Biostatistics (2009), **1**, 1, *pp.* 1–3
Printed in Great Britain

Supplementary material to Rank-based estimation in the ℓ_1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data

By BRENT A. JOHNSON

Department of Biostatistics, Emory University, Atlanta, GA 30322, USA
e-mail: bajohn3@emory.edu

1. Simulation Studies

Multiple simulation studies were conducted to assess the effect of model misspecification in rank-based variable selection for censored data. Specifically, we consider the effect of fitting the regularized Gehan estimator (Johnson, 2008; Cai and others, 2009) when the true underlying model is the partial linear model,

$$\log T_i = \phi(Z_i) + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i.$$

Our simulation study details are adapted from those given by Chen and others (2005) and all calculations were conducted in `R` using the `quantreg` package. We assume that the true regression coefficients are $\boldsymbol{\beta} = (3, 3/2, 0, 0, 2, 0, 0, 0)'$, $\varepsilon_i \sim N(0, \sigma^2)$ and mutually independent of $(Z_i, \mathbf{X}_i)$. The predictors $\mathbf{X}_i$ followed a standard normal with the correlation between the jth and kth components of $\mathbf{X}$ equal to $0{\cdot}5^{|j-k|}$. The random variable Z_i was correlated with $\mathbf{X}_i$ through the relation $Z_i = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + U_i$, where $(\gamma_1, \gamma_2, \gamma_3) = (1/4, 1/4, 1/2)$ and U_i is $\mathrm{Un}(-5, 5)$ and completely independent of all other random variables. As in Chen and others (2005), we considered quadratic and linear effects, $\phi(Z_i) = Z_i^2$ and $\phi(Z_i) = 2Z_i$, respectively. Finally, censoring random variables were simulated according to the rule, $C_i = \phi(Z_i) + \mathbf{X}_i'\boldsymbol{\beta} + U_i^*$, where U_i^* follows the uniform distribution $\mathrm{Un}(0, \tau)$ with $\tau = 8$. Our simulation results are displayed in Table 1.

Because the stratified estimator does not attempt to estimate the nonlinear effect ϕ, standard definitions of model error do not apply. Instead, we define the partial model error, $\mathrm{PME} \equiv (\widehat{\boldsymbol{\beta}}_{\mathrm{S}(1)} - \boldsymbol{\beta})' \left[E(\mathbf{X}\mathbf{X}')\right] (\widehat{\boldsymbol{\beta}}_{\mathrm{S}(1)} - \boldsymbol{\beta})$, which has the heuristic interpretation as additional variation in the model resulting from $\widehat{\boldsymbol{\beta}}_{\mathrm{S}(1)}$ after adjusting for Z_i. The median of PME over Monte Carlo data set is reported in Table 1. The relative median PME (RPME) is defined as the lasso median PME divided by the unpenalized median PME and summarizes the average improvement in PME due to lasso regularization and variable selection. In addition, we monitor the average number of correct (C) and incorrect (I) zeros over Monte Carlo data set. For comparison purposes, we also computed the Gehan and Gehan lasso estimates assuming a linear model in Z_i and $\mathbf{X}_i$. In each instance of lasso, five-fold cross-validation was used to tune the regularization parameter λ.

Table 1. *Simulation results comparing regularized Gehan and stratified estimator*

	Quadratic					Linear				
	PME					PME				
Method	Unpen	lasso	RPME	C	I	Unpen	lasso	RPME	C	I
$n = 75, \sigma = 1{\cdot}5$										
Gehan	7.49	4.73	63.15	1.39	0.20	0.30	0.24	80.00	1.32	0
Strat. ($K_n = 2$)	6.98	4.18	59.89	2.40	0.32	2.34	1.39	59.40	1.79	0.01
Strat. ($K_n = 4$)	1.81	1.46	80.66	3.08	0.08	0.83	0.51	61.45	2.51	0
Strat.* ($K_n = 8$)	1.03	0.70	67.96	3.85	0.02	0.51	0.24	47.05	3.58	0
$n = 75, \sigma = 3$										
Gehan	10.05	6.13	61.00	1.45	0.25	1.17	0.83	70.94	1.33	0
Strat. ($K_n = 2$)	9.66	5.23	54.14	2.48	0.38	3.03	1.75	57.76	2.25	0.01
Strat. ($K_n = 4$)	3.88	2.52	64.95	3.06	0.23	1.58	1.08	68.35	2.85	0.01
Strat.* ($K_n = 8$)	2.60	1.78	68.46	3.54	0.16	1.66	1.05	63.25	3.55	0.04
$n = 100, \sigma = 1{\cdot}5$										
Gehan	4.91	3.05	62.12	1.74	0.12	0.20	0.15	75.00	1.60	0
Strat. ($K_n = 2$)	4.67	2.74	58.67	2.56	0.17	1.66	0.94	56.63	2.23	0
Strat. ($K_n = 4$)	1.26	0.92	73.02	2.86	0.02	0.58	0.32	55.17	2.91	0
Strat.* ($K_n = 8$)	0.67	0.56	83.58	3.97	0.02	0.35	0.22	62.86	3.52	0
$n = 100, \sigma = 3$										
Gehan	7.20	4.45	61.81	1.73	0.20	0.87	0.61	70.11	1.44	0
Strat. ($K_n = 2$)	7.04	3.72	52.84	2.72	0.28	2.43	1.36	55.97	2.34	0
Strat. ($K_n = 4$)	2.53	1.92	75.89	3.10	0.08	1.35	0.78	57.78	2.65	0
Strat.* ($K_n = 8$)	1.69	1.08	63.91	3.62	0.02	1.03	0.65	63.11	3.38	0.01

* Tune λ using generalized cross-validation in Johnson (2008)

First, when the effect ϕ is linear, the Gehan lasso beats the stratified estimator which confirms our intuition. At the same time, it is interesting to note that stratified estimator gradually achieves similar operating characteristics as the unstratified estimator as the sample size increases and number of strata K_n increases. Nevertheless, the stratified estimator is far too cumbersome if the unknown function ϕ is indeed linear in Z_i. Now, the big improvements in the stratified estimator are seen when the unknown function ϕ is nonlinear. For example, when the sample size $n = 75$ and $\sigma = 1{\cdot}5$, the partial model error is 7.49 and 4.73 for the unpenalized and regularized Gehan, respectively. We compare this to the PME of the unpenalized and regularized stratified estimator, 1.03 and 0.70, respectively. Hence, there is an average seven-fold increase in PME if we fit the Gehan lasso when the true underlying model is a nonlinear (i.e. quadratic) function of Z_i. Finally, it comes as no surprise that the stratified estimator performs better as the number of

strata increase; such finding agrees with Chen and others (2005) and with stratified estimators, in general. Of course, the small sample improvement comes with added computational burden.

References

Cai, T., Huang, J. and Tian, L. (2009) Regularized estimation for the accelerated failure time model. *Biometrics* (In press).

Chen, K., Shen, J. and Ying, Z. (2005) Rank estimation in partial linear model with censored data. *Statistica Sinica* **15** 767–779.

Johnson, B. A. (2008) Variable selection in semiparametric linear regression with censored data. *J. R. Statist. Soc. Ser. B* **70** 351–370.

[*Received*]