Statistics 4100/5100 - Winter 2003 Survival Analysis
Course Outline

Instructor:	Dr. E. Susko
Time:	TR 8:30-10:30
Place:	Chase 227
Office hours:	M 10-12 TR 10-11
Textbook:	The Statistical Analysis of Failure Time Data by J.D. Kalbfleisch & R.L. Prentice

Marking Scheme

Assignments	35%
Midterm	25% (Feb 27, in class)
Final exam	40%

Course Outline

Basic concepts of failure time distributions
Kaplan-Meier estimation
Comparison of survivor curves
Parametric regression methods
Nonparametric regression methods: proportional hazards models
Competing risks
Additional Topics

Text Chapters: 1-4, parts of 6 and 8

There will be no supplemental examination in this course.

STATISTICS 5100 - Survival Analysis - Final Examination W2003

Name: _____

The multiple choice question 7 should be answered on the question sheet, all essay questions should be answered in the examination booklet.

(11) 1. Consider the exponential log-linear regression model for a two sample comparison

$$\log(T) = \alpha + \beta z + W$$

where e^W has an exponential distribution with mean 1 and $Z = 1$ if the individual is from group 1, and $Z = 0$ if the individual was from group 0.

(a) Derive the score statistic for the test of $H_0 : \beta = 0$. Your final expression should depend only on d_1, v_1, d_2, v_2, the numbers of deaths and the total survival ($v_i = \sum_j t_{ij}$) for the two groups. The overall total number of deaths is $d = d_1 + d_2$ and the total survival is $v = v_1 + v_2$.

(b) Suppose that the test statistic from (a) is stored in x. Indicate how you would calculate a p-value. In particular, indicate the distribution you would use and whatever parameters would be required for this calculation.

(16) 2. Consider the partial S-Plus output below. The variable `time` gives the survival times for patients with lymphoma, a cancer of the lymph system. The variable `delta` is 1 if the patient died during the study and 0 otherwise. The variable `wtime` is the waiting time for a bone marrow transplant and score is the Karnofsky score, a score for the fitness of the patient.

```
> hodg.surv <- Surv(time, delta)
> summary(survReg(hodg.surv ~ factor(score) + wtime, dist = "weibull",
          scale = 1))

                    Value Std. Error
   (Intercept)  4.9388496    0.30680
factor(score)1  1.5507769    0.61813
factor(score)2  0.5490331    0.39192
factor(score)3  0.3049390    0.19557
factor(score)4  0.2607690    0.12448
factor(score)5  0.2711884    0.10699
factor(score)6  0.5460322    0.09496
factor(score)7  0.3983366    0.06208
factor(score)8  0.3635853    0.11437
         wtime -0.0000306    0.00655

Weibull distribution
Loglik(model)= -166.5   Loglik(intercept only)= -196.7
        Chisq= 60.36 on 9 degrees of freedom, p= 1.1e-009
```

```
> summary(survReg(hodg.surv ~ factor(score) + wtime, dist = "weibull"))
                  Value Std. Error
   (Intercept) 4.949136    0.36887
factor(score)1 1.558632    0.73745
factor(score)2 0.553885    0.46744
factor(score)3 0.292488    0.23355
factor(score)4 0.259531    0.14834
factor(score)5 0.260362    0.12858
factor(score)6 0.550537    0.11344
factor(score)7 0.411523    0.07551
factor(score)8 0.397298    0.14124
         wtime 0.000831    0.00799
    Log(scale) 0.176086    0.17432
Weibull distribution
Loglik(model)= -166   Loglik(intercept only)= -183.3
        Chisq= 34.59 on 9 degrees of freedom, p= 7e-005

> summary(survReg(hodg.surv ~ wtime, dist = "weibull"))
             Value Std. Error      z         p
(Intercept) 6.1920     0.5851 10.583 3.56e-026
      wtime 0.0129     0.0133  0.971 3.32e-001
 Log(scale) 0.7159     0.1640
Weibull distribution
Loglik(model)= -182.7   Loglik(intercept only)= -183.3
        Chisq= 1.12 on 1 degrees of freedom, p= 0.29

> summary(survReg(hodg.surv ~ factor(score), dist = "weibull"))
               Value Std. Error
   (Intercept) 4.973     0.2923
factor(score)1 1.548     0.7290
factor(score)2 0.550     0.4653
factor(score)3 0.292     0.2330
factor(score)4 0.262     0.1462
factor(score)5 0.265     0.1210
factor(score)6 0.553     0.1111
factor(score)7 0.413     0.0747
factor(score)8 0.397     0.1410
    Log(scale) 0.174     0.1736
Weibull distribution
Loglik(model)= -166   Loglik(intercept only)= -183.3
        Chisq= 34.58 on 8 degrees of freedom, p= 0.000032
```

