SPC

International Journal of Advanced Statistics and Probability, 2 (2) (2014) 48-55
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJASP
doi: 10.14419/ijasp.v2i2.2291
Research Paper

Comparison of resampling method applied to censored data

**Claude Thiago Arrabal[1]*, Karina Paula dos Santos Silva[2], Ricardo Ferreira da Rocha[1]
Ricardo Nonaka[1], Silvana Aparecida Meira[3]**

[1] *Departament of Statistic, Federal University of São Carlos, Brazil*
[2] *Institute of Mathematic and Statistic, University of São Paulo, Brazil*
[3] *Federal University of Mato Grosso, Brazil*
**Corresponding author E-mail: claude.arrabal@gmail.com*

Abstract

This paper is about a comparison study among the performances of variance estimators of certain parameters, using resampling techniques such as bootstrap and jackknife. The comparison will be made among several situations of simulated censored data, relating the observed values of estimates to real values. For real data, it will be considered the dataset Stanford heart transplant, analyzed by Cho et al. (2009) using the model of Cox regression (Cox, 1972) for adjustment. It is noted that the Jackknife residual is efficient to analyze influential data points in the response variable.

Keywords: *bootstrap, Jackknife, simulation, Cox Regression Model, censored data.*

1. Introduction

Bootstrap resampling method researches began at the end of the 70s decade, although a lot of related projects can be verified before that period. The most theoretical development was elaborated after 1980, with Efron [7] and after been popularized by Miller [11]. We have several motivations to use these methods, for example when usual model assumptions can't be verified in certain dataset. For example non normal distribution with outliers or mixed distribution with errors. We can also point to asymptotic results when the number of available observations aren't enough to guarantee the asymptotic convergence. In these cases, methods based on simulation, more specifically resampling, can be useful to establish the uncertainty about estimation of the parameters. The Cox Regression model is widely used in Survival analysis for censored data [5].

1.1. Objectives

In this paper, we will present a comparison among performances of bootstrap and Jackknife methods using real and simulated censored data. For real data we will use five databases in presence of censor with different sample sizes, using Cox Regression model. For models comparison, asymptotic confidence interval using bootstrap and Jackknife methods will be used. The application to the real data is through Cox Regression Model [5], with Exponential link functions.

2. Methodology

We are interested in the uncertainty associated to the estimated values of certain parameters. In many cases, as there are certain difficulties in identifying the distribution associated to the estimator of the parameters, we use the bootstrap and the Jackknife methods, two simulation techniques to calculate deviation and to construct the confidence interval for the estimators. In this paper, we investigate models with simulated censored data and a real dataset using Cox Regression Model. For the real dataset, we compare the estimator with the results obtained by Cho [3]. We simulate the average time from an Exponential distribution and we estimate the parameters using maximum likelihood, aiming to verify the performance of the method. The criterion to evaluate the quality of the estimators is the mean squared error and the coverage probability. We evaluated the methods referred above for simulated data and for a real dataset, Stanford heart transplant data [12].

2.1. Bootstrap

Proposed by Efron [7], the bootstrap is a simulation technique to evaluate the standard error of the estimation of a parameter (Martinez et al. [10]). The idea is to perform various resamplings with replacement of the dataset. Let $\hat{\theta} = t(\mathbf{x})$ the estimator of θ calculated from a sample $(x_1, ..., x_n)$. A sample bootstrap $x^*_{(i)} = (x^*_1, ..., x^*_n)$ is a resample with replacement of size n from $(x_1, ..., x_n)$, the index $i = 1, ..., B$ refers to the number of wanted replicas B. This way, the bootstrap estimator of the variance of $\hat{\theta}$ is given by

$$\widehat{Var}(\hat{\theta}) = \frac{1}{B-1}\sum_{i=1}^{B}(t(x^*_{(i)}) - \hat{\theta}^*)^2, \tag{1}$$

where $t(x^*_{(i)})$ is the estimator of θ based on the i-th replica bootstrap and $\hat{\theta}^* = B^{-1}\sum_{i=1}^{B} t(x^*_{(i)})$. For a sufficiently large number of replicas, we can calculate confidence intervals (with significance level 2α) for $\hat{\theta}$ through normal approximation, which is given by

