Biocybernetics and Biomedical Engineering
2012, Volume 32, Number 3, pp. 69–78

Linear Regression Modeling of Interval-censored Survival Times Based on a Convex Piecewise-linear Criterion Function

PAWEŁ KAŁUŻNY[1,*], LEON BOBROWSKI[1,2]

[1]*Nałęcz Institute of Biocybernetics and Biomedical Engineering, Polish Academy Sciences, Warsaw, Poland*
[2]*Faculty of Computer Science, Białystok Technical University, Białystok, Poland*

Regression models of censored survival data are often required to handle the cases, where information on the dependent (response) variable is only available as intervals, within which the actual values are located. We report on implementation and some preliminary tests of a new general method for regression with an interval-censored response variable. This method is based on minimization of a convex piecewise-linear (CPL) criterion function introduced earlier for perceptron-type classifier design. The presented interval regression method (CPL-IR) can handle arbitrary pattern of exact and left-, right-, or interval-censored data in one flexible computational framework.

K e y w o r d s: interval regression, interval censoring, censored data, current-status data, survival time, CPL function

1. Introduction

Linear regression modeling with survival time as the dependent variable and some other variables as predictors frequently is required to handle censoring of survival times. The censored survival times are not known exactly but, instead, are known to be left-, right-, or interval-bounded by the time, depending on when the subject enters and leaves the study or depending on times between the consecutive examinations. By far the most common, and the most studied, are mixtures of the exact and right-censored survival times, arising if some subjects dropped-out or survived the termination of the study [1]. Classical survival analysis methods were primarily

* Correspondence to: Paweł Kałużny, Nałęcz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, ul. Księcia Trojdena 4, 02-109 Warsaw, Poland, e-mail: pkaluzny@ibib.waw.pl
Received 9 January 2012; accepted 14 May 2012

developed for this type of data and there exist mature commercial [2] and public domain software [3] for construction of nonparametric estimators of survival function, semiparametric Cox regression models of the hazard function and parametric maximum likelihood methods for regression analysis (we refer to [1] for general theoretical background).

Another, much less common, but important pattern, is encountered, if the dataset contains no exact observations of the survival time but only the left- and right- censored observations. This censoring type is called "current status" censoring, because each subject is submitted to a single (and often destructive) examination for presence or absence of the condition of interest (e.g. appearance of disease symptoms) and is analyzed mostly with maximum likelihood and nonparametric methods. Such data appear difficult for direct application of regression analysis techniques based on minimization of prediction error because no exact value of the dependent variable is known.

Interval censoring is the most general censoring pattern and typically arises in longitudinal studies or repetitive patient examinations, when the subject is known to survive certain point of time t^- and not survive beyond the other time point t^+. The interval-censored data are recently growing in importance in medical research and there is also increased interest in development of corresponding analysis methods [4].

A novel approach for solving the general interval regression problem was recently proposed by Bobrowski [5, 6] . This method formulates linear regression analysis of the interval-censored response variable as a geometrical problem of optimal linear separation of data points and is capable to handle arbitrary censoring patterns.

The purpose of this note is presentation of a version of this method and its demonstration in the context of a real survival dataset. We are interested in the regression coefficients and estimates of survival function obtained using the predicted survival times. We compare the CPL-IR with some results of parametric survival regression and the Cox model. The R system for statistical computing [7] was used for implementation of the CPL-IR method and as the resource of survival analysis data and techniques.

2. Interval Regression Based on CPL Error Criterion

The data for the interval regression analysis are given as a data matrix composed of M rows of the form

$$\boldsymbol{x}_j^T, y_j^-, y_j^+ \tag{1}$$

where $j = 1,...M$, $\boldsymbol{x}_j^T \in R^N$, is the N-dimensional vector of predictors and y_j^-, y_j^+ are the lower and the upper bound on the unobserved scalar dependent variable y_j, which satisfy the relations

