Published in final edited form as:

Test (Madr). 2009 November 1; 18(3): 463–467. doi:10.1007/s11749-009-0165-7.

Comments on: A review on empirical likelihood methods for regression

Gang Li and

Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095-1772, USA

Xuyang Lu

Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095-1772, USA

Gang Li: vli@ucla.edu

1 Introduction

Professors Chen and Keilegom have given a comprehensive and timely review on recent developments of empirical likelihood methods for regression. They have covered a wide array of topics including parametric regression, nonparametric regression, semiparametric regression, missing data, censored data regression, and goodness-of-fit tests. Our discussion will focus only on Sect. 6.1 regarding semiparametric linear regression for randomly right-censored data. Here we prefer to use the term “semiparametric regression” instead of “parametric regression” for Sect. 6.1 since the probabilistic model includes the completely unknown error distribution as a non-parametric component.

Professors Chen and Keilegom have discussed two popular empirical likelihood (EL) methods, a synthetic data EL method (cf. Qin and Jing 2001, Li and Wang 2003, and Qin and Tsao 2003), and a censored data EL method (cf. Zhou and Li 2008) for semiparametric linear regression with censored data. These two methods are based on the approaches of Koul, Susarla, and Van Ryzin (KSV) (1981) and Buckley and James (1979), respectively. Their properties have also been contrasted nicely in the paper. In particular, the synthetic data EL method requires a restrictive assumption that the censoring time C is independent of both the covariate X and the survival time Y , whereas the censored data EL method of Zhou and Li (2008) only assumes that the censoring time C is conditionally independent of Y given X . Below we present some numerical examples to illustrate that the synthetic data approach is very sensitive to the independence assumption of C and X . Hence caution should be exercised when using the synthetic data approach in practice. For simplicity, we only consider point estimation in the following discussion.

2 A simulation study

Table 1 presents results from a small simulation study to examine and compare the bias and variance of the KSV estimate and the Buckley–James estimate of the slope parameter under two different models. Both models assume that $Y = 1 + X + \varepsilon$, where X and ε are independent normal random variables with mean 0 and variance 0.25. Model 1 assumes that $C = a + 2X + \varepsilon$, where $\varepsilon \sim N(0, 0.25)$ is independent of X and ε . Hence C is not independent of X but is conditionally independent of Y given X . In model 2, we assume that $C \sim N(\mu, 16)$ is

independent of X and ε . The values of a and μ can be adjusted to give a prespecified censoring rate for each model.

It is seen from Table 1 that the KSV synthetic data estimate is seriously biased under model 1, where the censoring time C is not independent of the covariate X . The bias does not diminish as the sample size grows large. More specifically, the KSV method consistently overestimates the slope under model 1. To see why this happens, we recall that the KSV method first defines the synthetic dependent variable

$$Y_{iG} = \frac{\Delta_i T_i}{1 - \hat{G}(T_i-)},$$

where $\hat{G}(t)$ is the Kaplan–Meier estimate of the censoring distribution function G , and then regresses Y_{iG} on the covariate using the standard least squares estimation principle. It is easy to see that under model 1, the censoring rate decreases as X increases. This implies that there would be more zero Y_{iG} 's for small X values and more inflated nonzero Y_{iG} 's for large X values. Thus regressing Y_{iG} on X would lead to a larger slope estimate than regressing Y on X .

3 An example

We further illustrate the importance of the independence assumption of C and X to the KSV method using the heart transplant data of Miller (1976, Table 1). The data includes the lengths of survival (in days) after transplantation, ages at time of transplant, and T5 mismatch scores for 69 patients who received heart transplants at Stanford and were followed from 1 October 1967 to 1 April 1974. Twenty-four patients were still alive on 1 April 1974 and thus their survival times were censored. We consider model (6.1) of Koul et al. (1981), where the dependent variable Y is the logarithm to base 10 of the length of survival from transplantation, and the independent variable (covariate) is the mismatch score T5. As in Koul et al. (1981), regression of survival on the mismatch score T5 was performed with nonrejection related death being treated as censoring since the mismatch score is directed at the rejection phenomenon (cf. Miller 1976). A scatter plot of Y versus T5 is depicted in Fig. 1. There appears to be a negative correlation between Y and T5.