$$IC_B[\hat{\theta}, (1-2\alpha)] = [\hat{\theta}^* - z_\alpha[\widehat{Var}(\hat{\theta})]^{1/2}; \hat{\theta}^* + z_\alpha[\widehat{Var}(\hat{\theta})]^{1/2}], \tag{2}$$

where z_α corresponds to the α-th quantile of standard normal distribution. We can estimate the empiric density $\hat{F}$ as a normal approximation of real F. As such, we can estimate the confidence interval using $\hat{F}$ quantiles, so the interval is given by

$$IC_B[\hat{\theta}, (1-2\alpha)] = [\hat{F}(\alpha); \hat{F}(1-\alpha)]. \tag{3}$$

2.2. Jackknife

The Jackknife method is a resampling technique such as bootstrap. The idea consists of resampling the initial sample removing one or more observations in each replica. In this case, the number of replicas will be equal to the sample size, when we take off just one observation. The Jackknife samples $x_{(-i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ are the data with the $i-$th observation removed, the estimator is $\hat{\theta}^* = n^{-1}\sum_{i=1}^{n} t(x_{(-i)})$. We can calculate the Jackknife estimator of the bias $\widehat{vies}_{jack} = (n-1)(\hat{\theta}^* - \hat{\theta})$. So, the Jackknife estimator of the variance of $\hat{\theta}$ is

$$\widehat{Var}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(t(x_{(-i)}) - \hat{\theta}^*)^2, \tag{4}$$

where $t(x_{(-i)})$ is the estimator of θ based on i-th Jackknife replica.
Here we have the empiric density, so we will use the confidence interval for normal distribution approximation, that is constructed analogously to above mentioned.
As in bootstrap, we estimate the standard error using the standard deviation applied to Jackknife replicas.

2.3. Cox Regression Model

The Cox Regression model is a model for the hazard analysis which allows the analysis of survival data in the presence of covariables. Suppose we want to estimate the lifetime (t) of patients submitted to two types of treatment: x ($x = 0$ for the standard treatment and $x = 1$ for the new treatment). Thus, we define as $\lambda_0(t)$ the standard treatment failure rate and $\lambda_1(t)$ the new treatment failure rate. Therefore, the reason between these two failure rates (K) is represented by

$$\frac{\lambda_1(t)}{\lambda_0(t)} = K, \tag{5}$$

Considered constant for every time t. Considering that $K = exp(\beta x)$ we have

$$\lambda(t) = \lambda_0(t) exp\{\beta x\} \tag{6}$$

the Cox model for a single covariable.
A Cox model with p covariables, vector of covariables $x = (x_1, x_2, \cdots, x_p)^T$ and the vector of parameters associated to covariables $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$, is defined by

$$\lambda(t) = \lambda_0(t) g\left(x^T\beta\right), \tag{7}$$

where $\lambda_0(t)$ is the baseline hazard function and $g(\cdot)$ is the link function, which can be used in several ways, but Cox (1972) suggested

$$g\left(x^T\beta\right) = exp\left(x^T\beta\right) = exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}. \tag{8}$$

Generalizing the failure rate for the individuals i and j is given by

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) exp\left(x_i^T\beta\right)}{\lambda_0(t) exp\left(x_j^T\beta\right)} = exp\left\{x_i^T\beta - x_j^T\beta\right\}, \tag{9}$$

that doesn't depend on the time t.
The estimation of the parameters of the Cox model is made using the maximum likelihood method, where the likelihood function considering a random sample $t = (t_1, t_2, \cdots, t_n)$ is given by:

$$L(\beta) = \prod_{i=1}^{n} \left[f\left(t_i|x_i\right)\right]^{\delta_i} \left[S\left(t_i|x_i\right)\right]^{1-\delta_i} = \prod_{i=1}^{n} \left[\lambda\left(t_i|x_i\right)\right]^{\delta_i} S\left(t_i|x_i\right) \tag{10}$$

in which δ_i is the censoring indicator and $S(t)$ the survival function. In Cox model we have

$$S\left(t_i|x_i\right) = exp\left\{-\int_0^{t_i} \lambda_0(u) exp\left\{x_i^T\beta\right\} du\right\} = \left[S_0(t_i)\right]^{exp\left\{x_i^T\beta\right\}}, \tag{11}$$

thus, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} \left[\lambda_0(t_i) exp\left\{x_i^T\beta\right\}\right]^{\delta_i} \left[S_0(t_i)\right]^{exp\left\{x_i^T\beta\right\}}. \tag{12}$$

Later Cox proposed the partial maximum likelihood method (Cox [5]) to estimate the parameters of the model.

