Transform Both Sides Model: A Parametric Approach

A. Polpo[a,*], C. P. de Campos[b], D. Sinha[c], S. Lipsitz[d], J. Lin[c]

[a]*Department of Statistics, Federal University of São Carlos, Brazil*
[b]*Dalle Molle Institute for Artificial Intelligence, Switzerland*
[c]*Department of Statistics, Florida State University, USA*
[d]*Harvard Medical School, Brigham and Women's Hospital, USA*

Abstract

A parametric regression model for right-censored data with a log-linear median regression function and a transformation in both response and regression parts, named parametric Transform-Both-Sides (TBS) model, is presented. The TBS model has a parameter that handles data asymmetry while allowing various different distributions for the error, as long as they are unimodal symmetric distributions centered at zero. The discussion is focused on the estimation procedure with five important error distributions (normal, double-exponential, Student's t, Cauchy and logistic) and presents properties, associated functions (that is, survival and hazard functions) and estimation methods based on maximum likelihood and on the Bayesian paradigm. These procedures are implemented in `TBSSurvival`, an open-source fully documented `R` package. The use of the package is illustrated and the performance of the model is analyzed using both simulated and real data sets.

Keywords: Log-linear median, reliability, right-censored survival data, TBSSurvival.

*Correspondence to: Department of Statistics, Federal University of São Carlos, Rod. Washington Luiz, km 235, São Carlos-SP, Brazil, 13565-905. Tel: +55-16-33519384.

Email addresses: `polpo@ufscar.br` (A. Polpo), `cassio@idsia.ch` (C. P. de Campos), `sinhad@stat.fsu.edu` (D. Sinha), `slipsitz@partners.org` (S. Lipsitz), `jlin@stat.fsu.edu` (J. Lin)

The software is available at *http://cran.r-project.org/web/packages/TBSSurvival*.

1. Introduction

Regression models for reliability and survival censored data are widely used in many areas, for instance in engineering, medicine and biology. When choosing which model to apply to their data, users face (at least) two important questions: (i) are the theoretical properties of the model suitable for my domain? (ii) is it easy to produce and to interpret results with this model? The first question is essential, otherwise results are of no meaning. The second question relates in part to the availability and ease-of-use of software packages implementing the desired model. In this paper we propose a parametric regression model with log-linear median and tackle both aforementioned questions. The wide suitability of the model is addressed by its flexibility, allowing users to select one among many distributions for the error, in order to fit survival data with different shapes. The ease-of-use is tackled by devising the functions that characterize the model and methods to perform estimation of parameters, all implemented in a freely available open-source package.

There are many studies that deal with the quantile regression model under non-parametric and semi-parametric approaches for right-censored data (see, for example, BuHamra et al. (2004); Fung et al. (2012); Koenker (2008); Lin et al. (2012a,b)). The present work is a parametric extension of the Transform Both Sides (TBS) model of Lin et al. (2012b), a semi-parametric log-linear median regression model. Semi-parametric models such as Cox's (1972) proportional hazards model and linear transformation models (Cheng et al., 1995) are very popular for modeling effects of covariates on a survival response. Several authors, including Ying et al. (1995), gave compelling arguments in favor of focusing on the quantiles of the survival time for modeling and reporting data analysis results. The many semi-parametric and non-parametric approaches are mostly based on self-consistency and martingales, which estimate equations for the median regression (Cheng et al., 1997; Portnoy, 2003; Peng and Huang, 2008). Carroll and Ruppert (1984), and Fitzmaurice et al. (2007) propose parametric versions of a Box-Cox transform-both-sides regression model, considering only uncensored continuous responses, the original Box-Cox transformation, and the normal distribution for the error.

By using a parameter that handles the possible asymmetry in the distribution of the data, we allow the error distribution to be any zero-centered unimodal symmetric distribution. In this paper we especially consider five

important distributions: normal, double-exponential, Student's t, Cauchy and logistic, even if the model and our implementation allow the user to supply their own error distribution to be used. The choice of the error distribution strongly depends on the estimation problem being faced, and so TBS's flexibility stretches the use of the model to a wide range of problems. In other words, TBS provides a class of parametric models, according to the choice of the error distribution. We present the density function, the survival (distribution) function and the hazard function of the survival/failure time for each error distribution. We develop both maximum likelihood and Bayesian estimators, and their associated implementations. Because each model in our class has a parametric density function, algorithms for the maximum likelihood estimator and the Markov Chain Monte Carlo method for the Bayesian estimator are simple to implement, and take a reasonably short time to run. The computational tool is provided as an `R` package called `TBSSurvival`, which runs inferences with the proposed parametric regression model class. The aim of this package is to provide clean and fast procedures such that one can easily adapt their estimation methods to make use of the TBS model with minimum effort (ideally by just replacing a couple of function calls).

This paper is organized as follows. Section 2 proposes the parametric TBS model and discusses on its properties. The estimation procedures are presented in Section 3, along with an overview of the package's implementation. Section 4 contains the data examples and simulation studies, and finally the conclusions and future directions are presented in Section 5.

2. Transform-Both-Sides model

Let T_i be the survival time of subject $i = 1, \ldots, n$ and let $\boldsymbol{X_i}$ be the vector $(1, X_{1,i}, \ldots, X_{k,i})'$ of k time-constant corresponding covariates along with the intercept term. The Transform-Both-Sides (TBS) model (Lin et al., 2012b) assumes that

$$g_\lambda(\log(T_i)) = g_\lambda(\boldsymbol{\beta X_i}) + \varepsilon_i, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$ is a regression parameter, ε_i is an unspecified error with a common symmetric unimodal density f_ε free of covariates $\boldsymbol{X_i}$, and

$$g_\lambda(u) = \frac{\text{sign}(u)|u|^\lambda}{\lambda}, \tag{2}$$

with $\lambda > 0$, $\text{sign}(u) = 1$ if $u \geq 0$ and $\text{sign}(u) = -1$ if $u < 0$.

Table 1: Distributions of error that are explicitly covered in this work.