$$y_j^- \le y_j \le y_j^+. \tag{2}$$

In the context of the survival analysis we assume that the dependent variable is the logarithm of survival time; $y = \log(t)$. For data points with given exact values of the dependent variable $y_j^- = y_j^+ = y_j$. In this convention a right-censored response corresponds to the improper interval $[y_j^-, +\infty]$, and the left-censored response to $[-\infty, y_j^+]$, where $[-\infty, y_j^+]$, and $y_j^- = -\infty$, respectively.
Linear predictive model for the dependent variable $y \in R$ is a scalar product

$$y = [1, \mathbf{x}^T]\mathbf{v} = v_0 + v_1 x_1 +, ..., v_N x_N \tag{3}$$

where $[1,\mathbf{x}^T] \in R^{N+1}$ is the augmented data vector and $\mathbf{v} = [v_0, v_1, ..., v_N]^T \in R^{N+1}$ is the vector of adjustable parameters of the model. Determination of the optimal parameters $\mathbf{v}$ is achieved by minimization of a piecewise-linear criterion function constructed as follows. For an individual data vector $\mathbf{x}_j$ with corresponding bounds y_j^-, y_j^+ and for variable parameters $\mathbf{v} = [v_0, v_1, ..., v_N]$ of the model (3) we define two nonnegative error functions $\varphi_j^-(\mathbf{v})$ and $\varphi_j^+(\mathbf{v})$ (see Fig. 1) corresponding to the lower and upper bound, respectively. Each function is, by definition, equal to zero if the model output $\hat{y}_j = [1, \mathbf{x}_j^T]\mathbf{v}$ at the point j satisfies the corresponding constraint. Otherwise, the function $\varphi_j^+(\mathbf{v})$ is equal to the excess of model output behind the bound $\hat{y}_j - y_j^+$ if the output is too large, or the function $\varphi_j^-(\mathbf{v})$ is equal to $y_j - \hat{y}_j$ if the output is too small :

$$\varphi_j^+(\mathbf{v}) = \begin{cases} [1, \mathbf{x}^T]\mathbf{v} - y_j^+ & \text{if } [1, \mathbf{x}^T]\mathbf{v} > y_j^+ \\ 0 & \text{if } [1, \mathbf{x}^T]\mathbf{v} \le y_j^+ \end{cases}, \tag{4}$$

and

$$\varphi_j^-(\mathbf{v}) = \begin{cases} y_j^- - [1, \mathbf{x}^T]\mathbf{v} & \text{if } [1, \mathbf{x}^T]\mathbf{v} < y_j^- \\ 0 & \text{if } [1, \mathbf{x}^T]\mathbf{v} \ge y_j^- \end{cases}. \tag{5}$$

Sum of the two functions is a measure of the error of the model at the data point j, given the parameter vector $\mathbf{v}$, as shown on Fig. 1.

The total error function $\Phi(\mathbf{v})$ associated with the parameter vector $\mathbf{v}$ is the weighted sum of individual error functions $\varphi_j^-(\mathbf{v})$ and $\varphi_j^+(\mathbf{v})$ for each of data points

$$\Phi(\mathbf{v}) = \sum_{j=1}^{M} \alpha_j^- \varphi_j^-(\mathbf{v}) + \sum_{j=1}^{M} \alpha_j^+ \varphi_j^+(\mathbf{v}), \tag{6}$$

where $\alpha_j^- \ge 0$, $\alpha_j^+ \ge 0$ are weights, which determine relative influence of the individual data points on the regression hyperplane and may be used e.g. to compensate for outliers and asymmetry of the survival time distribution. In the current typical usage of the method we set $\alpha_j^- = \alpha_j^+ = 1$. The function is convex and piecewise lin-

ear (CPL) and can be effectively minimized with a basis exchange algorithm [8, 9] developed earlier for classification with formal neurons. See also [10] for geometrical analysis of the interval regression problem in terms of linear separability of datasets in the feature space of augmented data vectors $\mathbf{x}_j^T$.

