Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Likelihood Approximations and Discrete Models for Tied Survival Data

L. V. A. S. Chalita, E. A. Colosimo e
C. G. B. Demétrio

Relatório Técnico RTP-04/2000

Relatório Técnico
Série Pesquisa

Likelihood Approximations and Discrete Models for Tied Survival Data

LICIANA V. A. S. CHALITA
Depto. de Bioestatística
IB/UNESP, CP 510
18618-000, Botucatu-SP, Brazil
email:

ENRICO A. COLOSIMO
Departamento de Estatística
ICEx/UFMG, CP 702
30161-970, Belo Horizonte-MG, Brazil
email: enricoc@est.ufmg.br

CLARICE G. B. DEMÉTRIO
Departamento de Ciências Exatas
ESALQ/USP, Piracicaba-SP, Brazil
email: clarice@carpa.ciagri.usp.br

Abstract

Ties among event times are often recorded in survival studies. For example, in a two weeks laboratory study where event times are measured in days, ties are very likely to occur. Proportional hazards model might be used in these situations by constructing an approximated partial likelihood function and Breslow's (1974) is the most used one. This kind of approximation works well when the number of ties is small. On the other side, regression discrete models are suggested when there are many ties, that is, when data sets are broadly grouped. However, in many situations it is not clear which approach should be used in practice. In this work, empirical guidelines based on Monte Carlo simulations are proposed to help in this practical decision. A tie measure and the mean square error are employed for this purpose. A real example illustrates the proposed criterion.

Keywords: Breslow approximation; Cox model; Monte Carlo simulations; proportional hazards model; tied observations.

1. Introduction

The proportional hazards model (Cox, 1972) and its extensions have been established as the standard methodology for regression analysis of survival data. Partial likelihood (Cox, 1975) is the most used estimation method for the usual case of right-censoring and exact failure-time data. However, failure-time data are sometimes recorded in time intervals, or in other cases, time is measured in units that are too crude. This kind of data appear naturally in agronomy studies when the field visits are scheduled for specified dates (Colosimo, Chalita and Demétrio, 2000). It also appears in clinical trials or longitudinal studies when event occurrence is monitored at routine medical visits (Sun, 1996; Kim et al., 1993). Tied failure times are the result of these sort of situations.

Two approaches have been proposed in the literature for handling tied survival data: (1) partial likelihood approximations (Breslow (1974), Efron (1977), Farewell & Prentice (1980), Kalbfleish & Prentice (1973) and others) and (2) regression discrete models (Lawless, 1982). Exact partial likelihood that depends on the rank ordering of events can be calculated in the presence of ties as long as the number of distinct failure times remains small. Exact likelihood calculates all possible orderings at each failure time (Peto, 1972). However, calculations get burdensome when the number of ties at certain time gets large. Partial likelihood approximations have been proposed in the literature for handling ties in proportional hazards regression model. Approximations and the exact likelihood should be used only when the number of ties is small (Lawless, 1982). Hertz-Picciotto & Rockhill (1997) compared Breslow, Efron, and Kalbfleisch & Prentice's approximations using Monte Carlo simulations. Efron approximation performs better than the other two, particularly when sample size is small.

The discrete nature of the lifetimes should be explicitly recognized when there is a substantial number of ties and methods for handling discrete or grouped data should be used. Regression methods for grouped data are presented in Lawless (1982, pp. 372-390) and Collett (1994, Cap.8). They are simple to be understood and computational calculations are easily evaluated in a commercial computer package such as GLIM (Aitkin et al., 1989). The regression structure considers modeling the probability of a subject survival past a time visit conditional of his survival at the previous visit. Two approaches are presented: assuming that lifetimes come from a continuous proportional hazards model (Prentice & Gloeckler, 1978) and from a proportional odds model.

The motivation for this paper was a two weeks laboratory study related to a protective effect of the fungus *Saccharomycs boulardii* in debilitated mice. Ninety-three mice were selected from the same animal facility. Initially their immunologic system was weakened using a certain drug. Then four treatments (control and fungus in doses: 10mg, 1mg and 0.1mg) were randomly allocated to each animal and the mouse survival time was the measured response. The goal of the study was to compare the treatments taking into account the mouse initial weight. Response values were registered in days and there were some ties. There were 61 censoring observations and 13 distinct failures among the 32 observed failures during the period of study. It was difficult to classify the number of ties as being small or large. Therefore, there was not a clear indication which approach should be used for the analysis of this data set. In other words, it is not easy to decide if the number of ties is small enough to use likelihood approximations or large enough to use regression methods for grouped data. The aim of this paper is to propose some empirical guidelines for such kind of practical decision. The paper is outlined as follow. Likelihood approximations and discrete regression models are presented in Sections 2 and 3, respectively. Empirical guidelines are based on extensive Monte Carlo simulations that are presented in Section 4. A practical guideline based on simulation results is proposed in Section 5.

