson and standardized Pearson residues. The application of the technique was conducted through two data sets concerning agronomic experiments, in the first of which the response variable was the time until the blight of a flax cultivar susceptible to the pathogen *Fusarium oxysporum* and, in the second, the response variable was the time until the insect *Podysus nigrispinus* reached its adult phase. The bootstrap method indicated the evidence that the logistic model is more adequate to fit data for the experiment concerning flax. However, in the experiment concerning insects, the method indicated the Cox model as more adequate for the data, and the Pearson and standardized Pearson residue, in both cases, provided a greater indication for the choice of the considered models.

1 INTRODUÇÃO

Em muitas pesquisas científicas, nas mais diversas áreas do conhecimento, o interesse dos pesquisadores está em identificar fatores que expliquem o tempo até a ocorrência de um determinado evento como, por exemplo, o tempo até a falha de um componente, o tempo até a morte de um indivíduo, o tempo até a murcha de uma planta etc. Esse tempo é referido na literatura como tempo de vida ou tempo de falha e, uma característica importante nesses dados é a possibilidade de serem censurados, ou seja, por algum motivo não se observou o evento de interesse em alguns indivíduos, no período estudado. Nem sempre, porém, é possível observar o tempo exato de ocorrência do evento, conhecendo-se somente o intervalo em que o mesmo ocorreu. Dados desse tipo são conhecidos como agrupados ou de sobrevivência com censura intervalar e apresentam observações empatadas.

A análise de tais dados pode ser feita, no caso de um número pequeno de empates, pelo ajuste do modelo de Cox (Cox, 1972), através da função de verossimilhança parcial exata ou considerando-se aproximações, tais como as propostas por Breslow (1972, 1974), Peto (1972), Efron (1977) e Farewell e Prentice (1980). No caso de um número grande de empates considera-se o tempo como discreto e modela-se a probabilidade de ocorrência do evento num determinado intervalo, dado que ele não ocorreu no intervalo imediatamente anterior. Isso pode ser feito utilizando o modelo de riscos proporcionais de Cox para dados agrupados (Prentice & Gloeckler, 1978) ou o modelo logístico (Lawless, 1982). O ajuste dos modelos para tempos discretos na presença de censura intervalar pode ser realizado através dos modelos lineares generalizados (Nelder & Weddeburn, 1972).

Após o ajuste do modelo, existe o interesse em saber qual deles melhor explica os dados experimentais. Com essa finalidade, Chalita (1997) propôs dois testes escores a partir de uma distribuição mais geral, a família de transformação assimétrica de Aranda-Ordaz (1981), que tem como casos particulares as transformações *logit* (modelo logístico) e complemento log-log (modelo de riscos proporcionais de Cox). Os testes discriminam os modelos de Cox e logístico.

Este trabalho propõe uma nova metodologia para escolher entre os modelos de Cox e logístico para dados com censura intervalar. Esta metodologia é baseada no ajuste destes dois modelos e, através da simulação *bootstrap* dos resíduos, novas observações *bootstrap* são produzidas e novos ajustes dos modelos são feitos.

A decisão de escolha entre os modelos será tomada a partir da análise da distribuição empírica da diferença de *deviances* e do cálculo de um nível de significância empírico, baseado na distância entre as diferenças de *deviances* obtidas pelos ajustes dos modelos com base nas observações *bootstrap* e com os dados originais.

Como aplicação da metodologia serão utilizados dois conjuntos de dados de experimentos agronômicos, em que, no primeiro, a variável resposta é o tempo até a murcha de um cultivar de linho susceptível ao patógeno *Fusarium oxysporum*; no segundo, a variável resposta é o tempo até o inseto *Podisus nigrispinus* atingir a fase adulta. Ambos os conjuntos de dados apresentam alta proporção de empates, sendo necessário, portanto, a utilização dos modelos discretos. Os objetivos específicos deste trabalho são:

(i) ajustar os modelos logístico e de Cox para os dados dos dois experimentos em estudo, usando a teoria dos modelos lineares generalizados;

(ii) desenvolver programas computacionais para a reamostragem *bootstrap* dos três tipos de resíduos que serão considerados;

(iii) identificar os fatores que influenciaram no tempo de vida para cada experimento realizado;

(iv) analisar os resultados obtidos para cada tipo de resíduo considerado;

(v) discriminar entre os modelos de Cox e logístico;

(vi) avaliar a técnica *bootstrap* proposta neste trabalho.

2 REVISÃO DE LITERATURA

2.1 Análise de sobrevivência

A Análise de sobrevivência é uma das áreas da estatística na qual são desenvolvidos métodos para analisar dados provenientes de variáveis aleatórias tomando valores positivos. Essas variáveis correspondem ao tempo de ocorrência de um evento de interesse, tais como o tempo até a morte ou cura de um indivíduo, tempo até a falha de um componente eletrônico, dentre outros. Estes métodos foram originalmente desenvolvidos para estudos de mortalidade, explicando, portanto, o nome sobrevivência que, segundo Allison (1995), é um tanto inadequado por levar a uma visão restrita suas potenciais aplicações.

Existem duas características bastante importantes nos dados de sobrevivência. Uma, diz respeito à distribuição dos dados que, geralmente, é assimétrica à direita e a outra é a presença de censuras, que ocorre quando não é possível observar a resposta de algumas unidades que deverão ser incluídas na análise.

O principal interesse da área de análise de sobrevivência está na estimação da probabilidade de um indivíduo sobreviver até o tempo t, conhecida como função de sobrevivência e a razão instantânea de falha no tempo t, dado que ele sobreviveu até t, chamada de função de risco.

Dentre as técnicas utilizadas para estimar-se a função de sobrevivência destacam-se as paramétricas e as não-paramétricas. As técnicas não paramétricas são apenas descritivas e o estimador mais utilizado para a função de sobrevivência é o produto-limite de Kaplan-Meier (Kaplan & Meier, 1958), o qual é utilizado quando

se conhece o tempo exato de falha, embora também possa ser utilizado para um número pequeno de empates. Um outro estimador não-paramétrico para a função de sobrevivência é a tabela de vida ou o estimador atuarial, que considera os dados agrupados, ou seja, são conhecidos apenas os intervalos em que as falhas ou censura ocorreram.

Estas técnicas não levam em consideração as covariáveis relacionadas com o tempo de sobrevivência. Para considerar estas covariáveis devem-se utilizar os modelos paramétricos ou teste de vida acelerado para os quais se supõe uma distribuição de probabilidade conhecida para os tempos e o modelo semi-paramétrico de Cox (1972), para o qual não é necessário supor distribuição para a variável tempo.

O tempo de sobrevivência T tem função densidade de probabilidade, $f(t)$, definida como a probabilidade instantânea de o indivíduo falhar no intervalo $[t; t + \Delta t]$ por unidade de tempo, que pode ser expressa por

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0^+} \frac{P(\text{um indivíduo falhar em } [t; t + \Delta t])}{\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{dF(t)}{dt}. \end{aligned} \tag{1}$$

em que $F(t) = P(T \leq t)$ é a função de distribuição de T. Tem-se ainda que, se $f(t)$ define uma verdadeira função de densidade de probabilidade, então:

$$f(t) \geq 0 \qquad e \qquad \int_0^\infty f(t)dt = 1.$$

A função de sobrevivência, denotada por $S(t)$, representa a probabilidade de o indivíduo não falhar até um certo tempo t, isto é,

$$S(t) = P(T \geq t) = 1 - F(t) \tag{2}$$

e é uma função monótona decrescente que satisfaz a:

$$S(0) = 1 \qquad e \qquad \lim_{t \to \infty} S(t) = 0.$$

A função $S(t)$ é também conhecida como razão de sobrevivência acumulada e sua representação gráfica é chamada de *curva de sobrevivência*, que pode ser

usada para encontrar qualquer percentil de interesse e, também, como um primeiro elemento de comparação entre dois ou mais tratamentos.

A função risco, representada por *h(t)*, especifica a taxa de falha instantânea no tempo t e é dada por:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0^+} \frac{P(t \leq T \leq t + \Delta t / T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{P(t \leq T \leq t + \Delta t, T \geq t)}{\Delta t P(T \geq t)} \\ &= \lim_{\Delta t \to 0^+} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t P(T \geq t)}, \end{aligned} \tag{3}$$

pois $[t \leq T \leq t + \Delta t] \subset [T \geq t]$. Então,

$$h(t) = \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}.$$

As funções (1), (2) e (3) são matematicamente equivalentes, ou seja, dada uma delas as outras podem ser obtidas facilmente.

Relações entre as funções *f(t)*, *S(t)* e *h(t)*

1. A função de risco instantânea pode ser dada por:

$$h(t) = \frac{f(t)}{S(t)}. \tag{4}$$

2. A função de densidade de T pode ser dada por:

$$f(t) = \frac{dF(t)}{dt} = \frac{d[1 - S(t)]}{dt} = -S'(t). \tag{5}$$

3. Substituindo-se a equação (5) em (4), tem-se que:

$$h(t) = \frac{-S'(t)}{S(t)} = \frac{-d \log S(t)}{dt}. \tag{6}$$

4. Integrando-se (6) no intervalo $[0;t]$ e usando $S(0) = 1$, pode ser visto que:

$$-\int_0^t h(x)dx = \log S(t)$$

ou,

$$H(t) = -\log S(t),$$

ou ainda,

$$S(t) = exp[-H(t)] = exp\left[-\int_0^t h(x)dx\right]. \tag{7}$$

5. Das equações (4) e (7), obtém-se:

$$f(t) = h(t)exp[-H(t)].$$

O modelo de riscos proporcionais foi proposto por Cox (1972), sendo um dos modelos mais utilizados na prática pela sua flexibilidade, ou seja, não requer que se escolha alguma distribuição de probabilidade para os tempos de falha, além de levar em consideração as covariáveis relacionadas com esse tempo. Este modelo supõe independência entre os tempos observados e modela a função de risco através da função:

$$h(t \mid \mathbf{x}_l) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_l) \tag{8}$$

em que:

- $h_0(t)$ é a função risco de base (taxa de risco para $x_l = 0$);
- $\boldsymbol{\beta}$ é o vetor de dimensão $p \times 1$ de parâmetros desconhecidos;
- $\mathbf{x}_l$ é o vetor de dimensão $p \times 1$ de covariáveis observadas para o l-ésimo indivíduo.

O modelo dado por (8) é denominado modelo de riscos proporcionais pelo fato de a razão entre as taxas de falhas para dois indivíduos ser constante no tempo, ou seja, tomando-se a razão de riscos para dois indivíduos l e j e aplicando-se a equação (8), obtém-se o resultado abaixo, que não depende do tempo:

$$\frac{h_l(t \mid x_l)}{h_j(t \mid x_l)} = \frac{h_0(t)exp(\beta' x_l)}{h_0(t)exp(\beta' x_j)} = exp(\beta' x_l - \beta' x_j).$$

Devido a presença do componente não-paramétrico $h_0(t|\mathbf{x}_l)$ no modelo, o método da máxima verossimilhança usual não é apropriado para estimar os parâmetros. Desse modo, Cox(1975) formalizou o método da máxima verossimilhança parcial que consiste em condicionar a função de verossimilhança nos tempos de ocorrência do evento de modo a eliminar a função de perturbação $h_0(t|\mathbf{x}_l)$.

Ocorre, porém, que quando acontecem empates não se pode utilizar a função de verossimilhança parcial exata e, nesse caso, existem métodos alternativos tais como os apresentados por Breslow (1974) e Efron (1977).

Estes autores propuseram aproximações para a função de verossimilhança parcial e estas encontram-se implementadas em vários *softwares* estatísticos. Quando o número de empates é pequeno ambas as aproximações fornecem valores próximos, todavia, a aproximação de Efron (1977) apresenta resultados mais próximos daqueles encontrados pela forma exata. Autores como Kalbfleish & Prentice (1980), Lawless (1982), Cox & Oakes (1974), Collett (1994), Allisson (1995) e Chalita (1997) indicam que, no caso de muitos empates, deve-se considerar o tempo como discreto para modelar a probabilidade de o indivíduo falhar num determinado intervalo, dado que ele sobreviveu até o intervalo anterior, através dos modelos de riscos proporcionais de Cox para dados com censura intervalar ou o modelo logístico.

Kalbfleish & Prentice (1973) desenvolveram um modelo discreto para dados agrupados a partir do modelo de Cox, apresentando estimativas da função de sobrevivência para dados contínuos e discretos além de propor uma generalização do teste *log-rank* para comparação de várias curvas de sobrevivência. Enquanto Prentice & Gloeckler (1978) ajustaram o modelo de Cox para dados de câncer de mama

agrupados e apresentaram as propriedades assintóticas da função de verossimilhança para estimação do coeficiente de regressão e da função de sobrevivência. Nos estudos de Whitehead (1989) encontra-se um ajuste do modelo de riscos proporcionais de Cox a dados com censura intervalar através da utilização do modelo linear generalizado, considerando que a função de ligação é a *complemento log-log* e que a variável resposta é a indicadora de censura, que é binária.

Uma metodologia mais ampla para usar o modelo de riscos proporcionais na análise de dados com censura intervalar foi desenvolvida por Finkelstein (1986). Essa metodologia permite incorporar observações que, por algum motivo, se ausentaram do estudo em um determinado período ao acompanhamento realizado pelo pesquisador e depois voltem a ser incorporadas ao ensaio. Por exemplo, um indivíduo que foi monitorado semanalmente pode faltar por algumas semanas e depois retornar ao monitoramento. Neste caso podem ocorrer censuras em intervalos sobrepostos ou não disjuntos e os métodos utilizados por Prentice e Gloeckler (1978) não poderiam ser aplicados. A autora também apresenta um teste para verificar se há diferença entre as curvas de sobrevivência, mostrando a relação entre a estatística do teste e a estatística usual utilizada no teste *log-rank* para tempos exatos e censurados à direita.

O modelo logístico foi ajustado por Thompson (1977) a dados de remissão de leucemia agrupados em intervalos de tempo, considerando a censura uniforme no intervalo. Este modelo recai no modelo de Cox considerando tempos como contínuos quando os comprimentos dos intervalos aproximam-se de zero. Lawless (1982) discutiu com detalhes o caso de dados agrupados, ajustando o modelo de Cox e o logístico.

Collett (1994) trabalhou com dados de sobrevivência com censura intervalar mostrando três formas diferentes de análise, utilizando dados de recorrência de úlcera apresentados por (Whitehead, 1989). As formas consideradas para análise foram: sem considerar censura intervalar (considerando o tempo de recorrência como contínuo); sem considerar o tempo no qual uma recorrência foi detectada e con-

siderando censura intervalar juntamente com o período em que houve a recorrência. A conclusão do autor é que a terceira forma utilizada para ajustar esse conjunto de dados é, geralmente, a forma mais eficiente de análise, embora o resultado não fosse tão diferente da análise que assume o tempo como contínuo quando o número de intervalos não foi tão pequeno e o comprimento destes não foi grande.