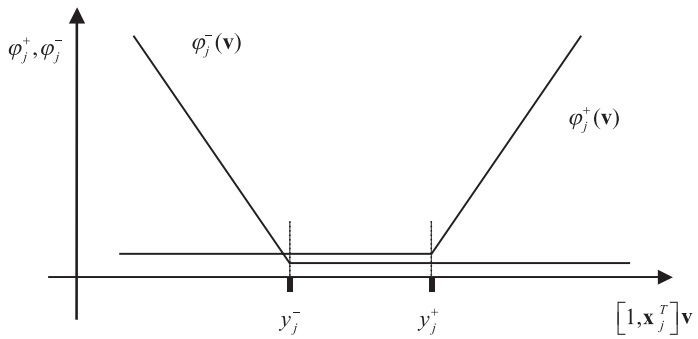

Fig. 1. Schematic drawing of the error functions $\varphi_j^-(\mathbf{v})$ and $\varphi_j^+(\mathbf{v})$ associated with interval-bounds y_j^-, y_j^+, respectively; all horizontal lines have vertical coordinate equal to 0

3. Computational Examples

For getting some idea how the CPL-IR method performs in practice we compare it with other two established approaches – the parametric survival regression and the Cox proportional hazards model. The parametric maximum likelihood method is implemented in function `survreg()` of the `R survival` package [3], and fits the maximum likelihood accelerated failure time model under selected parametric assumptions of the distribution of the survival times, including the Weibull distribution. The classical formulation of the Cox model does allow only for mixture of the exact and right-censored survival times [1]. However, the `intcox` package [11] provides the `R` function `intcox()`, which is a state-of-the-art implementation of a generalization of the Cox proportional hazards model to data with the interval-censored survival times, based on the optimization technique developed in reference [12].

Following the exposition of `intcox` in [13], we consider two datasets provided in this package. The `intcox.example` dataset has been generated using the Weibull distribution of the response times with shape parameter $\gamma = 0.75$ and scale $\exp(\beta_0 + \boldsymbol{\beta}^T\mathbf{x})^{-1/\gamma}$, where $\beta_0 = 0.1$ determines baseline hazard and $\boldsymbol{\beta} = [0.5, -0.5, 0.5, 0.5]^T$ are true coefficients of covariates' effects on the hazard function. The response times are interval-censored by the grid of 10 random times. There are two binary covariates `x.1` and `x.2`, and two continuous covariates `x.3` (sample from the uniform distribution on [–1, 1]) and `x.4` (sample from the standard Gaussian distribution).

The coefficients of linear models computed by the interval Cox method, the CPL interval regression and the parametric maximum likelihood method are given in Table 1. Note that, because the Cox model refers to the hazard function, the signs of the Cox model coefficients are opposite to the two other methods and the intercept term, which would refer to the baseline hazard, is not reported for the proportional hazard model. The minimum value of the criterion function was $\Phi_{min} = 118.59$.

The estimate of survival function for the interval censored survival times can be computed by function `survfit()` of the `survival` package, using the Turnbull iterative method for interval censored data ([1], p.129). For this data, another estimate of the survival function can be also obtained using the predicted values of the survival times computed with the CPL-IR method. Both survival curves are plotted on Fig. 2. The thicker line, corresponding to the estimate based on the predicted

Table 1. Coefficients of regression model for the survival times in `intcox.example` dataset, calculated by the interval-Cox method (intCox), the CPL-based interval regression (CPL-IR), and the parametric maximal likelihood method (SR). The parametric maximum likelihood model was calculated using function `survreg()` under the assumption of Weibull distribution of the response times. Intercept parameter is only reported for the accelerated time models (CPL-IR and SR) and corresponds to the baseline hazard in the proportional hazards (intCox) model

Method/Variable	intCox	SR	CPL-IR
(intercept)	NA	–0.227	–0.563
x.1	0.636	–0.712	–0.598
x.2	–0.435	0.740	0.554
x.3	0.330	–0.427	–0.580
x.4	0.425	–0.547	–0.344