Distribution	Parameter	Density function ($f_\varepsilon(\epsilon\|\xi)$)
Normal	$\xi = \sigma^2$	$(2\pi\sigma^2)^{-1/2} \exp\{-\epsilon^2/(2\sigma^2)\}$
DoubExp	$\xi = b$	$(2b)^{-1} \exp\{-\lvert\epsilon\rvert/b\}$
Student's t	$\xi = \eta$ $(d.f.)$	$\frac{\Gamma((\eta+1)/2)}{\Gamma(\eta/2)\sqrt{\pi\eta}} \left(1 + \frac{\epsilon^2}{\eta}\right)^{-(\eta+1)/2}$
Cauchy	$\xi = c$	$[\pi c\,(1 + (\epsilon/c)^2)]^{-1}$
Logistic	$\xi = s$	$\frac{\exp\{\epsilon/s\}}{s\left[(1+\exp\{\epsilon/s\})^2\right]}$

The parametric space for all parameters is $(0, +\infty)$.

The TBS model is an extension of the Box-Cox power family (Box and Cox, 1964), a popular transformation to obtain a symmetric and unimodal density for the (transformed) random variable. We assume that the density f_ε of errors ε_i in Equation (1) is centered at zero. Lin et al. (2012b) assume a non-parametric distribution for f_ε, and a semi-parametric model for the survival time. We instead propose a parametric distribution for f_ε to obtain a parametric regression model. Table 1 presents the error distributions that are explicitly considered in this work. For each distribution, ξ denotes its free parameter. Although we focus on these five distributions, we note that any unimodal symmetrical distribution centered at zero can be similarly used, and we explain how to do so later on.

The important characteristics of the model are obtained by the density, survival and hazard functions. Equation (1) can be rewritten as

$$\begin{aligned} \varepsilon_i &= g_\lambda(\log(T_i)) - g_\lambda(\boldsymbol{\beta X_i}), \text{ or equivalently as} \\ T_i &= \exp\left\{g_\lambda^{-1}\left[g_\lambda(\boldsymbol{\beta X_i}) + \varepsilon_i\right]\right\}, \end{aligned} \tag{3}$$

where the inverse function g_λ^{-1} is such that $g_\lambda^{-1}(u) = \text{sign}(u)|\lambda u|^{\frac{1}{\lambda}}$. Note that this formulation precludes negative values of T_i, which could occur in a Box-Cox transformation if estimates of $(\boldsymbol{\beta}, \lambda, \xi)$ have some (finite sample) bias (Fitzenberger et al., 2010). Considering fixed values for the parameters $(\boldsymbol{\beta}, \lambda, \xi)$, the distribution of T_i is a transformation of the error distribution.

By using Equation (3), we obtain density and survival functions of T_i as

$$f_T(t_i) = t_i^{-1}|\log(t_i)|^{\lambda-1} f_\varepsilon\left(g_\lambda(\log(t_i)) - g_\lambda(\boldsymbol{\beta X_i})|\xi\right), \quad (4)$$

$$S_T(t_i) = S_\varepsilon\left(g_\lambda(\log(t_i)) - g_\lambda(\boldsymbol{\beta X_i})|\xi\right). \quad (5)$$

For example, if one chooses the normal distribution for the error, that is, $\varepsilon_i \sim N(0, \sigma^2)$, then

$$f_T(t_i) = \frac{|\log(t_i)|^{\lambda-1}}{t_i\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-\left[g_\lambda(\log(t_i)) - g_\lambda(\boldsymbol{\beta X_i})\right]^2}{2\sigma^2}\right\},$$

$$S_T(t_i) = 1 - \Phi\left(\frac{g_\lambda(\log(t_i)) - g_\lambda(\boldsymbol{\beta X_i})}{\sigma}\right),$$

where $\Phi(\cdot)$ is the standard normal distribution function. Density and survival functions for the other error distributions in Table 1 can be easily obtained from Equations (4) and (5) as well.

Given that T_i is a continuous variable (we assume time to be continuous), we have that the hazard function is $h(t) = f(t)/S(t)$, with f and S as given by Equations (4) and (5), respectively. Because we can use distinct error distributions, the proposed TBS model has a great variety of hazard functions, which makes possible to fit many data types/shapes. Figure 1a presents hazard functions when the error is assumed to be from a normal distribution, while Figure 1b has hazard functions when the error distribution is defined as double-exponential (Laplace distribution). Both figures show that the TBS model can be adapted to increasing, decreasing, bathtub and other types of hazard functions.

Another relevant function in survival analysis is the quantile function. Define $\epsilon_{(\alpha)}$ as the α-th quantile of ε, that is, $S_\varepsilon(\epsilon_{(\alpha)}) = 1 - \alpha$. To obtain the α-th quantile of the survival time, that is, the function $t_{(\alpha)}$ such that $S_T(t_{(\alpha)}) = 1 - \alpha$, we simply substitute ε_i in Equation (3) by $\epsilon_{(\alpha)}$ and obtain

$$t_{(\alpha)} = \exp\left\{g_\lambda^{-1}\left[g_\lambda(\boldsymbol{\beta X}) + \epsilon_{(\alpha)}\right]\right\}. \quad (6)$$

The median survival time $t_{(0.5)}$ can be obtained from Equation (6) by using the fact that the error distribution is symmetrically centered at zero, which implies in $\epsilon_{(0.5)} = 0$. Thus,

$$t_{(0.5)} = \exp\left\{g_\lambda^{-1}\left[g_\lambda(\boldsymbol{\beta X})\right]\right\} \quad (7)$$

$$= e^{\boldsymbol{\beta X}} \ (= e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}).$$

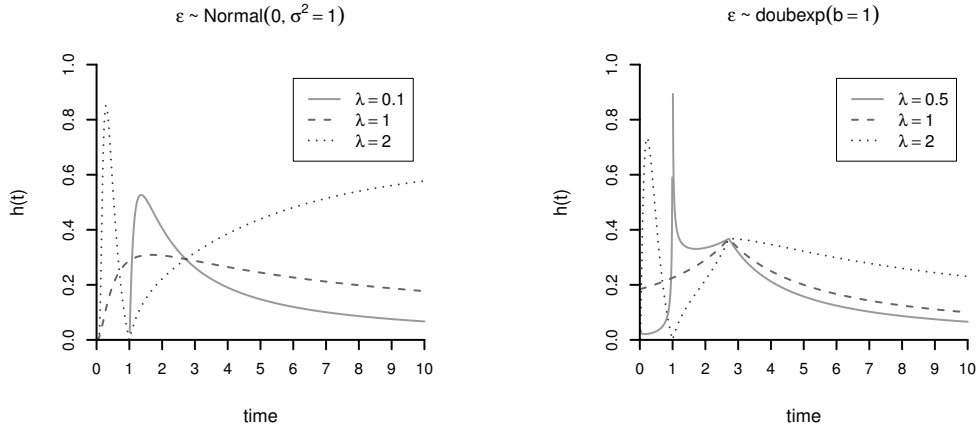

Figure 1: Hazard functions for error distribution defined as (a) normal and (b) double-exponential, no covariates and $\beta_0 = 1$.