We regressed Y on T5 using both the KSV synthetic data method and the Buckley–James method. The fitted models are given below.

$$Y = 1.311 + 0.258 * T5 \text{ (Koul et al. 1981),}$$

$$Y = 3.316 - 0.593 * T5 \text{ (Buckley and James 1979).}$$

The KSV fitted model suggests a positive correlation between the survival time and the mismatch score, which contradicts to the common belief that the lower the mismatch score, the longer the survival time. In contrast, the Buckley–James fitted model does suggest a negative correlation between the survival time and the mismatch score. So what might have gone wrong with the synthetic data method for this data?

Recall that Koul et al. (1981)'s method requires that the censoring time is independent of the covariate T5. Is this a reasonable assumption for the Stanford heart transplant data? To

answer this question, we compared the censoring time distributions between those patients with $T_5 \leq 0.7$ and those with $T_5 > 0.7$ in Fig. 2.

Figure 2 reveals that patients with $T_5 \leq 0.7$ tend to have shorter censoring times than those with $T_5 > 0.7$. The log-rank test of these distributions produced a p -value of 0.035. Thus, there is strong evidence that the censoring time is not independent of the mismatch score T_5 . The relatively large percentage of censored observations for patients with small T_5 scores leads to more zero values for the synthetic dependent variable Y_{iG} for small mismatch scores T_5 , thus artificially producing a positive slope estimate.

4 Concluding remarks

The independence assumption between C and X is crucial to the validity of the synthetic data approach of Koul et al. (1981). Our limited simulation shows that the synthetic data method can be seriously biased when this assumption is not met. The Stanford heart transplant data example demonstrates that the synthetic data method can produce a misleading conclusion when used inappropriately.

Finally, Table 1 shows that the Buckley–James method appears to be superior to the synthetic data method with smaller bias and smaller variance. On the other hand, the censored data EL confidence region of Zhou and Li (2008) is computationally much more demanding than the synthetic data EL confidence region. It is also difficult to extend the method of Zhou and Li (2008) to incorporate auxiliary information and to construct confidence regions for linear combinations of the regression coefficients, which are relatively easy using the synthetic data EL method (cf. Li and Wang 2003). Therefore, with large samples, the synthetic data EL method can still be useful if the independence assumption between C and X is met.

References

- Buckley J, James I. Linear regression with censored data. *Biometrika* 1979;66:429–436.
- Koul H, Susarla V, van Ryzin J. Regression analysis with randomly right-censored data. *Ann Stat* 1981;9:1276–1288.
- Li G, Wang Q. Empirical likelihood-based inference for a censored linear regression model. *Stat Sin* 2003;13:51–68.
- Miller RG. Least square regression with censored data. *Biometrika* 1976;63:449–464.
- Qin G, Jing B. Empirical likelihood for censored linear regression. *Scand J Stat* 2001;28:661–673.
- Qin G, Tsao M. Empirical likelihood inference for median regression models for censored survival data. *J Multivar Anal* 2003;85:416–430.
- Zhou M, Li G. Empirical likelihood analysis of the Buckley–James estimator. *J Multivar Anal* 2008;99:649–664. [PubMed: 19018294]

Fig. 1.
Scatter plot of log 10 (survival time) versus mismatch score T5

Fig. 2.
Estimated survival functions of the censoring time for patients with $T5 \leq 0.7$ and $T5 > 0.7$
(log-rank test p -value = 0.035)

Simulated bias and variance of the estimated slope using Koul et al. (1981)'s method and Buckley–James (1979)'s method

Table 1

Censoring rate	Sample size	Model 1			Model 2		
		Bias		Variance	Bias		Variance
		KSV	BJ	KSV BJD	KSV	BJD	KSV BJ
32%	50	0.9683	−0.0054	0.1226	0.0305	−0.0229	−0.0070 0.1149 0.0318
	100	1.0214	0.0005	0.0668	0.0148	−0.0087	0.0027 0.0561 0.0142
	500	1.0863	−0.0002	0.0164	0.0029	−0.0062	−0.0012 0.0105 0.0029
60%	50	1.7353	−0.0144	0.5080	0.0496	−0.0579	−0.0092 0.3254 0.0538
	100	1.9594	−0.0030	0.2926	0.0234	−0.0104	0.0014 0.1474 0.0247
	500	2.2859	−0.0018	0.0675	0.0043	−0.0135	−0.0002 0.0288 0.0048

Note: Each entry is based on 1,000 Monte Carlo samples. BJ refers to the Buckley–James (1979) method, and KSV refers to the synthetic data method of Koul et al. (1981)