Mais recentemente, Chalita (1997) fez comparações entre as aproximações da função de verossimilhança parcial para dados empatados e os modelos de Cox e logístico para tempos discretos. Para tal propósito, a autora fez simulações para os dados de tempo de falha utilizando a distribuição de Weibull, variando a proporção de censura e o tamanho da amostra, empregando 5, 10, 20 e 30 intervalos. Considerando o tempo como contínuo, utilizou o limite superior de cada classe e no caso discreto foi considerado o número de falhas e o número de indivíduos sob risco no intervalo. As conclusões obtidas mostraram que, quando o tempo é considerado como contínuo, a forma exata comporta-se melhor seguida pelas aproximações de Efron (1977) e Breslow (1974), independente da variação do número de intervalos, da proporção de censura, dos parâmetros e se o vetor de covariáveis é binário ou contínuo.

Para o tamanho de amostra $n = 20$, as diferenças entre os ajustes considerando-se o tempo como discreto e contínuo com aproximações são mais evidentes e, à medida que se aumenta o número de intervalos, o erro quadrático e o vício diminuem para os modelos contínuos e aumentam para os discretos. Isso porque a proporção de empates diminui e no caso discreto aumenta o número de parâmetros a serem estimados nos modelos. Com relação aos modelos para tempo discreto, o modelo de Cox ajustou-se melhor do que o logístico, embora esta conclusão possa ser explicada pelo fato de ter sido usada a distribuição de Weibull para gerar os dados, cujos riscos são proporcionais.

As Tabelas 1 e 2 mostram algumas sugestões empíricas que foram apresentadas por Chalita (1997) com relação ao uso dos modelos considerando os tempos discreto e contínuo com aproximações e levando-se em conta a proporção de

empates.

Tabela 1. Relação entre proporções de empates (pe) e o modelo a ser utilizado, para amostras de tamanho 20 e 50.

pe(%)	Modelos
< 20	Devem-se usar modelos para tempo contínuo com aproximações
20 a 30	Podem-se usar modelos para tempo contínuo com aproximações
> 30	Devem-se usar modelos para tempo discreto

Tabela 2. Relação entre proporções de empates (pe) e o modelo a ser utilizado, para amostras de tamanho 100 e 200.

pe(%)	Modelos
< 20	Devem-se usar modelos para tempo contínuo com aproximações
20 a 25	Podem-se usar modelos para tempo contínuo com aproximações
> 25	Devem-se usar modelos para tempo discreto

Fay (1999) apresentou um modelo geral para desenvolver um teste escore a fim de testar os parâmetros envolvidos nos modelos de riscos proporcionais de Cox, logístico e de chances proporcionais para dados com censura intervalar que tem como casos particulares os modelos apresentados por Finkelstein (1986) e Sun (1996). O teste proposto pelo autor considerou somente comparações entre k tratamentos e mostrou a equivalência entre a forma da estatística escore para o teste *log-rank* ponderado e a forma utilizada nos testes de permutação.

2.2 Modelos lineares generalizados

Os modelos lineares generalizados (MLG) foram desenvolvidos por Nelder & Wedderburn (1972) e constituem uma extensão dos modelos lineares clássicos de Gauss-Markov.

A definição de modelos lineares generalizados envolve três componentes:

1. Componente aleatório: representado pelas variáveis aleatórias independentes $Y_1, Y_2, \cdots, Y_n$ pertencentes à família exponencial na forma:

$$f_Y(y_i, \theta, \phi) = exp\left\{\frac{1}{a(\phi)}[y_i\theta - b(\theta)] + c(y_i; \phi)\right\}, \tag{9}$$

   em que $a(.), b(.)$ e $c(.)$ são funções conhecidas, sendo $\mu_i = E(Y_i) = b'(\theta)$ e $V(Y_i) = a(\phi)b''(\theta_i) = a(\phi)V(\mu)$, $V(\mu)$ é chamada de função de variância e, em geral, $a(\phi) = \phi/W$, em que ϕ é o parâmetro de dispersão e W são os pesos *a priori*. Quando se conhece ϕ, (9) pertence à família exponencial na forma canônica a um parâmetro e θ é chamado de parâmetro canônico ou natural.

2. Componente sistemático: um preditor linear dado por:

$$\eta_i = \sum_{j=1}^{\ell} \beta_j x_{ij} \qquad ou \qquad \eta = X\boldsymbol{\beta},$$

   em que X é uma matriz $n \times \ell$ do delineamento (covariáveis ou variáveis binárias) e $\boldsymbol{\beta}$ é um vetor $\ell \times 1$ de parâmetros desconhecidos.

3. Uma função de ligação $g(.)$ monótona e diferenciável que relaciona o componente aleatório com o componente sistemático:

$$g(\mu_i) = \eta_i = \sum_{j=1}^{\ell} \beta_j x_{ij}, \qquad i = 1, 2 \ldots, n.$$

Quando $g(\mu_i) = \theta_i$, a função de ligação é chamada de função de ligação canônica ou natural. As funções de ligação mais comuns para dados binomiais são a *logit*, *probit* e *complemento log-log.*

Os estimadores de máxima verossimilhança para os parâmetros do modelo, $\hat{\boldsymbol{\beta}}$, são obtidos pelo método escore de Fisher, implementado através do algoritmo de mínimos quadrados reponderados iterativamente, **IRLS** (McCullagh & Nelder, 1989).

Uma medida de discrepância chamada *deviance* foi proposta por Nelder & Wedderburn (1972), sendo a análise de *deviance* uma generalização da análise de variância. Essa medida é dada por

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2l(\mathbf{y}, \mathbf{y}) - 2l(\hat{\boldsymbol{\mu}}, \mathbf{y})$$

em que

$l(\mathbf{y}, \mathbf{y})$ é o logaritmo da função de verossimilhança para o modelo saturado, e

$l(\hat{\boldsymbol{\mu}}, \mathbf{y})$ é o logaritmo da função de verossimilhança para o modelo sob pesquisa,

servindo para avaliar a falta de ajuste de um modelo. No caso de modelos encaixados, a diferença de *deviances* mede o efeito de fatores, covariáveis e suas possíveis interações.

Cordeiro (1986) e Demétrio (1993) são duas referências em português que mostram detalhes do processo de estimação dos parâmetros e dão ênfase ao uso do *software* GLIM (Generalized Linear Iteractive Modelling). Dobson (1990) apresenta uma introdução aos modelos lineares generalizados de forma bastante didática sendo, portanto, recomendada para uma primeira leitura.

Um caso particular dos MLG's são os dados de proporções que podem ser analisados através da distribuição binomial. Collett (1991) descreveu detalhadamente os modelos para dados binários e binomiais apresentando várias aplicações

nos ensaios de Entomologia, em estudo de coorte, caso-controle e ensaios de dose-resposta.

Para se modelarem dados binomiais através dos modelos lineares generalizados considerem-se as variáveis aleatórias $Y_1, Y_2, \cdots, Y_n$ independentes, representando o número de sucessos em amostras de tamanho m_i, com probabilidade de sucesso π_i, $i = 1, 2, \cdots, n$. Se $Y_i \sim b(m_i, \pi_i)$ então, $E(Y_i) = \mu_i = m_i\pi_i$ e $V(Y_i) = m_i\pi_i(1 - \pi_i)$, podendo-se modelar π_i em termos das covariáveis ou fatores $X_i's$ como $g(\pi_i) = \boldsymbol{\beta}'\mathbf{x_i}$, onde $g(.)$ é uma função de ligação apropriada e $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos. As funções de ligação utilizadas neste trabalho serão *logit* (modelo logístico) e *complemento log-log* (modelo de Cox).

2.3 O método *bootstrap*

Nas duas últimas décadas, juntamente com os computadores de alta velocidade com grandes capacidades de armazenamento de dados a custos mais acessíveis, vem sendo cada vez mais utilizada uma metodologia para análise estatística conhecida como computacionalmente intensiva. Esses métodos começaram a se tornar ferramentas bastante úteis e alternativas para os métodos estatísticos tradicionais, pois visam substituir as complexidades analíticas ou uso de aproximações assintóticas para determinados problemas. Entre esses métodos, podem-se destacar o *Jackknife* e o *bootstrap*, apresentados por Efron (1979).

Algumas condições de regularidade sob as quais esse método é consistente foram discutidas por Bickel & Freedman (1981). Hall (1992) abordou as propriedades teóricas do método, enquanto os conceitos básicos e muitas aplicações práticas do método it bootstrap podem ser encontrados em Efron & Tibishirani(1993) e Davison & Hinkley (1997).

O *bootstrap* é um método que pode ser facilmente implementado tanto de forma não-paramétrica quanto paramétrica, dependendo do conhecimento do problema. No caso não-paramétrico, o método *bootstrap* reamostra os dados um

número **B** de vezes com reposição, em que os valores são escolhidos segundo uma função de distribuição empírica estimada, tendo em vista que, em geral, não se conhece a distribuição subjacente aos dados. No caso paramétrico, quando se tem informação suficiente sobre a forma da distribuição dos dados, a amostra *bootstrap* é formada realizando-se a amostragem diretamente nesta distribuição com os parâmetros desconhecidos substituídos por estimativas paramétricas. A distribuição da estatística de interesse aplicada aos valores da amostra *bootstrap*, condicional aos dados observados, é definida como a distribuição *bootstrap* desta estatística.

Considerando-se $Y_1, \ldots, Y_n$ uma amostra aleatória de tamanho n com distribuição de probabilidade desconhecida F que depende de um parâmetro θ, e sejam $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ os valores observados. Como a função F é desconhecida, pode-se estimá-la pela função de distribuição empírica, baseado na amostra de tamanho n que é dada por:

$$\hat{F}(y) = \frac{\#(y_j \leq y)}{n}.$$

A função $\hat{F}$ assume valores $1/n$ para cada valor da amostra y_i, $i = 1, \ldots, n$.

Supondo-se agora que se queira estimar o parâmetro θ que depende da distribuição F, isto é, supondo-se que $\theta = t(F)$ baseado em $\mathbf{Y}$. Um estimador de θ é $\hat{\theta} = s(\mathbf{Y})$ e é preciso saber qual a precisão deste estimador. Para isto, utiliza-se o método *bootstrap*, que retira amostras *bootstrap* com reposição $y_1^*, \ldots, y_n^*$ utilizando a função de distribuição empírica estimada $\hat{F}$. Assim, para a amostra *bootstrap* $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$, calcula-se

$$\hat{\theta}^* = s(\mathbf{y}^*).$$

Um diagrama esquemático para o método *bootstrap* é dado pela Tabela 3.

Ao se obter um grande número de replicações *bootstrap* de $y_1, \ldots, y_n$, obtem-se a distribuição empírica de $\hat{\theta}^*$, podendo-se calcular medidas de posição, dispersão, intervalos de confiança, dentre outras, de modo a analisar a precisão do estimador obtido inicialmente.

Tabela 3. Diagrama esquemático do método *bootstrap*

Mundo Real			Mundo *Boostrap*	
Distribuição de Probabilidade Desconhecida	Amostra Aleatória Observada		Distribuição Empírica	Amostra *Bootstrap*
$F \longrightarrow$	$\mathbf{y} = (y_1, \ldots, y_n)$	$\Longrightarrow$	$\hat{F} \longrightarrow$	$\mathbf{y} = (y_1^*, \ldots, y_n^*)$
	$\downarrow$			$\downarrow$
	$\hat{\theta} = s(\mathbf{y})$			$\hat{\theta}^* = s(\mathbf{y}^*)$
	Estatística de Interesse			Replicação *Bootstrap*

Quando se tem informações suficientes sobre F, o estimador *bootstrap* $\hat{\theta}^*$ de θ pode ser obtido da mesma forma que o já descrito anteriormente, apenas levando-se em conta a verdadeira distribuição subjacente aos dados.

Dentre as várias aplicações do método *bootstrap*, existe o interesse particular nos modelos de regressão. Freedman (1981) trabalhou com o método *bootstrap* em modelos de regressão e correlação mostrando a validade de sua aproximação para a distribuição dos estimadores dos parâmetros do modelo utilizando o método de mínimos quadrados. Esse assunto foi abordado por Hinkley (1988), Efron & Tibishirani (1993), Silva (1995) e Davison & Hinkley (1997) dentre outros. Silva (1995), por sua vez, apresentou uma descrição detalhada sobre o método *bootstrap* aplicado à regressão múltipla e mostrou a eficiência do método em um problema de regressão com restrição nos parâmetros utilizando dados de um experimento químico.

Para o modelo de regressão, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, em que $\mathbf{Y}$ é o vetor de respostas, $\mathbf{X}$ é a matriz de covariáveis, $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos e $\mathbf{e}$ é o vetor de erros independentes e identicamente distribuídos, com distribuição

F, média zero e variância σ^2, o plano de reamostragem pode ser feito através dos resíduos ou dos pares de observações (Davison & Hinkley, 1997).

A abordagem do método *bootstrap* também pode ser feita para os modelos lineares generalizados e, nesse caso, a reamostragem dos resíduos é realizada de forma análoga aos modelos lineares ordinários. Alguns dos resíduos discutidos na literatura usados para esses tipos de modelos, são: os resíduos de Pearson, os resíduos de Pearson padronizados, os resíduos componentes da *deviance*, os resíduos componentes da *deviance* padronizados e os resíduos padronizados na escala do preditor linear.

Moulton & Zeger (1989) utilizaram reamostragem dos vetores de observações com reposição em uma análise de medidas repetidas no contexto de modelos lineares generalizados seguindo a mesma metodologia utilizada em modelos lineares ordinários. Já em Moulton & Zeger (1991), foi sugerido o uso de reamostragem dos resíduos de Pearson padronizados obtidos do ajuste de um modelo linear generalizado.

2.4 Seleção de modelos

Ao se analisar um conjunto de dados, um modelo matemático é sempre pressuposto e, através do ajuste desse modelo e de testes apropriados, é possível verificar se existem efeitos significativos para as covariáveis envolvidas.

Em praticamente todos os *softwares* estatísticos, é possível fazer ajuste de modelos incluindo todos os efeitos de uma só vez ou um efeito por vez, verificando a qualidade do ajuste. Neste caso, diz-se que o ajuste do modelo é feito através de modelos encaixados e, uma medida utilizada para verificar o efeito de cada fator, é a diferença de *deviances*, como já citada na Seção 2.2. Quando os ajustes são feitos baseados em diferentes suposições de modelos, tem-se o caso de ajuste de modelos não encaixados.

O problema da seleção de modelos não encaixados foi apresentada em

Cox (1962), obtendo resultados assintóticos para testar duas hipóteses não encaixadas, considerando os modelos M_f como hipótese nula e M_g como a hipótese alternativa, baseado na estatística da razão de verossimilhança, dada por:

$$T_f = [L_f(\hat{\alpha}) - L_g(\hat{\beta})] - E_{\hat{\alpha}}[L_f(\hat{\alpha}) - L_g(\hat{\beta})]$$

em que $L_f(\alpha)$ é o logaritmo da função de verossimilhança de α sob o modelo M_f e $E_{\hat{\alpha}}[]$ é o valor esperado de $L_f(\alpha)$ com relação a a M_f para $\alpha = \hat{\alpha}$ (em que $\hat{\alpha}$ é o estimador de máxima verossimilhança de α). O autor mostrou que esta estatística é assintoticamente normal sob o modelo M_f.