Table 2 presents the quantile functions derived for the five error distributions that we discuss in this paper.

Table 2: Quantile functions for the error distributions.

Distribution	Parameters	$\epsilon_{(\alpha)}$
Normal	$\xi = \sigma^2$	$\sigma^2\Phi^{-1}(\alpha)$
DoubExp	$\xi = b$	$-b\ \text{sign}(\alpha - 0.5)\log(1 - 2\lvert\alpha - 0.5\rvert)$
Student's t	$\xi = \eta\ (d.f.)$	$\Psi_\eta^{-1}(\alpha)$
Cauchy	$\xi = c$	$c\tan(\pi(\alpha - 0.5))$
Logistic	$\xi = s$	$s\log(\alpha/(1-\alpha))$

Φ^{-1} is the inverse of the standard normal distribution function, Ψ_η^{-1} is the inverse of the t distribution (with η degrees of freedom) distribution function.

As a simple example, consider a binary covariate X_1 such that $X_1 = 1$ stands for the presence of some characteristic, while $X_1 = 0$ stands for its absence. In this case, we have $\boldsymbol{\beta}X = \beta_0 + \beta_1 X_1$, and the quantity so called

median odds O can be evaluated by

$$O = \frac{\text{median}(T|X_1 = 1)}{\text{median}(T|X_1 = 0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}, \tag{8}$$

which may be interpreted as "the median life time of subjects presenting the given characteristic is O times higher than the median life time of subjects without it" (note that life time is in fact lower if O is less than one). In this situation, the estimate of median(T) depends neither on the parameter λ nor on the parameter ξ of the error distribution. This is an important property of the TBS model, which implies that for any choice of error distribution and value of λ, the interpretation of the parameters $\boldsymbol{\beta}$ is directly related to the median survival time. In fact, the parameters $\boldsymbol{\beta}$ can be seen as logarithms of the median survival time, which facilitates inferences and helps with the elicitation of meaningful priors. For instance, it is easy to perform a hypothesis test for the difference in survival time caused by the characteristic $X_1 = 1$, by using the null hypothesis $\mathrm{H}_0 : \beta_1 = 0$; or to elicit subjective priors of a physician that has an opinion about the median survival time of patients with and without some characteristic.

3. Estimation Methods

Let the observed data be $\mathcal{D} = (\{y_i, \delta_i, \boldsymbol{x_i}\}, i = 1, \ldots, n)$, where $y_i = \min\{t_i, c_i\}$ is the observed survival time, $\delta_i = 0$ if $t_i \geq c_i$ (right censoring), $\delta_i = 1$ if $t_i < c_i$ (no censoring), and $\boldsymbol{x_i} = (1, x_{1,i}, \ldots, x_{k,i})'$ is the observed value of $\boldsymbol{X_i}$. As usual, we assume that the actual survival time T_i and the censoring time C_i are independent. This section presents the estimation methods based on maximum likelihood estimation (MLE) and on Bayesian estimation (BE) for the distribution of T_i.

3.1. Maximum likelihood estimation

Using the TBS model, the likelihood function for the given data is defined by the following equation:

$$L(\lambda, \boldsymbol{\beta}, \xi \mid \mathcal{D}) = \prod_{i=1}^{n} f_T(y_i \mid \lambda, \xi, \boldsymbol{\beta}, \boldsymbol{x_i})^{\delta_i} S_T(y_i \mid \lambda, \xi, \boldsymbol{\beta}, \boldsymbol{x_i})^{1-\delta_i}. \tag{9}$$

For the case with no censoring, we have $\delta_i = 1$ for every i, which is a particular case of the likelihood in Equation (9) where S_T disappears. For numerical

convenience, we maximize the log-likelihood function, which can be written as

$$l(\lambda, \boldsymbol{\beta}, \xi \mid \mathcal{D}) = \sum_{i=1}^{n} (\delta_i \log [f_T(y_i \mid \lambda, \xi, \boldsymbol{\beta}, \boldsymbol{x_i})] + (1 - \delta_i) \log [S_T(y_i \mid \lambda, \xi, \boldsymbol{\beta}, \boldsymbol{x_i})]) . \quad (10)$$

Unfortunately, there is no closed form solution for the (log-)likelihood maximization problem, and the numerical optimization of Equation (10) towards a global optimum solution is not in general an easy task. One important characteristic of it is the existence of partial derivatives with respect to the unknowns $(\lambda, \boldsymbol{\beta}, \xi)$, which helps numerical methods in finding good solutions for the maximization problem. In order to avoid some local maxima, we approach the problem by using a collection of different optimization methods, as well as multiple starting points (this should usually yield reasonable estimates).

Regarding the implementation, we develop `tbs.survreg.mle`, a function which acts as a wrapper to the TBS internal optimization procedure. The optimization is conducted by calling many different numerical methods of the `optim` function (R Development Core Team, 2012), namely Nelder-Mead, Broyden-Fletcher-Goldfarb-Shanno (BFGS), Conjugate Gradient (CG), and Simulated Annealing (SANN), and the augmented Lagrange multiplier method of `Rsolnp` (Ghalanos and Theussl, 2011). Finally, the best solution (in terms of likelihood) is returned. The TBS internal optimization procedure implements a multiple starting point idea, executing the maximization as many times as desired. It can also perform some fallbacks among methods in order to find feasible starting points, which are important for all of the previously mentioned numerical methods. The procedure makes use of the TBS density function `dtbs` and distribution function `ptbs`, which may vary according to the selection of the error distribution (this is implemented in a transparent manner to the user).

3.2. Bayesian estimation

In order to perform the Bayesian estimation, we need to compute the posterior distribution of $(\lambda, \boldsymbol{\beta}, \xi)$. We consider independent priors for the parameters, that is,

$$p(\lambda, \boldsymbol{\beta}, \xi | \mathcal{D}) \propto L(\lambda, \boldsymbol{\beta}, \xi \mid \mathcal{D}) p(\lambda) p(\boldsymbol{\beta}) p(\xi),$$

where $L(\lambda, \boldsymbol{\beta}, \xi \mid \mathcal{D})$ is the likelihood function given in Equation (9), and $p(\lambda)$, $p(\boldsymbol{\beta})$ and $p(\xi)$ are prior densities.

