A simple correction for ties when censoring times depend on covariates

Sven Ove Samuelsen

Department of Mathematics, University of Oslo

P.O. Box 1053 Blindern, 0316 Oslo, Norway

email: osamuels@math.uio.no,

Torbjørn F. Wisløff

Norwegian Centre for Health Services

Postboks 7004 St. Olavs plass 0130 Oslo, Norway

email: torbjorn.wisloff@nchs.no

and

Anders Skrondal

Division of Epidemiology, Norwegian Institute of Public Health

P.O.Box 4404 Nydalen, 0403 Oslo, Norway

email: anders.skrondal@fhi.no

Abstract

We point out that conventional methods for ties correction may be seriously biased when censoring times depend on covariates. We present a simple modification to the Efron correction method which appears to work remarkably well. The modified correction method is easy to implement and only slightly more computationally demanding than the Efron correction.

⁰*Key words:* Survival data, Proportional hazard model, Cox-regression, Ties, Efron correction, Peto-Breslow correction

1 Introduction

Tied survival data, or more generally tied time to event data, typically occurs when time is recorded coarsely. For instance, in a study of age at sexual debut (Sundet et al., 1992) age at event was recorded in whole years. For most individuals sexual debut takes place in their teens or early twenties and many individuals reported the same age at event. Thus, the data were heavily tied.

We will in this note assume that survival times are truly continuous. Ties then occur because the true times are grouped into disjoint intervals. For example, recorded age in whole years does not mean that the event took place at the birthday, but rather sometime between two birthdays. We will furthermore assume that censoring times are also continuous but recorded as grouped.

Correction for ties is a topic that has been taken very seriously for Cox-regression (Cox, 1972) with considerable theoretical developments (Peto, 1972, Kalbfleisch & Prentice, 1973, Kalbfleisch & Prentice, 1980, DeLong et al., 1994). However, these methods are very computationally demanding when the number of tied observations becomes large. In contrast, the most commonly used methods for handling ties such as the Peto-Breslow method (Peto, 1972 and Breslow, 1974) and the Efron correction (Efron, 1977) are conceptually and computationally simple although somewhat ad-hoc. For the Peto-Breslow method the same risk set, consisting of everybody at risk at the beginning of a grouping interval, is used for every event in that interval. The Efron method, on the other hand, mimicks that individuals are removed from the risk set after experiencing the events. This correction method can be viewed as down-weighting the contribution of individuals with an event in the interval to the risk set for events occurring late in the interval. Importantly, the individuals having experienced events typically have high risk covariates and the down-weighting alters the covariate distribution of individuals at risk which in turn affects the risk estimates.

When an event time and a censoring time are equal it has been a convention in survival analysis to let the event times precede the censoring times. This is the case for the counting process formulation (Andersen et al., 1993) where the indicator functions for being at risk are defined as continuous to the right. This convention is also reflected in the Peto-Breslow and Efron corrections, where all censorings in a grouping interval are effectively treated as if they occurred after all event times in the interval. If censoring times depend on covariates both these correction methods may then lead to biased estimation.

In the study of Sundet et al. (1992), which was based on a cross-sectional sample of the Norwegian population, an individual with a recorded age at say 18 would be between 18 and 19 years old. However, using the Efron correction this individual, if censored, would be treated as if her 19th birthday was imminent. Analysis of these data uncovered a strong cohort effect on age at sexual debut using year of birth as covariate. The covariate, however, is simply a linear transformation of the censoring time. This dependence produces a pronounced bias for the Efron correction method.

The purpose of this note is to present a simple modification of the Efron correction that accommodates tied censoring times. The basic idea is to down-weight the contribution to the risk set from individuals being censored as well as individuals experiencing events. In the next section we briefly review Cox-regression and the conventional correction methods and present our modified correction method. In Section 3 we apply different correction methods to the sexual debut data. In Section 4 we present a simulation study investigating the performance of the correction methods. Finally, we summarize our results and discuss implications for survival analysis in general.

