UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in STK4080 — Survival and event history analysis.

Day of examination: Friday December 10th 2010.

Examination hours: 14.30 – 18.30.

This problem set consists of 3 pages.

Appendices: None.

Permitted aids: Approved calculator.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

(a) Let $\alpha(t)$ be the hazard and $S(t)$ the survival function of a survival time T. Show the relationship $A(t) = \int_0^t \alpha(s)ds = -\log(S(t))$.

(b) Assume that $(\tilde{T}_i, D_i)$ are right censored survival data, i.e. $\tilde{T}_i$ is the observed length of follow-up and D_i the indicator of an event, where the uncensored survival times are independent and identically distributed with hazard $\alpha(t)$ and where the censoring times $c_1, \ldots, c_n$ are fixed values. Argue that the likelihood can be written

$$L = \prod_{i=1}^{n} \alpha(\tilde{T}_i)^{D_i} \exp(-A(\tilde{T}_i)).$$

Non-parametric estimation of the survival function $S(t)$ with right censored data is usually carried out by the Kaplan-Meier estimator

$$\hat{S}(t) = \prod_{s \leq t} [1 - \frac{dN(s)}{Y(s)}]$$

where $N(t)$ counts the number of observed events in $[0, t]$ and $Y(t)$ is the number at risk at time t. We then have the result that

$$\frac{\hat{S}(t)}{S^\star(t)} - 1 = -\int_0^t \frac{\hat{S}(s-)}{S^\star(s)} \frac{J(s)}{Y(s)} dM(s)$$

where $M(t) = N(t) - \int_0^t Y(s)\alpha(s)ds$, $S^\star(t) = \exp(-\int_0^t J(s)\alpha(s)ds)$ and $J(s) = I(Y(s) > 0)$ (You are not asked to show this result).

(Continued on page 2.)

(c) Explain why this result means that the Kaplan-Meier estimator is essentially unbiased and that $\frac{\hat{S}(t)}{S^\star(t)}$ has variance

$$\mathrm{Var}\left[\frac{\hat{S}(t)}{S^\star(t)}\right] = \mathrm{E}\left[\int_0^t \left(\frac{\hat{S}(s-)}{S^\star(s)}\right)^2 \frac{J(s)\alpha(s)ds}{Y(s)}\right]$$

where $J(s)/Y(s) = 0$ if $Y(s) = 0$. Suggest an estimator for the variance of the Kaplan-Meier estimator.

(d) Sketch a plot of $\hat{S}(t)$ with 95% confidence intervals of $S(t)$ and explain how you from such a plot may find estimates of percentiles with 95% confidence intervals.

Problem 2

Assume that the hazard of individual i is given by a proportional hazards model $\alpha_i(t) = \exp(\beta x_i)\alpha_0(t)$ where the x_i (for notational convenience) are one-dimensional covariates. The Cox-likelihood for β is then

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta x_i)}{\sum_{k\in\mathcal{R}(t_i)} \exp(\beta x_k)}$$

where the t_i are the D observed event times and $\mathcal{R}(t_i)$ is the risk set at time t_i (individuals are ordered so that the first D individuals experience the event and the rest are censored).

(a) Give an interpretation of the term

$$\frac{\exp(\beta x_i)}{\sum_{k\in\mathcal{R}(t_i)} \exp(\beta x_k)}$$

and explain why Cox-regression allows for estimating β with an arbitrary baseline hazard $\alpha_0(t)$.

In counting process notation we may write the score function $U(\beta) = \frac{\partial \log(L(\beta))}{\partial\beta}$ as

$$U(\beta) = \sum_{i=1}^{n} \int [x_i - \frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}]dN_i(t)$$

(you are not asked to derive this) where $N_i(t)$ is the indicator of an observed event for individual i before or at time t, $Y_i(t)$ the indicator of being at risk at time t,

$$S^{(0)}(\beta,t) = \sum_{k=1}^{n} Y_k(t)\exp(\beta x_k)$$

and

$$S^{(1)}(\beta,t) = \sum_{k=1}^{n} x_k Y_k(t)\exp(\beta x_k).$$

(Continued on page 3.)

(b) Show that

$$\sum_{i=1}^{n}[x_i - \frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}]Y_i(t)\alpha_i(t) = 0$$

and use this to derive

$$U(\beta) = \sum_{i=1}^{n}\int [x_i - \frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}]dM_i(t)$$

where $M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha_i(s)ds$.

(c) Use question (b) to show that $\mathrm{E}[U(\beta)] = 0$ and sketch a derivation of $\mathrm{Var}(U(\beta))$.

(d) With covariates x_i that can only take two values, 0 and 1, Cox-regression may also be used to test whether there is a difference between two groups and can in fact be considered equivalent to the log-rank test. This result follows from the identity

$$U(0) = N_{\bullet 1}(\tau) - \int Y_{\bullet 1}(t)\frac{dN_{\bullet 1}(t) + dN_{\bullet 0}(t)}{Y_{\bullet 1}(t) + Y_{\bullet 0}(t)}$$

where the left-hand side is the score function $U(\beta)$ evaluated at $\beta = 0$ and the right-hand side is (a version) of the log-rank test-statistic with $N_{\bullet j}(t)$ counting the number of events, $Y_{\bullet j}(t)$ is the number at risk in the group with $x_i = j$ and τ the largest event time. Show this identity.

Hint: First show that $S^{(0)}(0,t) = Y_{\bullet 1}(t)+Y_{\bullet 0}(t)$ and $S^{(1)}(0,t) = Y_{\bullet 1}(t)$.

Problem 3

A study of mortality among intravenous drugs users in Oslo was carried out in 1986-1991 with a focus on investigating whether the mortality among HIV-positive drug users was higher than among HIV-negative.

(a) In the table below it is reported, based on a standard Cox-regression model, the estimated regression coefficients $\hat{\beta}_j$ with standard errors se_j of covariates x_{i1} = indicator of being HIV-positive, x_{i2} = indicator of being a woman and x_{i3} = indicator of age above 29 years at inclusion in study. Calculate and interpret the estimated hazard ratios $\widehat{HR}_j$.

Calculate also 95% confidence intervals for the (true) hazard ratios and conclude about whether differences are significant.

Covariate	$\hat{\beta}_j$	se_j
HIV-positive (x_{i1})	0.79	0.22
Women (x_{i2})	0.02	0.22
Age $>$ 29 (x_{i3})	0.74	0.23

(b) Discuss other regression methods that could have been used for this data set.

END