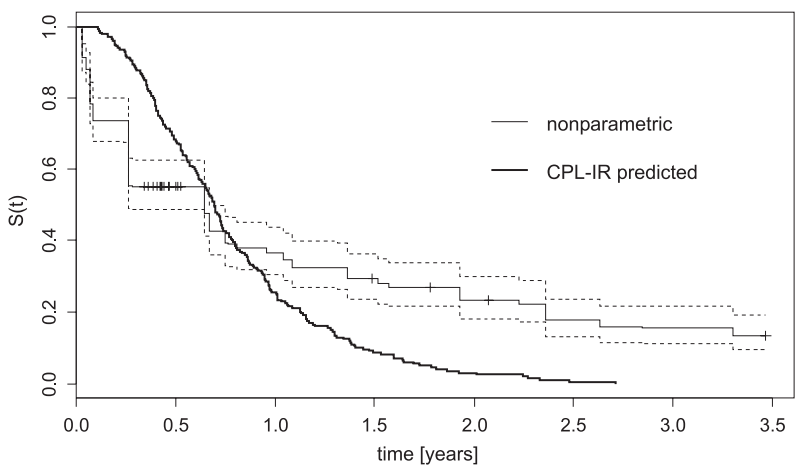

Fig. 2. Survival function $S(t)$ for simulated interval-censored data (the `intcox.example` dataset). The thin stair-like curve represents the nonparametric survival estimator with 95% confidence limits (dotted lines) determined by the `survfit()` function of the R `survival` package; the thicker and smoother line represents estimate of $S(t)$, computed with the survival times predicted by the model CPL-IR

survival times, indicates that the CPL-IR method tends to overestimate short survival times and underestimate long survival times. The crossing of both curves takes place approximately around the median survival time corresponding to $S(t) = 0.5$. This may be related to the fact that, for exact data without censoring, the CPL-IR method would produce predictions of the median of response variable.

The other dataset, `AA.data`, contains left- and right- censored (current status) data for $N = 149$ examinations of aneurisms in 83 patients with cerebral arteriovenous malformations (cAVM; a kind of abnormal, balloon-like swellings on weakened blood vessels; causing risk of rupture and hemorrage, commonly on basal arteries of the brain) after embolisation treatment (attempt to occlude of the swelling with purposefully induced blood clots). The variable `obs.t` is the inspection time for the current status examination. The degree of shrinkage of aneurisms following the embolism treatment is measured by a discrete, binary variable `mo` (equal 0 or 1; for shrinkage by less or more than 50%, respectively). This variable serves as a measure of success of the treatment. There are two relevant covariates to each observation: `lok` is the information on location of aneurism on midline (=0) or other (=1) arteries of the brain; `gr` is the integer-valued grouping variable (patient number) to which the individual observation refers (there may be multiple sites of aneurism per patient). The coefficients for this data are presented in Table 2.

Table 2. The coefficient of the CPL interval regression (CPL-IR) compared with the interval Cox method (intCox) and the parametric maximum likelihood regression under assumption of the Weibull (SR Weibull) and the exponential (SR Exp) survival time distributions for the aneursims dataset `AA.data`

Method /variable	intCox	SR Weibull	SR Exp	CPL-IR
(intercept)	NA	0.563	0.422	0.068
mo	-1.007	2.229	1.118	0.695
lok	-0.831	1.645	0.903	0.502

Here, both methods agree as to the direction of influence (positive coefficients in the survival-time models mean increase of the survival time, which is compatible with opposite-sign values in the proportional hazards models which quantify the influence of predictors on the hazard function), but differ more (by about 30%) in the magnitude. The minimum value of the CPL criterion function attained was $\Phi_{min} = 40.6$. The magnitude of model coefficients calculated by the parametric survival regression depends strongly on the assumed distribution of survival times, and e.g. the effect of `mo` variable is almost 2 times larger for the Weibull than for the exponential distribution (Table 2).

