CENSORED QUANTILE SURVIVAL WITH CURE

ROGER KOENKER AND NAVEEN NARISETTY

Abstract. We describe a quantile regression model for survival data that allows a positive proportion of subjects to become unsusceptible to recurrence of disease following treatment. Our approach follows recent work of Wu and Yin (2013, 2017). We compare two estimation strategies proposed by Wu and Yin with a third "data augmentation" approach based Yang et al. (2016). The methods are illustrated with data from a Lung Cancer survival study. Software and documentation are provided in R.

1. Introduction

Motivated to some degree by recent progress in cancer treatment there has been increasing interest in survival analysis models that accommodate a probability of "cure," that is a positive treatment effect that lengthens survival prospects to the extent that probability of recurrence or death from the original disease is reduced essentially to zero. Estimating such models from conventional survival data alone is obviously challenging since we must distinguish cured subjects from those merely censored by various aspects of the study design and still susceptible to the disease. Quantile regression models offer an attractive framework for such modeling since they provide a flexible, local specification of covariate effects in general, and treatment effects in particular.

In recent work Wu and Yin (2013, 2017) have proposed both estimating equation and multiple imputation methods for quantile regression models with cure. Building on their work this note describes some further developments of these methods along with testing and documentation of new software implementations in R.

2. The Model

The quantile regression survival model as introduced in Koenker and Geling (2001), Portnoy (2003) and Peng and Huang (2008) assumes

Version: May 6, 2017. This research was partially supported by Bristol-Myers Squibb. All computations were carried in the R language; code and data to replicate all the computations is available from the authors.

that the τth conditional quantile functions of the possibly transformed survival time T are given by,

$$Q_{T_i}(\tau|X = x_i) = x_i^\top \beta(\tau).$$

Typically, as in the classical accelerated failure time (AFT) model, survival times are log transformed and subject to censoring from the right. The possibility of cure is introduced via a latent variable, η, modeled as a binary response, so

$$\pi_i = \mathcal{P}(\eta_i = 1|z_i) = \pi(z_i^\top \gamma),$$

depends on covariates as mediated by the link function, π. In Wu and Yin, π is logistic, but we may consider other potential choices. When $\eta_i = 1$ we will say that subject i is susceptible to the event of interest, while if $\eta_i = 0$ they are unsusceptible, thus,

$$\tilde{Y} = \eta T + (1 - \eta)\infty,$$

subject to the usual constraints of censoring. We observe, $Y = \tilde{Y} \wedge C$, where C denotes the censoring time, and $\Delta = I(Y \leqslant C)$. We must assume, further, that Y and C are conditionally independent given the covariates X and Z.

Under these conditions we can define the counting process $N_i(t) = \Delta_i I(Y_i < t)$, and the cumulative hazard function,

$$\Lambda_Y(t|x_i, z_i) = -\log(1 - \pi(z_i^\top \gamma) F_T(t|x_i))$$

yielding the resulting martingale,

$$M_i(t) = N_i(t) - \Lambda_Y(t|x_i, z_i),$$

which can be exploited as in Peng and Huang (2008) to construct an estimating equation for the quantile regression process, $\beta(\tau)$, conditional on γ.

Wu and Yin (2013) initially proposed an estimation strategy that alternated between estimation of γ and β, but acknowledged that the procedure was unstable, and sometimes failed to converge. This was also our experience, so we have focused here on three alternative estimation strategies: one that replaces the parametric Peng-Huang estimation method for $\beta(\tau)$, by a nonparametric "local Nelson-Aalen" estimator, a second that relies on multiple imputation as proposed in Wu and Yin (2017), and a third that extends the data augmentation approach introduced in Yang et al. (2016). We will describe each approach briefly, before turning to a comparison of their performance.

3. The Estimators

An advantage of the parametric QR model is that it allows the researcher to be quite flexible about how many, and the way that covariates enter into the model, while maintaining the linear parametric structure familiar from regression modeling. From an asymptotic viewpoint this is reflected in parametric rates of convergence for the estimator of $\beta(\tau)$. The downside of this in the presence of censoring is that it requires a global (linear) specification of the covariates effects in order to justify the weighting schemes used to account for the censoring.

