REGRESSION WHEN THE PREDICTOR MAY BE CENSORED

David Oakes[*]

Abstract

There is a voluminous literature on regression models where the response variable is subject to right censoring. Situations where a predictor variable is subject to right censoring have received much less attention. However the increasing use of biomarkers as predictors in clinical studies has focussed attention on this area. Unlike in situations with censored response variables, simply omitting observations with the censored predictors typically does not lead to bias in the estimation of the regression coefficient. However there may be a substantial loss of efficiency. We explore some approaches to recovery of the available information from the censored observations using both parametric and nonparametric methods and report the results of an illustrative simulation.

Key Words: biomarkers; missing data; prediction; survival data

1. Introduction

There is a voluminous literature on regression models where the response variable is subject to right censoring. Situations where a predictor variable is subject to right censoring have received much less attention. One special case has been studied extensively, the limit of detection problem which arises when there is a prespecified upper or lower limit on values that can be recorded. However the increasing use of biomarkers as predictors in clinical studies - see, for example Brumback, Pepe and Alonzo (2006) and Cai et al. (2006) motivates us to consider more general patterns of censorship.

We study estimation of the regression coefficient β in the simple univariate linear regression model

$$Y_i = \alpha + \beta T_i + \epsilon_i,$$

where the ϵ_i are independent and identically distributed with mean zero and finite variance σ^2. The T_i are subject to right censoring, so that we observe $x_i = \min(T_i, c_i)$ and $\delta_i = 1(T_i \leq c_i)$. Note that Y_i is always observed, even when T_i is censored.

We shall assume that the T_i are independent and identically distributed with continous distribution $F(t) = \mathrm{pr}(T \leq t)$. We will usually assume that the c_i are also independent and identically distributed from a distribution with surivor function $G(c) = \mathrm{pr}(C > c)$. Typically both F and G will be unknown. We will investigate semiparametric and nonparametric approaches to analysis of this type of data.

2. An illustrative example

An example leading to simple calculations is to take $F(t) = 1 - \exp(-\lambda t)$ and $G(c) = \exp(-\mu c)$. Under these assumptions X and δ are independent, $\mathrm{pr}(\delta = 1) = \lambda/(\lambda + \mu)$ and X is exponential with parameter $\lambda + \mu$. To be specific, we will take $\lambda = \beta = \sigma = 1$ and $\mu = 0.5$, so that 1/3 of the observations on T will be censored. If there had been no censoring the asymptotic variance of $\hat{\beta}$ would have been

$$\mathrm{A.V.}(\hat{\beta}) \quad = \quad \lim_{n \to \infty} n \,\mathrm{var}(\hat{\beta})$$

[*] University of Rochester Medical Center, Rochester NY

$$
\begin{aligned}
&= \lim_{n\to\infty} \frac{\sigma^2}{\sum(t_i - \bar{t})^2/n} \\
&= \frac{\sigma^2}{\mathrm{var}(T)} \\
&= \frac{\sigma^2}{\lambda^2} = 1 \, .
\end{aligned}
$$

If simply omit the censored observations, by what factor is this increased? The answer is not 1.5 as might be expected from the reduced sample size (= 1/(2/3)). The censoring selectively eliminates observations that are more informative about β, those with greater T. As we noted δ and X are independent in this model and the X is exponentially distributed with parameter $(\lambda + \mu)$. So

$$
\begin{aligned}
\mathrm{A.V.}(\hat{\beta}_u) &= \lim_{n\to\infty} \frac{\lambda+\mu}{\lambda} \frac{\sigma^2}{\sum(x_i - \bar{x}_i)^2/n} \\
&= \frac{\lambda+\mu}{\lambda} \frac{\sigma^2}{\mathrm{var}(X)} \\
&= \frac{(\lambda+\mu)^3}{\lambda} \sigma^2 = 27/8 \, .
\end{aligned}
$$

The numerical value here depends on the parameter values used, but the phenomenon of variance inflation beyond that due to the reduced sample size is quite general.

3. First and Second Moment Properties

Write $e(t)$ and $v(t)$ for the mean and variance of the conditional distribution of the *residual lifetime* $T - t$ given $T > t$. These are finite for all t if $E(T) < \infty$ and $\mathrm{var}(T) < \infty$. Then,

$$
\begin{aligned}
E(Y|x,\delta) &= \delta(\alpha + \beta x) + (1-\delta)\{\alpha + \beta E(T|x, \delta = 0)\} \\
&= \alpha + \beta x + \beta(1-\delta)e(x),
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}(Y|x,\delta) &= E\{\mathrm{var}(Y|T)|x,\delta\} + \mathrm{var}\{E(Y|T)|x,\delta\} \\
&= \sigma^2 + \mathrm{var}(\alpha + \beta T|x,\delta) \\
&= \sigma^2 + (1-\delta)\beta^2 \mathrm{var}(T|x, \delta = 0) \\
&= \sigma^2 + (1-\delta)\beta^2 v(x).
\end{aligned}
$$

If $e(t)$ and $v(t)$ are known or can be estimated then β can be estimated by weighted least squares of Y on $E(T|X,\delta)$ or an estimate $\hat{E}(T|X,\delta)$. This approach does not require that the deviations ϵ_i around the true regression line be normally distributed.