The variability of coefficients obtained with the CPL-IR method can be estimated via the standard nonparametric bootstrap procedure i.e. by resampling (with replacement) cases from the original data set and repeating calculation of the regression coefficients. The result of such assessment for the CPL-IR applied to the `intcox.`

`example` dataset is plotted on Fig. 3, which summarizes the results of regression coefficients obtained from 100 bootstrap repetitions. The means of coefficients are close to original values in modulus (0.5) and mostly correct as to the sign of influence of predictor on the survival time. Compared with the results presented on an analogous (differing only by centering of the coefficient values) figure provided on the last page of reference [13], the variability of regression coefficients we obtained for the CPL-IR is, however, larger than the variability of the model calculated by the `intcox` procedure. We also observed for the CPL-IR some difference between the original coefficients for the full data set and the means of the bootstrap trials, similar to the bias reported in [13]. In general the magnitude of coefficients determined by the Cox model is not expected to be the same as the coefficients determined by the linear regression models with $y = \log(t)$ as dependent variable, except for the maximum likelihood estimation under assumption of exponential distribution of the survival times, if the survival times *are* in fact exponentially distributed.

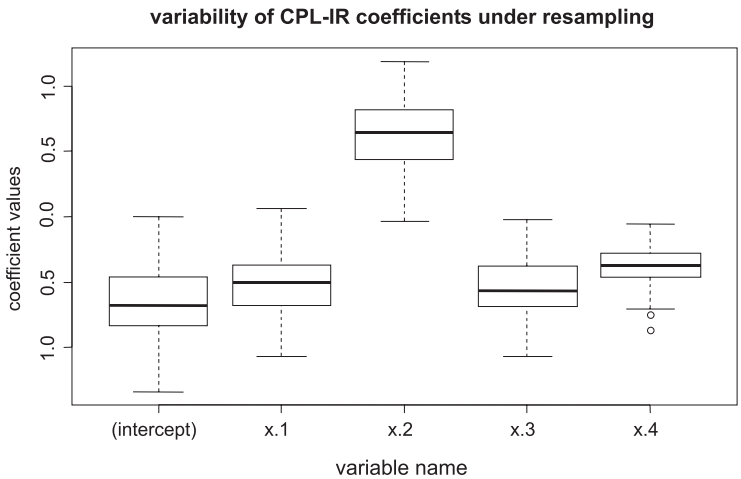

Fig. 3. The variability of coefficients of regression obtained with the CPL-IR method in a bootstrap resampling experiment with 100 repetitions on the `intcox.example` dataset. For each regression coefficient, the thick line represents the mean value for all bootstrap samples; the box extends over the range between the 1st and 3rd quartiles of the coefficient distribution; whiskers extend to the most extreme values, which are no more than 1.5 inter-quartile range distant from the box; isolated points represent the remaining values

4. Discussion

Estimation of the predictive regression models of the survival times is faced with two major difficulties: one is only partial knowledge of the survival times due to censoring and the other is the non-Gaussian and non-symmetric nature of the distributions of survival times. In the classical view, the estimation of regression

parameters would ideally take place in the maximum likelihood framework, which provide asymptotically the most effective and unbiased estimates. These theoretical advantage may, however, be hampered by practical difficulties in the construction of the likelihood function, which requires computation of probability of the data given the parameter vector. Choice of the parametric type of this distribution is, practically, not possible on theoretical grounds. The choice of the distribution influences the regression coefficients, particularly by large contribution of data points with the long survival times.

On the other hand, the prediction error based methods are influenced by asymmetry of the distribution of dependent variable, and some compensation may be needed for the distribution with heavy tail to reduce influence of the long survival times. In this context, the CPL-IR method appears to have a natural robustness property by the fact, that the contributions of the individual survival times to the overall model error are linear. This can be an advantage over least square error based procedures which were also extended to some cases of censored data [14].

The Weibull-distributed survival times in the simulated dataset considered as example are very favorable for the proportional hazard Cox model, because the model assumptions are fulfilled. They are also parametrically compatible with the parametric maximum likelihood method. On the other hand, they are not as favorable for the CPL-IR method because of the asymmetry in the distribution of survival times, which may result in some bias in the estimates of the coefficients. Some small bias with respect to the mean bootstrapped coefficients is, however, also reported for the `intcox` procedure [13]. The logarithmic transformation of the survival time, used as a dependent variable in the CPL-IR and in the parametric survival regression models tends to reduce asymmetry of the distribution of times.