3.1. **Local Nelson-Aalen.** Building on prior work of Beran (1981), Dabrowska (1987) and others, Wang and Wang (2009) proposed estimating censored QR models using a local, kernel weighted, version of the Kaplan-Meier estimator. Wu and Yin (2013) adapt this approach to construct a local Nelson-Aalen estimator; for cure applications this has the advantage that an estimating equation for γ can be constructed that avoids any global parametric specification of the quantile specific effects. The difficulty with this approach, of course, is that specification of kernel and its associated bandwidths becomes increasingly problematic as the dimension of the covariate space grows.

3.2. **Multiple Imputation.** Wu and Yin (2017) extend their prior approach by noting that conditional probabilities of subjects being uncured can be computed from the local Nelson-Aalen method and used to impute η's for the full sample. Of course, for subjects with $\Delta_i = 1$ these probabilities are necessarily one. The imputed η's could be used to generate an updated estimate of γ, leading to an updated estimate of Λ, and this process could be continued until some form of convergence is achieved. However, the Wu and Yin (2017) approach updates only the $\beta(\tau)$ parameters. Such imputation schemes can be expected to improve upon the earlier estimating equation equation method, but it still suffers from the inherent drawbacks of the local Nelson-Aalen approach.

3.3. **Data Augmentation.** The data augmentation estimator estends the approach introduced in Yang et al. (2016) and shares some features of the imputation method. In contrast to the two prior estimation methods, however, data augmentation relies on the linear parametric specification of the QR process allowing us to more easily accommodate several covariates in X. An initial estimator of the QR process, $\beta(\tau)$ on the grid $\tau_1, \ldots, \tau_M$ is obtained by simply computing the median regression estimator $\hat{\beta}(1/2)$, based on only the uncensored observations and

imposing the common slope assumption, so $\hat{\beta}(\tau) = \hat{\beta}(1/2) + \hat{\beta}_1(\tau)e_1$ where $\hat{\beta}_1(\tau)$ denotes the ordinary sample quantiles of the residuals from the median fit and e_1 is the first unit basis vector of $\mathbb{R}^p$. An initial estimator of γ is obtained by (naively) estimating the binary response model of δ on Z, i.e. assuming provisionally that all the censored subjects are cured. Given these initial estimators, we may begin the iteration:

- Generate η_i's,
- Reestimate γ,
- Generate the censored y_i's,
- Reestimate $\beta(\tau)$.

Accumulating the $\hat{\gamma}$'s and $\hat{\beta}(\tau)$'s from this iteration, point estimates can be obtained by simply averaging over the corresponding iterates. For both reestimation steps there is the option to resample with replacement from the relevant full sample as in the standard x, y bootstrap.

4. Software Implementation

In this section we will briefly describe the R implementation of the foregoing methods. The main function that provides a unified interface to all three estimation methods is `cqr()`, pronounced "cure." Not to be confused with `crq()`, which is the umbrella function for censored quantile regression applications in the R package **quantreg**, we expect eventually to try to fold the functionality of `cqr` into **quantreg** and perhaps even into `crq`, but for the moment it seems prudent to keep them separate.

The `cqr` function uses the extended formula interface of the package **Formula**, so one writes the model as `y | d ~ X | Z` where `y` denotes the observed response, `d` the censoring indicator, `X` the covariates of the QR model, and `Z` the covariates of the binary response model. The remaining arguments are standard, with the `method` argument taking one of three possible values, `LNA`, `Imp` or `DA` corresponding to the methods discussed in the previous section. Users have the option of specifying a vector of τ's of interest when evaluating $\hat{\beta}(\tau)$ as well as the grid of τ's used for the intermediate computations. The latter, by default, is set to the percentiles.