Based on the characteristics of the transformation g_λ, the most plausible values for λ lie on the interval $(0, 3)$, and thus we use a prior distribution with high density for values within that interval and a decreasing density for values larger than 3. Considering that the parameter ξ of the error distribution is directly related to its variance (according to the distributions used in this work), we suggest a prior that favors points in the interval $(0, 2)$, because we do not expect the error distribution to have too large a variance.

As mentioned before, an important property of the TBS model is the interpretation of the parameters $\boldsymbol{\beta}$ in terms of median survival time and median ratios. For this reason, the elicitation of priors for $\boldsymbol{\beta}$ is usually an easy task, for example by asking some quantiles to the specialist. Using Equations (7) and (8), it is possible to translate the specialist's prior median into values of $\boldsymbol{\beta}$. In the `TBSSurvival` package, we take the normal density for this prior, but we leave mean and variance to be chosen by the user (default values are set nevertheless).

As in the maximum likelihood estimation, there is no closed-form solution for the Bayesian estimators. In fact, in the case of Bayesian estimation, we cannot even find a closed-form solution for the posterior distribution. Hence, we are constrained to the use of simulation methods, such as the Metropolis-Hastings algorithm, to generate a sample from the joint posterior distribution in order to evaluate it. We implement the `tbs.survreg.be` function, which processes the data, calls the function `metrop` of the `mcmc` package (Geyer, 2010) to simulate from the posterior density, and consolidates the results to obtain the final estimation. More implementation details are available in the `TBSSurvival` documentation.

4. Experiments

In this section we present experiments with both simulated and real survival data. We show two real data examples to illustrate the use of the package and the performance of the TBS model. We also perform a simulation study to show the quality of the TBS model in a large amount of samples. Additional study cases, including examples with real data, are available and explained in the documentation of the estimation functions and in the test codes of the package. We start by a simulation study that compares TBS

Table 3: Bias and mean squared error (MSE) for the parameter estimates with MLE using simulated data (1000 copies) from the TBS model with multiple distinct values of (λ, ξ, β_0).

		$\widehat{\lambda}$		$\widehat{\xi}$		$\widehat{\beta_0}$	
Dist.	Cens.	Bias	MSE	Bias	MSE	Bias	MSE
Normal	0%	0.0010	0.0052	-0.0121	0.0520	0.0004	0.0025
	20%	0.0036	0.009	-0.0148	0.0728	0.0006	0.0025
	40%	0.0060	0.0149	-0.0213	0.1114	0.0003	0.0027
	60%	0.0111	0.0265	-0.0418	0.3344	-0.0037	0.0049
DoubExp	0%	0.0015	0.0051	0.0023	0.0066	-0.0009	0.0031
	20%	0.0038	0.0086	0.0036	0.0084	-0.0009	0.0031
	40%	0.0078	0.011	0.0043	0.0096	-0.0009	0.0031
	60%	0.0098	0.0144	0.0051	0.0122	-0.0020	0.0056
Student's t	0%	-0.0016	0.0028	-0.0156	0.0138	0.0013	0.0026
	20%	-0.0126	0.0032	-0.0256	0.0158	0.0034	0.0053
	40%	-0.0136	0.0027	-0.0361	0.0189	-0.0105	0.0048
	60%	-0.0190	0.004	-0.0365	0.0196	0.0001	0.0062
Cauchy	0%	-0.0197	0.0038	-0.0254	0.0091	0.0002	0.0027
	20%	-0.0094	0.0031	-0.0114	0.0080	0.0018	0.0064
	40%	-0.0109	0.0034	-0.0081	0.0090	0.0002	0.0065
	60%	-0.0128	0.0041	0.0016	0.0149	0.0075	0.0098
Logistic	0%	0.0008	0.003	-0.0008	0.0040	0.0033	0.0097
	20%	0.0018	0.0045	-0.0003	0.0048	0.0034	0.0097
	40%	0.0039	0.0066	0.0002	0.0059	0.0030	0.0100
	60%	0.0063	0.0105	0.0007	0.0098	0.0017	0.0193

with other available methods. A detailed description of the package usage is given in Section 4.2, together with the real data examples.

4.1. Simulation study

We analyze the proposed parametric TBS model in two distinct situations: (i) we simulate data in order to understand the TBS's ability to fit them; (ii) we compare estimation methods in the literature with the MLE and the Bayesian estimator of TBS (and themselves with each other).

The first simulation study has data generated from the TBS model for each of the five error distributions and each of the four censoring levels (0%, 20%, 40%, 60%), that is, 20 different combinations. Within each of these

combinations, we have simulated data according to different values of the TBS parameters λ, ξ, and β_0 (no covariates are used here) such that the experiment is not specific to a single choice of values. The used values are as follows: $\lambda = 0.5, 1, 2$, $\xi = 0.5, 1, 2$ and $\beta_0 = 1, 5$, which comprises $3 \times 3 \times 2 = 18$ scenarios. For each of these scenarios, we generate data with sample size of 1000 units, and we repeat this process with 1000 different copies, giving a total of 18 thousand replications for each censoring level and error distribution. Using MLE (and all numerical optimizers, from which we took the maximum likelihood one) in each of these 20×18 thousand copies, we produce the estimates $(\hat{\lambda}, \hat{\xi}, \hat{\beta}_0)$ whose bias and mean squared error are averaged and presented in Table 3. We see that bias and mean squared error of each estimated parameter is reasonably close to zero. The aim of this simulation study is to assert that the parameters of TBS model are correctly estimated when data are generated from the own model.

