Computational Statistics and Data Analysis 56 (2012) 1854–1868

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Inference in generalized linear regression models with a censored covariate

John V. Tsimikas [*], Leonidas E. Bantis, Stelios D. Georgiou

Department of Statistics and Actuarial–Financial Mathematics, University of the Aegean, Samos 83200, Greece

ARTICLE INFO

Article history:
Received 26 January 2011
Received in revised form 2 November 2011
Accepted 10 November 2011
Available online 22 November 2011

Keywords:
Accelerated failure time model
Censored covariate
Estimating functions
Generalized gamma distribution
Generalized linear model
Mean imputation
Quasi-likelihood

ABSTRACT

The problem of estimating the parameters in a generalized linear model when a covariate is subject to censoring is studied. A new method based on an estimating function approach is proposed. The method does not assume a parametric form for the distribution of the response given the regressors and is computationally simple. In the linear regression case, the proposed approach implies the use of mean imputation of the censored regressor. The use of flexible parametric models for the distribution of the covariate is employed. When survival time is considered as the covariate subject to censoring, the use of the generalized gamma distribution is explored, since it is considered as a platform distribution covering a wide variety of hazard rate shapes. The method can be further robustified by considering models of nonparametric nature typically used in survival analysis such as the logspline for the censored covariate. For models involving additional, fully observed, covariates the use of a generalized gamma accelerated failure time regression model is explored. In this setting, no parametric family assumption for the extra covariates is needed. The proposed approach is broader than likelihood based multiple imputation techniques. Moreover, even in cases with a known parametric form for the response distribution, the method can be considered a feasible alternative to likelihood based estimation. Simulation studies are conducted for continuous, binary and count data to evaluate the performance of the proposed method and to compare the estimates to standard ones. An application using a well known data set of a randomized placebo controlled trial of the drug D-penicillamine (DPCA) for the treatment of primary biliary cirrhosis (PBC) conducted at the Mayo Clinic is presented. Possible extensions of the method regarding the robustness as well as the type of censoring are also discussed.

1. Introduction

Parameter estimation and statistical inference in models where the response variable is subject to censoring have been thoroughly studied in the past with the Cox proportional hazards (PH) model and the accelerated failure time (AFT) model being the most celebrated models. Less attention has been paid to the situation where the covariate is subject to censoring. This situation arises, as for example, when considering the accuracy of a diagnostic marker as a function of the time lag between the measurement and the occurrence of disease (Cai et al., 2006). Gõmez et al. (2003) consider a regression model with an interval-censored covariate and develop an algorithm for the nonparametric maximum likelihood estimation of the regression coefficients. In their setting, no distributional form is assumed for the covariate. However, this is not the case for the response distribution. Pawitan and Self (1993) consider a repeated marker measurement setting and use Weibull regression models for infection and disease occurrence times that are subject to censoring. They also present arguments in

[*] Corresponding author. Tel.: +30 22730 82335.
E-mail address: tsimikas@aegean.gr (J.V. Tsimikas).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.11.010

favor of constructing models that consider modeling the disease marker process given the time variable (in their case, time to AIDS).

Another area where censored regressors are encountered frequently is econometrics where the covariate may include upper or lower cutpoints as discussed in Rigobon and Stoker (2007), or may be categorized into groups according to the value of the covariate as in Hsiao (1983). For instance, observed household income would have a *top coded* response. In survival studies, which is the focus of our paper, the time to event variable may play the role of a covariate, and it is obvious that random or *bound* censoring may be present in such a setting. Note that in survival studies bound censoring (or top coding) occurs due to the end of study (cutpoint). This type of censoring is also known as a ceiling effect in many scientific disciplines. Left censoring is somewhat less common in survival studies, but occurs frequently in econometrics and other scientific fields where it is typically referred to as bottom coding or the floor effect.

A related problem that has received attention in the past deals with surrogate predictors. Some strategies based on approximate quasi-likelihood techniques, including regression calibration, are discussed in Carroll and Stefanski (1990). Regression calibration in failure time regression models when some covariate values may be missing or mismeasured is addressed in Wang et al. (1997). More recently Wang and Pepe (2000) discussed the use of expected estimating equations in problems with measurement error in the covariate.