Uma abordagem para a escolha de dois modelos de regressão não lineares foi discutida por Williams (1970). Para selecionar qual desses modelos melhor se ajusta aos dados, o autor obteve distribuições para a estatística do teste através da simulação de cada um dos modelos ajustados, usando a especificação paramétrica do modelo e reajustando cada modelo para o conjunto de dados simulados. Em seu trabalho, foi sugerido também substituir o cálculo teórico de T_f e T_g, com T_g definido analogamente a T_f no parágrafo acima, pela simulação da razão de verossimilhança sob cada modelo estimado.

A comparação de modelos lineares não encaixados utilizando intervalos de confiança aproximados para a diferença no erro quadrático médio de previsão foi apresentada em Efron (1984). O autor mostrou que essa abordagem é relativamente próxima ao teste de Hotelling para comparar dois modelos de regressão linear simples. Os intervalos foram construídos através do método *bootstrap* paramétrico aplicado ao estimador C_p de Mallows's [1], citado por Efron(1984), da diferença entre os erros quadráticos médios.

Tomando como base o trabalho de Williams (1970), Hinde (1992) [2] substituiu a simulação paramétrica pela simulação *bootstrap*. Este procedimento tem a vantagem de a simulação *bootstrap* não depender das especificações do mo-

[1]MALLOWS, C. L. Some comments on C_p. **Technometrics**, v.15, p.661-675, 1973.

[2]HINDE, J. Choosing between non-nested models: a simulation aproach. 1992. (Trabalho não publicado).

delo, apenas envolvendo reamostragem de um conjunto finito de observações. O procedimento consiste no ajuste de dois modelos, M_1 e M_2, que podem ter diferentes especificações de erros e funções de ligação. Isso feito, retiram-se amostras *bootstrap*, reamostrando-se os resíduos de Pearson padronizados para o modelo ajustado M_1 e reajustando-se os dois modelos para cada amostra *bootstrap*. O mesmo foi feito considerando a reamostragem dos resíduos para o modelo M_2. As diferenças de *deviances* entre os modelos foram utilizadas para construir uma distribuição empírica e a comparação desta distribuição com a diferença de *deviances* indica a preferência de um modelo em relação ao outro, ou que não há diferença entre os dois modelos ou ainda que nenhum dos modelos se ajustou bem aos dados.

Shao (1996) propôs um método de seleção de variáveis baseado no *bootstrap*, enfatizando as boas propriedades do método e citando duas fortes razões para seu uso que são: i) a precisão dos procedimentos utilizados pelo *bootstrap* no contexto de regressão linear ser maior do que em outros métodos; ii) o procedimento de seleção *bootstrap* desenvolvido para o caso de regressão linear pode ser estendido, sem qualquer teoria adicional, para problemas mais complicados tais como, modelos de regressão não linear, modelos lineares generalizados e modelos auto-regressivos.

O autor considerou duas maneiras de gerar observações *bootstrap* que são a reamostragem através dos resíduos e dos pares de observações (x, y), mas incorporando uma modificação no tamanho da amostra *bootstrap* com a finalidade de tornar o procedimento de seleção de modelos mais consistente. O critério de seleção de modelos foi baseado no erro de predição agregado *bootstrap*, sendo escolhido como melhor modelo aquele que minimiza este erro.

Colosimo *et al.* (2000) desenvolveu testes escores para discriminar entre os dois modelos discretos estudados, modelos de riscos proporcionais de Cox e logístico, a partir da família assimétrica de Aranda-Ordaz (1981), dada por:

$$V_{\lambda}(p_i(\mathbf{x}_l)) = \begin{cases} \ln\left(\dfrac{[1-p_i(\mathbf{x}_l)]^{-\lambda}-1}{\lambda}\right) & \text{se } \lambda \neq 0 \\ \ln(-\ln(1-p_i(\mathbf{x}_l))) & \text{se } \lambda = 0 \end{cases} \tag{10}$$

em que $p_i(\mathbf{x}_l)$ é a probabilidade do l-ésimo indivíduo falhar até a_i dado que não falhou até a_{i-1}, considerando o vetor de covariáveis $\mathbf{x}_l$.

O teste tem como objetivo verificar qual função de ligação seria a mais adequada para um determinado conjunto de dados, uma vez que, esta família tem como casos particulares as transformações *logit* (modelo logístico) e *complemento log-log* (modelo de Cox).

Assim, pela inversão de (10) tem-se o modelo ajustado a $p_i(\mathbf{x}_l)$, que é dado por:

$$p_i(\mathbf{x}_l) = \begin{cases} 1-(1+\lambda\exp(\eta_{li}))^{-1/\lambda} & \text{se } \lambda\exp(\eta_{li}) > -1 \\ 1 & \text{se } \lambda\exp(\eta_{li}) \leq -1 \end{cases}$$

e tem como casos particulares as funções de ligações: *logit*,

$$\eta_{li} = \ln[p_i(\mathbf{x}_l)/(1-p_i(\mathbf{x}_l))]$$

quando $\lambda = 1$ (modelo logístico) e *complemento log-log*,

$$\eta_{li} = \ln[-\ln(1-p_i(\mathbf{x}_l))]$$

quando $\lambda \to 0$ (modelo de Cox).

A estatística do teste escore para o modelo de Cox foi obtida para a hipótese nula $H_0 : \lambda = 0$, sendo dada por:

$$Sr = \frac{U_\lambda^2(0,\hat{\beta},\hat{\gamma}^*)}{I_{\lambda\lambda}^{-1}(\hat{\theta}_0)},$$

em que $I_{\lambda\lambda}(\theta_0)$ é a matriz de informação observada.

De forma análoga à estatística escore para o modelo de Cox, pode-se obter Sr_1, para testar a hipótese $H_0 : \lambda = 1$ (modelo logístico), dada por:

$$Sr_1 = \frac{U_\lambda^2(\hat{\theta}_1)}{I_{\lambda\lambda}(\hat{\theta}_1) - I_{\lambda\phi}(\hat{\theta}_1)I_{\phi\phi}^{-1}(\hat{\theta}_1)I_{\phi\lambda}(\hat{\theta}_1)},$$

em que:

$$I_{\lambda\phi}(\theta_1) = \begin{bmatrix} I_{\lambda\gamma^*}(\theta_1) & I_{\lambda\beta}(\theta_1) \end{bmatrix}$$

e

$$I_{\phi\phi}(\hat{\theta}_1) = \begin{bmatrix} I_{\gamma^*\gamma^*}(\theta_1) & I_{\gamma^*\beta}(\theta_1) \\ I_{\beta\gamma^*}(\theta_1) & I_{\beta\beta}(\theta_1) \end{bmatrix}$$

em que, $I_{\lambda\lambda}(\theta_1)$, $I_{\lambda\gamma^*}(\theta_1)$, $I_{\lambda\beta}(\theta_1)$, $I_{\gamma^*\gamma^*}(\theta_1)$,$I_{\gamma^*\beta}(\theta_1)$ e $I_{\beta\beta}(\theta_1)$ são os limites das derivadas parciais de segunda ordem do logaritmo da função dada em (4).

As estatísticas Sr e Sr_1 têm distribuição aproximadamente qui-quadrado com 1 grau de liberdade.

Um procedimento para a escolha de modelos lineares não encaixados baseado no critério de Informação de Akaike (Akaike,1973) e no Critério Bayesiano de Informação (Schwartz, 1978) foi apresentado por Lindsey & Jones (1998). Os autores fizeram comparações entre modelos lineares generalizados utilizando dados de contagens de células no sangue em diagnóstico médico. Para pequenas amostras os resultados obtidos pelos dois critérios são próximos.

3 METODOLOGIA

3.1 Modelos para dados com censura intervalar

Considerando-se uma amostra de n indivíduos cujos tempos de vida, T_l, $l = 1, \ldots, n$, são agrupados em k intervalos, $I_i = [a_{i-1}, a_i)$, $i = 1, \ldots, k$, com $0 = a_0 < a_1 < \ldots < a_k = \infty$ e um vetor de covariáveis $\mathbf{x_l}$, $l = 1, \ldots, n$. Seja D_i o conjunto dos indivíduos que falharam no intervalo I_i; R_i, o conjunto dos indivíduos sob risco no início de I_i, e Δ_{li}, a variável indicadora de falha do l-ésimo indivíduo em I_i, que assume valor zero quando a l-ésima observação é censurada no i-ésimo intervalo, e um em caso contrário, para $i = 1, \ldots, k$, $l = 1, \ldots, n$. Assumindo-se que todas as censuras ocorrem no final do intervalo. A probabilidade do l-ésimo indivíduo falhar até a_i dado que não falhou até a_{i-1}, considerando o vetor de covariáveis $\mathbf{x}_l$ é expressa por:

$$p_i(\mathbf{x}_l) = P(T_l \leq a_i \mid T_l \geq a_{i-1}, \mathbf{x}_l). \tag{11}$$

Em termos da função de sobrevivência (11) pode ser escrita como:

$$\begin{aligned} p_i(\mathbf{x}_l) &= 1 - P(T_l \geq a_i \mid T_l \geq a_{i-1}, \mathbf{x}_l) \\ &= 1 - \frac{S(a_i \mid \mathbf{x}_l)}{S(a_{i-1} \mid \mathbf{x}_l)} \end{aligned} \tag{12}$$

em que $S(a_i \mid \mathbf{x}_l)/S(a_{i-1} \mid \mathbf{x}_l)$ é a probabilidade do l-ésimo indivíduo não falhar até a_i dado que ele não falhou até a_{i-1}.

Assim, a função de verossimilhança em termos da probabilidade condicional $p_i(\mathbf{x}_l)$ é dada por:

$$\begin{aligned} L &= \prod_{i=1}^{k} [\prod_{l \in D_i} p_i(\mathbf{x}_l) \prod_{l \in R_i - D_i} (1 - p_i(\mathbf{x}_l))] \\ &= \prod_{i=1}^{k} \prod_{l \in R_i} (p_i(\mathbf{x}_l))^{\Delta_{li}} (1 - p_i(\mathbf{x}_l))^{(1-\Delta_{li})}, \end{aligned} \tag{13}$$

que corresponde à função de verossimilhança para variáveis com distribuição de Bernoulli, pois a variável Δ_{li} é binária (isto é, ocorreu a falha, ou não, no intervalo I_i).

As probabilidades $p_i(\mathbf{x}_l)$ podem ser modeladas através dos modelos de riscos proporcionais de Cox ou logístico e as covariáveis associadas aos indivíduos podem ser fatores.

3.1.1 Modelo de riscos proporcionais de Cox

Como já visto em (8), a função de risco para T dado o vetor de covariáveis $\mathbf{x}_l$ é, de acordo com o modelo de Cox, dada por:

$$h(t \mid \mathbf{x}_l) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_l),$$

e a função de sobrevivência dado o vetor de covariáveis $\mathbf{x}_l$ é:

$$S(t \mid \mathbf{x}_l) = \exp(-\int_0^t h(u \mid \mathbf{x}) du) = [S_0(t)]^{\exp(\boldsymbol{\beta}' \mathbf{x}_l)}, \tag{14}$$

sendo $S_0(t)$ a função de sobrevivência de base.

Assim, a partir de (14), $p_i(\mathbf{x}_l)$ pode ser escrita como:

$$p_i(\mathbf{x}_l) = 1 - \left[\frac{S_0(a_i)|\mathbf{x}_l}{S_0(a_{i-1}|\mathbf{x}_l)} \right]^{\exp(\boldsymbol{\beta}' \mathbf{x}_l)},$$

ou ainda,

$$p_i(\mathbf{x}_l) = 1 - \gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)}, \tag{15}$$

em que $\gamma_i = S_0(a_i|\mathbf{x}_l)/S_0(a_{i-1}|\mathbf{x}_l)$.

Substituindo (15) em (13), a função de verossimilhança fica da seguinte forma:

$$L = \prod_{i=1}^{k} \prod_{l \in R_i} (1 - \gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)})^{\Delta_{li}} (\gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)})^{1-\Delta_{li}}.$$

O logaritmo da função de verossimilhança é então dado por:

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{k} \sum_{l \in R_i} (\Delta_{li} \ln(1 - \gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)}) + (1 - \Delta_{li}) \ln(\gamma_i^{\exp(\boldsymbol{\beta}'\mathbf{x}_l)})).$$

Como os $p_i(\mathbf{x})$ são restritos ao intervalo [0,1], é conveniente se utilizar a reparametrização $\gamma_i^* = \ln(-\ln(\gamma_i))$ sugerida por Prentice & Gloeckler (1978), sendo os $\gamma_i^* = \gamma_1^*, \ldots, \gamma_k^*$ irrestritos melhorando a convergência no processo iterativo para a estimação desses parâmetros. Assim, o logaritmo da função de verossimilhança fica:

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{k} \sum_{l \in R_i} [-(1 - \Delta_{li}) \exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l) + \Delta_{li} \ln(1 - \exp(-\exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l)))],$$

em que $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)'$.

As equações de verossimilhança são dadas por:

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \beta_j} &= \sum_{i=1}^{k} \sum_{l \in R_i} \left[\frac{-\Delta_{li} x_{lj} \exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l)}{\exp(\exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l)) - 1} - (1 - \Delta_{li}) x_{lj} \exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l) \right] \\
\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \gamma_i^*} &= \sum_{l \in R_i} \left[\frac{\Delta_{li} \exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l)}{\exp(\exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l)) - 1} - (1 - \Delta_{li}) \exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l) \right],
\end{aligned}$$

com $j = 1, \ldots, p$ e $i = 1, \ldots, k$, que quando igualadas a zero podem ser resolvidas por um método iterativo.

3.1.2 Modelo logístico

Adotando-se o modelo logístico, tem-se que:

$$p_i(\mathbf{x}_l) = 1 - (1 + \gamma_i \exp(\boldsymbol{\beta}'\mathbf{x}_l))^{-1}, \tag{16}$$

em que $\gamma_i = p_i(0)/(1 - p_i(0))$, $i = 1, \ldots, k$.

