Survival Data Analysis & Lab.
Assignment #3 SOLUTION

---

1. In a 10-year follow-up study conducted in Evans County, Georgia, involving persons 60 years or older, one research question concerned evaluating the relationship of social support to mortality status. A Cox proportional hazards model was fit to describe the relationship of a measure of social network to time until death. The social network index was denoted as SNI, and took on integer values between 0 (poor social network) to 5 (excellent social network). Variables to be considered for control in the analysis as either potential confounders or potential effect modifiers were AGE (treated continuously), Race (0,1), and SEX (0,1).

   (a) State an initial PH model that can be used to assess the relationship of interest, which considers the potential confounding and interaction effects of the AGE, RACE, and SEX (assume no higher than two-factor products involving SNI with AGE, RACE, and SEX).
   $h(t, \mathbf{X}) = h_0(t)\exp[\beta_1\text{SNI} + \beta_2\text{AGE} + \beta_3\text{RACE} + \beta_4\text{SEX} + \beta_5\text{SNI}\times\text{AGE} + \beta_6\text{SNI}\times\text{RACE} + \beta_7\text{SNI}\times\text{SEX}]$.

   (b) For your model in part 1(a), give an expression for the hazard ratio that compares a person with SNI=4 to a person with SNI=2 and the same values of the covariates being controlled.
   $HR = \exp[2\beta_1 + 2(\text{AGE})\beta_5 + 2(\text{RACE})\beta_6 + 2(\text{SEX})\beta_7]$

   (c) Describe how you would test for interaction using your model in part 1(a). In particular, state the null hypothesis, the general form of your test statistic, with its distribution and degrees of freedom under the null hypothesis.
   $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$.
   Likelihood ratio test statistics: $-2\ln L_R - (-2\ln L_F)$, which is approximately χ^2_3 under H_0, where R denotes the reduced model (containing no product terms) under H_0, and F denotes the full model (given in part 1(a) above).

   (d) Assuming a revised model containing no interaction terms, give an expression for a 95% interval estimate for the adjusted hazard ratio comparing a person with SNI=4 to a person with SNI=2 and the same values of the covariates in your model.
   95% CI for adjusted HR: $\exp\left[2\hat{\beta}_1 \pm 1.96 \times 2\sqrt{\text{var}(\hat{\beta}_1)}\right]$

   (e) For the no-interaction model described in part 1(d), give an expression (i.e., formula) for the estimated survival curve for a person with SNI=4, adjusted for AGE, RACE, and SEX, where the adjustment uses the overall mean value for each of the three covariates.
   $\hat{S}(t, \mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp[4\hat{\beta}_1 + \overline{(\text{AGE})}]\beta_2 + \overline{(\text{RACE})}\hat{\beta}_3 + \overline{(\text{SEX})}\hat{\beta}_4]}$

   (f) Using the no-interaction model described in part 1(d), if the estimated survival curves for persons with SNI=4 and SNI=2 adjusted for (mean) AGE, RACE, and SEX are plotted over time, will these two estimated survival curves cross? Explain briefly.
   The two survival curves will **not** cross, because both are computed using the same proportional hazards model, which has the property that the hazard functions, as well as their corresponding estimated survivor functions, will not cross.

2. For this question, we consider the survival data for 137 patients from the Veteran's Administration Lung Cancer Trial. The variables in this dataset are listed as follows:

Variable#	Variable name	Coding
1	Treatment	Standard=1, test=2
2	Cell type 1	Large=1, other=0
3	Cell type 2	Adeno=1, other=0
4	Cell type 3	Small=1, other=0
5	Cell type 4	Squamous=1, other=0
6	Survival time	(Days) integer counts
7	Performance status	0=worst, . . ., 100=best
8	Disease duration	(Months) integer counts
9	Age	(Years) integer counts
10	Prior therapy	None=0, some=10
11	Status	0=censored, 1=died

For these data, a Cox PH model was fitted yielding the following edited computer results:

```
Response:  survival time
```

Variable name	coef	se(coef)	Pr(>|z|)	exp(coef)	lower .95	upper .95
1 Treatment	0.290	0.207	0.162	1.336	0.890	2.006
3 Adeno cell	0.789	0.303	0.009	2.200	1.216	3.982
4 Small cell	0.457	0.266	0.086	1.579	0.937	2.661
5 Squamous cell	−0.400	0.283	0.157	0.671	0.385	1.167
7 Perf. status	−0.033	0.006	0.000	0.968	0.958	0.978
8 Disease dur.	0.000	0.009	0.992	1.000	0.982	1.018
9 Age	−0.009	0.009	0.358	0.991	0.974	1.010
10 Prior therapy	0.007	0.023	0.755	1.007	0.962	1.054

```
Likelihood ratio test = 24.920
```

(a) State the Cox PH model used to obtain the above computer results.
$h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10}]$