In a second simulation study, we compared the performance of the TBS model against the estimators of Portnoy (2003) and Peng and Huang (2008), using the implementation available in the `R` package `quantreg`. We have generated data from the TBS model with normal error distribution, one binary covariate, twenty different combinations of the parameters ($\lambda = \{0.5, 1.5\}$, $\xi = \{0.5, 1.5\}$, $\boldsymbol{\beta} = \{(-1, 1),\ (1, -1),\ (-0.5, 0.5),\ (0.5, -0.5),\ (0.5, 0.5)\}$), and two sample sizes ($n = 100, 1000$). The covariate X was generated from a Bernoulli distribution with probability of success 0.7. As for the censoring mechanism, we have considered a fixed value κ such that if $T_i > \kappa$, then the observed time was censored and its value set to κ (for example, this simulates a period of product testing within a factory; after that period, the survival time is censored). The value of κ was chosen such that approximately 20% of censored data were present in each case. Using these data, we have performed the estimation at five distinct quantiles (5%, 25%, 50%, 75% and 95%), evaluating the bias and the mean squared error (MSE). The TBS estimation with the Bayesian method was done with the default fixed values as defined in the `TBSSurvival` package for the MCMC procedure. This can lead to non-convergence of the MCMC chain, which we have not checked (it would be extremely time-consuming to manually verify the convergence of the MCMC in each case of these many simulations). Despite of that, the Bayesian estimator performed very well. For the MLE estimation, we have chosen to run three optimization methods with quite different internal characteristics (BFGS, Nelder-Mead and Rsolnp), from which the TBS estimation method automatically chooses the best (for each test case). The error distribution

has been defined as normal in all test cases, so we expected that estimated values from the TBS model should have lower bias and MSE when compared to values estimated with Portnoy (2003) and Peng and Huang (2008), which in fact has happened. The results are shown in Table 4. Because of the characteristics of these two other estimators, it was not possible to perform the estimation at the 95% quantile of the survival time (some of the cases for the 75% quantile and $n = 100$ were also impossible to be estimated). This can be seen by the percentage of successful estimations (lines marked with a (%) in Table 4). In one hand, all methods presented similarly good results for small quantiles (5%, 25% and 50%). On the other hand, the TBS model has estimated well even for high quantiles, although we notice that the MSE has considerably increased in such cases. These results suggest that the estimation with the non-parametric quantile regression model is as good as the parametric TBS model apart from estimations at high quantiles, and thus should be preferred at low quantiles. This is however dependent on having relatively small amount of censored data, otherwise quantile regression models may face some estimation problems, which is later discussed in our real data applications. We note that a comparison against the model of Lin et al. (2012b) was not possible because their method showed poor MCMC convergence. In fact, the chain converged, however we obtained high autocorrelation, so we could not guarantee independence of the generated points from the posterior. We have also noticed that the model of Lin et al. (2012b) seems to work better for small sample sizes ($n \leq 100$).

In our experiments, we cannot report a single best continuous optimization method to be used inside the MLE. We suggest the user to run at least BFGS, Nelder-Mead and Rsolnp, since CG is similar in essence to BFGS and SANN is a very slow procedure. According to our experiments, CG and SANN are recommended only when the others fail. Thus, we have implemented this idea in our package: if the chosen method fails to find even a feasible solution, SANN is automatically called too. One can also think of how to choose an error distribution. This again depends on the problem instance, and it is not possible to tell in advance which distribution will perform the best. For example, we might say that the parameter of the Student's t distribution approximates very well both the normal and the Cauchy distributions, so it should be preferred to the other two. However, because of the distinct nature of the parameters of each distribution (degree of freedom for Student's t, which controls its tails, while for normal and Cauchy the parameter is about their scales), we would have different models, so nor-

Table 4: Comparison of estimation methods for 20% right-censored data generated from the TBS model with normal error distribution. Columns show the estimation error in some specific quantiles. For the estimations of Portnoy (2003) and Peng and Huang (2008), it is also shown the percentage of cases where the method has found a solution.

	n		5%	25%	50%	75%	95%
Portnoy	100	Bias	-0.038	-0.017	-0.051	-0.390	–
		MSE	0.008	0.026	0.120	0.646	–
		(%)	1.000	1.000	1.000	0.434	0.000
	1000	Bias	-0.019	-0.012	-0.031	-0.074	–
		MSE	0.001	0.002	0.014	0.056	–
		(%)	1.000	1.000	1.000	0.976	0.000
Peng-Huang	100	Bias	-0.001	0.019	0.026	-0.099	–
		MSE	0.007	0.030	0.128	0.410	–
		(%)	1.000	1.000	1.000	0.724	0.000
	1000	Bias	0.003	0.007	0.017	0.036	–
		MSE	0.001	0.002	0.014	0.062	–
		(%)	1.000	1.000	1.000	0.981	0.000
TBS MLE	100	Bias	0.006	0.014	0.011	-0.004	-0.116
		MSE	0.002	0.018	0.091	0.475	3.110
	1000	Bias	0.000	0.002	0.005	0.006	-0.002
		MSE	0.000	0.001	0.010	0.045	1.586
TBS BE	100	Bias	0.000	0.014	0.032	0.111	0.186
		MSE	0.003	0.019	0.101	0.631	3.150
	1000	Bias	-0.002	0.001	0.006	0.024	0.119
		MSE	0.000	0.002	0.009	0.056	1.562

mal and Cauchy distributions shall be considered as reasonable options too. The `TBSSurvival` package allows us to evaluate and compare different error distributions in order to take an informed decision.

4.2. Real data sets

Two study cases with real data sets are considered: one for the reliability of equipments and another with the survival time of patients with colon cancer. In the first data set, we perform an inferential analysis of the *Alloy T7987* data set (Meeker and Escobar, 1998, pp. 130–131) using the estimation of the TBS model with the five described error distributions (of Table 1) using both MLE and BE. Throughout the analysis, we provide the `R` code that illustrates the use of the `TBSSurvival` package. The data consist in a sample of 67 specimens of the equipment Alloy T7987 that failed before having accumulated 300 thousand cycles of testing and 5 specimens that survived at least 300 thousand cycles without failure (this was the censoring time). In Table 5, some summary statistics of the failure time are presented.

Table 5: Summary statistics of the failure time in thousand cycles for the Alloy T7987 data set.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
94.0	135.8	168.0	175.4	198.5	300.0

The function `tbs.survreg.mle` is used to perform the MLE, here assuming no covariates. The only necessary argument to the estimation function is the survival formula with the *time* of events and the censoring indicator *delta* from the Alloy data set as arguments. The usage is very similar to the usage of the estimation functions in the `survival` package (Therneau, 2012). This is convenient if one wants to adapt code that already contains those estimation functions to be used with TBS. For the Alloy T7987 data set, the MLE can be performed using the following `R` code:

```
library("TBSSurvival")
data(alloyT7987)
tbs.mle <- tbs.survreg.mle(Surv(alloyT7987$time,
                                alloyT7987$delta) ~ 1,
                           dist=dist.error("all"))
```

The above code runs the MLE estimation using all available optimization methods, then chooses automatically the best one for the data. Furthermore, it automatically estimates with all the five error distributions discussed previously, returning the estimation for each one of them, as well as an indication of the model with best Akaike information (accessible through the element `tbs.mle$best`).

