New Approaches in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 16, Ljubljana: FDV, 2000

On the Use of Buckley and James Least Squares Regression for Survival Data

Janez Stare[1], Harald Heinzl[2], and Frank Harrell[3]

Abstract

The method of Buckley and James (1979) for fitting linear regression models to censored data has been shown to have good statistical properties under usual regularity conditions. Nevertheless, even after 20 years of its existence, it is almost never used in practice. We believe that this is mainly due to lack of software and we state three reasons for using the method. We argue against some findings by Heller and Simonoff (1992) and in remainder of the paper briefly explore the method's behaviour under model misspecification. We conclude that at present there are no good procedures for checking the model's assumption of homoscedasticity under censoring and therefore do not recommend using this method with censoring higher than 20%.

1 Introduction

The field of Survival analysis is dominated by the Cox proportional hazards model (Cox, 1972). While there are many good reasons for this, at least part of the domination is due to the fact that modelling via the Cox model is easily accessible to practitioners since almost every statistical package has the procedure implemented. Still, other methods exist, and it can be argued that they could often be used (Aalen, 1989). For example, there is probably little doubt that the Cox model would rarely be used if there were no censored data and that linear regression would then prevail.

In this paper we focus our attention to the method of Buckley and James (1979), which is the usual least squares regression adapted for censored data. The method has been shown to be consistent under usual regularity conditions and superior to other least squares approaches to censored data (Miller and Halpern, 1982; Heller and Simonoff, 1990). Heller and Simonoff (1992) also compared the

[1] Dept of Biomedical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia.
[2] Dept of Medical Computer Sciences, Vienna University, Austria.
[3] Divis. of Biostatistics and Epidemiology, School of Medicine, University of Virginia, USA.

method to the Cox model. In Section 3 we critically review this comparison and argue that the guidelines given by Heller and Simonoff can not be used.

The rest of the paper deals with the problem of model assumptions, certainly the most critical point in using the method. Surprisingly, there has been little work in this area, although a sensible usage of the model crucially depends on the assumptions of linearity and homoscedasticity. We conclude that the method of Buckley and James can safely be used only when the amount of censoring is small.

2 The method of Buckley and James

The model assumes that time *T*, or some monotone transformation of it, is linearly related to the covariate vector **x**, say

$$T_i = \beta_0 + \beta' \mathbf{x}_i + \varepsilon_i \qquad (1)$$

where ε_i are iid with $E(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \delta^2$ and the distribution of ε_i is independent of **x**. Since under censoring we only observe $Y_i = \min(T_i, C_i)$, where C_i are censoring times, and equation (1) does not hold for Y_i, the usual least squares regression approach is not applicable. Buckley and James define

$$Y_i^* = Y_i \delta_i + E(T_i | T_i > Y_i)(1 - \delta_i),$$

where $\delta_i = I(T_i \le C_i)$, the censoring indicator. Considering the above equation for $\delta_i = 1$ and $\delta_i = 0$, one easily verifies that $E(Y_i^*) = E(T_i)$. The idea then is to replace Y_i for censored observations with Y_i^*. To do this we need to estimate the quantity $E(T_i | T_i > Y_i)$ for such Y_i. Now

$$E(T_i | T_i > Y_i) = E(\beta_0 + \beta' \mathbf{x}_i + \varepsilon_i | \beta_0 + \beta' \mathbf{x}_i + \varepsilon_i > Y_i) = \beta_0 + \beta' \mathbf{x}_i + E(\varepsilon_i | \varepsilon_i > Y_i - (\beta_0 + \beta' \mathbf{x}))$$

and it remains to estimate $E(\varepsilon_i | \varepsilon_i > Y_i - (\beta_0 + \beta' \mathbf{x})) = \int_{Y_i - (\beta_0 + \beta' \mathbf{x})}^{\infty} \varepsilon \frac{dF}{1 - F(Y_i - (\beta_0 + \beta' \mathbf{x}))}$,

where *F* is the distribution function of ε. After substituting for *F* its Kaplan-Meier (1958) estimate $\hat{F}$ (one minus the usual Kaplan-Meier survival function estimate), we have

$$y_i^* = y_i\delta_i + \left(\beta'\mathbf{x}_i + \frac{\sum_{\varepsilon_j > \varepsilon_i} w_j \varepsilon_j}{1 - F(\varepsilon_i)} \right)(1 - \delta_i).$$

where w_j are steps of $\hat{F}$.

If we could observe y_i^*, a reasonable estimate would be

$$\hat{\beta} = \frac{(\mathbf{X} - \overline{\mathbf{x}})' \mathbf{y}^*(\hat{\beta})}{(\mathbf{X} - \overline{\mathbf{x}})'(\mathbf{X} - \overline{\mathbf{x}})}$$

Replacing y_i^* by their estimates and taking into account that estimates depend on β, we need iterations. Finally we have $\beta_0 = \overline{y}^* - \beta'\overline{\mathbf{x}}$.
The reader is advised to convince himself that the method ensures that no censored observation is shortened in the process.

3 Why use the Buckley and James method?

Given that Cox model is currently the method of choice for survival data, maybe the question is ‘Why not use the Cox model?’. We give here three reasons:

1. The basic assumption of Cox model, proportionality of hazards, is not always met. This fact is often overlooked, again probably because there are sometimes no alternatives (software!).
2. Published results of fits with Cox model do not allow their usage for prediction purposes. To be able to predict one needs to estimate the baseline hazard which can only be done if one has data available. This is not the case with linear regression where only coefficients are needed for prediction.
3. Results of Cox model fit are difficult to explain to non-statisticians and give less information than results of linear fits.

Once we decided not to use the Cox model, there are still many options which might be considered. We consider the Buckley and James method in this paper because it has been shown to be superior to other semiparametric least squares approaches (Heller and Simonoff, 1990) and better than parametric versions of (1) if distribution of ε_i is misspecified.

Heller and Simonoff (1992) compared the predictive ability of the Buckley and James method to the Cox model and gave a decision table for choice between the two models, based on censoring proportion and a certain R^2 calculated for the Cox model. Their simulations also involved different censoring failure time distributions, but these were found not to be very important in judging the predictive ability of Buckley and James method. After a reader’s reaction

(Feingold, 1993) the table was corrected to give even more credit to the Buckley and James method. We reproduce the corrected version in Table 1. In our opinion such a table is misleading and since we feel that there is some potential danger in such a table actually being used by practitioners to decide which model to use, we give here some warnings concerning Heller and Simonoff's comparison.

Table 1: A 'decision' table of Heller and Simonoff. B&J stands for the Buckley and James method.

Censoring proportion	Strength of Regression	Model choice
< 40%	All	B&W
40% - 60%	.00 - .25 .25 - .65 .65 - 1.00	B&J Cox B&J
> 60%	All	Cox
correction > 60%	.00 - .55 .55 - 1.00	Cox B&J

First, it is conceptually wrong to base a decision on any results obtained after fitting the Cox model. Usage of an R^2 measure would make sense only if it could be calculated on both models and the values compared. If we are deciding between Buckley and James and Cox, we can rarely expect both to be correct, so that some measure calculated on one of them should have no influence on the other.

The results of Heller and Simonoff should be understood in the following way: if both models are correct, than in most situations predictions based on Buckley and James method are better than predictions with Cox model. This is not surprising. If Buckley and James method is consistent and unbiased and the underlying model is truely linear, then there is no reason why any other method would predict better. Given that predictions with Cox model in Heller and Simonoff 's paper were based on medians and the criterion used was relative squared error, it is in fact surprising that the Cox model performed better in any situation at all. It is even more surprising that the Cox model was favoured more often with high censoring, when medians sometimes, depending on the strength of effect, can not be estimated (the authors used the highest observed value in such situation[4]). It should also be noted that two thirds of Heller and Simonoff's simulations were based on non-proportional hazard models (log-normal and gamma errors), which further favoured Buckley and James.

Table 2: Illustration of bias when fitting a linear model to nonlinear data. See text for details. *B&J* stands for Buckley and James method.

Proportion censored	B&J	Fit to noncensored data	Bias
10%	4.98	5.06	0.08
30%	3.94	5.07	1.13
50%	2.88	5.05	2.17
70%	2.06	5.03	2.97

4 When should one use the Buckley and James method?

The main requirement for usage of any statistical model is that it's assumptions are met. From derivations in Section 2 it follows that two requirements must be met if one wants to use the Buckley and James method:

1. the model must be linear (in the parameters),
2. the distribution of residuals must not depend on the values of covariates (homoscedasticity).

Checking these assumptions under censoring is difficult, if at all possible. We report here on some results of our simulations. We first illustrate the *effect* of not meeting the above assumptions, and then we discuss the problems encountered in *checking* the assumptions.

Table 2 gives results of model fitting when data were generated as $Y = X^2 + \varepsilon$, X was uniform on (0,5) and ε followed extreme value distribution with shape parameter 2. A misspecified model in the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ was fitted to data sets with 100 subjects and different censoring proportions. Listed coefficients are averages of 100 runs for each censoring proportion. For comparison, coefficients of fits to noncensored data are given.

[4] Personal communication.

It is evident that bias increases with censoring, almost linearly. While this is only one example of violating the linearity assumption, it shows quite clearly that we can expect biased results with Buckley and James method, if this assumption is not met.

To study effects of heteroscedasticity we again used a simple linear model with X uniform on (1,5) and normally distributed error. The standard deviation of the error distribution was proportional to values of X. For example, degree 5 of heteroscedasticity means that standard deviation of the error distribution increased 5 times (from σ to 5σ) when X increased from 1 to 5. Of course, if the increase in variability had been observed on a larger interval of X, heteroscedasticity and bias would have been smaller.

Our simulations show that bias increases with:

- degree of heteroscedasticity,
- censoring proportion,
- error variability (σ),
- effect size (β).

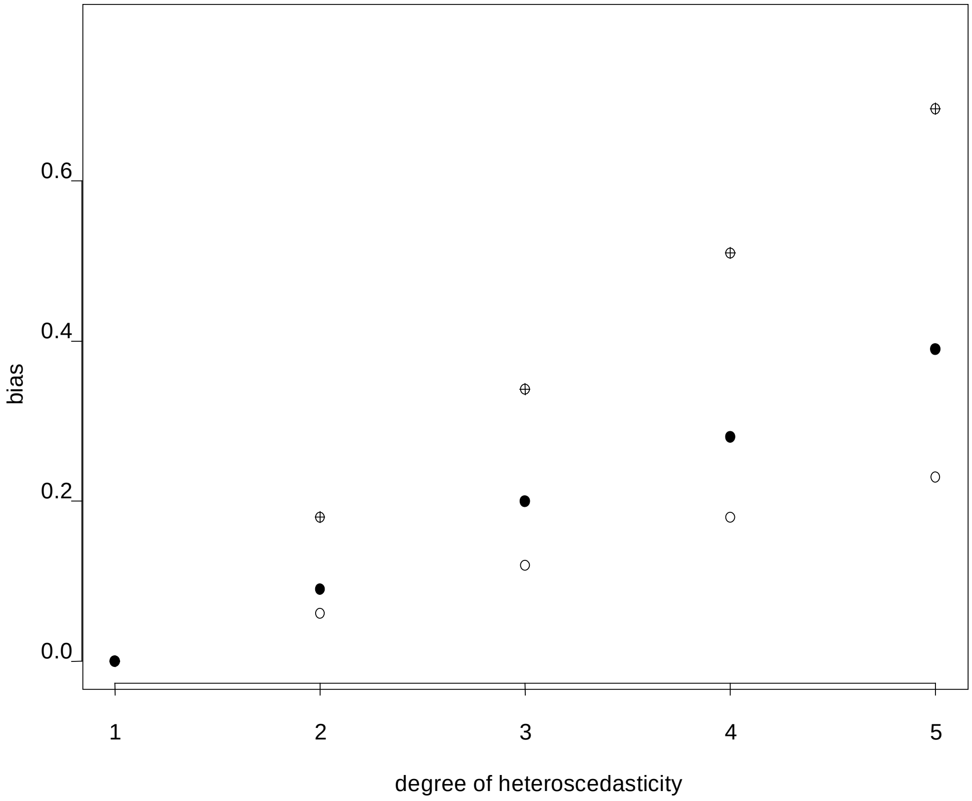

Figure 1: Bias for 20 (symbol o), 30 (symbol •) and 50% censoring with β = 1.