2. Approximations for the partial likelihood

Let T be a continuous random variable representing an individual's lifetime and let $x = (x_1, \ldots, x_p)$ be a known vector of covariates associated with the individual. Cox (1972) proposed a proportional hazards model to analyze censored data specified by

$$h(t; x) = h_0(t) \exp(\beta' x),$$

where $\beta = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients and $h_0(t)$ is an arbitrary unspecified baseline hazards function. Partial likelihood proposed by Cox (1975) is used to estimate β without specifying the baseline hazard function $h_0(t)$. It is written as

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta' x_i)}{\sum_{l \in R_i} \exp(\beta' x_l)}, \qquad (1)$$

where r is the number of distinct failures in the sample and R_i is the risk set of individuals at time t_i.

Tied failure times pose special problems on β estimation because partial likelihood depends just on the rank ordering of events and possible tied failure times. It is reasonable to suppose that ties are merely the result of imprecise measurement of time and that there is a true ordering. If the ties ordering were known it would be possible to build the partial likelihood (1) in the usual way. However, in the absence of any knowledge, all possible orderings have to be considered. Several methods have been proposed for handing ties in the Cox proportional hazards model. In this section, exact likelihood, Breslow (1974) and Efron (1977) approximations are presented.

Consider that d_i individuals are observed to fail at t_i, $i = 1, ..., r$, where $t_1 < \ldots < t_r$ and $\sum_{i=1}^{r} d_i = n$. Although the ranks of individuals failing at t_i are known to be less than those failing at t_j $(i < j)$, the arrangement of the ranks of the d_i individuals is unknown. Let's denote each of those possibilities by A_s, where $s = 1, \ldots, d_i!$ and $G_i = Pr(A_1 \cup A_2 \ldots \cup A_{d_i}) = \sum_{s=1}^{d_i!} Pr(A_s)$. Then, the exact partial likelihood for handling ties is

$$L(\beta) = \prod_{i=1}^{r} G_i = \prod_{i=1}^{r} [\exp(\beta' S_i) \sum_{P \in Q_i} \prod_{r=1}^{d_i} \{ \sum_{l \in R(t_i, p_r)} \exp(\beta' x_l) \}^{-1}], \qquad (2)$$

where Q_i is the set of permutations of the symbols $i_1, \ldots, i_{d_i}$; $P = (p_1, \ldots, p_{d_i})$ is an element in Q_i; $R(t_i, p_r)$ is the set difference $R_i - (p_1, \ldots, p_{r-1})$; $S_i = \sum_{l \in D_i} x_l$ and D_i is the set of individuals failing at t_i.

Computationally, the Equation (2) is extremely awkward if the number of ties is large at any failure time. Then approximations have been proposed in the literature and Breslow's is the most popular one. It was first proposed by Peto (1972) and generalized by Breslow for the partial likelihood as

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta' S_i)}{\{\sum_{l \in R_i} \exp(\beta' x_l)\}^{d_i}}. \quad (3)$$

Efron (1977) suggested an alternative approximation given by

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta' S_i)}{\prod_{r=1}^{d_i} \left[\{\sum_{l \in R_i} \exp(\beta' x_l)\} - (r-1) d_i^{-1} \{\sum_{l \in D_i} \exp(\beta' x_l)\} \right]}. \quad (4)$$

The exact likelihood, Breslow's and Efron's approximations are compared in Section 4 in the presence of various amounts of ties.

3. Regression Discrete Models

Consider that lifetimes are grouped into k intervals, $I_i = [a_{i-1}, a_i)$, $i = 1, \ldots, k$, where $0 = a_0 < a_1 < \ldots < a_k = \infty$ and assume that every censoring takes place at the end of the interval. Let Δ_{li} be an indicator variable for the subject's lifetime l in interval I_i ($\Delta_{li} = 0$ if it is censored and $\Delta_{li} = 1$ otherwise). Usually, the likelihood function is written in terms of the probability of l-th subject death at I_i conditional on being alive at a_{i-1} and the covariates values x_l, that is,

$$p_i(x_l) = P[T_l \leq a_i \mid T_l \geq a_{i-1}, x_l]. \quad (5)$$

Thus, the likelihood function can be obtained by considering covariates x_l such that: (1) the contribution of a uncensored observation (at I_i) for the likelihood is $[\{1 - p_1(x_l)\} \ldots \{1 - p_{i-1}(x_l)\}] p_i(x_l)$ and (2) the contribution of a censored observation (at a_i) for the likelihood is $[\{1 - p_1(x_l)\} \ldots \{1 - p_i(x_l)\}]$. The likelihood function is then given by

$$\prod_{i=1}^{k} \prod_{l \in R_i} \{p_i(x_l)\}^{\Delta_{li}} \{1 - p_i(x_l)\}^{(1-\Delta_{li})}. \quad (6)$$

Equation (6) is the likelihood function from a Bernoulli distribution with response Δ_{li} and success probability $p_i(x_l)$. The regression structure represented by the probability in Equation

(5) can be modeled adopting the proportional hazards model or the proportional odds model (Collett, 1991).

Proportional hazards approach takes $p_i(x_l)$ of the form

$$p_i(x_l) = 1 - \gamma_i^{\exp(\beta' x_l)}, \tag{7}$$

where $\gamma_i = S_0(a_i)/S_0(a_{i-1})$, $i = 1, \ldots, k$ and $S_0(.)$ is the baseline survival function. Proportional odds approach takes $p_i(x_l)$ of the form

$$p_i(x_l) = 1 - \{1 + \gamma_i \exp(\beta' x_l)\}^{-1}, \tag{8}$$

where $\gamma_i = p_i(0)/\{1 - p_i(0)\}$, $i = 1, \ldots, k$. Plugging in Equations (7) and (8) in the likelihood function (6), proportional hazards and proportional odds models are obtained, respectively, for grouped survival data.

Model (7) can be linearized by using a complementary log-log transformation, that is,

$$\log[-\log\{1 - p_i(x_l)\}] = \gamma_i^* + \beta' x_l = \eta_{li}$$

where $\gamma_i^* = \log(-\log \gamma_i)$ is the interval effect and $\eta_{li}, i = 1, \ldots, k, l = 1, \ldots, n$, is called linear predictor. Model (8) can be linearized by using a logistic transformation such as

$$\log\left\{\frac{p_i(x_l)}{1 - p_i(x_l)}\right\} = \gamma_i^* + \beta' x_l = \eta_{li},$$

where $\gamma_i^* = \log(\gamma_i)$ is the same interval effect. Therefore, these models can be fitted using standard methods for modeling binary data.

4. Simulation Description

Monte Carlo simulations were performed in order to evaluate the partial likelihood approximations and two discrete models. Survival time T_l and censoring time C_l, $l = 1, 2, \ldots, n$, were assumed to follow a Weibull and uniform on $(0,\theta)$ distributions, respectively. The observed response is $Y_l = \min(T_l, C_l)$ and a failure indicator. Three different values for the Weibull distribution shape parameter were used, 0.5, 1 and 2. Some values for the parameter θ were fixed to produce 0, 35%, 50% and 65% of censoring observations in the simulated data. A covariate, x_l, was generated following a uniform distribution on (0,1). The Weibull's scale parameter was modeled as $\exp(\beta x_l)$ and β was fixed equal to 1. The sample sizes generated were 50 and 100.

For each combination of shape parameter, proportion of censoring and sample size, 500 replications were generated. Ties in y were produced by grouping the data as follows for each simulated data set. A total of k time intervals of equal length was created by splitting the range of Y ($y_{(n)} - y_{(1)}$, where $y_{(i)}$ is the ith observed order statistics) using $k-1$ cutoff points. The values used for k were set to 5, 10, 20 and 30. The observations falling into the same interval are assumed to be tied.

In order to measure the proportion of ties (pt) in a particular survival data set the following quantity is proposed

$$pt = \frac{nf - r}{n}$$

where nf is the number of failures. This measure is the key quantity to evaluate the appropriateness of the discrete models and the likelihood approximations. Mean squared error and pt were calculated for each replication. Maximum likelihood estimates and their standard errors were obtained by using PROC PHREG in SAS program, version 6.12.

5. Results and Discussion

The results of the simulations are presented in Figures 1 to 6. The following conclusions can be obtained from these figures:

(i) when the number of intervals increases, the mean square error decreases for continuous models and increases for discrete models. This fact can occur because increasing the number of intervals increases also the number of parameters to be estimated in the discrete models and decreases the proportion of ties;

(ii) comparisons among models does not appear to be affected by the shape parameter of the Weibull distribution for the same sample size, proportion of censoring and number of intervals;

(iii) for continuous models, the exact form fits best following by Efron's and Breslow's approximations, independently of the number of intervals and the proportion of censoring. This result is in agreement with those obtained by Hertz-Picciotto & Rockhill (1997);

(iv) for the discrete models, the Cox's model seems to fit better than the logistic one. This fact might be explained because the Weibull's distribution is one member of proportional hazard family.