Table 6 presents some characteristics of the estimated model: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the parameter estimates and (in parenthesis) the standard error of the parameter estimates for the MLE. The results in Table 6 were especially formatted for this paper, and they can be obtained using the function

```
print(tbs.mle)
```

We see that the TBS model, for this data set, has better fit (lowest AIC) when the error distribution is normal. Figure 2 presents the reliability (that is, survival) and hazard functions for the estimated TBS model. Figure 2a also shows the Kaplan-Meier estimated curve for comparison. These figures are obtained using the following code:

```
### Survival plot
plot(tbs.mle$norm)
km <- survfit(formula = Surv(alloyT7987$time,
              alloyT7987$delta == 1) ~ 1)
lines(km) # add the Kaplan-Meier estimates to the
          # survival plot.

### Hazard plot
plot(tbs.mle$norm,plot.type="hazard")
```

Another aspect that can be used to verify the goodness of fit is to check whether the errors ε_i follow the estimated error distribution and whether their values are close to zero. Table 7 gives the summary statistics of the error, which shows a reasonably good fit of the model to the data. Table 7 and some other results, such as the Wald's test for the parameters $\boldsymbol{\beta}$, are obtained using

```
summary(tbe.mle$norm)
```

Table 6: Some quantities of the TBS Model for the Alloy T7987 data set using MLE: AIC, BIC, parameter estimates and their standard errors in parenthesis.

Error Distribution	AIC	BIC	$\widehat{\lambda}$	$\widehat{\beta_0}$	$\widehat{\xi}$
Normal	737.95	744.78	0.0301	5.1214	0.0044
			(0.1161)	(0.0384)	(0.00142)
DoubExp	740.25	747.08	0.0021	5.1240	0.0506
			(0.4751)	(0.0114)	(0.03883)
Student's t	741.76	748.59	1.6855	5.1311	51.6520
			(0.1994)	(0.0739)	(1051.07353)
Cauchy	751.71	758.54	0.0028	5.0879	0.0368
			(0.9660)	(0.0346)	(0.05759)
Logistic	738.39	745.22	0.0068	5.1066	0.0373
			(1.0979)	(0.0388)	(0.06671)

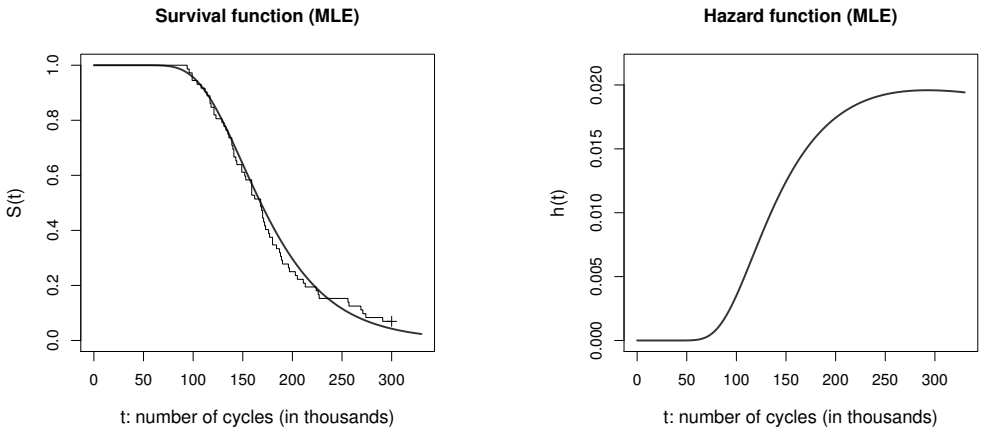

Figure 2: Alloy T7987 data set (MLE): (a) Survival function (continuous line regards the estimated TBS model with normal distribution for the error; the step function regards the Kaplan-Meier estimates) and (b) hazard function (estimated by the TBS model with normal distribution for the error).

Table 7: Summary statistics of the error with the TBS model, using MLE and the normal distribution for the error.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.12690	-0.04809	-0.01204	-0.01223	0.02267	0.10650

As explained in Section 2, the median survival/failure time can be estimated by $e^{\widehat{\beta_0}}$ ($\widehat{\beta_0}$ is the usual notation for an estimate of β_0), and the 95% confidence interval for the median can be obtained by

$$\left(\exp\left\{\widehat{\beta_0} + \phi_{(0.025)}\mathrm{sd}(\widehat{\beta_0})\right\}, \exp\left\{\widehat{\beta_0} + \phi_{(0.975)}\mathrm{sd}(\widehat{\beta_0})\right\}\right), \tag{11}$$

where $\phi_{(\alpha)}$ is the α quantile of a random variable with standard normal distribution. The `R` code to perform these calculations is:

```
### Quantile estimates
# Median survival/failure time:
exp(tbs.mle$norm$beta)

### 95% I.C. for the median failure time
# lower bound:
exp(tbs.mle$norm$beta+qnorm(0.025,0,1)*tbs.mle$norm$beta.se)
# upper bound:
exp(tbs.mle$norm$beta+qnorm(0.975,0,1)*tbs.mle$norm$beta.se)
```

Using these formulas, we obtain the median as 167.57 and the 95% confidence interval as $(155.41, 180.68)$. Note that the estimated median failure time is very close to the median given in Table 5, probably because there are only 7% of cases with censored data.

Now we turn our attention to the Bayesian estimation. In order to perform BE, we employ the function `tbs.survreg.be`. It receives as arguments the same formula as the one used for the MLE, plus the initial guess values and some usual control parameters for the Markov Chain Monte Carlo (MCMC) procedure, which follow the convention of the function `metrop` of the `mcmc` package (Geyer, 2010) (for more details on the control parameters, we refer to Gamerman and Lopes (2006), Chapter 6). It also allows the user to define `prior.mean` and `prior.sd`, the hyper-parameters for the prior distribution, as described in Section 3.2.