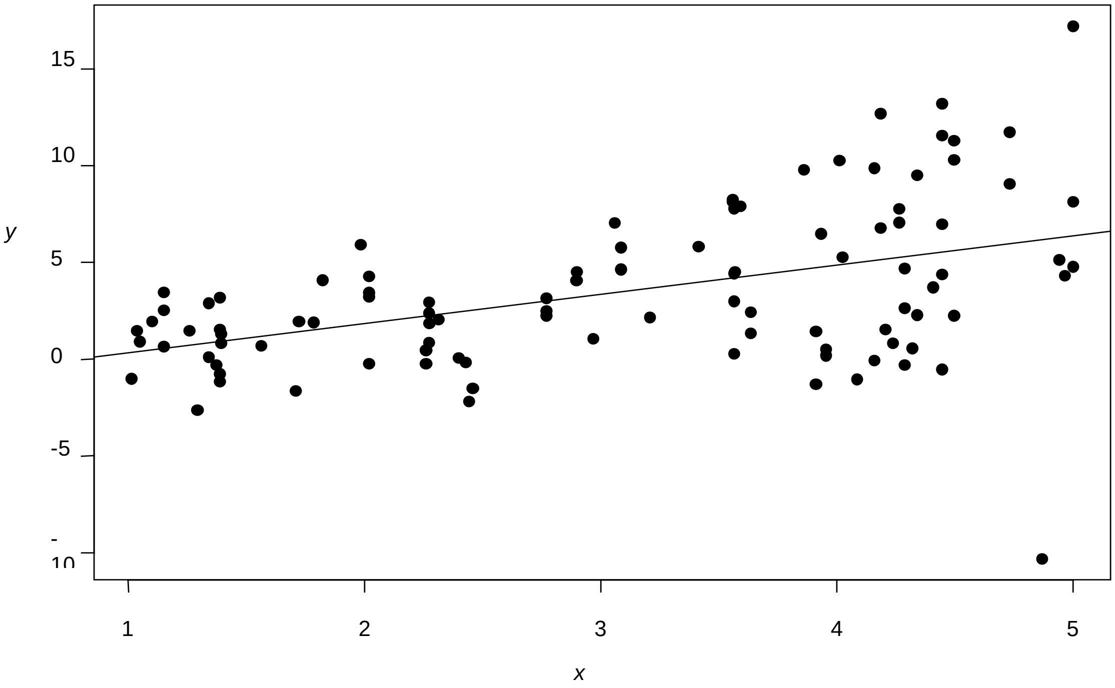

Figure 2: Non-censored heteroscedastic data.

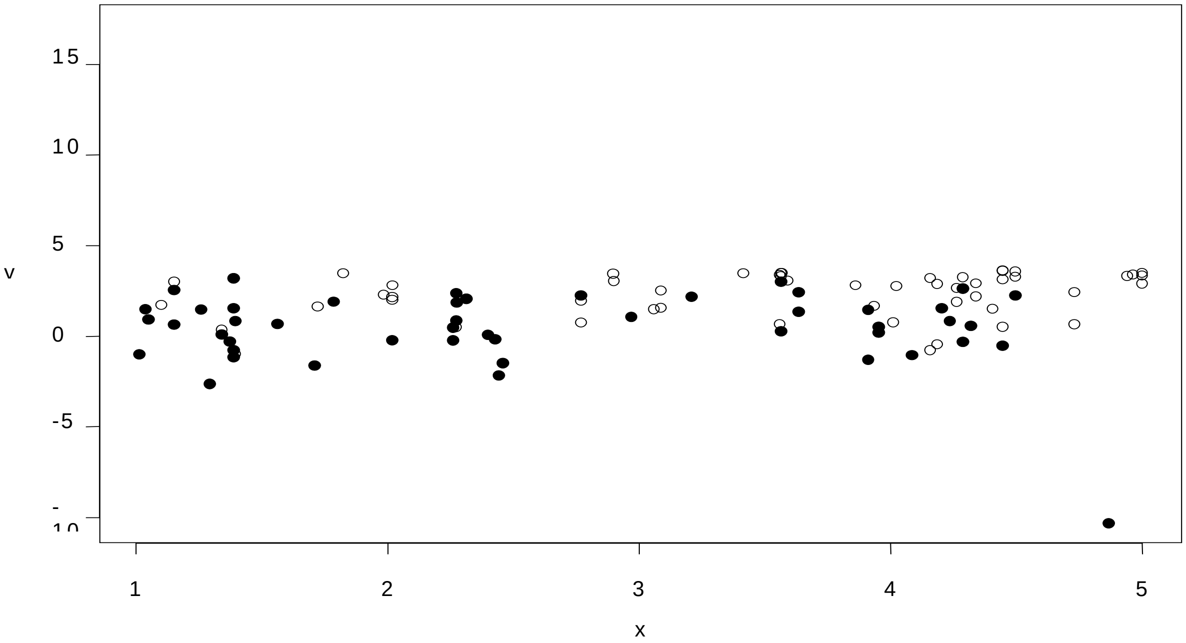

Figure 3: Data of Figure 2 after 50% censoring. Symbol o denotes censored observations.

