Final Project (Please type your report and hand in before **March 20, Friday**)

Consider the Veterans' administration lung cancer data. In the trial, male patients with advance inoperable lung cancer were randomized to either a standard or test chemotherapy. The primary endpoint of the the therapy comparison was time to death. Only 9 of the 137 survival times were censored. The data consist of variables

```
trt: 1=standard 2=test
celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
time: survival time
status: censoring status
karno: Karnofsky performance score (100=good)
diagtime: months from diagnosis to randomisation
age: in years
prior: prior therapy 0=no, 1=yes
```

The complete data set is available in R. You can access it using the R commands

```
library(survival)
veteran
?veteran
```

1. Conduct the logrank test for the treatment effect of new-therapy and report your findings (hint: use the R function "survdiff").

2. Estimate the hazard ratio of the test versus the standard chemotherapy and construct the associated confidence interval based on the Cox regression model. Report your findings (hint: use the R function "coxph").

3. You may also fit a parametric Weibull regression to estimate the hazard ratio of the test versus the standard chemotherapy and the associated confidence interval. Is the estimate similar to that obtained under Cox regression? Report your findings (hint: use the R function "survreg").

4. It is known that the baseline Karnofsky performance score oftentimes is associated with the survival time. Estimate the hazard ratio of the test versus the standard chemotherapy but adjusting for the baseline "Karnofsky performance score ", using the multivariate Cox regression model. Report your findings.

5. Given the results in testing the treatment effect, the clinical investigator decided to develop a prognostic regression model for predicting the survival time using the baseline "Karnofsky performance score" only. To this end, one may build a Cox regression model with the dichotomized baseline performance score as a single covariate with all 137 patients (pooled from both arms) to predict the survival probability of future patients.

(a) Assuming a Cox regression model

$$h(t|Z) = h_0(t)e^{\beta_0 Z}$$

with $Z = I(\text{score} > 60)$, the survival function for a patient with covaraite $Z = z_0$ is

$$\hat{S}(t|Z = z_0) = \exp\left\{-\hat{H}_0(t)e^{\hat{\beta}z_0}\right\},$$

where $H_0(t)$ is the Breslow estimator for the baseline cumulative hazard function and $\hat{\beta}$ is the maximum partial likelihood function. Using this fact, obtain and plot the estimated survival functions for two patients: patient A with a baseline performance score of 70 and patient B with a baseline performance score of 50.

(b) When $z_0 = 0$, i.e., the baseline performance score is ≤ 60, $\hat{S}(t|Z = 0) = \exp\{-\hat{H}_0(t)\}$. In this case, one can show that

$$\sqrt{n}\left\{\hat{S}(t|Z = 0) - S(t|Z = 0)\right\}$$
$$= \frac{e^{-H_0(t)}}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \left[\int_0^t \frac{S^{(1)}(\beta_0, s)}{S^{(0)}(\beta_0, s)^2} dN(s)\right]' \left[n^{-1}I(\beta_0)\right]^{-1} \int_0^\tau \left[Z_i - \frac{S^{(1)}(\beta_0, s)}{S^{(0)}(\beta_0, s)}\right] dM_i(s) - \int_0^t \frac{n}{S^{(0)}(\beta_0, s)} dM_i(s) \right\} + o_p(1),$$

where $N(s) = \sum_{i=1}^n N_i(s)$, $I(\beta)$ is the second derivative of

$$-\log\{PL(\beta)\} = \sum_{i=1}^{n} \delta_i \left[\beta Z_i - \log\{S^{(0)}(\beta, U_i)\}\right],$$

$M_i(s) = N_i(s) - \int_0^s I(U_i \geq x)e^{\beta_0 Z_i}h_0(x)dx$ and

$$S^{(j)}(\beta, s) = \sum_{i=1}^{n} I(U_i \geq s)e^{\beta Z_i}Z_i^j.$$

Here we assume that the observed data consist of $\{(U_i, \delta_i, Z_i), i = 1, \cdots, n\}$. Use this expansion to design a resampling procedure for constructing the 95% confidence interval for $S(t|Z = 0)$. Describe your procedure and construct the 95% confidence interval for $S(100|Z = 0)$ based on veteran data.

(c) How can you construct a 95% confidence interval for the corresponding median survival time for patient B? Construct the 95% confidence interval using the veteran data. (This is a bonus question.)

(d) Since Z only takes value of 0 or 1, one may simply compute the KM estimator based on all patients whose baseline performance score is > 60 to estimate the survival function for patient A. Similarly, one may use the KM estimator based on all patients whose baseline performance score is ≤ 60 to estimate the survival function of patient B. Will these estimates be different from those based on the Cox regression model? Why?