An empirical suggestion can be proposed observing the behavior of the mean square error in Figures 1 to 6. The mean values of pt and number of intervals are showed in these figures for the studied conditions in the simulations. As expected, the proportion of ties decreases as the of number of intervals increases and proportion of censoring and sample size decreases. It can be observed that the mean square error of the discrete models gets larger than the square error of the approximations as pt decreases. In this way the empirical guide presented in Table 1 seems to be in agreement with these conclusions.

Table 1: **Empirical guide for the decision between models.**

pt (%)	Models
< 20	**Should be** used the continuous models with likelihood approximation
20 a 25	**Might be** used the continuous models with likelihood approximation
> 25	**Should be** used discrete models

In the application described in Section 1 pt is equal to 0.204. According to the empirical guidelines in Table 1, likelihood approximations might still be a reasonable way to analyze this data set.

6. Final Remarks

An empirical guideline was proposed for deciding between discrete models and likelihood approximations in the presence of ties. This decision was based on a quantity pt that measures the proportion of ties in a sample. A simulation study investigated the performance of the exact partial likelihood, Breslow's and Efron's approximations and two discrete models for tied failure times. Based on the simulation results the following guideline was suggested: discrete models should be used for $pt > 25\%$ and otherwise, proportional hazards model with likelihood approximation might be the used in the survival analysis.

References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM* Clarendon Press, Oxford.

Breslow, N. (1974). Covariance analysis of censored survival data *Biometrics*, **30**, 89-99.

Collett, D. (1991). *Modelling Binary Data.* Chapman and Hall, London.

Collett, D. (1994). *Modelling Survival Data in Medical Research.* Chapman and Hall, London.

Colosimo, E. A., Chalita, L. V. A. S. and Demétrio, C. G. B. (2000). Tests of proportional hazards and proportional odds models for grouped survival data. *Biometrics*, (to appear).

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, London.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal os the American Statistical Association*, **83**, 414-425.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**, 845-854.

Hertz-Picciotto and Rockhill (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, **53**, 1151-1156.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihood based on Cox's regression and life model. *Biometrika*, **60**, 267-279.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley, New York.

Kim, M. Y., De Gruttola, V. G. and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application of AIDS. *Biometrics*, **49**, 13-22.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley, New York.

Peto, R. (1972). Contribution to discussion of paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B*, **34**, 205-207.

Peto, R. (1973). Experimental survival curves for interval censored data. *Applied Statistics*, **22**, 86-91.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analyis of grouped survival data with application to breast cancer data. *Biometrics*, **34**, 785-793.

Rabinowitz, D., Tsiatis, A. and Aragon, J. (1995). Regression with interval-censored data. *Biometrika*, **82**, 501-513.

Rao, C. R. (1973). *Linear Statistical Inference and Its Application.* John Wiley, New York.

Sun, J. (1996). A nonparametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine*, **15**, 1387-1395.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored, and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290-295.

Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics*, **53**, 1199-1211.