2 Correction methods for ties

Under the proportional hazard assumption the hazard of an event for individual i with covariate vector z_i is given as

$$\lambda_i(t) = \exp(\beta' z_i) \lambda_0(t),$$

where β is a vector of regression coefficients and $\lambda_0(t)$ a baseline hazard function. When there are no tied survival times Cox (1972) suggested that β could be estimated by maximizing the partial likelihood

$$L(\beta) = \prod_{\tilde{t}_i} \frac{\exp(\beta' z_i)}{\sum_{k \in \mathcal{R}(\tilde{t}_i)} \exp(\beta' z_k)},$$

where $\tilde{t}_i$ is an observed event time and $\mathcal{R}(\tilde{t}_i)$ the set of individuals at risk at this time.

If the data are tied the true $\tilde{t}_i$ are not observed but instead only known to lie in some interval $I_j = [t_j, t_{j+1})$. The Peto-Breslow method for handling ties consists of estimating β by maximizing

$$L_{PB}(\beta) = \prod_j \prod_{\tilde{t}_i \in I_j} \frac{\exp(\beta' z_i)}{\sum_{k \in \mathcal{R}(t_j)} \exp(\beta' z_k)}.$$

Thus the same risk set $\mathcal{R}(t_j)$, consisting of all individuals at risk at the beginning of the interval, is used for all events in I_j . The above way of expressing $L_{PB}(\beta)$ highlights that it is a product over all events. Often it is alternatively written as a product over all grouping intervals by the expression

$$L_{PB}(\beta) = \prod_j \frac{\exp(\beta' Z_j)}{[\sum_{k \in \mathcal{R}(t_j)} \exp(\beta' z_k)]^{d_j}}.$$

Here $Z_j = \sum_{i \in \mathcal{D}_j} z_i$ where $\mathcal{D}_j$ is the set of individuals with events in I_j and d_j is the number of individuals in $\mathcal{D}_j$.

The Efron correction adjusts for removal of individuals in $\mathcal{D}_j$ from the risk set by maximizing

$$L_E(\beta) = \prod_j \prod_{\tilde{t}_i \in I_j} \frac{\exp(\beta' z_i)}{S_E(\beta, j, l)} = \prod_j \frac{\exp(\beta' Z_j)}{\prod_{l=0}^{d_j-1} S_E(\beta, j, l)}$$

where we arbitrarily attach indices $l = 0, \dots, d_j - 1$ to the d_j events in I_j and define the expression in the denominator as

$$S_E(\beta, j, l) = \sum_{k \in \mathcal{R}(t_j)} \exp(\beta' z_k) - \frac{l}{d_j} \sum_{k \in \mathcal{D}_j} \exp(\beta' z_k).$$

Thus the contribution to $S_E(\beta, j, l)$ from individuals with events in I_j is down-weighted as time elapses.

It is now quite straightforward to modify the Efron correction to also allow for down-weighting of censored individuals from the risk set. Let $\mathcal{C}_j$ be the set of individuals censored in I_j . We suggest a modified estimator maximizing

$$L_M(\beta) = \prod_j \prod_{\tilde{t}_i \in I_j} \frac{\exp(\beta' z_i)}{S_M(\beta, j, l)} = \prod_j \frac{\exp(\beta' Z_j)}{\prod_{l=0}^{d_j-1} S_M(\beta, j, l)},$$

where

$$S_M(\beta, j, l) = \sum_{k \in \mathcal{R}(t_j)} \exp(\beta' z_k) - \frac{l}{d_j} \sum_{k \in \mathcal{D}_j} \exp(\beta' z_k) - \frac{l + 0.5}{d_j} \sum_{k \in \mathcal{C}_j} \exp(\beta' z_k).$$

Note that both censored individuals as well as individuals experiencing events in I_j are down-weighted in $S_M(\beta, j, l)$.

In $S_M(\beta, j, l)$ we suggest to down-weight the censored individuals by the factor $(l + 0.5)/d_j$ to allow for censoring before the first event and after the d_j -th event. Alternatively, the factors l/d_j and $(l + 1)/d_j$ were considered but differences were negligible.