Assim, escrevendo-se a função de verossimilhança (13) considerando-se (16) tem-se:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{k} \prod_{l \in R_i} \left(\frac{\gamma_i \exp(\boldsymbol{\beta}'\mathbf{x}_l)}{1 + \gamma_i \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right)^{\Delta_{li}} \left(\frac{1}{1 + \gamma_i \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right)^{1-\Delta_{li}}.$$

Usando-se a reparametrização $\gamma_i^* = \ln(\gamma_i)$, sugerida por Lawless (1982) para melhorar a convergência e aproximar a distribuição assintótica dos estimadores de máxima verossimilhança para a distribuição normal, o logaritmo da função de verossimilhança fica:

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \ln(L(\boldsymbol{\beta}, \boldsymbol{\gamma})) = \sum_{i=1}^{k} \sum_{l \in R_i} \Delta_{li}(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l) - \ln(1 + \exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l))$$

e as equações de verossimilhança são dadas por:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \beta_j} &= \sum_{i=1}^{k} \sum_{l \in R_i} x_{lj} \Delta_{li} - \left[\frac{x_{lj} \exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l)}{1 + \exp(\alpha_i + \boldsymbol{\beta}'\mathbf{x}_l)} \right], \\ \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \gamma_i} &= \sum_{i=1}^{k} \sum_{l \in R_i} \Delta_{li} - \left[\frac{\exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l)}{1 + \exp(\gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l)} \right], \end{aligned}$$

em que $j = 1, \ldots, p$ e $i = 1, \ldots, k$, que quando igualadas a zero podem ser resolvidas usando um processo iterativo.

3.2 Utilização dos modelos lineares generalizados para o ajuste dos modelos logístico e de Cox

Whitehead (1989) ajustou o modelo de riscos proporcionais de Cox para dados com censura intervalar utilizando a transformação *complemento log-log* a $1 - p_i(\mathbf{x}_l)$, dado por (14), e considerando a variável indicadora de falha, Δ_{li}, como a variável resposta.

Assim, aplicando-se a transformação *complemento log-log* em (15), tem-se:

$$\ln(-\ln(1 - p_i(\mathbf{x}_l))) = \boldsymbol{\beta}'\mathbf{x}_l + \ln(-\ln(\gamma_i)) \qquad i = 1, \ldots, k \qquad l = 1, \ldots, n.$$

ou ainda,

$$\ln(-\ln(1 - p_i(\mathbf{x}_l))) = \boldsymbol{\beta}'\mathbf{x}_l + \gamma_i^* \qquad i = 1, \ldots, k \qquad l = 1, \ldots, n$$

em que $\gamma_i^* = \ln(-\ln(\gamma_i))$ são parâmetros incorporados ao modelo para ajustar os termos correspondentes aos k-níveis do fator associado ao intervalo de tempo.

No caso do ajuste do modelo logístico, a transformação utilizada é a *logit* e a variável resposta também é a variável indicadora de censura, Δ_{li}.

Assim, a transformação *logit* aplicada em (16), resulta em:

$$\ln\left(\frac{p_i(\mathbf{x}_l)}{1 - p_i(\mathbf{x}_l)}\right) = \ln(\gamma_i) + \boldsymbol{\beta}'\mathbf{x}_l = \gamma_i^* + \boldsymbol{\beta}'\mathbf{x}_l = \eta_{li},$$

em que $\gamma_i^* = \ln(\gamma_i)$ corresponde ao efeito de intervalo de tempo.

3.3 Planos de reamostragem

Em modelos de regressão ou em modelos seguindo algum tipo de delineamento experimental, uma maneira prática de se fazer replicações *bootstrap* é reamostrar os resíduos, de acordo com Davidson & Hinkey (1997).

No presente trabalho, inicialmente, será feito o ajuste dos modelos de Cox e logístico para dados com censura intervalar com o auxílio dos modelos lineares generalizados. A fim de verificar a possiblidade de selecionar um dos modelos citados, o procedimento a ser seguido, considerando M_c e M_l como os modelos de riscos proporcionais de Cox e o logístico, respectivamente, será:

1. Ajustam-se os modelos M_c e M_l com os dados da amostra original e calcula-se a diferença de *deviances* entre os dois modelos, dada por:

$$d_{obs} = D_c - D_l$$

em que D_c e D_l são as *deviances* referentes aos ajustes dos modelos M_c e M_l, respectivamente. Com os parâmetros estimados através do ajuste inicial de M_c, obtêm-se os resíduos e reamostram-se esses resíduos um número B de vezes, obtendo-se B resíduos *bootstrap*. Com esses B resíduos *bootstrap*, geram-se as novas observações *bootstrap* como descritas em **(3.3.1)**. Novamente, faz-se o ajuste dos modelos M_c e M_l aos dados das amostras *bootstrap*, calculando-se as diferenças de *deviances*, dada agora por:

$$d_{cb} = D_c - D_l$$

para $b = 1, \ldots, B$.

2. De modo análogo, com os parâmetros estimados através do ajuste inicial de M_l, obtêm-se os resíduos e reamostram-se esses resíduos um número B de vezes, obtendo-se B resíduos *bootstrap*. Com esses B resíduos, obtêm-se as

novas observações *bootstrap* de forma análoga ao item anterior. A seguir, faz-se o ajuste para M_c e M_l com os novos dados obtidos da amostra *bootstrap*, calculando-se as diferenças de *deviances*, agora dada por:

$$d_{lb} = D_c - D_l$$

para $b = 1, \ldots, B$.

3. O valor de d_{obs} será, então, comparado aos valores obtidos das diferenças de *deviances* $\{d_{cb}\}$ e $\{d_{lb}\}$ obtidos com a simulação *bootstrap*. Com o objetivo de verificar a evidência de um modelo em relação a outro, serão construídos histogramas para verificar a distribuição de $\{d_{cb}\}$ e $\{d_{lb}\}$ e serão propostos os respectivos níveis de significância empíricos, dados por:

   - nível de significância $= \frac{1}{B}\sum_{b=1}^{B} I(d_{cb} > d_{obs})$ para o modelo de Cox
   - nível de significância $= \frac{1}{B}\sum_{b=1}^{B} I(d_{lb} > d_{obs})$ para o modelo logístico

   em que $I(.)$ é a função indicadora do número de vezes que as diferenças de *deviances* de ambos os modelos supera a d_{obs}.

Caso a diferença de *deviances* d_{obs} seja negativa, os níveis de significância serão calculados apenas trocando o sinal de $>$ (maior) pelo sinal de $<$ (menor), ou seja, nesse caso deverá ter uma maior concentração de valores negativos na distribuição empírica da diferença de *deviances* para que o modelo considerado como verdadeiro (hipótese nula) não seja rejeitado.

A simulação *bootstrap* será realizada com a reamostragem dos seguintes resíduos:

- **Resíduo simples**

$$r_i = y_i - \hat{\mu}_i, \qquad i = 1, 2, \ldots, n,$$

em que $\hat{\mu}_i$ são os valores ajustados;

- **Resíduo de Pearson**

$$r_{Pi} = \frac{y_i - \hat{\mu}_i}{[V(\mu_i)]^{1/2}}, \qquad i = 1, 2, \ldots, n,$$

em que $V(\mu_i)$ é a função de variância;

- **Resíduo de Pearson padronizado**

$$r_{PPi} = \frac{y_i - \hat{\mu}_i}{[c_i \hat{k} V(\hat{\mu}_i)(1 - h_i)]^{1/2}}, \qquad i = 1, 2, \ldots, n.$$

em que os valores h_i são os elementos da diagonal da matriz H dada por $H = W^{(1/2)} X (X'WX)^{-1} X' W^{1/2}$, c_i são pesos conhecidos e k é o parâmetro de dispersão que pode ser desconhecido. Para dados binomiais com probabilidade $\pi(x_i)$ e denominador m_i, $c_i = 1/m_i$ e k é constante e igual a 1.

Os planos de reamostragem serão, então, elaborados com base nesses resíduos e, para medir a qualidade da discrepância entre os dois modelos, será analisada a diferença de *deviances*, observando-se em direção a qual dos modelos ajustados seu valor está situado. Programas em S-Plus, V. 3.3, serão elaborados para a simulação *bootstrap*.

3.3.1 Obtenção das amostras *bootstrap*

A obtenção das amostras *bootstrap* está baseada em Davison & Hinkley (1997), considerando cada tipo de resíduo e está definida como segue:

- **Resíduo simples**

$$y^*_{boot} = \hat{y} + \varepsilon^*_i \tag{17}$$

em que $\hat{\mu}$ é o valor ajustado e ε_i^* é uma amostra *bootstrap* dos resíduos simples sob os modelos a serem considerados (M_c ou M_l).

- **Resíduo de Pearson**

$$y^*_{bootPi} = \hat{\mu} + V(\hat{\mu}_i)\varepsilon^*_{Pi} \tag{18}$$

em que ε^*_{Pi} é uma amostra *bootstrap* dos resíduos de Pearson sob os modelos a serem considerados (M_c ou M_l).

- **Resíduo de Pearson Padronizado**

$$y^*_{bootPPi} = \hat{\mu} + c_i\hat{k}V(\hat{\mu}_i)(1-h_i)]^{1/2}\varepsilon^*_{PPi} \tag{19}$$

em que ε^*_{PPi} é uma amostra *bootstrap* dos resíduos de Pearson padronizados sob os modelos a serem considerados (M_c ou M_l).

3.4 Aplicações

Como aplicação da metodologia proposta neste trabalho, serão utilizados dois conjuntos de dados resultantes de dois experimentos agronômicos, sendo o primeiro da área de Fitopatologia e o segundo da Entomologia e, pelo fato dos dois conjuntos de dados apresentarem alta proporção de empates, serão analisados pelos modelos de tempo de vida discretos de Cox e logístico.

3.4.1 Aplicação em Fitopatologia

O experimento foi realizado com um cultivar de linho, susceptível ao patógeno *Fusarium oxysporum*, plantado sobre quatro substratos: M: solo natural

(controle), MI:M + ilita, MK:M + caolinita e MM:M + montmorilonita. Este conjunto de dados foi estudado e apresentado em Chalita (1997). Neste experimento à suscetibilidade dos cultivares à murcha causada pelo patógeno, foi observada através do tempo até a planta murchar. As avaliações foram feitas aos 21, 24, 28, 31, 35, 38, 42, 45 e 49 dias gerando uma proporção de empates de 27%. Portanto, o tempo pode ser considerado como uma variável discreta e os modelos de Cox e logístico podem ser ajustados.

3.4.2 Aplicação em Entomologia

O experimento foi realizado pelas pesquisadoras Elisabeth de Nardo e Maria Aico Watanabe no laboratório de Entomologia da EMBRAPA/Meio ambiente - Jaguariúna - SP com o objetivo de avaliar o efeito de uma formulação da *Anticarsia gemmatalis Nuclear Polyhedrosis Virus* (AgNPV) sobre o predador *Podysus nigrispinus Dallas*. A *Anticarsia gemmatalis* é uma praga que ataca a soja e o seu principal predador é o inseto *Podysus nigrispinus*, que realiza um controle natural desta praga. Alternativamente, agentes entomopatogênicos ou patógenos (virus, bactérias, fungos e nematóides) estão sendo bastante utilizados no controle biológico de pestes e são considerados menos agressivos para o ambiente do que pesticidas químicos, embora tenham a desvantagem de infectarem outros organismos que são úteis ao ambiente. O risco associado com esses agentes de biocontrole tem aumentado consideravelmente e, portanto, estão sendo motivo de estudo em todo o mundo.

Neste experimento foram utilizados dois tratamentos, um com a larva infectada com AgNPV (grupo tratado) e o outro com a larva sadia (grupo controle), o interesse é avaliar se o virus atinge o predador influenciando no seu tempo de vida. Outro fator considerado foi o sexo dos insetos *Podysus nigrispinus*. A variável observada foi o tempo em que o predador atingia a fase adulta e os tempos avaliados foram 24, 25, 26, 27 e 28 dias. Devido aos poucos intervalos e aos empates, o tempo

pode ser considerado como uma variável discreta e novamente os modelos de Cox e logístico podem ser ajustados.

Em ambos os casos serão ajustados os modelos de Cox e logístico, realizar-se-á uma simulação *bootstrap*, além de uma discussão sobre a escolha do modelo e da metodologia empregada.

4 RESULTADOS E DISCUSSÃO

4.1 Aplicação com cultivar de linho

A fim de avaliar a sobrevivência do cultivar de linho nos quatro substratos, foi realizada uma análise descritiva dos dados através do teste *Log-rank* e construiu-se as curvas de sobrevivência para cada susbtrato considerando o estimador tabela de vida. O teste *log-rank* forneceu um $p-valor = 0{,}0002$, evidenciando assim, diferença significativa entre os tipos de substratos. Pela análise da Figura 1, pode-se verificar que o substrato M (solo natural) é o que apresenta menor resistência à murcha causada pelo patógeno.

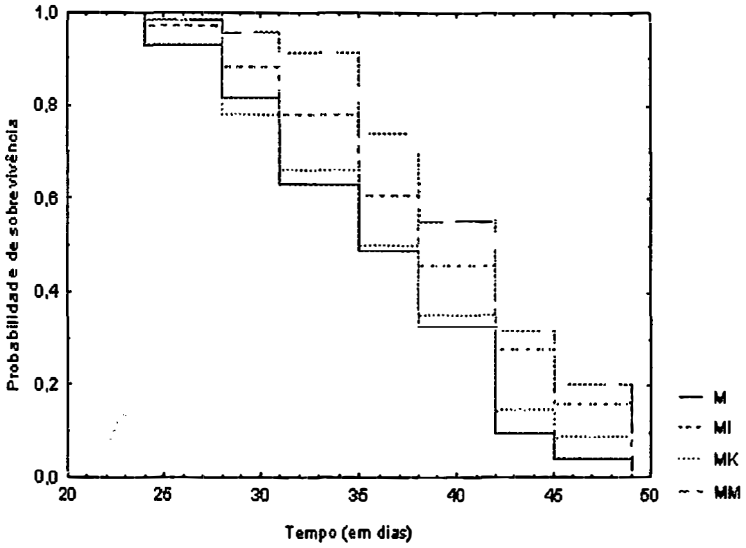

Figura 1 - Curvas de sobrevivência estimadas pela tabela de vida versus tempo de vida em dias para cada substrato.

A seguir, foram ajustados a esses dados os modelos de Cox e logístico. A Tabela 4 mostra a análise de *deviances* obtida para o ajuste dos dois modelos.

Tabela 4. Análise de *deviances* do experimento com o cultivar de linho.

Causas de Variação	GL	*Deviances*	
		Modelo de Cox	Modelo Logístico
Nula	34	225,894	225,894
Intervalo	8	188,427**	188,427**
Solo	3	17,378**	18,253**
Residual	23	20,089	19,214

** - significativo ao nível de 1%

Pela análise da Tabela 4, pode-se comprovar que tanto o efeito de intervalo quanto o de solo foram significativos para os dois modelos considerados. Assim, para o modelo de Cox tem-se:

$$\ln[-\ln(1-p_i(\mathbf{x}_l))] = \gamma_i^* + \alpha_j,$$

e, portanto,

$$p_i(\mathbf{x}_l) = 1 - [exp(-exp(\gamma_i^*))]^{exp(\alpha_j)},$$

e, para o modelo logístico, tem-se:

$$ln\left[\frac{p_i(\mathbf{x}_l)}{1-p_i(\mathbf{x}_l)}\right] = \gamma_i^* + \alpha_j,$$

e, assim

$$p_i(\mathbf{x}_l) = 1 - [1 + exp(\gamma_i^* + \alpha_j)]^{-1}$$

em que:

- $p_i(\mathbf{x}_l)$ é a probabilidade da l-ésima planta murchar no i-ésimo intervalo, dado que não murchou até o intervalo anterior, considerando cada tipo de solo.
- γ_i^* é o efeito do i-ésimo intervalo, $i = 1, \ldots, 10$;
- α_j é o efeito do j-ésimo substrato, $j = 1, \ldots, 4$.