4. Exponential Distribution for T

We now return to the situation where T has an exponential distribution with known parameter λ. By the lack of memory property, $e(x) = 1/\lambda$ and $v(x) = 1/\lambda^2$, so

$$
\mathrm{var}(Y|x, \delta = 0) = \sigma^2 + \beta^2/\lambda^2,
$$

free of x. For the $n - d$ censored observations the predictor x_i is replaced by $x_i + 1/\lambda$. Adding the same value to every x does not affect $\hat{\beta}$ or its precision, so these observations provide an independent estimate $\hat{\beta}_c$ with

$$\mathrm{var}(\hat{\beta}_c) = \frac{\sigma^2 + \beta^2/\lambda^2}{\sum_c (x_i - \bar{x}_c)^2}.$$

By the independence of δ and X, $n^{-1}\sum_c (x_i - \bar{x}_c)^2 \to \mu/(\lambda+\mu)^3$, and

$$\mathrm{A.V.}(\hat{\beta}_c) = \Big(\frac{\lambda+\mu}{\mu}\Big)\Big(\sigma^2 + \frac{\beta^2}{\lambda^2}\Big)(\lambda+\mu)^2.$$

With our numerical values

$$\mathrm{A.V.}(\hat{\beta}_c) = 27/2\ .$$

A third independent estimator $\hat{\beta}_d$ of β may be obtained by considering the difference $\bar{y}_c - \bar{y}_u$ between the mean of y for the censored and uncensored observations, since

$$E(Y|X, \delta = 0) - E(Y|X, \delta = 1) = \frac{\beta}{\lambda}.$$

Setting $\hat{\beta}_d = \lambda(\bar{y}_c - \bar{y}_u)$ gives

$$\mathrm{A.V.}(\hat{\beta}_d) = \lambda^2\sigma^2\Big(\frac{\lambda+\mu}{\lambda}\Big) + \lambda^2\Big(\sigma^2 + \frac{\beta^2}{\lambda^2}\Big)\Big(\frac{\lambda+\mu}{\mu}\Big).$$

With our numerical values,

$$\mathrm{A.V.}(\hat{\beta}_d) = 15/2\ .$$

The optimal linear combination $\hat{\beta}_{lc}$ of $\beta_u, \beta_c, \beta_d$ weights each inversely as its variance and is (essentially) just the weighted least squares estimator of y on $x+(1-\delta)e(x)$. Writing $I_. = \mathrm{var}(\hat{\beta}_.)^{-1}$ for the information from each estimator our numerical values give

$$I_{lc} = I_u + I_c + I_d = 8/27 + 2/27 + 2/15 = 68/135 \approx 0.5,$$

and

$$I_{lc}/I_u = 1.7\ .$$

so that appreciable recovery of information from the censored observations is possible.

Further recovery of information is possible from second moments. For the residual variation of the y_i around $E(y_i|x_i, \delta_i)$ is σ^2 when $\delta = 1$ and $\sigma_c^2 = \sigma^2 + \beta^2/\lambda^2$ when $\delta = 0$. Writing s_u^2 and s_c^2 for the corresponding estimates gives a fourth estimate of β namely

$$\hat{\beta}_v = \lambda(s_c^2 - s_u^2)^{\frac{1}{2}}.$$

The variability of this estimate will depend on the distribution of the errors ϵ_i. With normally distributed errors $(d-2)s_u^2/\sigma^2$ will follow a chi-square distribution with $d-2$ degrees of freedom, so that

$$\mathrm{A.V.}(s_u^2) = 2\sigma^4\frac{\lambda+\mu}{\lambda}.$$

The distribution of s_c^2 is more complicated as the errors here are the sums of a normal $epsilon_i$ and a centered exponential variable with parameter λ/β. However a straightforward calculation involving the fourth moments of these two distributions gives

$$\mathrm{A.V.}(s_u^2) = \Big(2\sigma^4 + 4\sigma^2\frac{\beta^2}{\lambda^2} + 8\frac{\beta^4}{\lambda^4}\Big)\Big(\frac{\lambda+\mu}{\mu}\Big),$$

so that the asymptotic variance of $\hat{\beta}_v$ can be calculated using the delta method.

5. Fully Parametric Approach

Tsimikas et al., (2012) in one of the few papers on this topic, consider the estimation of β under a fully parametric model when the ϵ_i are normally distributed and the distribution of T takes various parametric forms. In particular they point out that when T follows an exponential distribution, the conditional distribution of T given $Y = y$ and $T > x$ is a truncated normal distribution. The full log-likelihood in $(\alpha, \beta, \lambda, \sigma)$ can then be maximized directly or by the EM algorithm, (Dempster, Laird and Rubin, 1977). The latter requires the only computation of the conditional expectations $E(T|T > x, Y = y)$ and $E(T|T^2 > x)$, which are expressible in terms of the normal c.d.f.