3. Implementation and Simulation

3.1. Generation of the survival times

Considering the survival function in (11) we can find the distribution function of the Cox regression model.

$$F\left(t_i|x_i\right) = 1 - S\left(t_i|x_i\right) = 1 - \left[S_0(t_i)\right]^{exp\left\{x_i^T\beta\right\}}. \tag{13}$$

According to Bender et al. [1] in order to generate the survival times of the Cox regression model, we can use the inverse function method in (13) for the exponential, Weibull and Gompertz distribution.

1. Exponential Distribution

   Considering the exponential link function, we have the generation of time given by

   $$T = -\frac{log(u)}{\lambda exp\left\{x_i^T\beta\right\}} \tag{14}$$

   in which $u \sim U(0,1)$. Considering that the survival times are exponentially distributed with scale parameters $\lambda exp\left\{x_i^T\beta\right\}$, that are dependent on the regression coefficients and on the considered covariables.

2. Weibull Distribution

   Considering the Weibul link function, we have the survival time given by

   $$T = -\left(\frac{log(u)}{\lambda exp\left\{x_i^T\beta\right\}}\right)^{\frac{1}{\nu}} \tag{15}$$

   in which $u \sim U(0,1)$. In Weibull distribution, the survival times ate distributed with scale parameters $\lambda exp\left\{x_i^T\beta\right\}$ and fixed parameter ν, when $(\nu = 1)$ we have a exponential distribution.

3. Gompertz Distribution

   Considering the Gompertz link function we have the survival time generation given by

   $$T = \frac{1}{\alpha} log\left[1 - \frac{\alpha log(u)}{\lambda exp(x_i^T\beta)}\right] \tag{16}$$

   in which $u \sim U(0,1)$. In Gompertz distribution, the survival times are distributed with scale parameters $\lambda exp\left\{x_i^T\beta\right\}$ and fixed parameters α, so, when $(\alpha = 0)$ we have an exponential distribution.

3.2. Resampling Process Method

The resampling using the bootstrap method was made in two ways. The first way was to resample every line of the data matrix. The second way was a stratified sample using the censored and the uncensored data.
The resampling using the Jackknife method was made taking the lines off the data matrix one at a time.
Two datasets were generated using the following process

- In the first dataset we generate the response variable only, i.e., the survival time after the transplant, according to the exponential model with rate $\lambda exp\left\{x_i^T\beta\right\}$, in which we used the dependent variables of the real dataset.
- The second dataset was generate in a stratified way based on the censors, considering the time with exponential distribution with rate $\lambda exp\left\{x_i^T\beta\right\}$ and for the ages we considered a Poisson distribution with parameter μ equal to the mean of the ages of the patients. The censor was random distributed with the same proportion as the real data.

4. Results and Discussions

The real dataset was used to illustrate the proposed method. From 184 cases, 71 are censored and the response variable is survival time after heart transplant, the covariables are the average age of the patients and its square root value. The average age of the patients is 41.09 years, with a dispersion around the mean of 11.04. The covariable T5 mismatch wasn't used because it isn't significant. The Jackknife residual was calculates for the influential data points analysis and we found 4 outliers, such as the obtained in Cho et al. [3]. The influential data points were removed from the data.