The CPL criterion function (3) does not incorporate explicit information on the distribution of the survival times, and more detailed analysis would be necessary to detect if this can cause practically important bias in the estimated regression coefficients. However, the asymmetry of the distribution function may be well compensated by the logarithmic transformation of the time and the robustness of the criterion functions. The CPL function grows linearly and therefore is (relatively to e.g. least squares) not oversensitive to the cases which lay far from the regression hyperplane. Qualitatively, on the analyzed test data, the CPL model coefficients appear comparable to the coefficients of the maximum likelihood model and compatible with the Cox model.

An additional advantageous feature of the CPL-IR method is that the complexity of the criterion function minimization task, apart of the number of inequality constraints, does not significantly depend on the pattern of censoring (left-, right- or interval). In contrast, the logarithmic likelihood criterion used in the parametric regression involves more complex functional form for interval censoring than for one-sided censoring ([1], section 3.5).

The formulation of the CPL-IR regression does not require special treatment of ties (several patients have the same survival times). The tied observations in the survival times may be abundant if time is measured as a coarse discrete variable, like months or years. Theoretically, such ties introduce large number of combinations of terms in the partial likelihood function and require special treatment in the classical Cox semiparametric methods ([1], section 8.3) to reduce computing time.

5. Conclusion

The interval regression method based on the CPL criterion function minimization provides a unified treatment of exact and left-, right-, or interval-censored data in an arbitrary pattern using computational framework of the CPL function minimization. The estimated model coefficients are similar to those obtained with other methods and appear acceptable for data mining and exploratory purposes.

Acknowledgments

The work was supported by the grant N R13 0014 04 from the NCBiR and the statutory research funds of the IBIB (Nr. 4.2/12) and the TU Białystok.

References

1. Klein J.P., Moeschberger M.L.: Survival analysis: techniques for censored and truncated data. Springer, 1997.
2. Allison P.D.: Survival Analysis using the SAS System. A Practical Guide, SAS Institute, 1995.
3. Therneau T.: A Package for Survival Analysis in S, R package "survival" v. 2.36-5, http://cran.r-project.org/package=survival, 2011.
4. Lesaffre E., Komarek A., Declerk D.: An overview of methods for interval-censored data with an emphasis on applications in dentistry. Statistical Methods in Medical Research 2005, 14, 539–552.
5. Bobrowski L.: Linear prognostic models based on interval regression with CPL functions (in Polish). Symulacja w Badaniach i Rozwoju 2010, 1, 109–117.
6. Bobrowski L.: Prognostic Models Based on Linear Separability. In: P. Perner (Ed.) Advances in Data Mining, Springer Verlag, Berlin 2011, 11–24.
7. R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, http://www.r-project.org, 2010.
8. Bobrowski L., Niemiro W.: A method of synthesis of linear discriminant function in the case of nonseparability. Pattern Recognition 1984, 17, 177–273.
9. Bobrowski L.: Design of piecewise linear classifiers from formal neurons by some basis exchange technique. Pattern Recognition, 1991, 24, 863–870.
10. Bobrowski L.: Interval Uncertainty in CPL Models for Computer Aided Prognosis, In: Z. Hippe, J.L. Kulikowski, T. Mroczek (Eds) Human-Computer Systems Interaction. Backgrounds and Applications 2, Springer, Berlin 2012, 443–461.
11. Henschel V., Heiss C., Mansmann U.: intcox: Iterated Convex Minorant Algorithm for interval censored event data, R package v. 0.9.2, http://cran.r-project.org/package=intcox, 2009.

12. Pan W.: Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval-Censored Data. Journal of Computational and Graphical Statistics 1999, 78, 109–120.
13. Henschel V., Heiss C., Mansmann U.: Intcox: Compendium to apply the iterative convex minorant algorithm to interval censored event data, http://cran.r-project.org/web/packages/intcox/vignettes/intcox.pdf, 2009.
14. Buckley, J., James, I.: Linear regression with censored data. Biometrika 1979, 66, 429–436.