6. Fully Nonparametric Approach

We now consider weighted least squares estimation when the distribution of T is unknown. This approach requires only the usual second moment assumptions on the ϵ_i, normality is not required. We propose to estimate $F(\cdot)$ by the usual Kaplan-Meier estimator $\hat{F}(\cdot)$ and use this to estimate $x + e(x)$ and $v(x)$ by

$$x + \hat{e}(x) = \sum_{x_i > x, \delta_i = 1} \frac{x_i d\hat{F}(x_i)}{1 - \hat{F}(x)},$$

$$\hat{v}(x) = \sum_{x_i > x, \delta_i = 1} \frac{x_i^2 d\hat{F}(x_i)}{1 - \hat{F}(x)} - \hat{e}(x)^2.$$

It appears to be helpful to force $\hat{F}(\cdot)$ to be a proper distribution, by setting $\hat{F}(x) = 1$ for $x \geq \max(x_1, \ldots, x_n)$.

Estimates of α and β may now be obtained by iteratively weighted least squares. The asymptotic distributions of these estimates remains to be investigated. The general context is that of a regression problem where measurement of the predictor variable is subject to small but highly correlated errors. The limiting limiting behavior of the process $\hat{e}(x)$ was presented in Yang (1978).

Bantis (2013) presents an alternative approach that uses splines for nonparametric estimation of the distribution of T.

7. An Illustrative Simulation

We performed a simple simulation under our selected model using 100 repetitions with a sample sample size of 1000. Values of the mean and standard deviation of the estimated slope were as follows:

(i)Regression on all the t_i.......................................	1.022(0.0343)
(ii)Regression on the uncensored t_i.........................	1.004(0.0600)
(iii)Weighted Least Squares : True Weights.............	1.007(0.0423)
(iv)Weighted Least Squares : Estimated Weights.....	1.020(0.0516)

In (iii) the true distributions and parameter values were used to calculate the conditional expectations $e(x)$ and the variances $v(x)$. In (iv) $e(x)$ and $v(x)$ were estimated nonparametrically and a preliminary estimate of β was calculated from the observations with t uncensored.

The numerical values of the standard deviations for (i), (ii) and (iii) are consistent with the calculated asymptotic variance calculations.

The results suggest that all methods yield consistent estimates of β, that appreciable recovery of information from observations with censored predictors is possible when the distribution $F(\cdot)$ of these is known and that some information recovery is possible even when $F(\cdot)$ must be estimated nonparametrically.

8. Discussion

Gomez et al. (2003) present a joint estimation procedure in the context of interval censored data which, in our notation, combines nonparametric estimation of the distribution $F(t)$ of T, with a parametric model for the regression of Y around T. The investigators propose an EM type of approach to maximize the *joint* (semi-parametric) likelihood in F and the regression parameters. They did not specifically address the special case of right-censored data.

Tsimikas et al. (2012) and Bantis (2013) discuss the extension to logistic regression and other generalized linear models (GLM's). Iteratively weighted least squares estimators are derived which reduce to the usual ones in the absence of censoring. The computations involve calculation of the conditional expectations such as

$$E\left\{g^{-1}(\alpha+\beta T)\middle|T>x\right\},$$

and

$$E\left\{\frac{dg^{-1}(\alpha+\beta T)}{d\beta}\middle|T>x\right\},$$

where $g(\cdot)$ is the link function for the GLM. Parametric models will rarely yield explicit forms for these conditional expectations, but there appears to be no special difficulty in applying the nonparametric approach.

Situations with multiple predictors are more difficult to address, for example with two predictors T_1, T_2, both subject to censoring, the weighted least squares approach requires calculation of the conditional expectations $E(T_1^p T_2^q|T_1 > x_1, T_2 = t_2)$ and $E(T_1^p T_2^q|T_1 > x_1, T_2 > x_2)$ for $(p,q) = (1,0),(2,0),(0,1),(0,2),(1,1)$. A fully non-parametric approach would require use of a bivariate survivor function estimator, for example that of Dabrowska (1978).

REFERENCES

Bantis, L.E. (2013), *Statistical methods for the Evaluation of Diagnostic Biomarkers in the Presence of Censoring*, Ph.D. Thesis, University of the Aegean, School of Sciences, Department of Statistics and Actuarial-Financial Mathematics, Salmos Island, Greece.

Brumback, L.C., Pepe, M.S. and Alonzo, T.A. (2006), "Using the ROC curve for gauging treatment effect in clinical trials," *Statistics in Medicine*, 25, 575–590.

Cai, T., Pepe, M.S., Lumley, T., Zheng, Y., Jenny, N.S. (2006), "The sensitivity and specificity of markers for event times," *Biostatistics*, 7, 187–197.

Dabrowska, M.D. (1988), "Kaplan-Meier estimate on the plane," *Annals of Statistics*, 16, 1475–1489.

Dempster, A.P., Laird, N.M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Tsimika, J.V., Bantis, L.E., and Georgiou, S.D. (2011), "Inference in generalized linear regression models with a censored covariate," *Computational Statistics and Data Analysis*, 56, 1854–1868.

Yang, G.L. (1978), "Estimation of a Biometric Function," *Annals of Statistics*, 6, 112–116.