A Tabela 5 apresenta as estimativas dos parâmetros com seus respectivos erros padrões, segundo os dois modelos considerados.

Tabela 5. Estimativas dos parâmetros obtidas através dos ajustes dos modelo de Cox e logístico para os dados do experimento com cultivar de linho.

Parâmetros	Modelo de Cox		Modelo Logístico	
	estimativas	erro padrão	estimativa	erro padrão
γ_1^*	-3,1356	0,3455	-3,0650	0,3567
γ_2^*	-1,8503	0,3799	-1,7335	0,3902
γ_3^*	-1,6775	0,3784	-1,5489	0,3898
γ_4^*	-1,1102	0,3655	-0,9108	0,3787
γ_5^*	-0,8603	0,3663	-0,6244	0,3825
γ_6^*	-0,1089	0,3594	0,2950	0,3851
γ_7^*	-0,4282	0,3931	-0,1052	0,4265
γ_8^*	0,2280	0,3942	0,7773	0,4577
γ_9^*	0,2118	0,4831	0,7513	0,6002
α_2^*	-0,5342	0,1748	-0,6327	0,2092
α_3^*	-0,1758	0,1735	-0,2075	0,2088
α_4^*	-0,6375	0,1734	-0,7845	0,2080

A razão de risco é obtida pelo cálculo da $exp(\alpha_j), j = 1, \ldots, 4$ e encontra-se na Tabela 6 juntamente com o $p - valor$ correspondente para cada modelo considerando os substratos M, MI e MM. Na comparação entre M e MI, tem-se que o risco de uma planta no substrato MI murchar é 0,9311 vezes o das plantas do substrato M considerando o modelo de Cox. Como o valor da razão de risco é inferior a 1, é conveniente calcular o inverso deste valor. Assim, o valor do risco de uma planta do substrato M murchar é 1,88 vezes o das plantas do substrato MI. Já para o modelo logístico, o risco obtido foi de 1,71, bastante próximo ao obtido pelo ajuste do modelo de Cox. O risco para os substratos M e MM foi de 2,19 (modelo de Cox) e 1,89 (modelo logístico), e a interpretação desses resultados é análoga aos anteriores. Com relação aos substratos M e MK, não foi constatada diferença significativa entre eles, uma vez que foi encontrado um $p - valor$ de 0,3204 e 0,3109 para os modelos de Cox e logístico, respectivamente.

Tabela 6. Razão de risco e $p - valor$ segundo o modelo adotado - Experimento com o cultivar de linho.

Parâmetros	Razão de risco		$p - valor$	
	Modelo de Cox	Modelo Logístico	Modelo de Cox	Modelo Logístico
M/MI	1,88	1,71	0,0025	0,0022
M/MM	2,19	1,89	0,0002	0,0002

O critério de informação de Akaike (AIC) foi utilizado para comparar os modelos de Cox e logístico e os valores encontrados foram, respectivamente, 159,2 e 160,1, ou seja, os valores obtidos foram muito próximos, indicando que ambos os ajustes podem ser considerados razoáveis, não evidenciando um melhor ajuste de um modelo em relação a outro.

As curvas de sobrevivência estimadas para cada modelo considerado são mostradas nas Figuras 2 e 3. Novamente, pode-se notar que os gráficos são bastante similares.

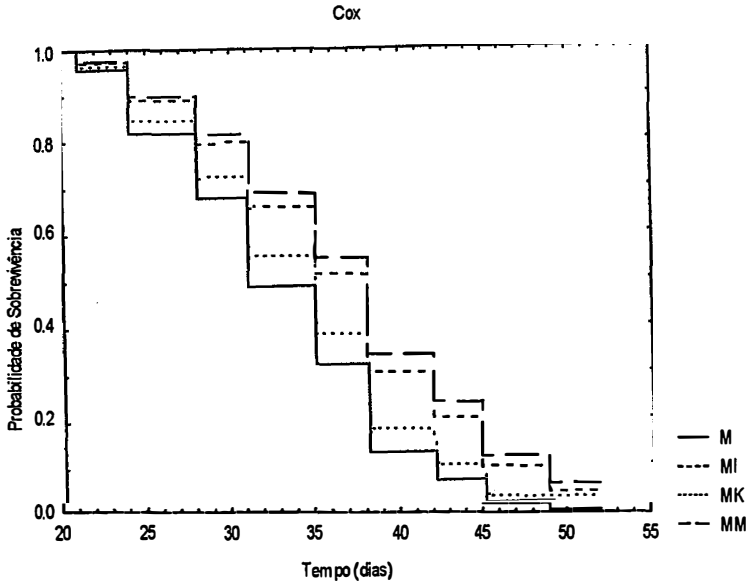

Figura 2 - Curva de sobrevivência segundo o tipo de solo ajustada para o modelo de Cox.

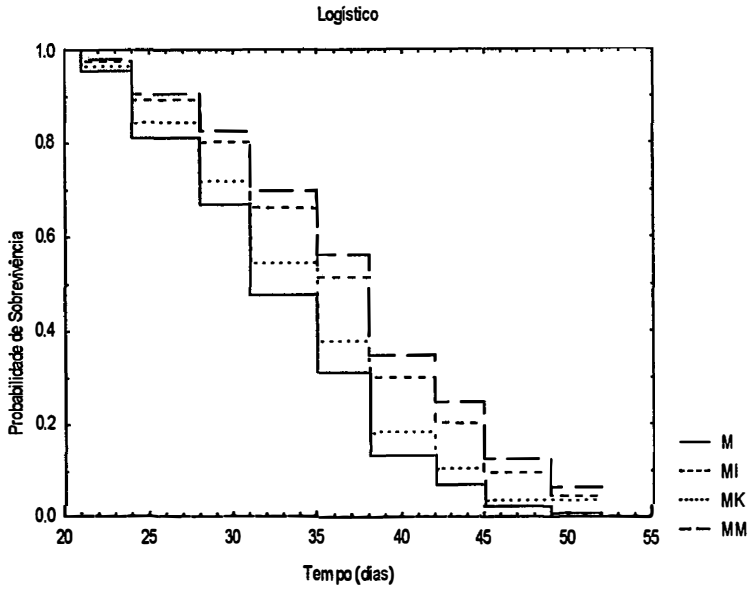

Figura 3 - Curva de sobrevivência segundo o tipo de solo ajustada para o modelo logístico.

Desse modo, analisando os ajustes dos modelos de Cox e logístico através das *deviances*, das estimativas dos parâmetros e respectivos erros padrões dados na Tabela 5, das razões de risco (Tabela 6), das curvas de sobrevivência estimadas e apresentadas nas Figuras 2 e 3 e pelo Critério de Informação de Akaike, praticamente não se pode dizer que o modelo logístico forneceu um melhor ajuste em relação ao modelo de Cox, uma vez que todos os resultados obtidos foram muito similares. Neste caso, qualquer um dos dois modelos poderia ser adotado para evidenciar a significância dos fatores considerados (intervalo e solo).

Apesar disso, é sempre interessante eleger um dos modelos como o que melhor se adequa aos dados e, em termos da *deviance* residual, o modelo logístico forneceu o menor valor, indicando um melhor ajuste.

Assim, como uma outra forma de evidenciar este fato, foi aplicada a metodologia proposta, calculando-se a diferença de *deviances* entre os modelos, $D_{obs} = D_c - D_l = 0,87$, que é um valor positivo e baixo. Desse modo, supondo-se que o modelo logístico é o verdadeiro, fez-se o ajuste para os dois modelos considerados e foram extraídos três tipos de resíduos: simples, de Pearson e de Pearson padronizados, cujas fórmulas foram apresentadas na seção anterior. O mesmo foi feito considerando o modelo de Cox como verdadeiro.

De posse desses resíduos, foram feitas 1000 replicações *bootstrap* recompondo as observações, agora chamadas de observações *bootstrap*, e novamente os dois modelos foram ajustados para cada caso e os níveis de significância foram calculados. Como o valor da D_{obs} foi positivo, para que o modelo suposto seja verdadeiro é de se esperar que o número de vezes que as *deviances* calculadas utilizando as observações *bootstrap* supere a *deviance* observada seja grande, evidenciando que a *deviance* do modelo suposto verdadeiro seja cada vez menor em relação a do outro modelo. Os valores dos níveis de significância para cada modelo são mostrados na Tabela 7.

Tabela 7. Níveis de significância obtidos através de 1000 replicações *bootstrap* segundo o tipo de resíduo e os modelos de Cox e logístico para os dados do experimento com o cultivar de linho.

Tipos de resíduos	Modelos	
	Modelo de Cox	Modelo Logístico
Simples	0,245	0,281
Pearson	0,195	0,328
Pearson padronizado	0,186	0,313

Analisando a Tabela 7, nota-se que, para os três resíduos considerados , o modelo logístico sempre apresenta um nível de significância maior, evidenciando o resíduo de Pearson dentre os demais, embora a diferença entre este e o resíduo de Pearson padronizado seja pequena, o que mostra que a ponderação pela *leverage* praticamente não alterou os resultados. As Figuras 4, 5 e 6, mostram a distribuição empírica da d_{obs} após 1000 replicações *bootstrap*, considerando os resíduos simples, de Pearson e Pearson padronizados, para os modelos de Cox e logístico, respectivamente. Em todas as três figuras nota-se que a área acima da d_{obs} é maior para o modelo logístico.

Desse modo, através desta análise, tem-se novamente a indicação de que o modelo logístico é o mais adequado para ajustar esses dados.

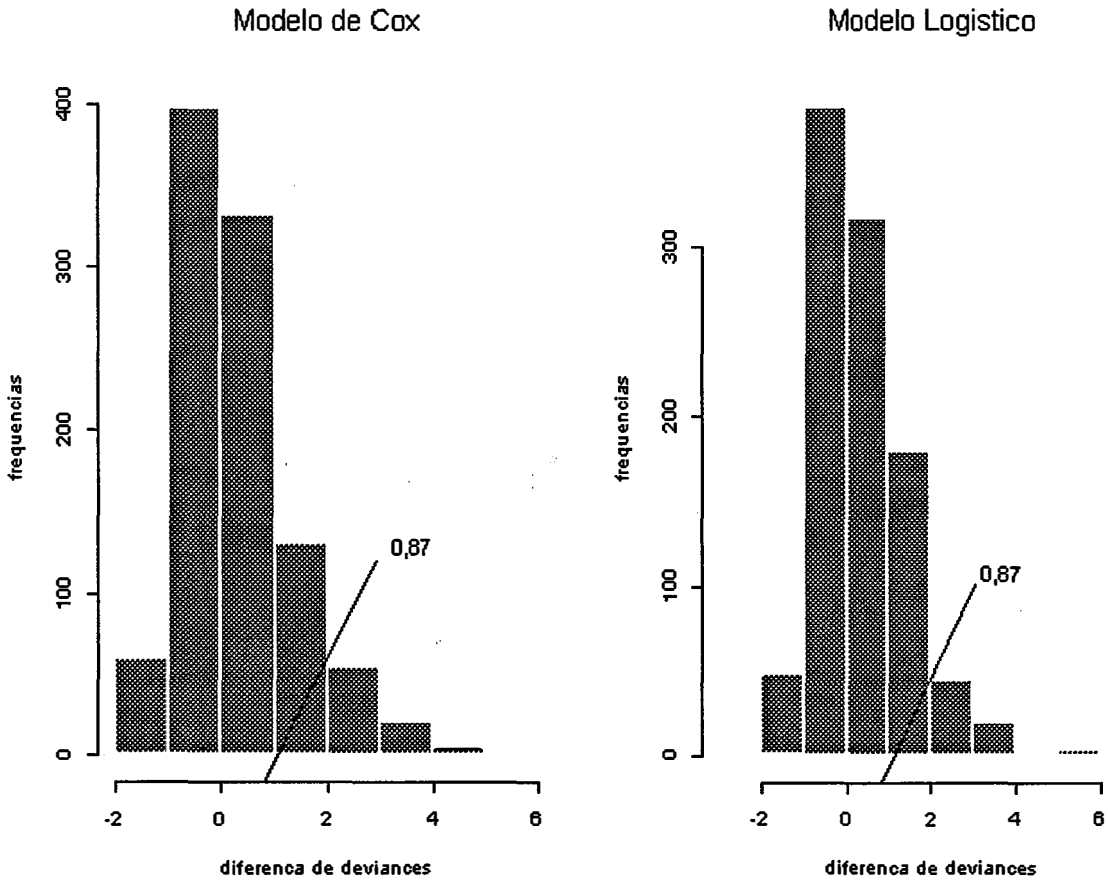

Figura 4 - Histograma da distribuição da diferença de *deviances* obtidas através de 1000 replicações *bootstrap* considerando os resíduos simples no ajuste dos modelos de Cox e logístico para os dados do experimento com cultivar de linho.

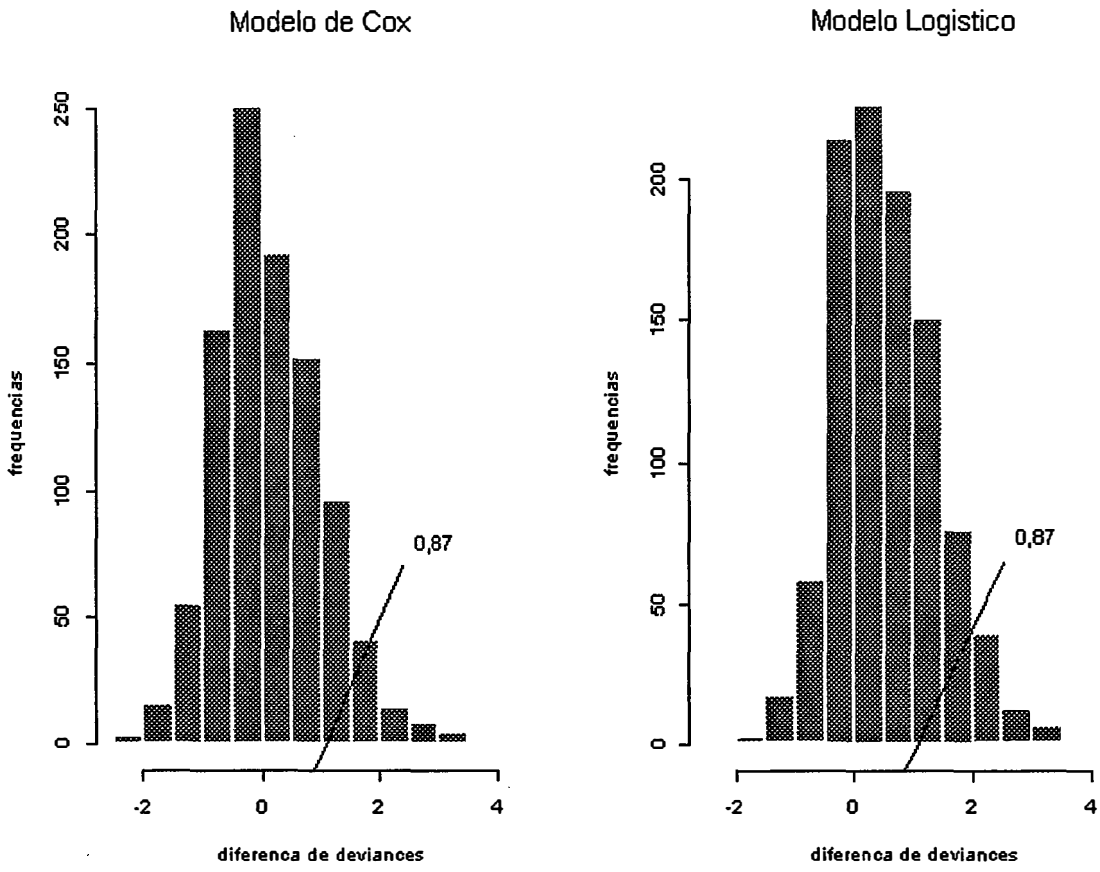

Figura 5 - Histograma da distribuição da diferença de *deviances* obtidas através de 1000 replicações *bootstrap* considerando os resíduos de Pearson no ajuste dos modelos de Cox e logístico para os dados do experimento com cultivar de linho.

