of the survival time of an individual as $\lambda(t|z) = \lambda_0(t)\,\mathrm{e\ \ p}(\beta_0 z)$, where z is a covariate, λ_0 is an un nown baseline hazard function and β_0 is a regression parameter. (For notational simplicity, we assume that the covariate is one-dimensional and non-time dependent). Grouped data in this setting are occurrence/e posure data for cells determined by time intervals and covariate strata, see, e.g., Breslow (1986), Preston et al. (1987) and Selmer (1990).

Our main result, stated in Section 2, shows how grouping disturbs the asymptotic behavior of the ma imum partial li elihood estimator of β_0. n estimator of the Sheppard correction is provided in Section 3, and its performance is assessed through a simulation study in Section 4. The proof of the main result is given in Section 5.

2 Correction for grouping

Let (X, C, Z) be random variables such that the survival time X and the censoring time C are conditionally independent given the covariate Z. Denote $\delta = 1_{\{X \leq C\}}$ and $T = X \wedge C$. The ungrouped data consist of n independent replicates (T_i, δ_i, Z_i) of (T, δ, Z). Co 's ma imum partial li elihood estimator $\hat{\beta}$ is obtained by ma imizing

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta Z_i}}{\sum_{k \in \mathcal{R}_i} e^{\beta Z_k}} \right\}^{\delta_i}$$

where $\mathcal{R}_i$ is the set of individuals observed to be at ris at time T_i-. Under suitable regularity conditions (see ndersen and Gill, 1982), $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, V)$, where V^{-1} is consistently estimated by $-n^{-1}\ \ U(\hat{\beta})/\ \ \beta$ and U is the partial li elihood score function $U(\beta) = \ \ \log L(\beta)/\ \ \beta$.

The grouped data based estimator $\hat{\beta}_g$ is obtained by ma imizing the following appro imation to the partial li elihood:

$$L_g(\beta) = \prod_{r,j} \left\{ \frac{e^{\beta z_j}}{\sum_k Y_{rk} e^{\beta z_k}} \right\}^{N_{rj}}$$

where the product is over the grouping cells, the sum is over the covariate strata, and z_j is the midpoint of the jth covariate stratum. Here Y_{rj} and N_{rj} are, respectively, the total time at ris (e posure) and the number of observed failures (occurrence) in the rjth grouping cell $\mathcal{C}_{rj} = \ \ _r \times \mathcal{I}_j$. We assume that the time intervals } $\sqrt{}$} | | |ı| {} $\sqrt{}$}

where the double integral is over the region covered by the cells used in grouping the data, $\bar{z}(\beta, t) = s^{(1)}(\beta, t)/s^{(0)}(\beta, t)$, $Y(t) = 1_{\{T \geq t\}}$ and $(t, z) = P(T \geq t, Z \leq z)$. Here $s^{(k)}(\beta, t) = E\{Y(t)Z^k e^{\beta Z}\}$, and $\dot{}$, $'$ denote the partial derivatives of with respect to t and z, respectively. The various derivatives implicit in Δ are assumed to e ist and to be continuous. Two mild conditions, (C1) and (C2) in Section 5, are also assumed to hold.

25Dfl(()Tjfl/T8fflTffl500TDfl(to)Tjfl130TDflfl(()Tjfl/T8020TDfl(hold0Ti228.9ttI03())T-2192fl2230TDfl(c and

$$\psi(z) = \int_0^1 \{z - \bar{z}(\beta_0, t)\} e^{\beta_0 z} \lambda_0(t) \quad '(t, z)\, dt.$$

It follows from the e pression for Δ_1 that if there is only minor variation in the baseline hazard λ_0 over the follow-up period, then a correction for grouping in the time domain would not be necessary. Use Holford's (1976) grouped data based estimator of λ_0:

$$\hat{\lambda}_0(t) = \frac{\sum_j N_{rj}}{\sum_j Y_{rj} e^{\hat{\beta}_g z_j}} \quad \text{for } t \in \quad {}_r.$$

We recommend inspection of a plot of $\hat{\lambda}_0$ to assess the variation in λ_0 over the follow-up period.

grouped data based estimator of $s^{(k)}(\beta, t)$ is given by $S_g^{(k)}(\beta, t) = n^{-1} \sum_j z_j^k Y_{rj} e^{\beta z_j}$ at $t \in \quad {}_r$, see Lemma 5.1(ii). We may estimate $\quad '(t, z)$, at $(t, z) \in \mathcal{C}_{rj}$, by $Y_{rj}/(nwl)$. These estimators can be plugged into Δ_1 and ψ, replacing each integral by a sum of terms, where for Δ_1 the terms involve the increment in $\hat{\lambda}_0^2$ from one time interval $\quad {}_r$ to the ne t. The last term in Δ_2 is consistently estimated by $\int_0^1 S_g^{(0)}(\beta_g, t)\lambda_0(t)\, dt$. consistent grouped data based estimator of V^{-1} is given by $\hat{V}$g

esTheTjfl1760TD61Tffl

columns of Table 1). lthough the effect of the grouping in this e ample is modest—less than half a standard error—the Sheppard correction is e pected to continue to perform adequately in cases where the bias is more pronounced.