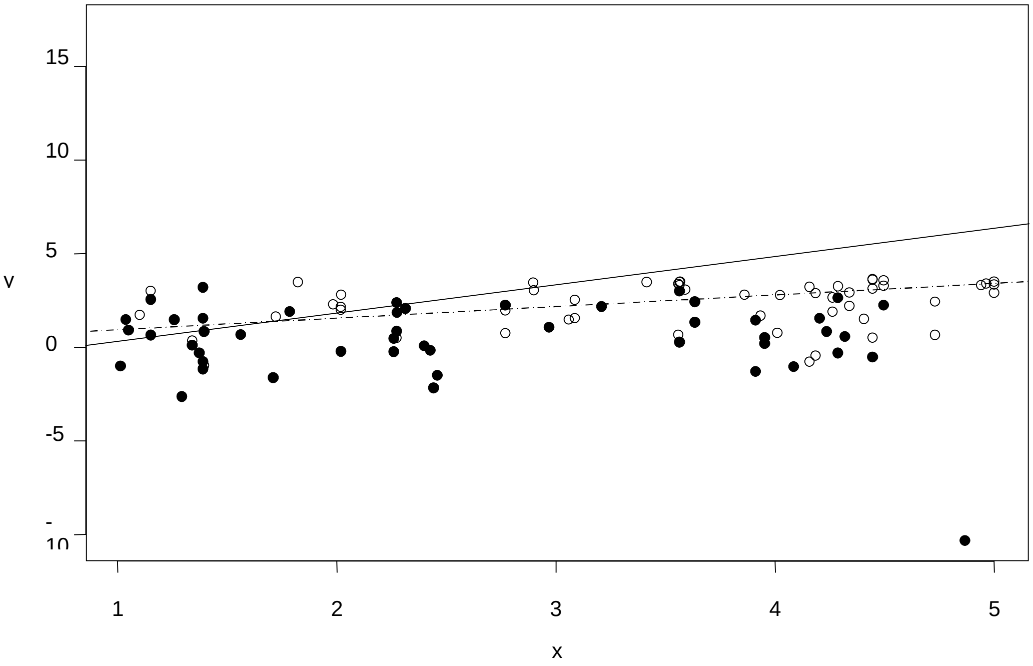

Figure 4: Buckley and James fit to data in Figure 3 (dashed line). The solid line shows the unbiased fit.

The first two features are illustrated in Figure 1. Results shown are σ=1 at X=1 and β=1. Effect size does not play an important role, in our simulations bias was no different for betas from 0.2 to 2. This means (see Figure 1) that for β = 0.2 and high heteroscedasticity for example, one can get an estimate of β that is negative even with censoring proportion of 20%.

While probably obvious, we nevertheless stress here the following fact: if the model is misspecified, the Buckley and James method will *not* reproduce the underlying best fit, as the method of least squares does when there is no censoring.

Given that violation of assumptions gives biased results, the question is how we can check the model fit. Surprisingly, almost nothing has been done in this direction, one exception being the paper by Hillis (1995).

In Figures 2–4 we illustrate the difficulties encountered in checking the heteroscedasticity assumption. Figure 2 shows uncensored data obtained under the model described above, with degree of heteroscedasticity equal to 5. In Figure 3 the same data are presented under 50% censoring. It is obvious that despite high heteroscedasticity of the underlying model, nothing can be discerned from such a graph. Figure 4 illustrates the bias. The same problem arises with checking the linearity assumption. Hillis proposed to change the censored residuals by randomly changing their values, but this change occurs under the assumptions of correctly specified model, and therefore doesn't really help.