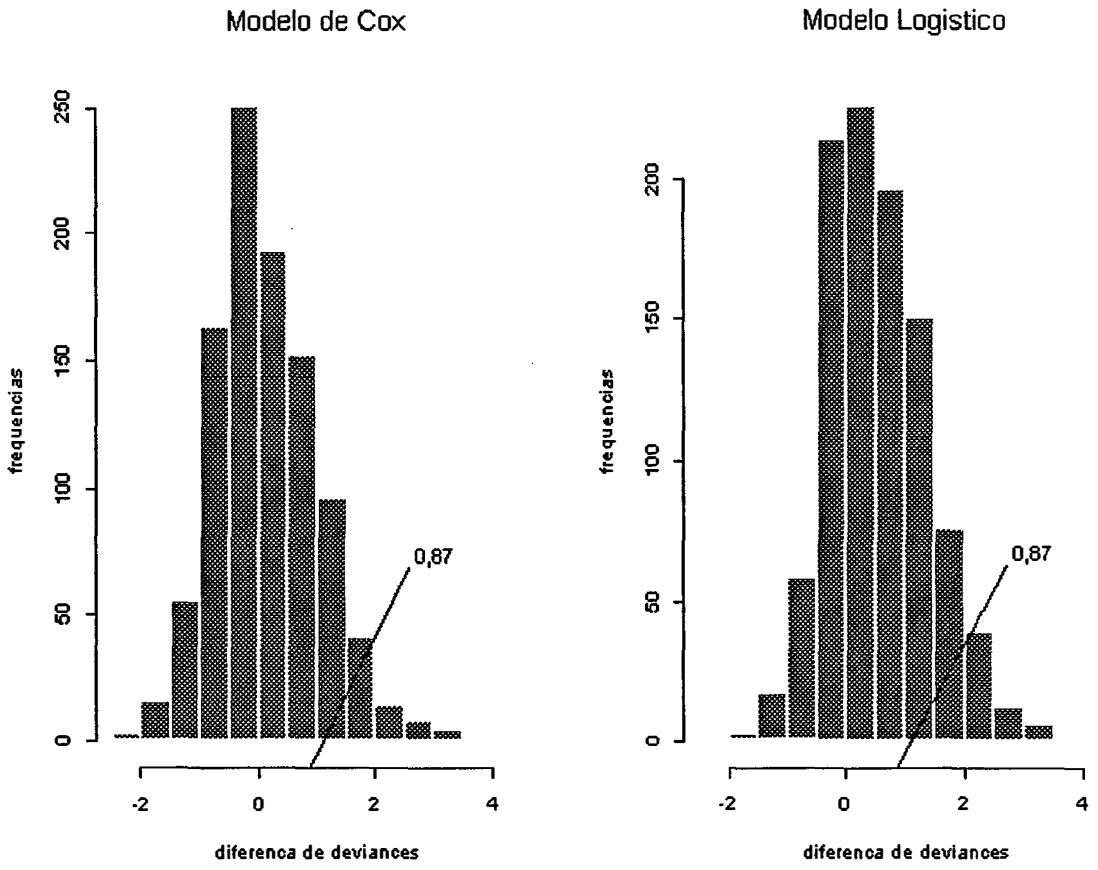

Figura 6 - Histograma da distribuição da diferença de *deviances* obtidas através de 1000 replicações *bootstrap* considerando os resíduos de Pearson padronizados no ajuste dos modelos de Cox e logístico para os dados do experimento com cultivar de linho.

4.2 Aplicação com inseto

Neste experimento os fatores envolvidos foram *intervalo*, *grupo*, que consistia de duas dietas dadas ao inseto *Podysus nigrispinus*, sendo uma constituída da larva infectada com AgNPV (grupo tratado) e a outra com a larva sadia (grupo controle) e *sexo*. O interesse foi avaliar se esses fatores influenciaram na probabilidade do inseto *Podysus nigrispinus* chegar a fase adulta em determinado intervalo, dado que não chegou no intervalo anterior. Devido aos poucos intervalos e aos empates, o tempo pode ser considerado como uma variável discreta e novamente os modelos de Cox e logístico podem ser ajustados. Analisando-se os resultados dos ajustes dos modelos para este experimento (Tabela 8), conclui-se que os fatores *intervalo* e *sexo* foram significativos para os dois modelos, mas o fator *grupo* foi significativo apenas para o modelo de Cox. A interação *sexo*grupo* não foi significativa em nenhum dos dois modelos considerados.

Tabela 8. Análise de *deviances* do experimento com inseto para os modelos de Cox e logístico.

Causas de Variação	GL	*Deviances*	
		Modelo de Cox	Modelo Logístico
Nula	15	162,373	162,373
Intervalo	4	140,738**	140,738**
Sexo	1	8,463**	9,612**
Grupo	1	4,104*	$2{,}006^{ns}$
Sexo*Grupo	1	$0{,}725^{ns}$	$0{,}411^{ns}$
Residual	8	8,343	9,606

** - significativo a 1%, * - significativo a 5% , ns - não significativo

Portanto, as *deviances* residuais para os ajustes finais dos modelos de Cox (retirando-se a interação *sexo*grupo*) e logístico (retirando-se a interação *sexo*grupo* e o fator *grupo*) são 9,068 e 12,023, respectivamente. Portanto a diferença

de *deviances* observada, neste caso, foi de $d_{obs} = -2,955$. Os valores dos AIC para os modelos de Cox e logístico foram respectivamente 59,829 e 60,784. A comparação entre as *deviances* e o AIC indicam que o modelo de Cox se adequa melhor aos dados.

Com base nestes resultados pode-se escrever $p_i(\mathbf{x}_l)$ para o modelo de Cox como:

$$p_i(\mathbf{x}_l) = 1 - [exp(-exp(\gamma_i^*))]^{exp(\alpha_j + \tau_r)},$$

em que:

- $p_i(\mathbf{x}_l)$ é a probabilidade do l-ésimo inseto chegar a fase adulta no i-ésimo intervalo, dado que não chegou até o intervalo anterior, considerando-se os fatores *sexo* e *grupo*.
- γ_i^* é o efeito do i-ésimo intervalo, $i = 1, \ldots, 5$;
- α_j é o efeito do j-ésimo *sexo*, $j = 1, 2$;
- τ_r é o efeito do r-ésimo *grupo*, $r = 1, 2$,

e, para o modelo logístico:

$$p_i(\mathbf{x}_l) = 1 - [1 + exp(\gamma_i^* + \alpha_j)]^{-1}$$

sendo:

- $p_i(\mathbf{x}_l)$ é a probabilidade do l-ésimo inseto chegar a fase adulta no i-ésimo intervalo, dado que não chegou até o intervalo anterior, considerando-se o fator *sexo*;
- γ_i^* é o efeito do i-ésimo intervalo, $i = 1, \ldots, 5$;
- α_j é o efeito do j-ésimo *sexo*, $j = 1, 2$.

As Figuras 7 e 8 apresentam as curvas de sobrevivência estimadas para o fator *sexo* segundo os modelos de Cox e logístico e a Figura 9 mostra a curva de sobrevivência estimada para o fator *grupo* considerando apenas o modelo de Cox, uma vez que este fator não foi significativo para o modelo logístico.

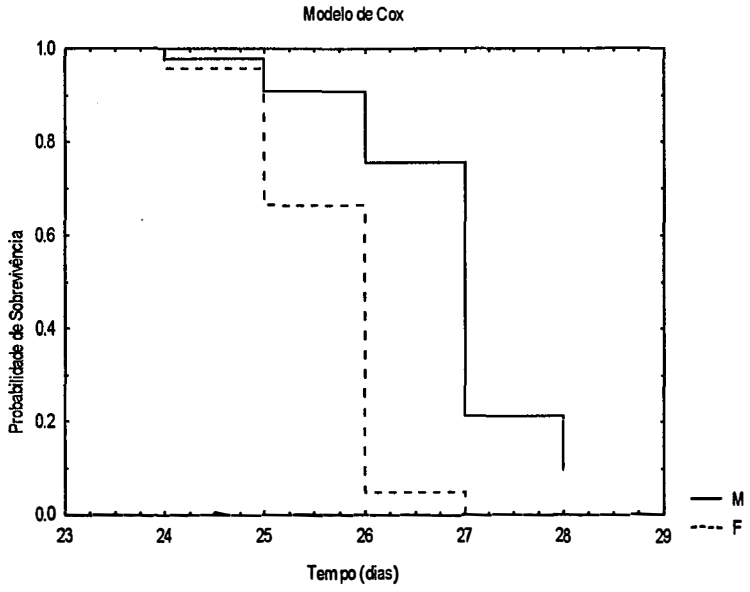

Figura 7 - Curva de sobrevivência estimadas segundo o sexo - Modelo de Cox.

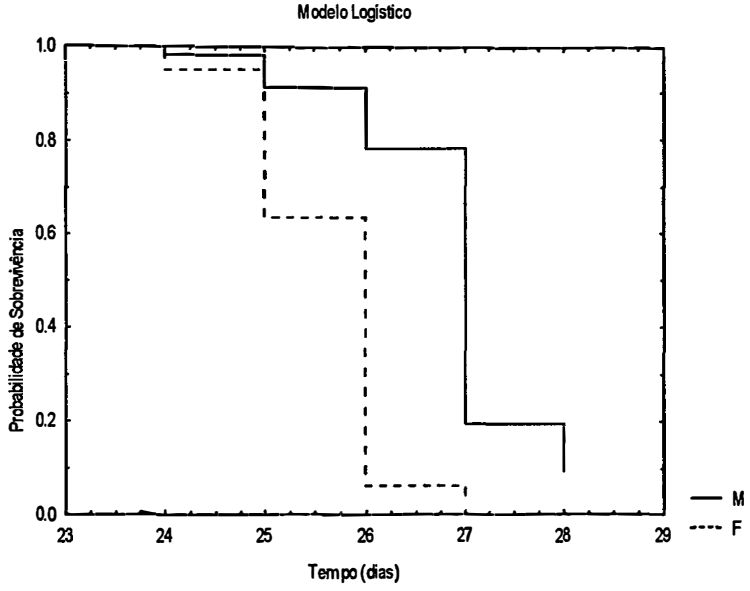

Figura 8 - Curva de sobrevivência segundo o sexo - Modelo logístico.

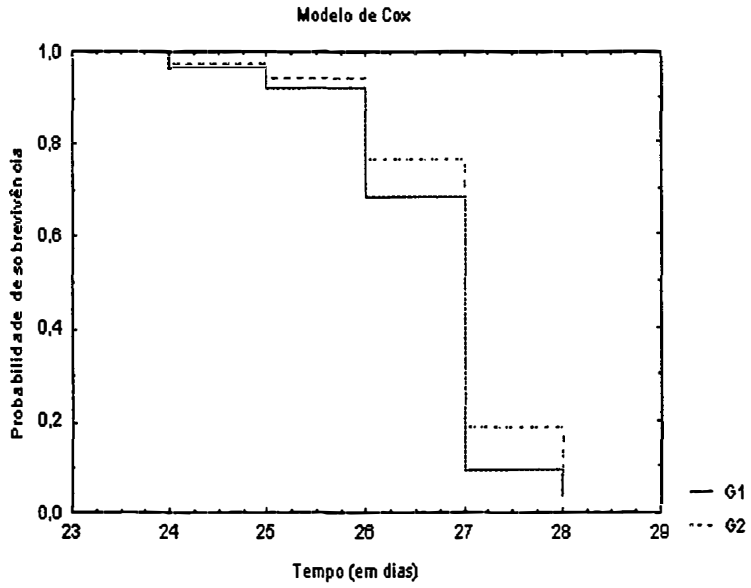

Figura 9 - Curva de sobrevivência segundo o grupo - Modelo de Cox.

As Tabelas 9 e 10 apresentam as estimativas dos parâmetros e seus erros padrões, segundo os modelos estudados.

Tabela 9. Estimativas dos parâmetros obtidas através do ajuste do modelo de Cox para os dados do experimento com insetos.

Parâmetros	estimativas	erro padrão
γ_1^*	-2,4260	0,6231
γ_2^*	-1,6228	0.7718
γ_3^*	-0,7478	0.6233
γ_4^*	1,3073	0.5990
γ_5^*	0,8512	0,6862
α_2	-0.7942	0,2551
τ_2	-0,4820	0,2380

Tabela 10. Estimativas dos parâmetros obtidas através do ajuste do modelo logístico para os dados do experimento com insetos.

Parâmetros	estimativas	erro padrão
γ_1^*	-2,5273	0,6227
γ_2^*	-1,4278	0.8112
γ_3^*	-0,6629	0.6458
γ_4^*	2,2519	0.6755
γ_5^*	1,2225	0.8210
α_2	-1,1562	0,3876

Com base nesses resultados pode-se calcular as razões de risco para os fatores considerados, de acordo com o modelo adotado. A Tabela 11 apresenta as razões de risco e o $p - valor$ correspondente. De acordo com os modelos adotados tem-se, por exemplo, que o risco de uma fêmea atingir a idade adulta é 2,21 vezes o risco dos machos, quando considerado o modelo de Cox. Já para o modelo logístico esse mesmo risco é de 3,18.

Com relação ao fator *grupo* (tratamento), que foi significativo apenas no modelo de Cox, tem-se que o risco de um inseto que se alimentou da larva sadia (grupo controle) atingir a fase adulta é de 1,62 vezes mais que o dos insetos alimentados da larva infectada com AgNPV (grupo tratado).

Tabela 11. Razão de risco e $p - valor$ segundo o modelo adotado.

Parâmetros	Razão de risco		$p - valor$	
	Modelo de Cox	Modelo Logístico	Modelo de Cox	Modelo Logístico
F/M	2,21	3,18	0,0018	0,0028
C/T	1,62	-	0,0428	-

Para discriminar entre os modelos, trabalhou-se inicialmente com o modelo logístico incluindo o fator grupo, embora este fator não tenha sido significativo para este modelo. Neste caso a $d_{obs} = -0,949$ e, portanto, se espera que o modelo mais adequado apresente uma maior concentração de valores à esquerda da d_{obs} na distribuição empírica da diferença de *deviances* do que o outro modelo. Pela análise das Figuras 10, 11 e 12 percebe-se que o modelo de Cox é então o mais adequado.

Na Tabela 12 encontram-se os níveis de significância empíricos considerando os tipos de resíduos e os modelos estudados. De acordo, com os resultados obtidos há uma indicação de que o modelo de Cox se adequa melhor aos dados. Com relação aos tipos de resíduos considerados, verifica-se que o resíduo simples é o único que apresenta níveis de significância próximos para os dois modelos, já os resíduos de Pearson e Pearson padronizados sugerem o modelo de Cox como mais adequado.

Tabela 12. Níveis de significância obtidos através de 1000 replicações *bootstrap* segundo o tipo de resíduo e os modelo de Cox e logístico (com grupo) para o experimento com inseto.

Tipos de resíduos	Modelos	
	Cox	Logístico
Simples	0,505	0,439
Pearson	0,479	0,148
Pearson padronizado	0,505	0,154

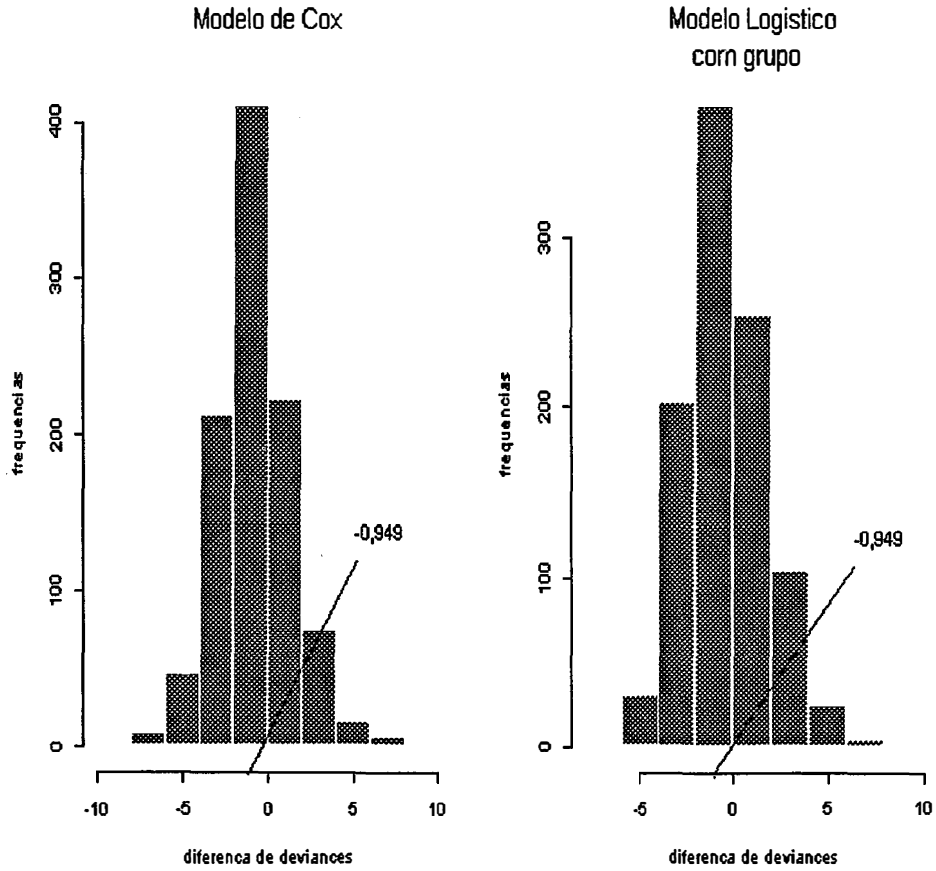

Figura 10 - Histograma das diferenças de *deviances* obtidas através de 1000 replicações *bootstrap* dos resíduos simples no ajuste dos modelos de Cox e logístico (com o fator grupo) para os dados do experimento com inseto.

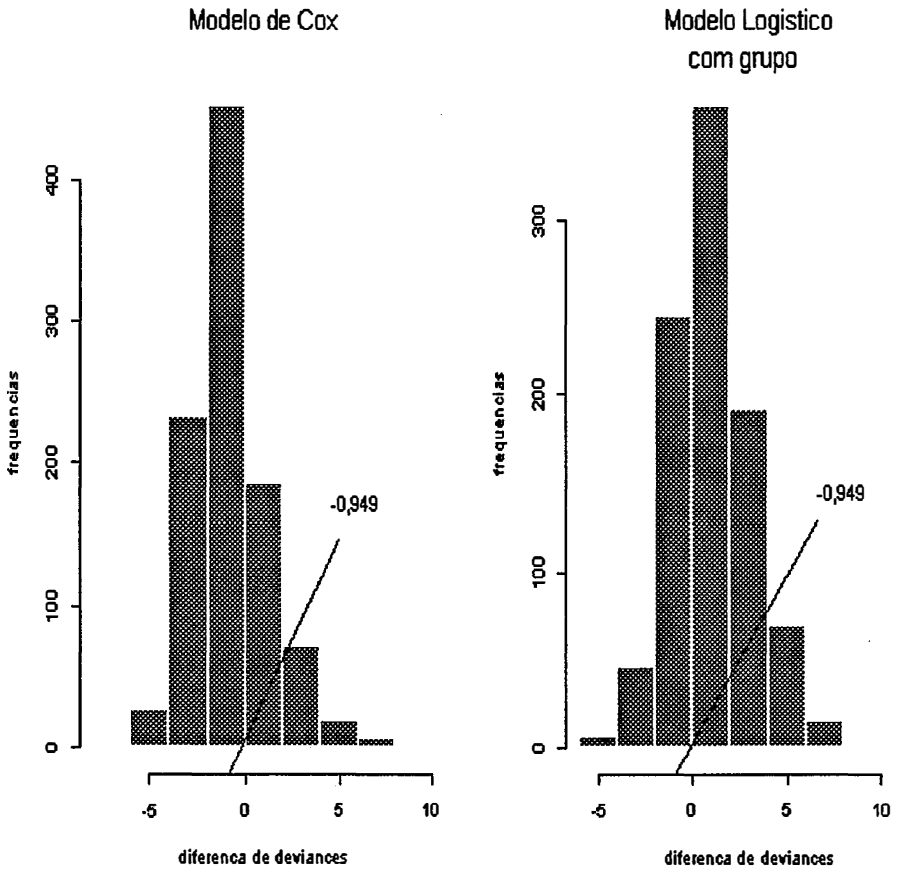

Figura 11 - Histograma das diferenças de *deviances* obtidas através de 1000 replicações *bootstrap* dos resíduos de Pearson no ajuste dos modelos de Cox e logístico (com o fator grupo) para os dados do experimento com inseto.

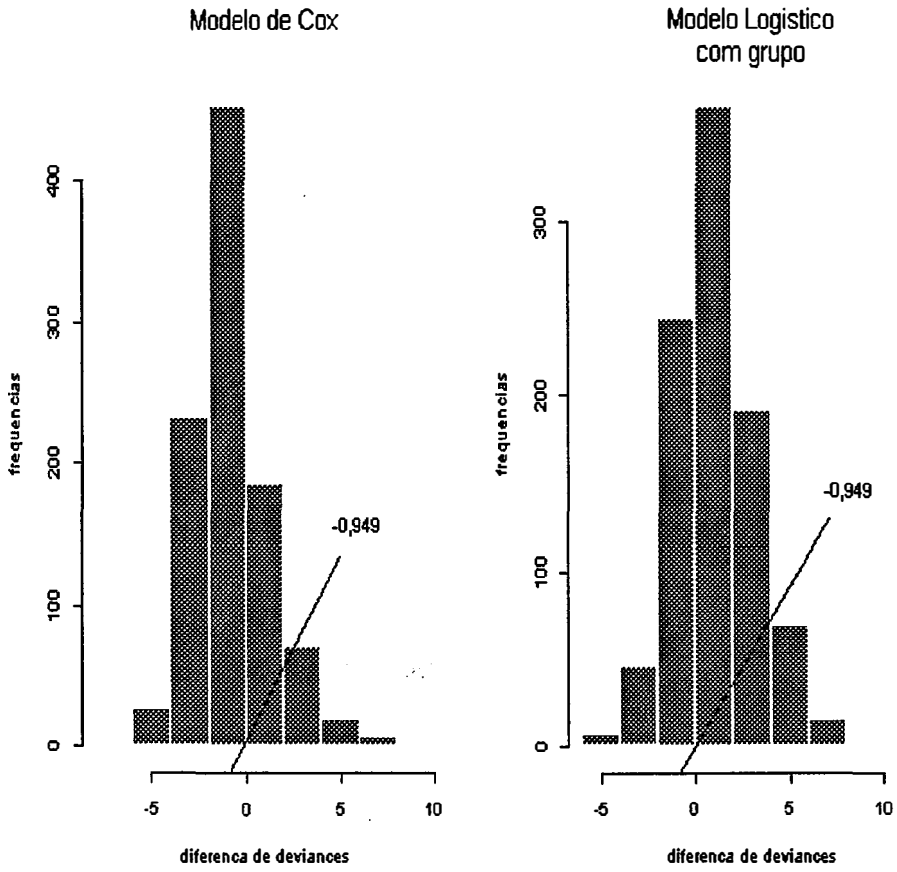

Figura 12 - Histograma das diferenças de *deviances* obtidas através de 1000 replicações *bootstrap* dos resíduos de Pearson padronizados no ajuste dos modelos de Cox e logístico (com o fator grupo) para os dados do experimento com inseto.

As Figuras 13, 14 e 15 apresentam as distribuições empíricas da diferença de *deviances* excluindo o fator grupo do modelo logístico. Nesse caso a diferença de *deviances* observada foi $d_{obs} = -2,955$, ou seja, espera-se que o melhor modelo apresente maior concentração de valores à esquerda da d_{obs}. A discriminação entre os modelos foi bem mais evidente quando utilizou-se o resíduo de Pearson, embora os outros dois tipos de resíduos também tenham apresentado um maior nível de significância para o modelo de Cox.

Tabela 13. Níveis de significância obtidos através de 1000 replicações *bootstrap* segundo o tipo de resíduo e os modelo de Cox e logístico (sem grupo) - Experimento com inseto.

Tipos de resíduos	Modelos	
	Cox	Logístico
Simples	0,711	0,639
Pearson	0,694	0,103
Pearson padronizado	0,374	0,100

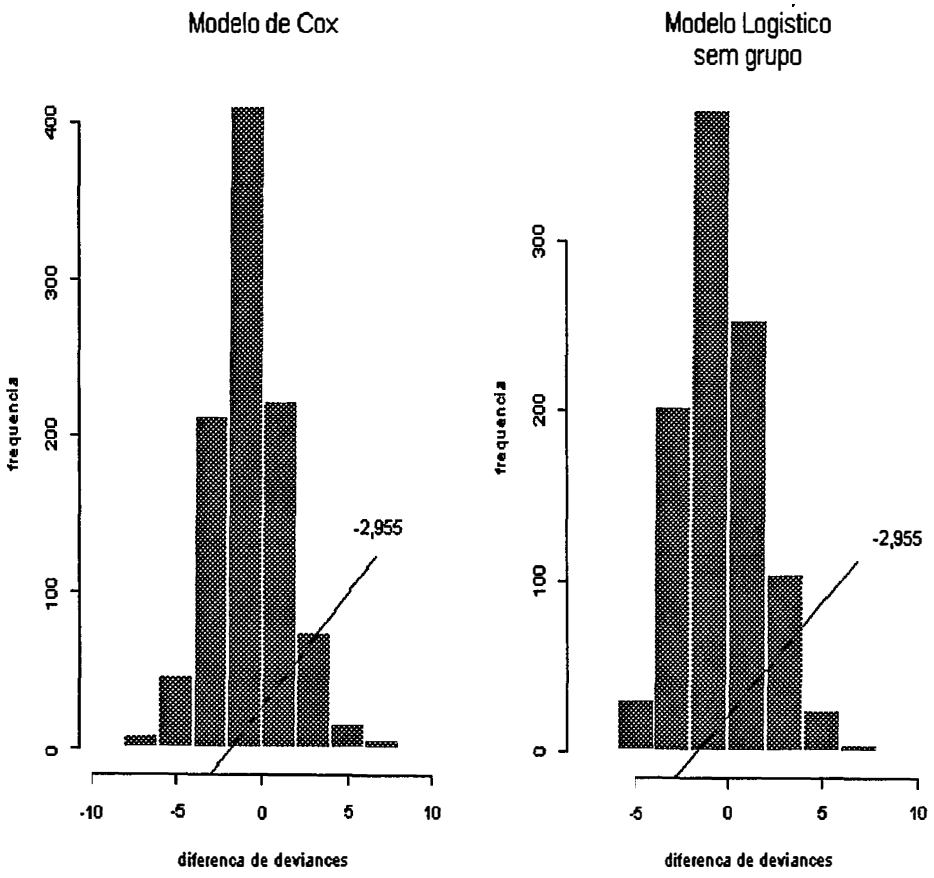

Figura 13 - Histogramas das diferenças de *deviances* obtidas através de 1000 replicações *bootstrap* considerando-se os resíduos simples no ajuste dos modelos de Cox e logístico para os dados do experimento com inseto (sem o fator grupo).