Table 1: Monte Carlo estimates of the mean Sheppard correction and the the (normalized) mean difference between $\hat{\beta}$ and $\hat{\beta}_g$; observedand$\hat{\beta}$ti

We shall e amine the various terms in (5.1) through a series of lemmas.

dopting the notation of G, let $S^{(k)}(\beta,t) = n^{-1}\sum_{i=1}^{n} Z_i^k Y_i($

Lemma 5.3 $A = \{U(\beta_0) - U_g(\beta_0)\}/n = \Delta V^{-1} + \quad {}_P\{l^3 + w^3 + (l + w + c_n)n^{-1/2}\}$.

Proof In terms of the martingales $M_i(t) = N_i(t) - \int_0^t Y_i(\ )\lambda_0(\ )e^{\beta_0 Z_i}\, d\ $ and $\bar{M} = \sum_{i=1}^n M_i$ we write A as

$$\frac{1}{n}\sum_{i,j}\int_0^1 (Z_i - z_j)\, 1_{\{Z_i \in \mathcal{I}_j\}} dM_i(\ ) \tag{5.2}$$

$$+\frac{1}{n}\int_0^1 \left\{ \frac{S_g^{(1)}(\beta_0,\ )}{S_g^{(0)}(\beta_0,\ )} - \frac{S^{(1)}(\beta_0,\ )}{S^{(0)}(\beta_0,\ )} \right\} d\bar{M}(\ ) \tag{5.3}$$

$$-\frac{1}{n}\sum_{r,i,j}\int_{\ r} z_j e^{\beta_0 Z_i} 1_{\{Z_i \in \mathcal{I}_j\}} Y_i(\ )\lambda_0(\ )\, d \tag{5.4}$$

$$+\frac{1}{n}\sum_r \frac{S_g^{(1)}(\beta_0, t_r)}{S_g^{(0)}(\beta_0, t_r)} \sum_{i,j}\int_{\ r} e^{\beta_0 Z_i} 1_{\{Z_i \in \mathcal{I}_j\}} Y_i(\ )\lambda_0(\ )\, d\ \ , \tag{5.5}$$

where t_r is the

There is

References

ndersen, P. K. and Gill, R. D. (1982). Co 's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

Bic el, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Statist.* **42** 1656–1670.

Breslow, N. E. (1986). Cohort analysis in epidemiology, in . C. t inson and S. E. Fienberg, eds., *A C l bration of Statistics: th ISI C nt nary Volum* . Springer-Verlag, New Yor , 109–143.

Co , D. R. (1972). Regression models and life tables (with discussion). *. Roy. Statist. Soc. B.* **34** 187–220.

Dempster, . P. and Rubin, D. B. (1983). Rounding error in regression: The appropriateness of Sheppard's corrections. *. Roy. Statist. Soc. B.* **45** 51–59.

Don, F. J. H. (1981). note on Sheppard's corrections for grouping and ma imum li elihood estimation. *. Mult. Anal.* **11** 452–458.

Hahn, M. G. (1978). Central limit theorems in $D[0,1]$. *Z. Wahrsch. V rw. G bi t* **44** 89–102.

Haitovs y, Y. (1973). *R gr ssion stimation from group d obs rvations.* Hafner Press, New Yor .

Haitovs y, Y. (1983). Grouped data. *Encyclop dia of Statistical Sci nc s* **3** 527–536 (1983), Eds. N.L. Johnson and S. Kotz, John Wiley, New Yor .

Hoem, J. M. (1987). Statistical analysis of a multiplicative model and its application to the standardization of vital rates: a review. *Int. Statist. R v.* **55** 119–152.

Holford, T. R. (1976). Life tables with concomitant information. *Biom trics* **32** 587–597.

Huet, S. and Kaddour, . (1994). Ma imum li elihood estimation in survival analysis with grouped data on censored individuals and continuous data on failures. *Appl. Statist.* **43** 325–333.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal li elihoods based on Co 's regression and life model. *Biom trika* **60** 267–278.

Kolassa, J. E. and McCullagh, P. (1990). Edgeworth series for lattice distributions. *Ann. Statist.* **18** 981–985.

Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *. Am r. Statist. Assoc.* **76** 231–240.

Lindley, D. V. (1950). Grouping corrections and ma imum li elihood equations. *Proc. Camb. Phil. Soc.* **46** 106–110.

Prentice R. L. and Gloec ler L. . (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biom trics* **34** 57–67.

Preston, D. L., Kato, H., Kopec y, K. J., and Fujita, M. S. (1987). Life Span Study Report 10, Part I, Cancer Mortality among -Bomb Survivors in Hiroshima and Nagasa i, 1950–82. *Radiat. R s.* **111** 151–178.

Selmer, R. (1990). comparison of Poisson regression models fitted to multiway summary tables and Co 's survival model using data from a blood pressure screening in the city of Bergen, Norway. *Statistics in M dicin* **9** 1157–1165.