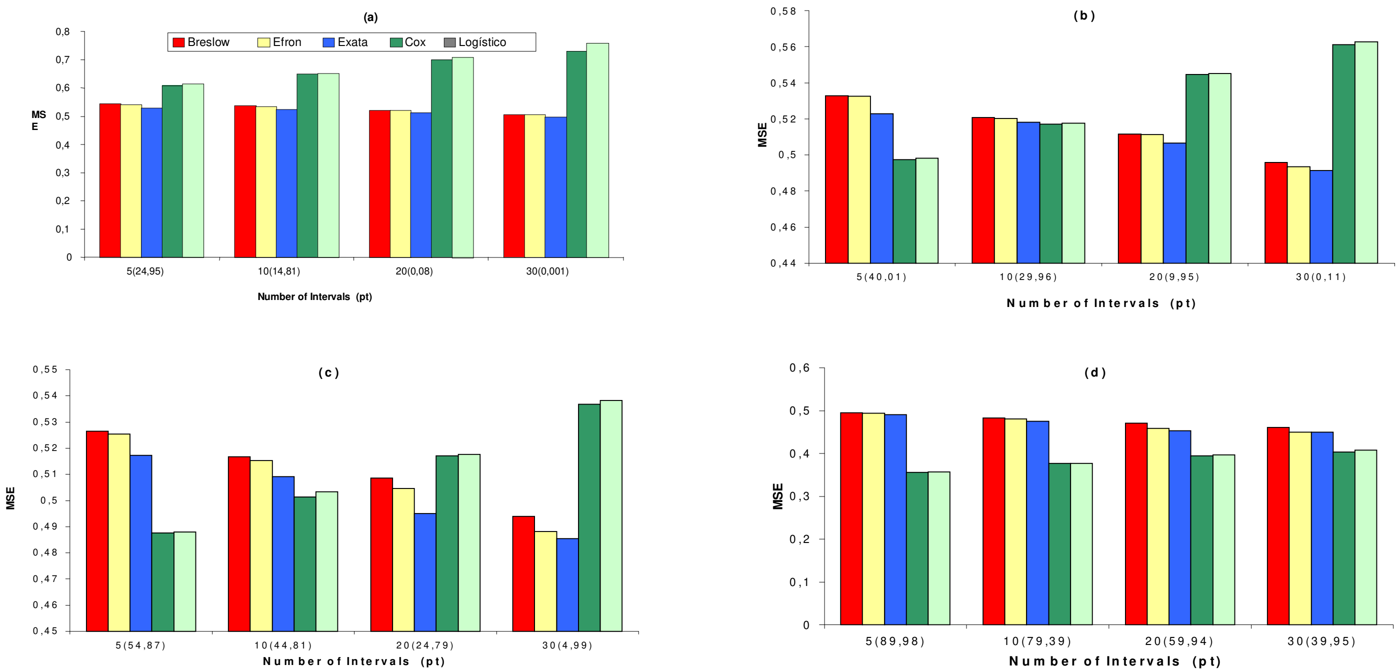

Figure 1: Plots of MSE for n=50, shape parameter=0.5 and : (a) pc=0.65; (b) pc=0.5; (c) pc=0.35; (d) pc=0.

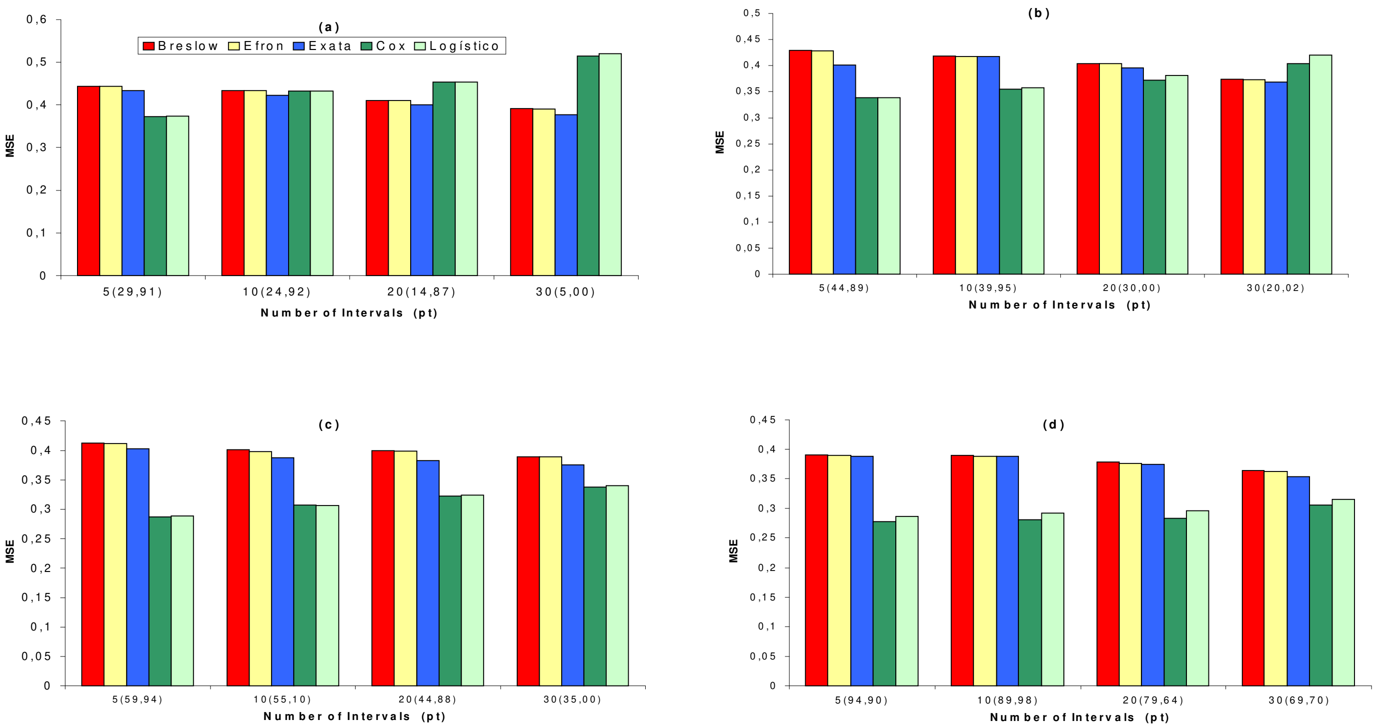

Figure 2: Plots of MSE for n=100, shape parameter=0.5 and : (a) pc=0.65; (b) pc=0.5; (c) pc=0.35; (d) pc=0.