Table 8 presents the deviance information criterion (DIC), parameters' point estimates (posterior mean) and their standard deviations from the posterior distribution of the parameters. For the Bayesian estimation, we see that the TBS model has best fit when the error distribution is the logistic (according to the DIC), which differs from the MLE case, where the normal distribution has had the best fit. The Gelman-Rubin statistics for the parameter estimates $\widehat{\lambda}$, $\widehat{\xi}$, and $\widehat{\beta_0}$ are, respectively, 1.006, 1.002, and 1.000, which indicates that the convergence of MCMC was considerably good. We have used four chains with different initial values to evaluate the Gelman-Rubin Statistic. The burn-in used was 500 thousands, the jump between observations was two thousands, and the sample size of the posterior was one thousand (these values are greater than the usually required ones, so the convergence results should be accurate enough). Since the convergence of MCMC sampling from the posterior distribution depends on user verification, it is not possible to implement an automatic method to chose the best model (distribution error) in the case of the Bayesian analysis. For instance, the BE with logistic distribution for the error is achieved using the code:

```
tbs.be.logistic <- tbs.survreg.be(Surv(alloyT7987$time,
                                  alloyT7987$delta) ~ 1,
                                  dist=dist.error("logistic"),
                                  burn=500000,jump=2000,
                                  size=1000,scale=0.07)
print(tbs.be.logistic)
summary(tbs.be.logistic)
```

For the estimation using other error distributions one can just change the `dist.error()` call within the arguments of `tbs.survreg.be` and check the convergence of the MCMC samples.

We also computed the high posterior density (HPD) interval with 95% of credibility, obtaining median equal to 166.75 and HPD interval equal to $(153.64, 179.82)$. A possible advantage of BE over MLE is that one can easily evaluate the HPD boundaries for the survival function, as shown in Figure 3. The `R` code to build the HPD for the median survival/failure time and the code to generate the graphs are:

```
### Quantile estimates
# Median failure time (post[,3] contains the info about beta):
median(exp(tbs.be.logistic$post[,3]))
```

Table 8: Some quantities of the TBS model for the Alloy T7987 data set using BE: DIC, parameter estimates and their standard deviations in parenthesis.

Error Distribution	DIC	$\widehat{\lambda}$	$\widehat{\beta_0}$	$\widehat{\xi}$
Normal	727.55	1.4743	5.1294	0.994
		(0.4311)	(0.0392)	(1.0008)
DoubExp	729.43	1.4822	5.106	0.9014
		(0.6198)	(0.0365)	(0.74612)
Student's t	739.74	1.6742	5.132	72.37
		(0.0549)	(0.0399)	(19.74458)
Cauchy	745.83	1.8302	5.0822	1.0267
		(0.5929)	(0.0351)	(0.7146)
Logistic	719.39	1.5247	5.1165	0.7392
		(0.6756)	(0.041)	(0.65695)

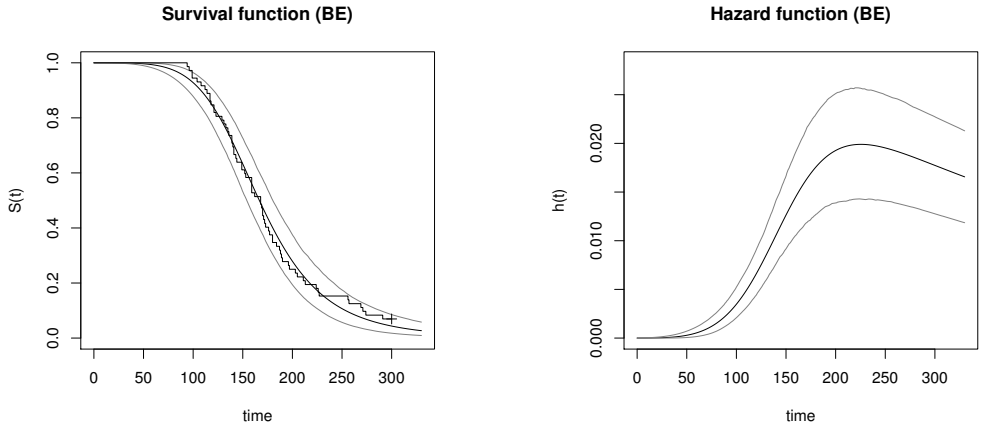

Figure 3: Alloy T7987 data set (BE): (a) Survival function (continuous black line regards the estimated TBS model with logistic distribution for the error; gray lines show the 95% HPD credible interval; the step function regards the Kaplan-Meier estimates) and (b) hazard function (black line is for the estimated TBS model with logistic distribution for the error; gray lines show the 95% HPD credible interval).

```
### 95% I.C. for the median failure time
HPDinterval(as.mcmc(exp(tbs.be.logistic$post[,3])),0.95)

### Survival plot
plot(tbs.be.logistic)

### Hazard plot
plot(tbs.be.logistic,plot.type="hazard")
```

In the second study case, we use the well-known *colon* data set from the package `survival` (Therneau, 2012), with right-censored survival data and covariates for patients with colon cancer. Our goal is to demonstrate the use of the TBS model for survival analysis with covariates. For that purpose, we use a covariate, available in the data set, which indicates whether the number of cancer-affected lymph nodes is greater than 4. This variable is known to split the patients in groups of distinct survival outcome. The survival data correspond to *Progression Free Survival* (PFS) time, that is, time until disease recurrence. The function call for the TBS estimation procedure in the presence of covariates follows the same standard as in `R` formulas, for example `Surv(colon$time,colon$status) ~ colon$node4`, using the covariates on the right-hand side. We omit further implementation details, because they follow the very same structure as the one presented for the Alloy data set. Figure 4 shows the estimated survival curves using MLE and BE. The Kaplan-Meier estimation is also presented for comparison. We see that the HPD credible intervals nicely encompass the Kaplan-Meier curves. Using the TBS model with the Bayesian estimation, the median PFS time for patients with more than four affected lymph nodes is 841.08 days, with 95% credible interval (c.i.) equal to $(695.37, 982.69)$, while for patients with less than or equal to four lymph nodes is 3372.81 days, c.i. $(2920.62, 3831.99)$. The median odds O, as described in Section 2, is 0.25, c.i. $(0.2, 0.31)$, which means that the median PFS time for patients with at most four affected lymph nodes is four times the median PFS time of the others. Note that, by using the median odds with the TBS model, we are able to obtain confidence intervals for O. Results using the MLE are very similar, and so omitted.