Neither $L_{PB}(\beta)$, $L_E(\beta)$ nor $L_M(\beta)$ are partial likelihoods. Thus it does not follow that variances can be estimated by inverting their respective information matrices. However, software implementations of the Peto-Breslow and Efron corrections typically base variance estimates on the inverse information. In the simulations in Section 4 we have adopted the same approach and investigate the performance by comparing the mean of the model based variances with the empirical variances of the parameter estimates.

Other methods for handling ties have been suggested. A simple ad-hoc approach, particularly useful when working with a Cox-regression program that has only the Peto-Breslow method implemented, consists of ordering the event and censoring times in I_j randomly. This is accomplished by adding random numbers to the times, for instance from the uniform distribution over $[0, t_{j+1} - t_j]$. A drawback is the extra variation induced by the random ordering which may be pronounced for small sample sizes.

The rather ad-hoc procedures discussed so far does not take into account that individuals with high risk covariates will tend to have shorter times to events, which is still the case within each grouping interval. In fact, the usual Efron correction is consistent with assuming that all orderings of the event times within I_j are equally likely whereas the modified Efron correction is consistent with the notion that all orderings of both event times and censoring times within I_j are equally likely.

In contrast, the partial likelihood for tied data derived by Kalbfleisch & Prentice (1973, 1980) is based on the likelihood of the different orderings of event times. A numerical approximation was considered by DeLong et al. (1994) and is implemented in SAS and STATA. However, this partial likelihood is derived under the assumption that all censored survival times in I_j occur after the event times in I_j . More generally for interval censored data, simulation and imputation methods have been suggested that take orderings of event times but not censoring times into account (Sihna et al., 1994, Satten, 1996, Satten et al., 1998, Datta et al., 2000).

For heavily tied data, discrete time failure methods are often recommended (e.g. Kalbfleisch & Prentice, 1980). The discrete time model corresponding to the proportional hazard model uses a complementary log-log link for a discrete time hazard, i.e. $\log(-\log(1 - P(\tilde{t}_i \in I_j | \tilde{t}_i \geq t_j, z_i))) = \alpha_j + \beta' z_i$. This model is easily fitted with software for generalized linear models. However, the typical application of the discrete time models also assume that censored times in I_j occur after all events in I_j .

3 Application to sexual debut data

In 1987 the National Institute of Public Health carried out a survey of sexual habits in Norway. The survey was a cross-sectional random sample of the population at age 18-60. Sundet et al. (1992) analyzed time to sexual debut and demonstrated a strong cohort effect. The effect was stronger among women and we will only present results for the 3107 female respondents here.

Following Samuelsen (2003) we analyzed time to first intercourse using a proportional hazard model with the covariate $z = (\text{year of birth}-1900)/30$. Using the modified Efron correction proposed here produced a regression parameter estimate of 0.791. In comparison, Samuelsen (2003) reported the estimate 0.695 for the Peto-Breslow method, 0.752 for the usual Efron correction and 0.755 for a discrete time model with a complementary log-log link.

Adding uniformly distributed numbers to the censored event times gave an average estimate of 0.792 with a range from 0.782-0.801 when refitting the model 100 times with different random numbers. This method was thus in good agreement with the modified Efron method.

All methods suggest a strong cohort effect on age at sexual debut and the differences between estimates from different methods may not be important from a subject matter point of view. However, the difference between the modified Efron correction

and the usual Efron correction is of the same magnitude as the difference between the usual Efron correction and the Peto-Breslow method which is often deemed important. Furthermore, only 9.4% of the event times were censored due to the adult study population. One might expect that differences would be larger if a larger proportion of the event times were censored. In the next section this is investigated in a simulation study.