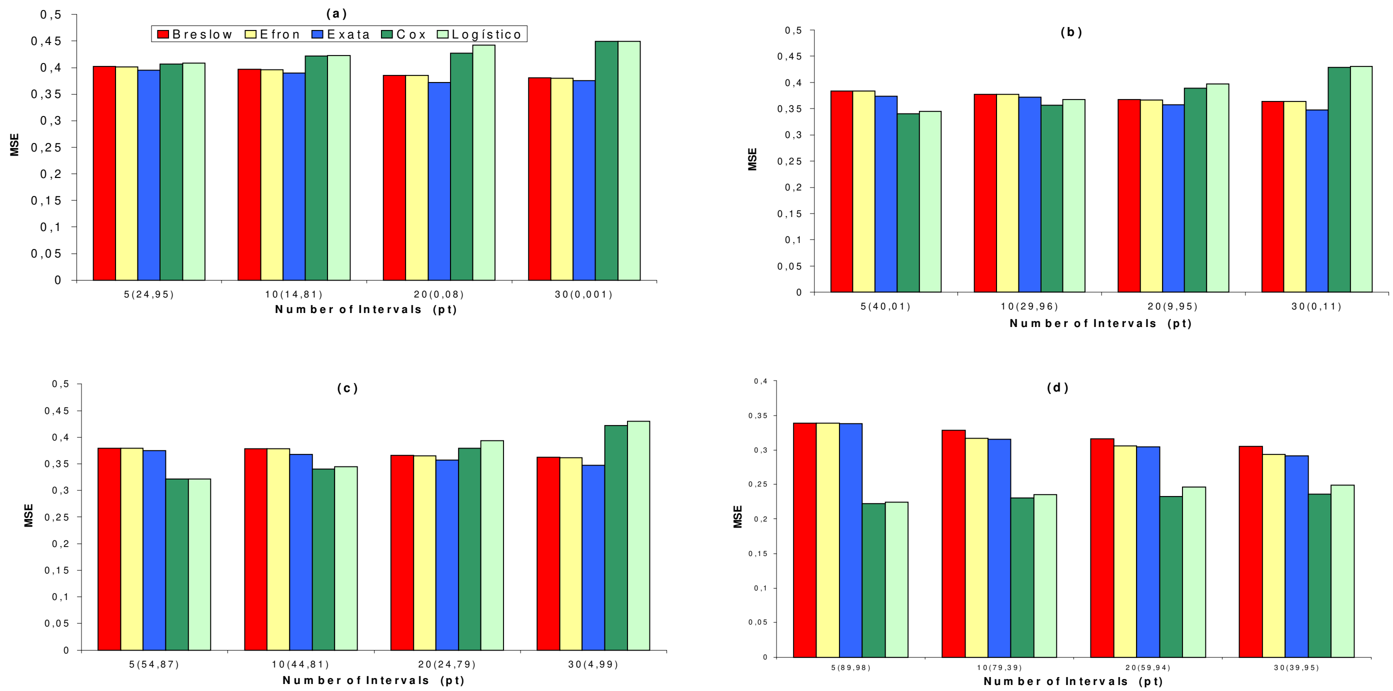

Figure 3: Plots of MSE for n=50, shape parameter=1 and : (a) pc=0.65; (b) pc=0.5; (c) pc=0.35; (d) pc=0.