Table 1: Real Data

	assint.	bootstrap	boot. stratified	Jackknife
β_0	-0,143379	-0,132657	-0,133409	-0,143369
β_1	0,002243	0,002106	0,002122	0,002243
CI(0,95;β_0) inf	-0,264084	-0,243093	-0,242222	-0,248490
CI(0,95;β_0) sup	-0,022675	-0,022221	-0,024597	-0,038247
CI(0,95;β_1) inf	0,000726	0,000727	0,000770	0,000942
CI(0,95;β_1) sup	0,003759	0,003485	0,003474	0,003543
emp. CI(0,95;β_0) inf	0	-0,241779	-0,245448	-0,248490
emp. CI(0,95;β_0) sup	0	-0,019797	-0,025882	-0,038247
emp. CI(0,95;β_1) inf	0	0,000664	0,000814	0,000942
emp. CI(0,95;β_1) sup	0	0,003473	0,003473	0,003543

In Table (1) we can observe that, for this dataset, the results using Jackknife are better than using bootstrap, however, we observe that the results are very close.
In Fig. (1) we can observe the asymptotic confidence intervals for the parameters β_0 and β_1, respectively.

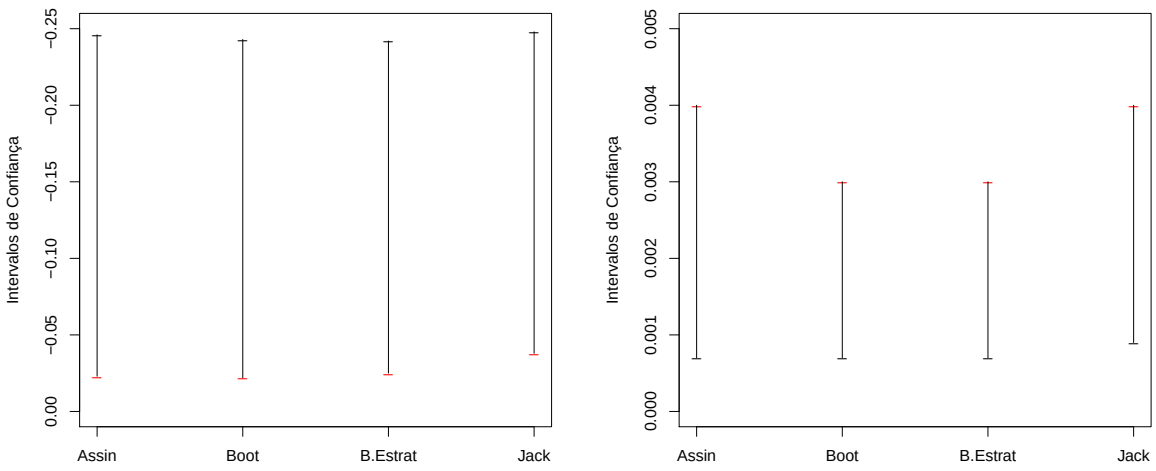

Fig. 1: Asymptotic confidence interval for β_0 and β_1, considering exponential link function

For Table (2) the response variable, i.e., the survival time after transplant, was simulated. We considered as true the value of the parameter $\beta_0 = -0,150792$ and $\beta_1 = 0,002632$, so we replicated 1000 times the simulation to calculate the coverage probability through asymptotic confidence interval and using the percentile $0,025$ and the $0,975$.
Using asymptotic confidence interval, we observe that the coverage probability based on Jackknife presents a better result than the bootstrap. However, when we compare the coverage probability using the percentile we observe that the bootstrap presents a better result.
In Fig. (2) we observe the asymptotic confidence intervals for the parameters β_0 and β_1, respectively. In Fig. (3) we observe the coverage probability the the respective parameters.