4 Simulation study

We first simulated survival times from an exponential distribution with hazard $\lambda_z(t) = \exp(z)$, in other words from a proportional hazards model with constant baseline $\lambda_0(t)=1$ and regression parameter $\beta=1$. The covariates z were drawn from a uniform distribution over the interval $[0, 1]$. We used two censoring schemes, in the first the censoring time equals $c = 1 - z$ and in the second $c = z$. Censoring is hence linearly dependent on the covariate in accordance with the application in the previous section. Censoring occurred for 51% of the individuals under the first censoring scheme and for 47% under the second scheme. Tied data were created by grouping the censored survival times into the 10 intervals $[0, 0.1 >, [0.1, 0.2 >, \dots, [0.9, 1.0 >$.

Both censoring schemes were simulated 5000 times. In each simulation we first recorded the Cox-regression estimates based on untied data and subsequently estimates from the Peto-Breslow method, the usual Efron correction, the modified Efron correction and estimates obtained by adding random numbers to the tied data. In addition, we recorded the variance estimates for all methods.

Results from the simulations are given in Tables 1 and 2. Evidently the modified

Table 1
Results with censoring scheme 1 ($\beta=1$).

Correction method	Ave. β -est.	Ave. variance-est.	Emp. variance
Untied data	1.001	0.042	0.041
Peto-Breslow	0.433	0.036	0.029
Efron	0.463	0.036	0.033
Add random number	0.955	0.041	0.040
Modified Efron	0.954	0.041	0.039

Efron method performs well under both censoring schemes although there is some bias under the first scheme. The method of adding a random number seems to work equally well in these cases. In contrast, both the Peto-Breslow and the Efron methods are substantially biased under both censoring schemes. Note that the Efron correction is even more biased than the Peto-Breslow method under the second censoring scheme.

Table 2
Results with censoring scheme 2 ($\beta=1$).

Correction method	Ave. β -est.	Ave. variance-est.	Emp. variance
Untied data	0.996	0.038	0.037
Peto-Breslow	1.270	0.035	0.027
Efron	1.377	0.035	0.033
Add random number	1.007	0.038	0.036
Modified Efron	1.007	0.038	0.036

We also estimated the discrete time model with a complementary log-log link for the simulated data. The results were almost identical to those of the Efron correction and thus substantially biased. For the Cox-estimator based on untied data there is no apparent bias implying that our results can not be attributed to small sample bias of the Cox-estimator for untied data.

Regarding variance estimation, the suggestion of simply using the the inverse information matrix worked quite well except for the Peto-Breslow method. Surprisingly, we obtain a slightly smaller variance for the modified Efron than for the Cox-estimator with untied data. That the variance estimates with the Peto-Breslow method is smaller than the empirical variances is in accordance with theory (see e.g. Oakes, 1981 or Samuelsen, 2003).

The linear dependency between covariates and censoring times in these simulations is as extreme as it can be. Hence, we have also carried out simulations in order to investigate bias when there is less severe dependency between the censoring time and covariate and for stronger dependency between time to event and covariate. In these simulations the covariate was still uniformly distributed over $[0, 1]$. The event times were drawn from a proportional hazards model with $\lambda_z(t) = \exp(\alpha + \beta z)$ for several values of the regression parameter β and baseline hazard $\exp(\alpha)$. The dependency between the censoring time and covariate was quantified through the correlation ρ . Specifically, with u uniformly $[0, 1]$ and independent of z we drew censoring time c as $c = \sqrt{1 - \rho^2} u + \rho(1 - z)$ for $\rho < 0$, as $c = \sqrt{1 - \rho^2} u + \rho z$ for $\rho > 0$ and as $c = u$ for $\rho = 0$. The value of α was determined to produce 50% censoring. The censored survival data were then grouped into 10 intervals of equal length. In each simulation the total sample size was $n = 100000$ and for each value of ρ , β and α the simulations were repeated 5 times.

Table 3 gives the average estimates of β with both the conventional and modified Efron correction for some values of β and ρ . To ease interpretation, we also report the empirical correlation r between the uncensored event times and the covariate and the empirical correlation r' between log uncensored event time and the covariate. Note that a model with parameters β and ρ is equivalent to a model with parameters $-\beta$ and $-\rho$. There is thus symmetry in the table, apart from estimation error, around

Table 3

Average estimates of regression parameter β with 50% censoring using the Efron correction (E) and the modified Efron correction (M) for different values of β and ρ .