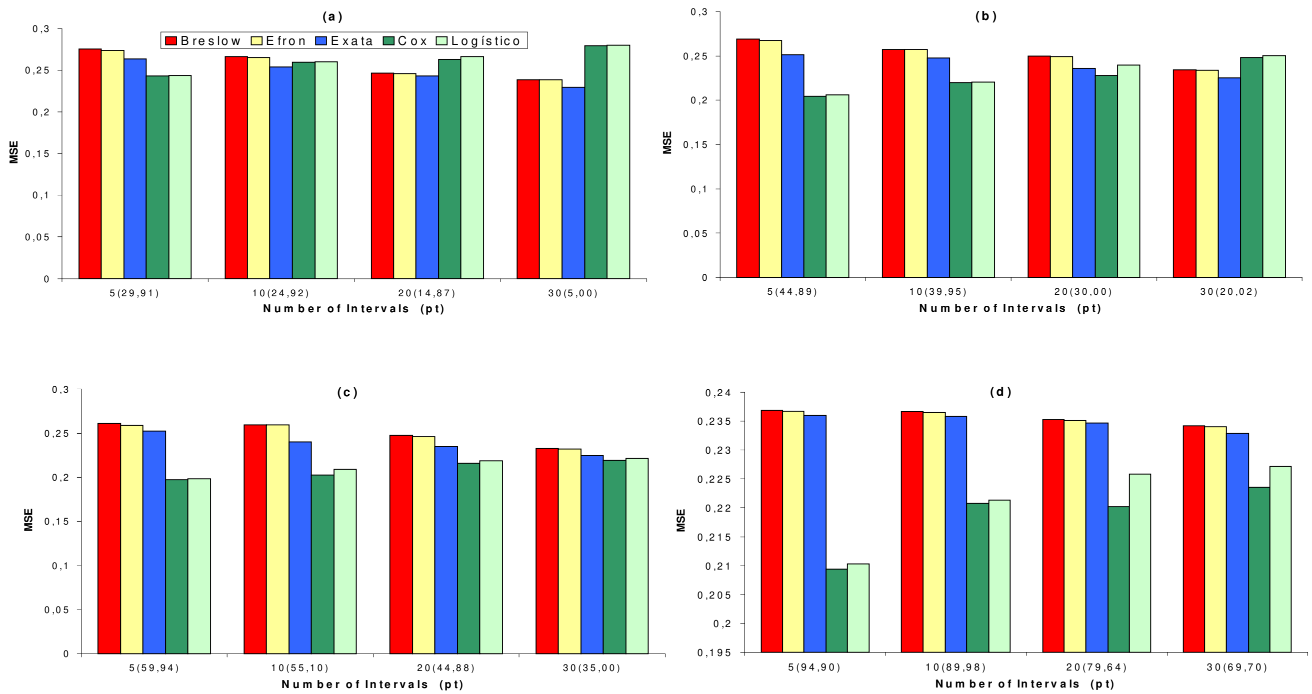

Figure 4: Plots of MSE for n=100, shape parameter=1 and : (a) pc=0.65; (b) pc=0.5; (c) pc=0.35; (d) pc=0.