Table 2: Data with simulated response

	assint.	bootstrap	boot. stratified	Jackknife
β_0	-0,150792	-0,155002	-0,154611	-0,150821
β_1	0,002632	0,002704	0,002699	0,002632
CI(0,95;β_0) inf	-0,274769	-0,287244	-0,286260	-0,283435
CI(0,95;β_0) sup	-0,026816	-0,022760	-0,022962	-0,018206
CI(0,95;β_1) inf	0,001083	0,001058	0,001064	0,000978
CI(0,95;β_1) sup	0,004180	0,004349	0,004333	0,004287
Coverage β_0	0,986	0,982	0,984	0,985
Coverage β_1	0,957	0,956	0,955	0,960
emp. IC(0,95;β_0) inf	0	-0,290283	-0,288953	-0,159213
emp. IC(0,95;β_0) sup	0	-0,024211	-0,024259	-0,140802
emp. IC(0,95;β_1) inf	0	0,001109	0,001109	0,002516
emp. IC(0,95;β_1) sup	0	0,004418	0,004395	0,002738
emp. Coverage β_0	0	0,981	0,980	0,152
emp. Coverage β_1	0	0,949	0,949	0,112
Bias β_0	0,007413	0,011623	0,011231	0,007442
Bias β_1	-0,000389	-0,000461	-0,000456	-0,000389

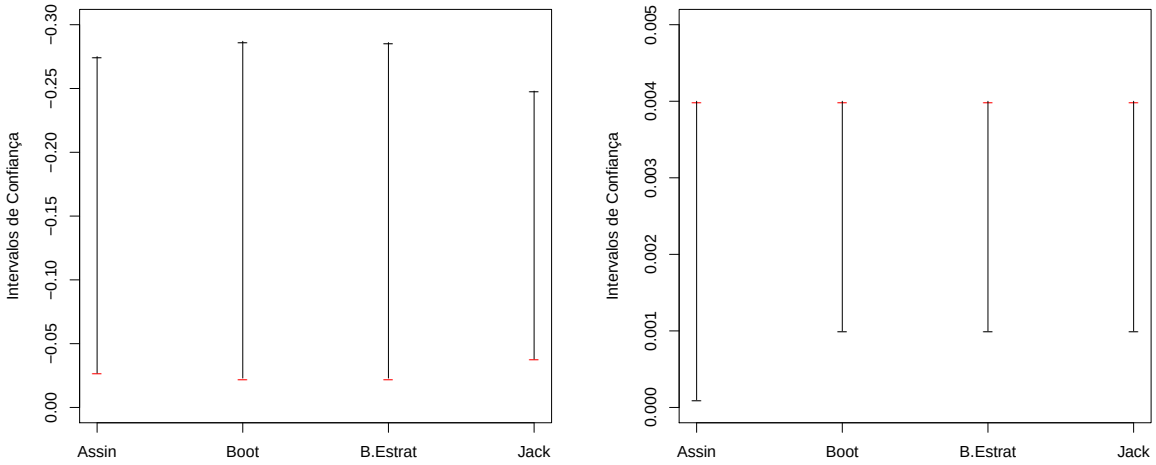

Fig. 2: Asymptotic confidence interval for β_0 and β_1, considering the simulation of the survival time with exponential link function

Table 3: Simulated Data

	assint.	bootstrap	boot. stratified	jackknife
β_0	-0,150322	-0,152268	-0,152263	-0,150347
β_1	0,002334	0,002369	0,002368	0,002335
IC(0,95;β_0) inf	-0,396869	-0,423179	-0,422669	-0,418898
IC(0,95;β_0) sup	0,096224	0,118643	0,118143	0,118204
IC(0,95;β_1) inf	-0,000575	-0,000832	-0,000827	-0,000836
IC(0,95;β_1) sup	0,005244	0,005569	0,005563	0,005506
Coverage β_0	0,949	0,956	0,954	0,957
Coverage β_1	0,954	0,957	0,957	0,959
emp. CI(0,95;β_0) inf	0	-0,421859	-0,421631	-0,166832
emp. CI(0,95;β_0) sup	0	0,124203	0,123091	-0,131870
emp. CI(0,95;β_1) inf	0	-0,000889	-0,000879	0,002123
emp. CI(0,95;β_1) sup	0	0,005568	0,005563	0,002530
emp. Coverage β_0	0	0,948	0,952	0,097
emp. Coverage β_1	0	0,957	0,952	0,105
Bias β_0	0,006943	0,008888	0,008884	0,006968
Bias β_1	-0,000373	-0,000126	-0,000125	-0,000379