β		-4	-3	-2	-1	0	1	2	3	4
r		0.52	0.50	0.43	0.26	0.00	-0.26	-0.43	-0.50	-0.52
r'		0.67	0.56	0.41	0.22	0.00	-0.22	-0.41	-0.56	-0.67
ρ										
-1.0	E	-4.21	-3.27	-2.32	-1.37	-0.45	0.47	1.36	2.25	3.10
	M	-3.97	-2.99	-1.99	-1.00	-0.02	0.96	1.92	2.89	3.82
-0.8	E	-4.10	-3.16	-2.18	-1.22	-0.21	0.75	1.75	2.72	3.68
	M	-3.95	-2.98	-1.98	-1.01	0.02	0.99	2.00	2.97	3.94
-0.6	E	-4.10	-3.11	-2.13	-1.15	-0.17	0.83	1.81	2.80	3.77
	M	-3.99	-2.98	-1.99	-1.00	-0.01	1.00	1.98	2.98	3.95
-0.4	E	-4.04	-3.07	-2.08	-1.11	-0.11	0.88	1.86	2.84	3.84
	M	-3.96	-2.98	-1.98	-1.01	0.00	0.99	1.97	2.95	3.96
-0.2	E	-4.00	-3.04	-2.05	-1.07	-0.05	0.95	1.91	2.92	3.87
	M	-3.95	-3.00	-2.00	-1.01	0.01	1.01	1.97	2.98	3.94
0.0	E	-3.94	-2.97	-1.99	-0.98	0.00	0.99	2.00	2.97	3.93
	M	-3.95	-2.98	-2.00	-0.99	0.00	1.00	2.00	2.98	3.95
0.2	E	-3.87	-2.92	-1.91	-0.92	0.06	1.06	2.04	3.04	4.02
	M	-3.94	-2.98	-1.97	-0.98	0.01	1.00	1.99	2.99	3.98
0.4	E	-3.84	-2.86	-1.85	-0.88	0.11	1.10	2.07	3.07	4.05
	M	-3.96	-2.98	-1.97	-0.99	0.00	1.00	1.97	2.98	3.97
0.6	E	-3.77	-2.81	-1.82	-0.81	0.16	1.14	2.13	3.11	4.08
	M	-3.95	-2.98	-1.99	-0.98	0.00	0.99	1.99	2.98	3.97
0.8	E	-3.69	-2.72	-1.75	-0.75	0.23	1.21	2.19	3.16	4.12
	M	-3.95	-2.98	-2.00	-0.99	0.00	0.99	1.99	2.99	3.97
1.0	E	-3.08	-2.24	-1.36	-0.47	0.44	1.37	2.32	3.28	4.19
	M	-3.79	-2.88	-1.92	-0.96	0.02	0.99	2.00	2.99	3.94

r is correlation between z and survival time

r' is correlation between z and log of survival time

the entry $\beta = \rho = 0$.

In Figure 1 we display the bias $\hat{\beta} - \beta$ for the Peto-Breslow method, the Efron method, the discrete time model with complementary log-log link as well as our modified Efron method from the same simulations, augmented with some additional values of β and ρ .

Figure 1: Bias of correction methods with 50% censoring

Table 3 and Figure 1 show that the modified Efron correction is practically unbiased for these models except for a small bias for $\rho = \pm 1$ and large absolute values of β . The usual Efron correction is clearly biased for strong correlations between covariate and censoring time. For $\beta = 0$ one might erroneously conclude that there is an effect with this method. However, the bias is rapidly reduced as the correlation between covariate and censoring time decrease. Nevertheless, there is a pronounced bias with as small a correlation as $\rho = \pm 0.4$.

Estimation of β by the discrete time method produces almost the same bias as the conventional Efron correction. The maximum bias of the Peto-Breslow correction is larger than for the Efron correction although there are situations when the bias produced by the Peto-Breslow correction is smaller. This is because the dependence of the bias on β is in the opposite direction to the dependence on ρ , thus cancelling out in some cases.