(a) For the Weibull regression model, what is the likelihood ratio statistic for a test of the null hypothesis that the score variable is unimportant for prediction of survival times? In answering the question, indicate what a covariate vector z would be for a given individual and and state the null hypothesis in terms of the coefficient vector β. Indicate how a p-value would be calculated with an expression involving the likelihood ratio statistic; be sure to give a clear indication of the distribution used for the p-value calculation and any parameters required.

(b) In the full Weibull regression model, what is an appropriate test statistic for the null hypothesis that the coefficient for `wtime` is 0. How would you

calculate a p-value for this test.

(c) Indicate how you would calculate a p-value for a test that the Weibull distribution is a more appropriate distribution for $e^{\sigma W}$ than the exponential. Express the null hypothesis in terms of the parameters of the model and calculate an appropriate test statistic; be sure to give a clear indication of the distribution used for the p-value calculation and any parameters required.

(6) 3. Suppose $T_1, \ldots, T_n$ is a random sample from a distribution with cumulative hazard $\Lambda(t)$. Show that $\Lambda(T_1), \ldots, \Lambda(T_n)$ are independent and identically distributed from the exponential distribution with mean 1.

Explain how you can use the about result and the estimated $\hat{\beta}$ and $\hat{\Lambda}_0(t)$ from a Cox proportional hazards analysis to check the proportional hazards assumption.

(5) 4. You are involved in a study comparing a new and an old treatment for a disease. A researcher claims that a proportional hazards models for the data is inappropriate. Instead the researcher claims that the new treatment may initially benefit the patient with the disease but will eventually increase the risk of death relative to a person taking the old treatment.

Explain how you might use time-dependent covariates to model this with a proportional hazards model. Include in your explanation a definition of a reasonable time dependent covariate and a statement about parameter values that agree with the researchers explanation.

(9) 5. Consider the proportional hazards model

$$\lambda(t; z) = \lambda_0(t) e^{z^T \beta}$$

where

$$\lambda_0(t) = \lambda_j, \qquad a_{j-1} \le t < a_j, \quad j = 1, \ldots, k$$

and $0 = a_0 < a_1 < \cdots < a_k = +\infty$. Let d_j be the number of deaths in $[a_{j-1}, a_j)$. Let R_j be the risk set immediately prior to a_j, C_j the set of individuals censored in $[a_{j-1}, a_j)$ and D_j is the set of individuals failing in $[a_{j-1}, a_j)$. For a given value of β, derive the maximum likelihood estimate of λ_j. Hint: the estimate should be of the form $\hat{\lambda}_j = d_j / S_j$ for some S_j.

(5) 6. Suppose T can take on values only in $\{a_1, a_2, \ldots\}$. Let

$$\lambda_{j|z} = P(T = a_j | T \ge a_j; z)$$

One possibile discrete analogue of the proportional hazards model is the model with

$$F(t; z) = F_0(t)^{\exp[z^T \beta]}$$

where $F_0(t)$ is the survivor for an individual with $z = 0$. Show that in this case

$$\lambda_{j|z} = 1 - [1 - \lambda_{j|0}]^{\exp[z^T \beta]}$$

(16) 7. For each of the questions below give the answer that is most appropriate.