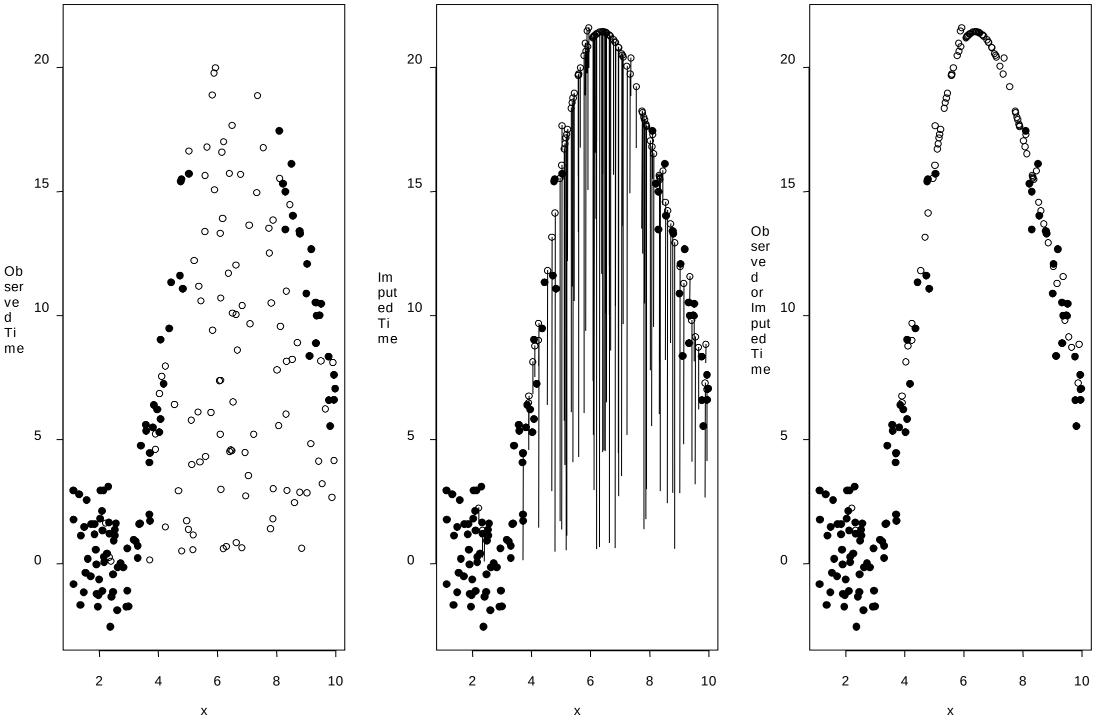

Figure 5: Splines fit to highly curved and censored data.

In our experience usage of splines was very useful in dealing with non-linearity, and one example is given in Figure 5. At this point, we can not report on any positive finding regarding homoscedasticity check. Adjustments of methods known from ordinary linear regression all fail and we can only suggest that Buckley and James method can safely be used with low proportions of censoring when such an assumption can still be checked using the usual graphical methods.

5 Conclusions

We have shown that the Buckley and James' least squares regression method is biased under misspecified model, so that its sensible usage depends crucially on the assumptions. While linearity assumption can be satisfactorily dealt with using splines, there is at present no method for checking the homoscedasticity assumption. Given that heteroscedasticity may introduce considerable bias, one should take great care in using the method with higher censoring proportions. In extreme cases, even 20% censoring could be dangerous.

With survival data heteroscedasticity will of course be common. On the other hand, logarithm of time will usually be modeled, which may already solve the problem. Probably, one will have to rely to some extent on experience.

6 Software

We wrote a S-plus program called *bj*, which has been included in Frank Harrell's library *Design*, available at http://hesweb1.med.virginia.edu/biostat/s/win/. Besides the fit, the program will allow easy usage of splines, produce renovated scatterplots as suggested by Smith and Zhang (1995), and plot Hillis' residuals besides the ordinary residuals.

References

[1] Aalen, O.O. (1989): A linear regression model for the analysis of life times. *Statistics in Medicine,* **8,** 907-25.

[2] Buckley, J. and James, I. (1979): Linear regression with censored data. *Biometrics*, **66**, 429-436.

[3] Feingold, M. (1993): Choice of prediction estimator in censored regression models. Biometrics, **49**, 661-664.

[4] Heller, G. and Simonoff, J.S. (1990): A comparison of estimators for regression with a censored response variable. *Biometrics*, **77**, 515-520.

[5] Heller, G. and Simonoff, J.S. (1992): Prediction in censored survival data: a comparison of the proportional hazards and linear regression models. *Biometrics*, **48**, 101-115.

[6] Hillis, S.L. (1995): Residual plots for the censored data linear regression model. *Statistics in Medicine*, **14**, 2023-2036.

[7] James, I.R. and Smith P.J. (1984): Consistency results for linear regression with censored data. *The Annals of Statistics*, **12**, 590-600.

[8] Kaplan, E.L. and Meier, P. (1958): Nonparametric estimation from incomplete observations. *J. American Statistical Association*, **53**, 457-481.

[9] Lai, T.L. and Ying, Z. (1991): Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *The Annals of Statistics*, **19**, 1370-1402.

[10] Miller, R. and Halpern, J. (1982): Regression with censored data. *Biometrika,* **69**, 521-531.

[11] Smith, P.J. and Zhang, J. (1995): Renovated Scatterplots for Censored Data. *Biometrika*, **82**, 447-452.