We have also conducted some simulations with 90% censoring. The bias with the modified Efron correction was again very small. The maximum bias with the conventional Efron correction was a little larger than for 50% censoring and approximately of the same magnitude as the bias with the Peto-Breslow correction.

5 Discussion

We have proposed a modification to the Efron correction for tied survival data that takes into account that censoring times typically do not occur after all event times in intervals of grouped times. We have shown in the application to the sexual debut data that estimates based on the conventional and modified Efron corrections may be considerably different. The modified method worked remarkably well for all situations considered in our simulations, whereas the conventional Efron correction may perform very badly.

The modified Efron correction is as easy to implement as the usual Efron correction. Furthermore it is computationally almost as simple as the traditional Efron correction or the Peto-Breslow method. Thus, we recommend that the modified Efron correction should be implemented in software packages.

In our application and in some of the simulations the censoring time was an exact linear function of the covariate. We expect much smaller bias for the traditional methods when the dependence is weaker, indeed when censoring is independent of covariates the modified method should not be less biased than the traditional Efron method. However, considerable dependence is sometimes expected in practice as reflected in the sex debut application.

We have not derived variance estimators with the modified Efron correction. Instead we followed the usual approach for variance estimation for the conventional Efron correction, basing the variance estimates on the inverse information. Our simulations indicate that this may work well. However, since the regression parameter estimates are solutions to estimating equations it may be prudent to use robust variance estimates akin to Lin & Wei (1989).

Correction for ties has received much attention for Cox-regression. However, tied survival data is a general problem affecting many other methods such as Aalen's estimator of cumulative regression functions under an additive hazard model with time-varying effects (see e.g. Andersen et al., 1993). The basic idea behind both the usual and our modified Efron-correction is just that different orderings of event (and censoring) times are often almost equally likely. It is thus our conjecture that such correction methods may be applied successfully for many methods in survival analysis.

References

- [1] Andersen, PK, Borgan, Ø, Gill, RD, and Keiding, N (1993). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.
- [2] Breslow, NE (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89-99.
- [3] Cox, DR (1972). Regression models and life tables (with Discussion). *J. Roy. Statist. Soc. B* **74**, 187-200.
- [4] DeLong DM, Guirguis, GH and Ying, CS (1994) Efficient computation of subset selection probabilities with application to Cox regression. *Biometrika* **81**, 607-11.
- [5] Efron, B (1977). The efficiency of Cox's likelihood function for censored data. *J. Am. Statist. Assoc.* **72**, 557-65.
- [6] Kalbfleisch, JD and Prentice, RL (1973) Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**, 267-278.
- [7] Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical analysis of failure time data*. Wiley, New York.
- [8] Lin, D.Y and Wei, L.J. (1989) Robust inference for the the Cox proportional hazards model. *J. Am. Statist. Assoc.* **84**, 1074-1078.
- [9] Oakes, D (1981) Survival analysis: aspects of partial likelihood (with Discussion). *Int. Statist. Rev.* **49**, 235-64.
- [10] Peto, R (1972) Discussion of the paper by DR Cox.
- [11] Samuelsen, SO (2003) Mimicking Cox-regression. *Communications in Statistics. Theory and Methods* **32**, 151-69.
- [12] Satten, GA (1996) Rank-based inference in the proportional hazard model for interval censored data. *Biometrika* **83**, 355-370.
- [13] Satten, GA, Datta, S, Williamson, JM (1998) Infernce based on imputed failure times for the proportional hazard model with interval censored data. *J. Amer. Statist. Assoc.* **93**, 318-327.
- [14] Sinha, D, Tanner, MA and Hall, WJ (1994) Maximization of the marginal likelihood of grouped survival data. *Biometrika* **81** 53-60.
- [15] Sundet, JM, Magnus, P, Kvale IL, Samuelsen, SO and Bakketeig, LS (1992) Secular trends and socioeconomic regularities of coital debut age in Norway. *Archives of Sexual Behavior* **22**, 241-252.