The default link function for the binary response cure component of the model is logistic, but other link functions compatible with the R `glm` function are easily available. These include probit and cauchit, but one could also use one of the parametric links available from the package **glmx**, Zeileis et al. (2015).

gtab	**Bias**			**MSE**		
	LNA	Imp	DA	LNA	Imp	DA
Intecept	−0.007	−0.007	−0.001	0.102	0.102	0.115
Slope	0.060	0.060	0.065	0.296	0.296	0.321

btab	**Bias**			**MSE**		
	LNA	Imp	DA	LNA	Imp	DA
Intecept	0.063	0.009	−0.047	0.103	0.099	0.103
Slope	0.107	0.071	0.083	0.454	0.441	0.434

5. Simulations

We consider two distinct simulation settings, one that reproduces that in Wu and Yin (2017) and the other that is based on the Lung Cancer data also analyzed in Wu and Yin (2017). In the former case we have a single, uniformly distributed covariate, x, that determines the cure proportion,

$$\log(\pi_i/(1-\pi_i)) = \gamma_0 + \gamma_1 x_i,$$

and the event time model,

$$Y_i = \log T_i = \beta_0 + \beta_1 x_i + (1 + x_i)u_i$$

where $u \sim \mathcal{N}(0,1)$. Censoring is determined by x and a random uniform, $R \sim U[0, L)$ as,

$$C_i = I(x_i < 1/2)R_i + I(x_i \geqslant 1/2)(R_i + 1).$$

As in Wu and Yin (2017), we set $\gamma = (1, -1)$, $\beta = (2, 1)$ and $L = 40$. Note, however, that this L, which represents the duration of the study in clinical trial applications corresponds to a rather unrealistic, essentially infinite value. We report results for our three estimators for both bias and mean squared error (MSE) in Tables **??** and **??**; in the latter we aggregate over the nine deciles. The experiment is based on 1000 replications. Performance of the three methods both in terms of bias and MSE are quite similar.

We now briefly reconsider the lung cancer data considered in Wu and Yin (2017).

Employing the same model as Wu and Yin, we report results from all three fitting methods. The data consists of 280 observations with 64% censoring. There are three covariates: tumor histology, patient age and patient gender. All three are used in both the logistic cure model and the QR survival model. Although we have used the same bandwidth parameters for the local Nelson-Aalen estimation for the "LNA" and

	LNA	Imp	DA
Intercept	1.069 (0.646)	1.069 (0.646)	0.273 (0.285)
Histology	−0.506 (0.563)	−0.506 (0.563)	−0.471 (0.326)
Age	0.731 (0.312)	0.731 (0.312)	0.591 (0.187)
Sex	−0.686 (0.568)	−0.686 (0.568)	−0.258 (0.376)

TABLE 1. Estimates of the γ parameters for the logistic cure model, bootstrap standard errors in parentheses

"Imp" estimators, our estimates differ slightly from those reported in Wu and Yin (2017). Table 1 reports γ estimates for the three methods, while Figure 1 depicts $\beta(\tau)$ estimates. Standard errors and pointwise confidence bands are based on 200 replications.

Again we see that the three methods produce similar conclusions. In our judgement the data augmentation approach is preferable for several reasons. It is less sensitive to the upper tail of quantile regression model, it is more easily adaptable to several covariates, and avoids inherently delicate bandwidth selection issues.

6. Conclusion

Quantile regression methods offer an attractive approach to estimating survival models with a positive cure proportion. Covariate effects are flexibly modeled in the upper tail where the cure effect is most salient. Here, we have adopted the modeling strategy of Wu and Yin (2013) and Wu and Yin (2017), however their estimation methods, which are based on the local Nelson-Aalen approach of Wang and Wang (2009) are compared with an alternative data augmentation approach proposed recently by Yang et al. (2016). The latter approach has a number of advantages, and it is the approach we would recommend for most applications.

References

Beran R. 1981. Nonparametric regression with randomly censored survival data. Technical report: University of California, Berkeley.

Dabrowska DM. 1987. Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics* **14**: 181–197.

Koenker R, Geling O. 2001. Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association* **96**: 458–468.

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

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

Wang HJ, Wang L. 2009. Locally weighted censored quantile regression. *Journal of the American Statistical Association* **104**: 1117–1128.

Wu Y, Yin G. 2013. Cure rate quantile regression for censored data with a survival fraction. *Journal of the American Statistical Association* **108**: 1517–1531.

Wu Y, Yin G. 2017. Multiple imputation for cure rate quantile regression with censored data. *Biometrics* **73**: 94–103.

Yang X, Narisetty NN, He X. 2016. A new approach to censored quantile regression estimation. University of Michigan Technical Report.