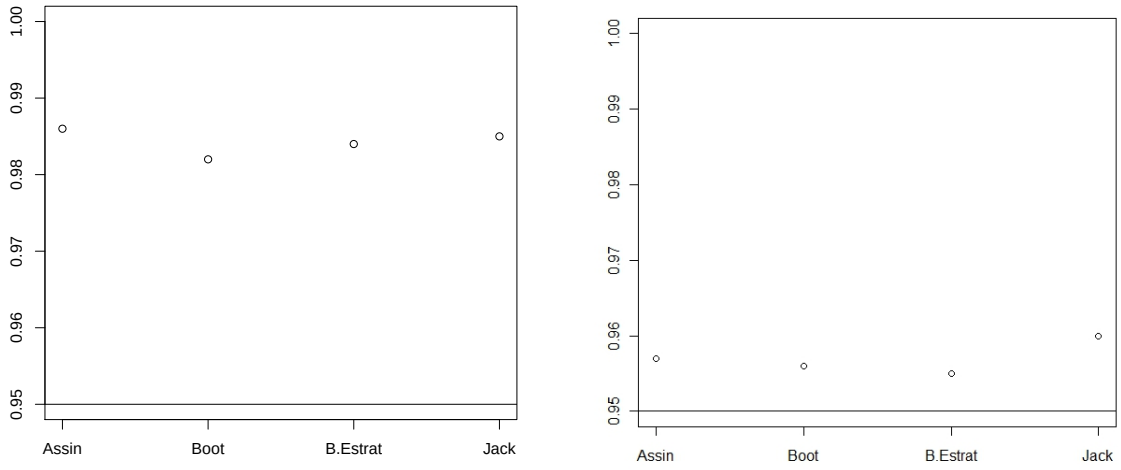

Fig. 3: Coverage for β_0 and β_1n considering the simulation of the survival time with exponential link function

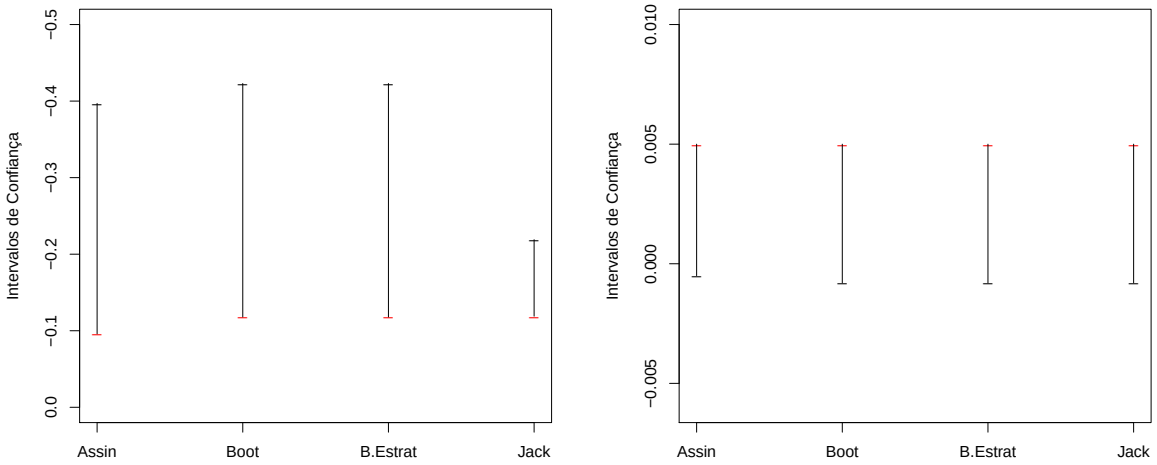

Fig. 4: Asymptotic confidence interval for β_0 simulating data, considering exponential link function

For the Table (3), we generated the all dataset. We considered as true the value of the parameter $\beta_0 = -0,150792$ and $\beta_1 = 0,002632$, so we replicated 1000 times the simulation to calculate the coverage probability through asymptotic confidence interval and using the percentile $0,025$ and the $0,975$.
Using asymptotic confidence interval we observe that the coverage probability based on Jackknife presented a better result than the bootstrap. However, when we compare the coverage probability using the percentile we observe that the bootstrap presented a better result, as Table (2).
In the Fig. (4) we observe the asymptotic confidence intervals for the parameters β_0 and β_1, respectively. And Fig. (5) we observe the coverage probability for respective parameters.