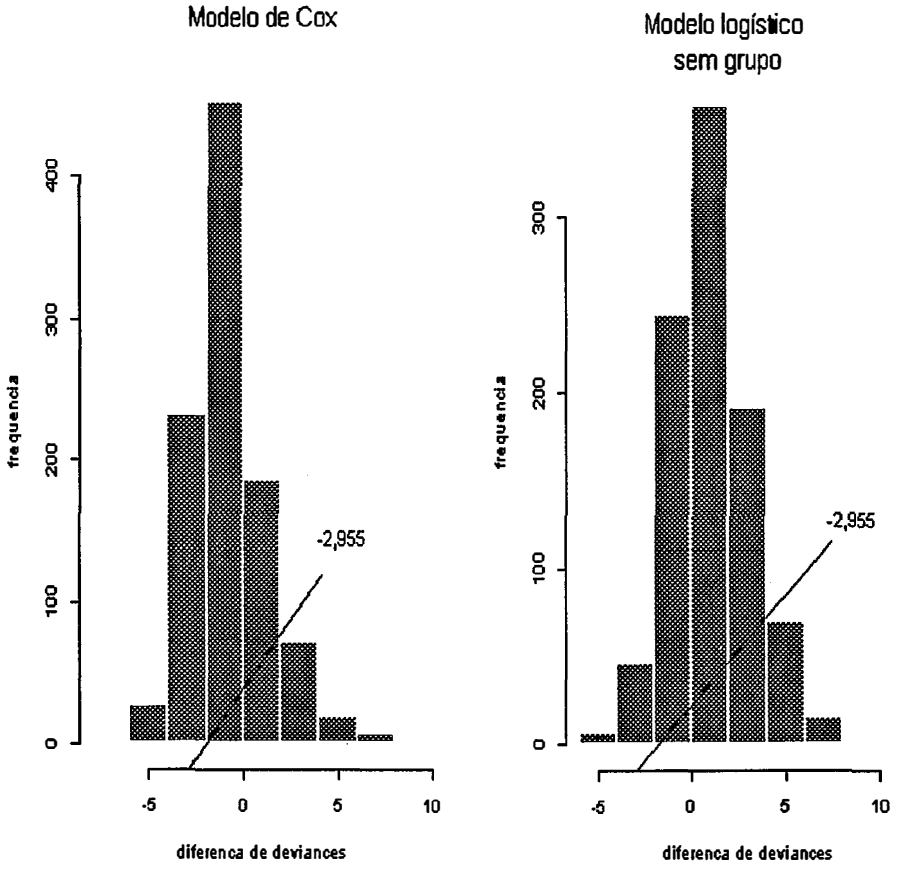

Figura 14 - Histogramas das diferenças de *deviances* obtidas através de 1000 replicações *bootstrap* considerando-se os resíduos de Pearson no ajuste dos modelos de Cox e logístico para os dados do experimento com inseto (sem o fator grupo).

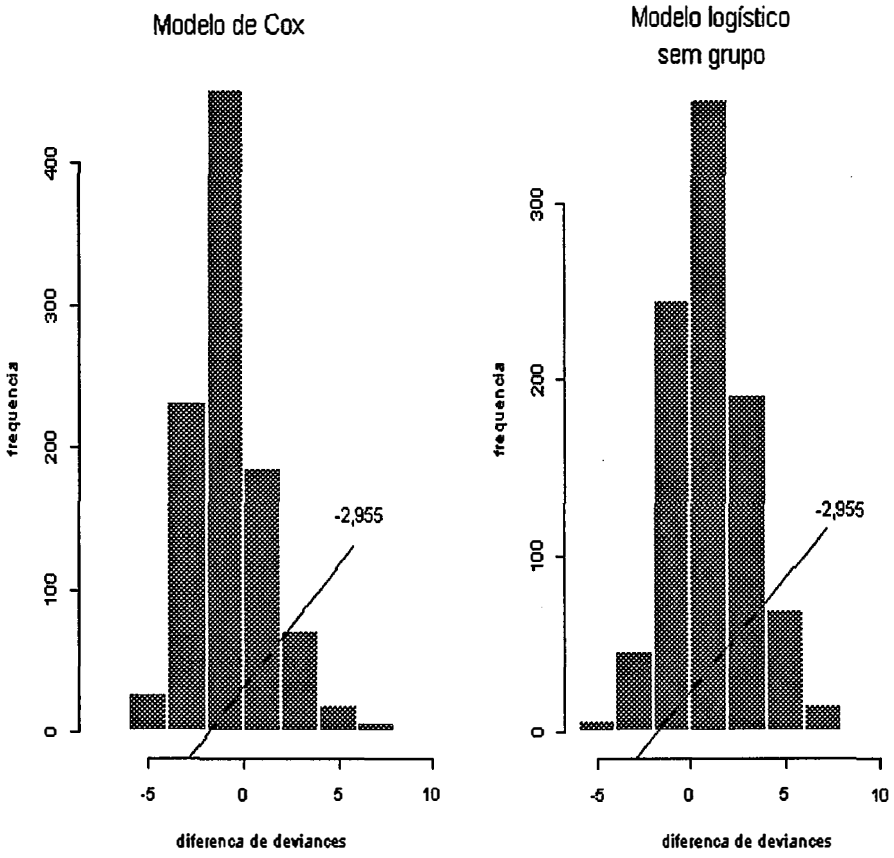

Figura 15 - Histogramas das diferenças de *deviances* obtidas através de 1000 replicações *bootstrap* considerando-se os resíduos de Pearson padronizados no ajuste dos modelos de Cox e logístico para os dados do experimento com inseto (sem o fator grupo).

5 CONCLUSÕES

De acordo com a metodologia desenvolvida e com as análises dos dados dos experimentos considerados, pode-se concluir que:

- quando se tem dados com censura intervalar e o número de empates é grande, o ajuste dos modelos de Cox e logístico se torna simples através dos modelos lineares generalizados, facilitando a obtenção de resíduos e, consequentemente, do processo de reamostragem;

- na análise dos dados do experimento com o linho, observou-se um pequeno valor para a diferença de *deviances*, e ainda assim, através da metodologia *bootstrap* proposta, foi possível evidenciar uma indicação do modelo logístico como o que melhor se adequa aos dados;

- com relação a análise dos dados do experimento com o inseto, novamente se observou um pequeno valor para a diferença de *deviances* na presença do fator *grupo* no modelo logístico, mas que ainda assim, o método apontou o modelo de Cox como o que melhor se adequa aos dados; sem considerar o fator *grupo* no modelo logístico, a diferença de *deviances* foi maior, e novamente houve a indicação do modelo de Cox como o mais adequado, através da metodologia proposta;

- comparando-se os níveis de significância obtidos para os três tipos de resíduos considerados (resíduos simples, de Pearson e Pearson padronizados), nota-se que o resíduo de Pearson e o de Pearson padronizado produzem valores próximos, sendo os responsáveis na indicação do melhor modelo;

- pode-se notar que, quando a diferença de *deviances* é um valor pequeno, os níveis de significância empíricos calculados mostram poucas evidências na indicação do melhor modelo, o que já não acontece quando esse valor é grande;

Como sugestão para a continuidade deste estudo, simulações poderiam ser feitas, considerando-se um dos modelos como o verdadeiro, e verificando-se, portanto, a eficiência do método *bootstrap* proposto na escolha do modelo. Além disso, seria interessante considerar vários valores de diferenças de *deviances* para se ter uma regra empírica de discriminação através do método sugerido aqui, ou seja, considerar diversos valores nas diferenças de *deviances* e verificar o comportamento do método com relação a essa variação.

Outro ponto a ser levado em conta seria utilizar o esta técnica considerando outros tipos de resíduos como, por exemplo, o resíduo de *deviance*, o resíduo de *deviance* padronizado e o resíduo padronizado na escala do predito linear, e verificar a eficiência do método na discriminação de modelos.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: KOTZ, S.; JOHNSON, N.L. (Ed.) **Breakthroughs in statistics**. New York: Springer, 1973. v.1, p.610-624.

ALLISON, P. D. **Survival analysis using the SAS system**: a pratical guide. Cary: SAS Institute, 1995. 292p.

ARANDA-ORDAZ, F. J. On two families of transformations to additivity for binary response data. **Biometrika**, v.68, n.2, p.357-364, 1981.

BICKEL. J.; FREEDMAN, D.A. Some asymptotic theory for the bootstrap. **The Annals of Statistics**, v.9, p.1196-1217, 1981.

BRESLOW, N. Contribution to discussion of paper by D. R. Cox. **Journal of Royal Statistical Society**. Series B, v.34, p.216-217, Mar. 1972.

BRESLOW, N. Covariance analysis of censored survival data. **Biometrics**, v.30, n.1, p.89-99, Mar. 1974.

CHALITA, L.V.A.S. Modelos para dados agrupados e censurados. Piracicaba, 1997. 135p. Tese (Doutorado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

COLLETT, D. **Modelling binary data**. London: Chapman & Hall, 1991. 369p.

COLLETT, D. **Modelling survival data in medical research**. London: Chapman & Hall, 1994. 347p.

COLOSIMO, E.A.; CHALITA, L.V.A.S.; DEMÉTRIO, C.G.B. Tests of proportional hazards and proportional odds models for Grouped survival data. **Biometrics**, v.56, n.4, p.1233-1240, Dec. 2000.

CORDEIRO, G.M. **Modelos lineares generalizados**. Campinas: ABE, 1986. 286p.

COX, D.R. Further results on tests of separate families of hipotheses. **Journal of the Royal Statistical Society**. Series B, v.24, p.406-424, 1962.

COX, D.R. Regression models and life-tables (with discussion). **Journal of Royal Statistical Society**. Series B, v.34, p.187-220, Mar. 1972.

COX, D.R. Partial likelihood. **Biometrika**, v. 62, n.2, p.269-276, Mar. 1975.

COX, D.R.; OAKES, D. **Analysis of survival data**. London: Chapman & Hall, 1984. 195p.

DAVISON, A.C.; HINKLEY, D.V. **Bootstrap methods and their application**. Cambridge: University Press, 1997. 575p.

DEMÉTRIO, C.G.B. **Modelos lineares generalizados na experimentação agronômica**. Porto Alegre: UFRGS, Depto. de Estatística, 1993. 125p.

DOBSON A.J. **An introduction to generalized linear models.** London: Chapman & Hall, 1990. 174p.

EFRON, B. The efficiency of Cox's likelihood function for censored data. **Journal of the American Statistical Association**, v.72, n.359, p.557-565, 1977.

EFRON, B. Bootstrap methods: another look at jackknife. **Annals of Statistics**, v.7, p-1-26, 1979.

EFRON, B. Comparing non-nested linear models. **Journal of the American Statistical Association**, v.79, p.791-803,1984.

EFRON, B. ; TIBISHIRANI, R.J. **An introduction to the bootstrap**. New York: Chapman & Hall, 1993. 436p.

FAREWELL, V. T.; PRENTICE, R. L. The aproximation of partial likelihood with emphasis on case-control studies. **Biometrika**, v.67, n.2, p.273-279, 1980.

FAY, M. P. Comparing several score tests for interval censored data. **Statistics in Medicine**, v.18, p.273-285, 1999.

FINKELSTEIN, D. M. A Proportional hazards model for interval-censored failure time Data. **Biometrics**, v.42, n.4, p.845-854, 1986.

FREEDMAN, D.A. Bootstrapping regression models. **Annals of Statistics**, v.9, p.1218-1228, 1981.

HALL, P. **The bootstrap and edgeworth expansion**. New York: Springer, 1992. 352p.

HINKLEY, D.V. Bootstrap methods. **Journal of the Royal Statistical Society**. Series B, v.50, n.3, p.321-337, 1988.

KALBFLEISH, J.D.; PRENTICE, R.L. Marginal likelihoods based on Cox's regression and life model. **Biometrika**, v.60, p.267-279, Mar. 1973.

KALBFLEISH, J.D.; PRENTICE, R.L. **The Statistical analysis of failure time data**. New York: John Wiley, 1980. 321p.

KAPLAN, E.L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of American Statistical Association**, v.53, p.457-481, May 1958.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. New York: John Wiley, 1982. 579p.

LINDSEY, J.K.; JONES, B. Choosing among generalized linear models applied to medical data. **Statistics in Medicine**, v.17, p.59-68, 1998.

McCULLAGH, P.; NELDER, J.A. **Generalized linear models**. 2.ed. London: Chapman & Hall, 1989. 511p.

MOULTON, L.H.; ZEGER, S.L. Analyzing repeated measures on generalized linear Models via the Bootstrap. **Biometrics**, v.45, p. 381-394, 1989.

MOULTON, L.H.; ZEGER, S.L. Bootstrapping generalized linear models. **Computational Statistics and Data Analysis**, v.11, p.53-63, 1991.

NELDER, J.A.; WEDDERBURN, R.W.M. Generalized linear models. **Journal of the Royal Statistical Society**. Series A, v.135, p.370-384, 1972.

PETO, R. Contribution to dicussion of paper by D.R. COX. **Journal of the Royal Statistical Society**. Series B, v.34, p.205-207, 1972.

PRENTICE, R. L.; L. A. GLOECKLER. Regression analysis of grouped survival data with application to breast cancer data. **Biometrics**, v.34, n.1, p.57-67, 1978.

SCHWARTZ, G. Estimating the dimensions of a model. **Annals of Statistics**, v.6, n.2, p.461-463, 1978.

SHAO, J. Bootstrap model selection. **Journal of the American Statistical Association**, v.91, p.655-665, 1996.

SILVA, D.N. O método bootstrap e aplicações à regressão múltipla. Campinas, 1995. 158p. Dissertação (Mestrado) - Instituto de Matemática Estatística e Computação, Universidade Estadual de Campinas.

STATISTICAL SCIENCE. **S-Plus for Windows**: user's manual. Seattle, 1993. 2v.

SUN, J. A non-parametric test for interval-censored failure time data with applications to AIDS studies. **Statistics in Medicine**, v.15, p.1387-1395, 1996.

THOMPSON, JR.,W.A. On the treatment of grouped observations in life studies. **Biometrics**, v. 33, p. 463-470, June 1977.

WHITEHEAD, J. The analysis of colapse clinical trials, with application to a comparison of two ulcer treatments. **Statistics in Medicine**, v.8, p.1439-1454, 1989.

WILLIAMS, D.A. Discrimination between regression models to determine the patern of enzyme synthesis in cell cultures. **Biometrics**, v.28, n.1 p.23-32, 1970.

Apêndice: Dados

Tabela 1. Dados obtidos do experimento com cultivar de linho.

Tempo	Solo	Número de plantas murchas	Número de censuras
21	M	5	66
24	M	8	58
28	M	13	45
31	M	10	35
35	M	11	24
38	M	16	8
42	M	4	4
45	M	2	2
49	M	1	1
21	MI	2	70
24	MI	6	64
28	MI	7	57
31	MI	12	45
35	MI	10	35
38	MI	12	23
42	MI	8	15
45	MI	8	7
49	MI	3	4
21	MK	1	70
24	MK	14	56
28	MK	8	48
31	MK	11	37
35	MK	10	27
38	MK	14	13
42	MK	4	9
45	MK	6	3
21	MM	1	71
24	MM	2	69
28	MM	3	66
31	MM	12	54
35	MM	13	41
38	MM	16	25
42	MM	8	17
45	MM	9	8
49	MM	5	3

Tabela 2. Dados obtidos do experimento com inseto *Podysus nigrispinus.*

Tempo	Sexo	Grupo	Número de adultos	Número de censuras
24	M	1	1	29
25	M	1	1	28
26	M	1	4	24
27	M	1	21	3
28	M	1	1	2
26	F	1	8	12
27	F	1	12	0
24	M	2	1	28
25	M	2	3	25
26	M	2	3	22
27	M	2	14	8
28	M	2	4	4
24	F	2	1	20
26	F	2	6	14
27	F	2	11	3
28	F	2	3	0