Zeileis A, Koenker R, Doebler P. 2015. *glmx: Generalized Linear Models Extended*. R package version 0.1-1.
URL `https://CRAN.R-project.org/package=glmx`

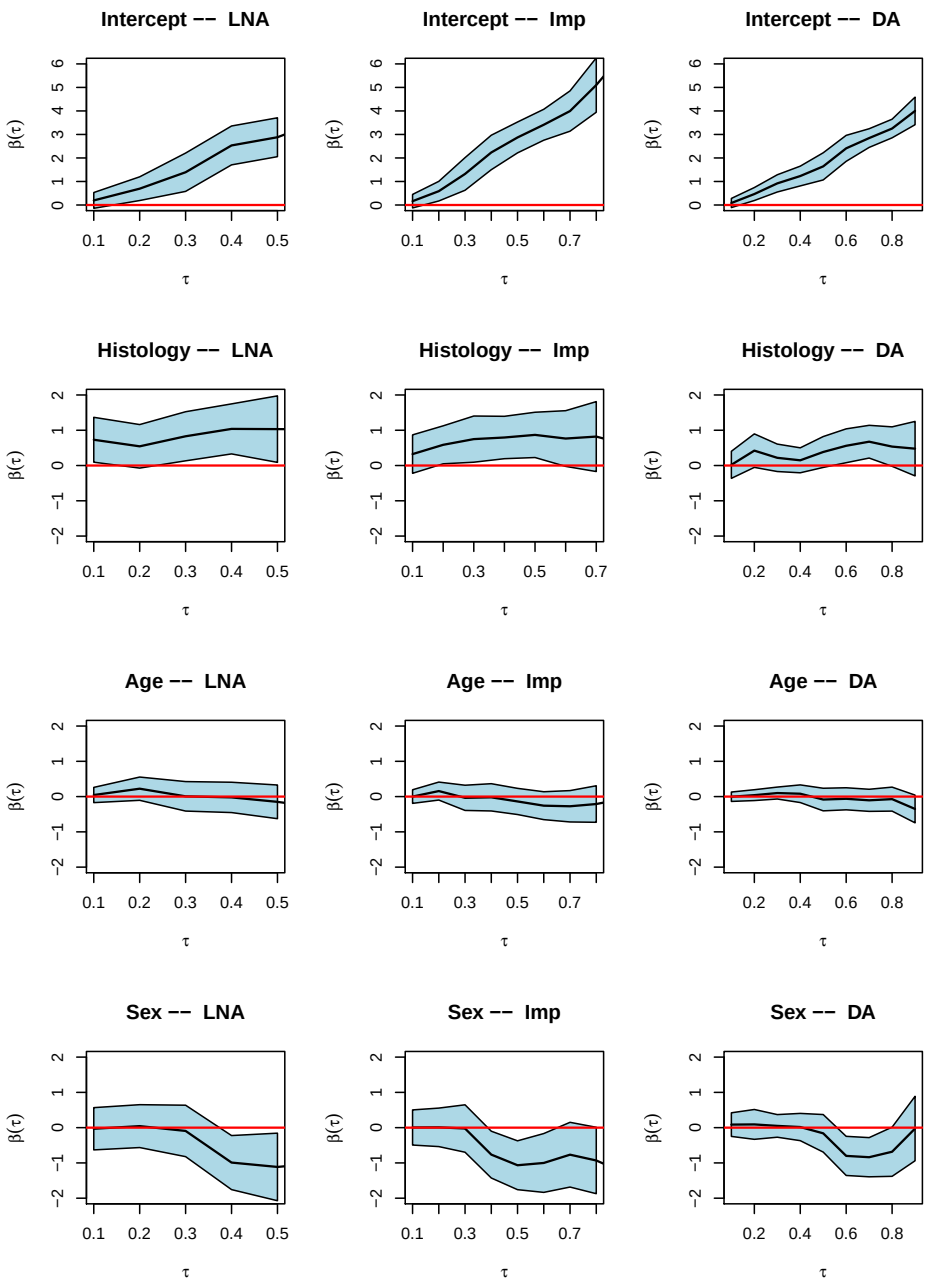

FIGURE 1. Estimated Quantile Regression Coefficients for the Lung Cancer Model: Comparison of three estimators, the blue pointwise bands are based on 200 bootstrap replications.