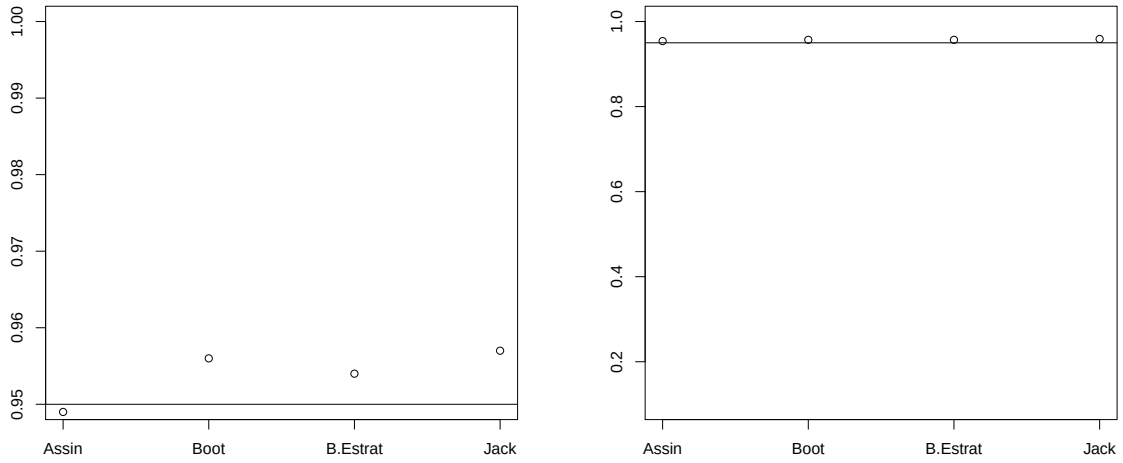

Fig. 5: Coverage for β_0 simulating data, considering exponential link function

5. Final Considerations

In this paper we could obtain information about the performance of the evaluated methods, that may support an decision in which technique simulation to use.
We observe that the Jackknife residual is efficient to analyze influential data points in the response variable.

References

[1] Bender, R., Augustin, T., Bletter, M.,"Generating Survival Times to Simulate Cox Proportional Hazards Models", *Sonderforschungsbereich* Vol.386, (2003), pp.1-19.

[2] Chernick, M. R., "Bootstrap Methods: A Guide for Practitioners and Researchers", *Wiley Series in Probability and Statistics*, New York, (2007).

[3] Cho, H., Ibrahim, J. G., Sinha, D., Zhu, H., "Bayesian Case Influence Diagnostics for Survival Models", *Biometrics*, Vol.65, (2009), pp.116-124.

[4] Colosimo, E. A., Giolo, S. R., "Análise de sobrevivência aplicada", *Edgard Blucher*, (2006).

[5] Cox, D. R., "Regression models and life-tables (with discussion)", *Journal of the Royal Statistical Society, Series B: Methodological*, Vol.34, (1972), pp.187-220.

[6] Cox, D. R., "Partial Likehood", *Biometrika*, Vol.62, (1975), pp.269-276.

[7] Efron, B., "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics*, (1979), Vol.7

[8] Efron, B., "The Jackknife, the Bootstrap and Other Resampling Plans", Department of Statistics, Stanford University, (1994).

[9] Efron, B., Tibshirani, R.J., "An Introduction to the Bootstrap", *Monographs ond Statistics and Applied Probability*, Vol.57 (1993).

[10] Martinez, E. Z., Louzada-Neto, F., "Bootstrap confidence interval estimation", *Rev. Mat. Estat.*,(Sã Paulo), Vol.19, (2001), pp.217-251.

[11] Miller, R. G., "The Jackknife - A review", *Biometrika*, Vol.61, (1974), pp.1-15.

[12] Miller, R., Halpern, J., "Regression with censored data", *Biometrika*, Vol.69, (1982), pp.521-531.

[13] Shao, J.; TU, D., "The Jackknife and Bootstrap". *Springer-Verlag*, 281, (1995).