Math 473 Exam 2 Spring 2021 Name_____

```
       variables              AIC
treat age serum size gleas   32.17
treat age       size gleas   30.17
treat           size gleas   28.57
                size gleas   27.53
                size         31.04
         coef exp(coef) se(coef)     z     p
treat -1.1127    0.3287   1.2031 -0.92 0.355
size   0.0826    1.0861   0.0475  1.74 0.082
gleas  0.7102    2.0345   0.3379  2.10 0.036

Likelihood ratio test=13.8  on 3 df, p=0.00323
n= 38, number of events= 6
```

1) Data is from Collett (2003, p. 10) regarding survival of prostate cancer patients. The predictors are *treat* (1 if placebo, 2 if DES), *age* = patient age, *serum* (a prognastic variable), *size* of the primary tumour, and *gleas* (Gleason index: the more advanced the tumour, the higher the index). Results from backward elimination are shown.

a) What is the best starting submodel I_I?

b) Are there any other candidate submodels? Explain briefly.

c) Consider the model containing x_1 = *treat*, x_2 = *size* and x_3 = *gleas*. Find the $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ if $x_1 = 2$, $x_2 = 20$, and $x_3 = 10$.

	M1	M2	M3	M4
# of predictors	7	4	2	1
# with $0.01 \leq$ p-value ≤ 0.05	1	1	1	0
# with p-value > 0.05	6	3	0	0
$-2\log(L)$	175.22	175.63	177.52	181.12
$AIC(I)$	189.22	183.63	181.52	183.12
p-value for change in PLR test	1.0	0.936	0.806	0.434

2) The above table gives summary statistics for 4 PH regression models considered as final submodels after performing variable selection. Assume that the PH assumptions hold for all 4 models. The full model was M1. Which model should be considered as the first starting submodel I_I? Explain briefly why each of the other 3 submodels should not be used as the starting submodel.

```
Full model
 variable    coef  std._err.  z     pval
      age   -0.029 0.008     -3.53  0.000
  bectota    0.008 0.005      1.68  0.094
  ndrugtx    0.028 0.008      3.42  0.001
  herco_2    0.065 0.150      0.44  0.663
  herco_3   -0.094 0.166     -0.57  0.572
  herco_4    0.028 0.160      0.18  0.861
   ivhx_2    0.174 0.139      1.26  0.208
   ivhx_3    0.281 0.147      1.91  0.056
     race   -0.203 0.117     -1.74  0.082
    treat   -0.240 0.094     -2.54  0.011
     site   -0.102 0.109     -0.94  0.348

Likelihood ratio test = 24.436 on 11 df, p = 0.011

Reduced model
 variable    coef  std._err.  z     pval
      age   -0.026 0.008     -3.25  0.001
  bectota    0.008 0.005      1.70  0.090
  ndrugtx    0.029 0.008      3.54  0.000
   ivhx_3    0.256 0.106      2.41  0.016
     race   -0.224 0.115     -1.95  0.051
    treat   -0.232 0.093     -2.48  0.013
     site   -0.087 0.108     -0.80  0.422

Likelihood ratio test = 21.038 on 7 df, p = 0.004
```

3) The data studies time until illegal drug use relapse. Variables were *age*, *becktota*, *ndrugtx*, $herco_2 = 1$ if heroin user and 0 else, $herco_3 = 1$ if cocaine user and 0 else, $herco_4 = 1$ if used neither heroin nor cocaine and 0 else, $ivhx_2 = 1$ if previous but not recent IV drug use and 0 else, $ivhx_3 = 1$ if recent IV drug use and 0 else, $race = 1$ for white and 0 else, $treat = 1$ for short treatment and 0 for long and *site*.

Using the output for the full and reduced model above, test whether the reduced model is good.

```
                coef exp(coef) se(coef)     z      p
X1           -0.0148    0.9853   0.0470 -0.32 0.7526
X2            0.0516    1.0530   0.0394  1.31 0.1905
factor(X3)2   0.2197    1.2457   0.7489  0.29 0.7692
factor(X4)3 -2.4437    0.0868   0.9606 -2.54 0.0110
X5           -0.2436    0.7838   0.8859 -0.27 0.7833
X6           -0.0579    0.9437   0.1612 -0.36 0.7193
X7            2.6127   13.6357   0.9176  2.85 0.0044

Likelihood ratio test=22.3  on 7 df, p=0.00223
n= 37, number of events= 17
```

4) The above output is for the Collett (2003, p. 367) data for leukemia patients who received a bone marrow transplant. Y is the survival time in days, and the predictors are $x_1 = page$ = age of patient, $x_2 = dage$ = age of donor, factor *type* of leukemia where the three types were coded as indicator variables x_3 and x_4, $x_5 = preg$ = donor pregnancy (0 for no, 1 for yes), $x_6 = index$, and $x_7 = Gvhd$ = graft-versus host-disease (0–no,1–yes) which can cause transplanted cells to attack host cells (potentially fatally). A PH model was used.

a) Test $\boldsymbol{\beta} = \mathbf{0}$.

b) Find a 95% CI for β_7.

c) Do a 4 step test for $Ho : \beta_7 = 0$.

```
        coef exp(coef) se(coef)     z Pr(>|z|)
treat 0.5964    1.8155   0.5870 1.016     0.31

Likelihood ratio test= 1.05  on 1 df,   p=0.3
```

5) The data from Collett (2003, p. 335-6) gives censored survival times for patients with ovarian cancer. The indicator variable for treatment $x = 0$ treatment is type 0 (cyclophosphamide and adriamycin) and $x = 1$ if treatment = 1 (cyclophosphamide).

a) Do a PLR test for $\beta = 0$.

b) The solid line in each plot is the estimated PH survival function. There are 3 bands of vertical circles. The middle circles correspond to the Kaplan Meier estimator. The outer circles are pointwise confidence interval bands for the Kaplan Meier estimator. The CI for $S(0)$ is [1,1]. For group 0, the CI for S(1200) is approximately [0.35,0.95]. Is the proportional hazards model reasonable? Explain briefly.

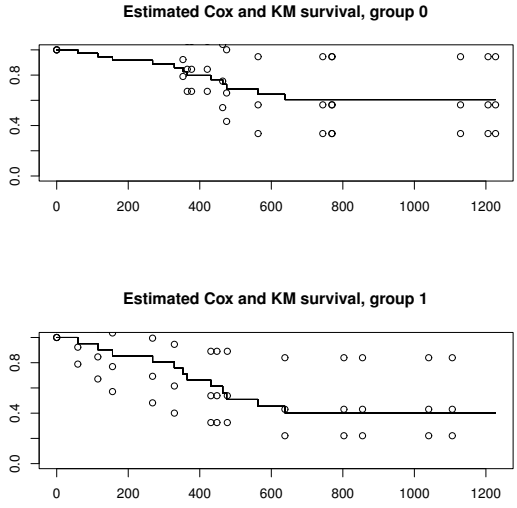

Figure 1: