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

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