Finally, we point out that methods available in the `R` package `quantreg` for non-parametric quantile regression models have also produced good estimates when targeting low quantiles of the Colon data set, but were not able to

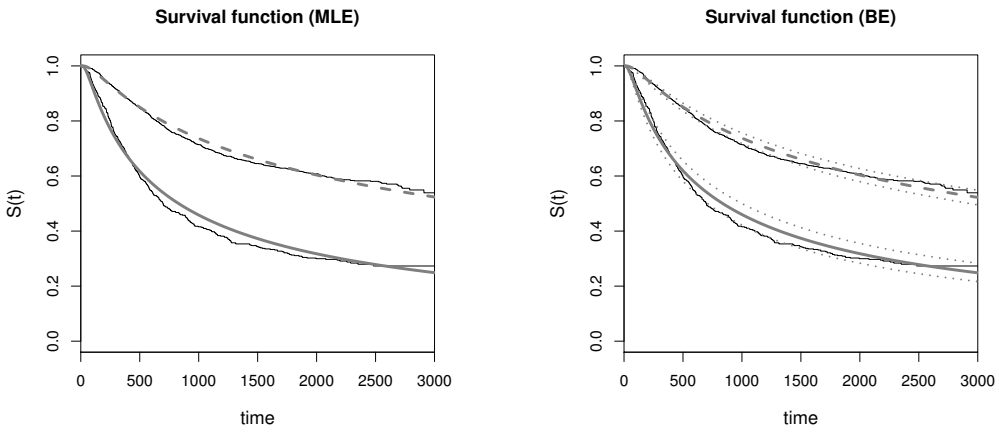

Figure 4: *Colon* data set: (a) MLE estimation and (b) BE estimation using the number of lymph nodes greater than four as discriminant between the two groups of patients (dashed gray line is the curve for patients with at most four affected lymph nodes; solid gray line is for patients with more than four affected lymph nodes; black step function regards the Kaplan-Meier estimates; in the BE plot, the dot lines show the 95% HPD credible interval).

estimate the median (more specifically, they have failed with quantiles greater than 45%, approximately), likely because of the considerably high amount of censoring in these data.

5. Closing Remarks

We presented the parametric regression model based on the Transform-Both-Sides (TBS) with log-linear median, where a parameter handles the data asymmetry, while the error distribution can be any unimodal symmetric distribution centered at zero. We explicitly worked with five error distributions, namely normal, double-exponential, Student's t, Cauchy and logistic. We presented the relevant functions that characterize the model (density, survival distribution, hazard, and quantile) and developed the maximum likelihood estimation and the Bayesian estimation, along with some additional methods for confidence/credible intervals.

The TBS model has been illustrated with applications in both simulated and real data sets, and has been compared to quantile regression models. When one has to deal with the estimation of low quantiles and/or data without much censoring, quantile regression models are a very good option. However, they may fail to estimate high quantiles when there is a considerable amount of censored data, in which case the TBS model seems to be more suitable. Besides quantile estimation, we also discuss on how TBS can be used to compare treatments and to help in eliciting subjective priors.

These procedures have been implemented using the `R` language in the `TBSSurvival` package, which we made freely available. The main procedures work similarly to other widely used packages for survival analysis, which allows the user to promptly replace their estimation methods by the estimation with TBS methods, if they will. The implementation is already mature, even though we will continue to enhance it, specially towards new distribution functions and faster and more accurate numerical optimization methods for the estimation, which we acknowledge to be an important point that can be improved. Other extensions to the package are also planned, including the semi-parametric estimation, and the treatment of other censoring types.

Acknowledgments

The research work for this paper has been partially supported by grants from the São Paulo Research Foundation (FAPESP) of Brazil, from the Hasler

Foundation grant n.10030, from the Swiss National Supercomputing Centre (*small development project*), and from the National Cancer Institute (NCI) of the USA. The authors would like to thank the editorial board and reviewers for their helpful and constructive comments.

References

Box, G. E. P., Cox, D. R., 1964. An analysis of transformations. Journal of the Royal Statistical Society, Series B 26, 211–243.

BuHamra, S. S., Al-Kandari, N. M., Ahmed, S., 2004. Inference concerning quantile for left truncated and right censored data. Computational Statistics & Data Analysis 46 (4), 819–831.

Carroll, R. J., Ruppert, D., 1984. Power-transformations when fitting theoretical models to data. Journal of the American Statistical Association 79, 321–328.

Cheng, S., Wei, L., Ying, Z., 1995. Analysis of transformation models with censored data. Biometrika 82, 835–845.

Cheng, S., Wei, L., Ying, Z., 1997. Predicting survival probabilities with semiparametric transformation models. Journal of the American Statistical Association 92 (437), 227–235.

Cox, D., 1972. Regression models and life-table. Journal of the Royal Statistical Society, Series B 34, 187–202.

Fitzenberger, B., Wilke, R. A., Zhang, X., 2010. Implementing Box-Cox Quantile Regression. Econometric Reviews 29(2), 158–181.

Fitzmaurice, G. M., Lipsitz, S. R., Parzen, M., 2007. Approximate median regression via the Box-Cox transformation. The American Statistician 61, 233–238.

Fung, W.-K., He, X., Hubert, M., Portnoy, S., Wang, H. J., 2012. Editorial for the special issue on quantile regression and semiparametric methods. Computational Statistics & Data Analysis 56 (4), 753–754.

Gamerman, D., Lopes, H. F., 2006. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman and Hall/CRC.

Geyer, C. J., 2010. Markov Chain Monte Carlo. R package version 0.8.
URL http://cran.r-project.org/web/packages/mcmc

Ghalanos, A., Theussl, S., 2011. Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. R package version 1.11.
URL http://cran.r-project.org/web/packages/Rsolnp

Koenker, R., 2008. Censored quantile regression redux. Journal of Statistical Software 27 (6), 1–25.

Lin, G., He, X., Portnoy, S., 2012a. Quantile regression with doubly censored data. Computational Statistics & Data Analysis 56 (4), 797–812.

Lin, J., Sinha, D., Lipsitz, S., Polpo, A., 2012b. Semiparametric Bayesian survival analysis using models with log-linear median. Biometrics 68(4), 1136–1145.

Meeker, W., Escobar, L., 1998. Statistical Methods for Reliability Data. John Wiley & Sons, New York.

Peng, L., Huang, Y., 2008. Survival analysis with quantile regression models. Journal of the American Statistical Association 103 (482), 637–649.

Portnoy, S., 2003. Censored regression quantiles. Journal of the American Statistical Association 98, 1001–1012.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.r-project.org/

Therneau, T., 2012. A Package for Survival Analysis. R package version 2.36-12.
URL http://cran.r-project.org/web/packages/survival

Ying, Z., Jung, S.-H., Wei, L., 1995. Survival analysis with median regression models. Journal of the American Statistical Association 90, 178–184.