When dealing with a censored covariate the simplest approach is to discard the censored data and perform the analysis only with the observed data. This is called a Complete Case (CC) analysis. The CC method provides consistent parameter estimates under noninformative censoring. Obviously the CC method suffers from low efficiency which can be dramatic when heavy censoring is involved. In the case where both the distribution of the response given the covariates and the covariate distribution are assumed to lie within known parametric families, one can estimate the parameters via maximum likelihood. This is the approach taken by Austin and Hoch (2004) in the simple fully parametric setting where a ceiling effect is present on the covariates and where the joint distribution of the response and covariates is a multivariate normal distribution. However, when dealing with distributions other than the normal, computational issues arise. The method proposed in this paper is both computationally simple and can be used without assumptions about the response distribution given the covariates. In many cases a parametric model, regarding the censored covariates, may be justified. Consider, for example, a situation where a new time dependent biomarker is to be evaluated. It is not uncommon to have historical data that allow us to use a parametric model for the distribution of the time to event covariate. Moreover, other fully observed covariates of interest may be utilized via an accelerated failure time (AFT) model. A flexible parametric model is the AFT generalized gamma regression model, which is considered as a platform for parametric analysis for survival data (see Cox et al., 2007).

The paper is organized as follows. In Section 2, we discuss the simple linear regression problem with the covariate being censored. In Section 3, we present our method, based on estimating equation theory that deals with the generalized linear model. We give details regarding the three most common settings, a continuous, a binary and a count response. We show how our method can be extended to accommodate other observed covariates, via the use of an AFT parametric model and discuss the use of the generalized gamma distribution for the censored covariate. In Section 4, we present the results from simulation studies. Finally, in Section 5 we apply our method on real data from a randomized placebo controlled trial of the drug D-penicillamine (DPCA) for the treatment of primary biliary cirrhosis (PBC) conducted at the Mayo Clinic (Fleming and Harrington, 1990). We conclude the article by discussing possible generalizations of our methods.

2. Simple linear regression with a censored covariate

Consider the case of the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

with the covariate subject to random right censoring, Type I censoring, or both. The data for the i-th subject consist of (Y_i, T_i, Δ_i), where Y_i is the response $T_i = \min(X_i, C_i)$, C_i being the censoring variable and Δ_i is the indicator variable that informs us whether the i-th subject's covariate value is censored ($\Delta_i = 0$) or observed ($\Delta_i = 1$).

We assume that $E(\epsilon_i|X_i) = 0$. If we want to apply the Complete Case (CC) analysis, then we use only data with $\Delta_i = 1$. For the regression model to be well specified and for the CC to yield unbiased estimators, we assume that the mean of ϵ_i does not vary with Δ_i, that is $E(\epsilon_i|\Delta_i) = 0$. Given the Δ_i's, a parametric approach is to assume that both $f_{Y|X}(y_i|x_i)$ (the conditional distribution of Y_i given $X_i = x_i$) and $f_X(x_i)$ (the marginal distribution of X_i) lie within known parametric families. Given that the censoring and event times are independent we have

$$f_{Y|T,\Delta=1}(y_i|t_i, \delta_i = 1) = \frac{f_Y(y_i)}{f_X(t_i)S_C(t_i)} \int_{t_i}^{\infty} f_{X,C|Y}(t_i, c|y_i)dc,$$

where $S_C(t) = P(C > t)$, and y_i, t_i, δ_i are the realizations of the random variables Y_i, T_i, Δ_i respectively. Furthermore, if we assume conditional independence of censoring and event times given the response value $Y = y$ we obtain

$$f_{Y|T,\Delta=1}(y_i|t_i, \delta_i = 1) = \frac{f_{Y|X}(y_i|t_i)}{S_C(t_i)} \int_{t_i}^{\infty} f_{C|Y}(c|y_i)dc.$$