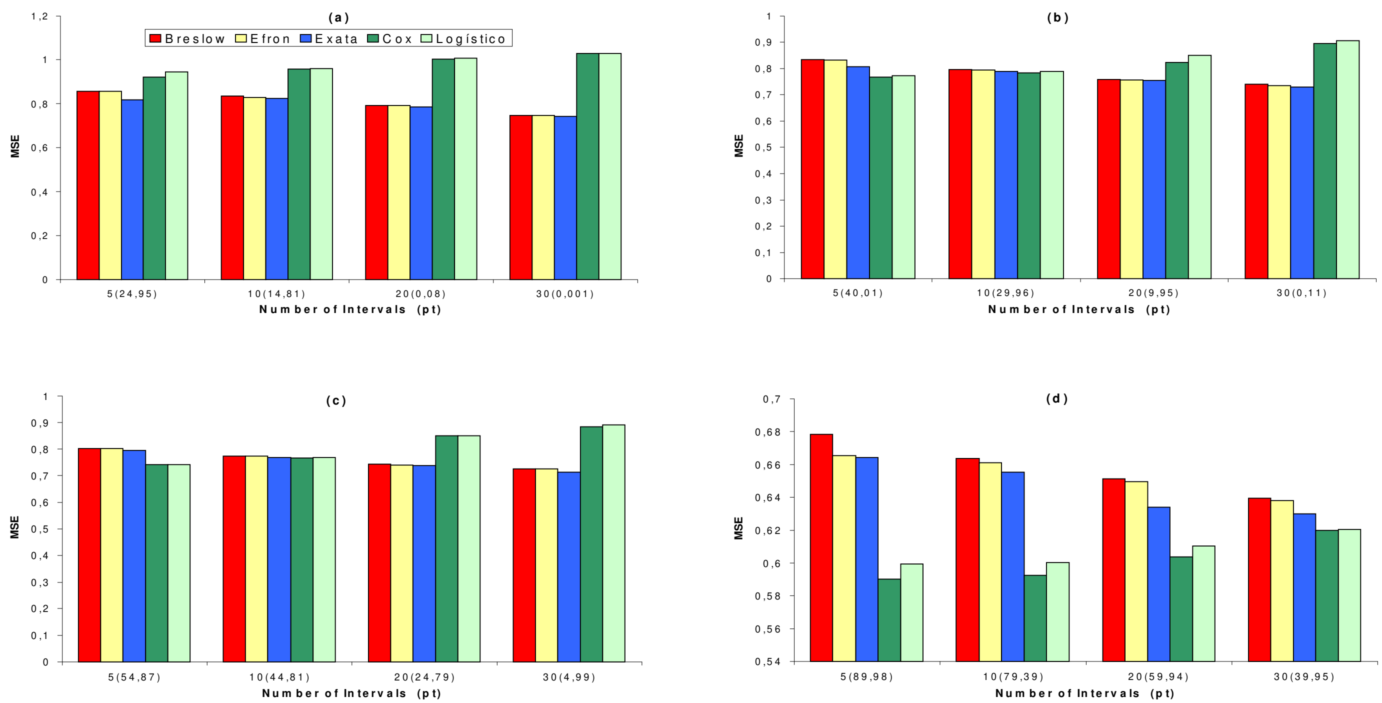

Figure 5: Plots of MSE for n=50, shape parameter=2 and : (a) pc=0.65; (b) pc=0.5; (c) pc=0.35; (d) pc=0.

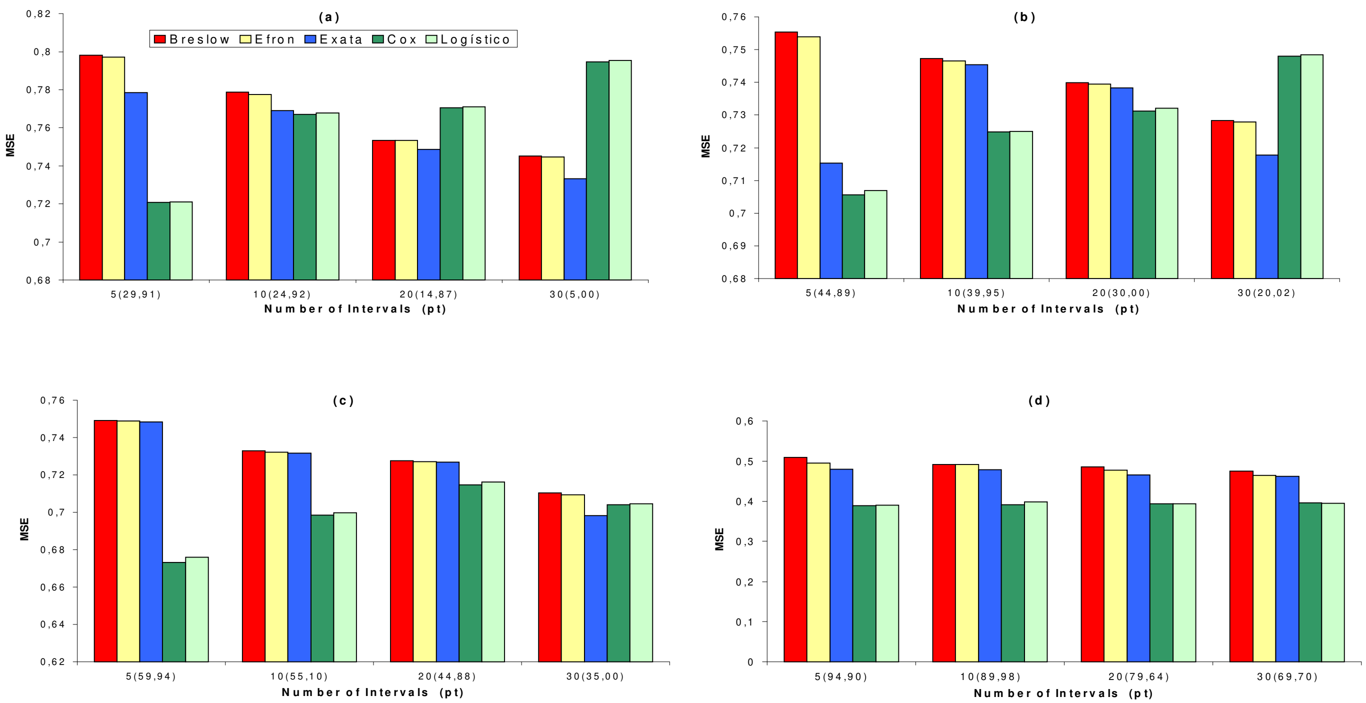

Figure 6: Plots of MSE for n=100, shape parameter=2 and : (a) pc=0.65; (b) pc=0.5; (c) pc=0.35; (d) pc=0.