(a) Consider a log-linear model for a single group

$$\log(T) = \beta_0 + W$$

where $\exp(W)$ has an exponential distribution with mean 1. If all of the observations were censored, β_0 would be estimated as

i. _____$-\infty$
ii. _____$+\infty$
iii. _____0
iv. _____need more information to know

(b) We fit a log-linear model

$$\log(T) = \alpha + z^T\beta + W$$

where $\exp(W)$ has an exponential distribution with mean 1. Values $\hat{t}_i = t_i \exp(-z_i^T\beta)$ are then calculated. If the exponential distribution is an appropriate distribution for $\exp(W)$ a plot of the Kaplan-Meier estimate of $\hat{t}_i$ against the exponential distribution should look like

i. _____a line with slope approximately equal to 1 and approximate 0 intercept.
ii. _____a line with slope possibly very different from 1 but 0 intercept.
iii. _____a line with slope approximately equal to 1 but with an intercept possibly very different from 0.
iv. _____a flower.

(c) Recall that the Weibull survivor function is $F(t) = \exp[-(\lambda t)^\gamma]$ and the logistic survivor is $F(t) = [1 + (\lambda t)^\gamma]^{-1}$. If a plot of $\log[-\log[\hat{F}(t)]]$ against t was approximately linear with slope 1, where $\hat{F}$ is the Kaplan-Meier estimate,

i. _____the Weibull distribution would be an appropriate model for the data but not the exponential distribution.
ii. _____the logistic distribution would be appropriate model for the data.
iii. _____the exponential distribution would likely be appropriate.
iv. _____none of the above.

(d) If an exponential log-linear model was appropriate and came up with an estimated coefficient of -2 we would expect the coefficient from `coxph` to be

i. _____approximately equal to -2.
ii. _____approximately equal to 2.
iii. _____might not be close to either of these values.

(e) If the Kaplan-Meier estimate $\hat{F}(t) = 0.65$ and $\text{se}[\hat{F}(t)] = 0.03$ an estimate of $\text{se}[\log(\hat{F}(t))]$ is

i. _____21.67
ii. _____0.046
iii. _____0.071
iv. _____0.023

(f) If the Kaplan-Meier estimate $\hat{F}(t) = 0.65$ and $\text{se}[\hat{F}(t)] = 0.03$ and a 95% confidence interval for $F(t)$ is $[0.5912, 0.7088]$, then an approximate 95% confidence interval for $\hat{F}(t)^{-1/3}$ is

i. _____$[1.122, 1.191]$
ii. _____$[1.012, 1.170]$
iii. _____$[1.413, 1.223]$
iv. _____all of the above
v. _____none of the above

(g) We can fit a competing risks model with a proportional hazards specification for the type-specific hazard: $\lambda_j(t; z) = e^{z^T\beta}\lambda_0(t)$, by using `coxph` and

i. _____successively fitting models in which deaths correspond to specific types and all other types of death are treated as censored observations.
ii. _____stratifying for the type-of-death variable.
iii. _____considering a model in which there are interactions between the type-of-death variable and every other variable.

(h) Suppose that for a discrete failure time T, the model

$$\text{logit}(\lambda_{j|z}) = \text{logit}(\lambda_{j|0}) + z^T\beta$$

Here $\text{logit}(u) = \log[u/(1-u)]$. Suppose the above was actually a discretization of a continuous random variable T^*. Specifically, suppose that $T = a_j$ whenever $T^* \in (a_j, a_{j+1}]$. Consider now what would occur as as the discrete time points get closer and closer to each other. The model above implies

i. _____a proportional hazards model for T^*.
ii. _____an accelerated life model for T^*.
iii. _____none of the above.