(b) Using the printout above, what is the hazard ratio that compares persons with adeno cell type with persons with large cell type? Explain your answer using the general hazard ratio formula for the Cox PH model.
Adeno cell type: $\mathbf{X}^* =$ (treatment,1,0,0,perfstat,disdur,age,prther)
Large cell type: $\mathbf{X} =$ (treatment,0,0,0,perfstat,disdur,age,prther)
$HR = \frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} = \exp\left[\sum_{i=1}^{p} \beta_i (X_i^* - X_i)\right] = \exp[0 + \hat{\beta}_3(1-0) + \hat{\beta}_4(0-0) + \hat{\beta}_5(0-0) + 0 + 0 + 0 + 0]$
$= \exp[\hat{\beta}_3] = \exp[0.789] = 2.20$

(c) Using the printout above, what is the hazard ratio that compares persons with adeno cell type with persons with squamous cell type? Explain your answer using the general hazard ratio formula for the Cox PH model.
Adeno cell type: $\mathbf{X}^* =$ (treatment,1,0,0,perfstat,disdur,age,prther)
Squamous cell type: $\mathbf{X} =$ (treatment,0,0,1,perfstat,disdur,age,prther)
$HR = \frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} = \exp\left[\sum_{i=1}^{p} \beta_i (X_i^* - X_i)\right] = \exp[0 + \hat{\beta}_3(1-0) + \hat{\beta}_4(0-0) + \hat{\beta}_5(0-1) + 0 + 0 + 0 + 0]$
$= \exp[\hat{\beta}_3 - \hat{\beta}_5] = \exp[0.789 - (-0.400)] = \exp[1.189] = 3.28$

(d) Based on the computer results, is there an effect of treatment on survival time? Explain briefly.
There does not appear to be an effect of treatment on survival time, adjusted for the other variables in the model. The hazard ratio is 1.3, which is close to the null value of one, the p-value of 0.162 for the Wald test for treatment is not significant, and the 95% confidence interval for the treatment effect correspondingly includes the null value.

(e) Give an expression for the estimated survival curve for a person who was given the test treatment and who had a squamous cell type, where the variables to be adjusted are performance status, disease duration, age,

and prior therapy.
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{\exp[2\hat{\beta}_1+\hat{\beta}_5+\overline{(\text{perfstat})}\hat{\beta}_7+\overline{(\text{disdur})}\hat{\beta}_8+\overline{(\text{age})}\hat{\beta}_9+\overline{(\text{prther})}\hat{\beta}_{10}]}$

(f) Suppose a revised Cox model is used which contains, in addition to the variables already included, the product terms: treatment×performance status; treatment×disease duration; treatment×age; and treatment×prior therapy. For this revised model, give an expression for the hazard ratio for the effect of treatment, adjusted for the other variables in the model.
$HR = \frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} = \exp[\beta_1 + (\text{perfstat})\hat{\beta}_{11} + (\text{disdur})\hat{\beta}_{12} + (\text{age})\hat{\beta}_{13} + (\text{prther})\hat{\beta}_{14}]$ where β_1 is the coefficient of the treatment variable and β_{11}, β_{12}, β_{13}, and β_{14} are the coefficients of product terms involving treatment with the four variables indicated.

3. The data for this question contain survival times of 65 multiple myeloma patients. A partial list of the variables in the dataset is given below:

Variable 1: observation number
Variable 2: survival time (in months) from time of diagnosis
Variable 3: survival status (0=alive, 1=dead)
Variable 4: platelets at diagnosis (0=abnormal, 1=normal)
Variable 5: age at diagnosis (years)
Variable 6: sex (1=male, 2=female)

Below, we provide edited computer results for several different Cox models that were fit to this dataset. A number of questions will be asked about these results.

Model 1:

Variable name	coef	se(coef)	Pr(>|z|)	exp(coef)	lower .95	upper .95
Platelets	0.470	2.854	.869	1.600	0.006	429.689
Age	0.000	0.037	.998	1.000	0.930	1.075
Sex	0.183	0.725	.801	1.200	0.290	4.969
Platelets×age	−0.008	0.041	.850	0.992	0.915	1.075
Platelets×sex	−0.503	0.804	.532	0.605	0.125	2.924

Likelihood ratio test = 3.920

Model 2:

Variable name	coef	se(coef)	Pr(>|z|)	exp(coef)	lower .95	upper .95
Platelets	−0.725	0.401	.071	0.484	0.221	1.063
Age	−0.005	0.016	.740	0.995	0.965	1.026
Sex	−0.221	0.311	.478	0.802	0.436	1.476

Likelihood ratio test = 3.494

Model 3:

Variable name	coef	se(coef)	Pr(>|z|)	exp(coef)	lower .95	upper .95
Platelets	−0.706	0.401	.078	0.493	0.225	1.083
Age	−0.003	0.015	.828	0.997	0.967	1.027

Likelihood ratio test = 2.982

Model 4:

Variable name	coef	se(coef)	Pr(>|z|)	exp(coef)	lower .95	upper .95
Platelets	−0.705	0.397	.076	0.494	0.227	1.075
Sex	−0.204	0.307	.506	0.815	0.447	1.489

Likelihood ratio test = 3.384

Model 5:

Variable name	coef	se(coef)	Pr(>|z|)	exp(coef)	lower .95	upper .95
Platelets	−0.694	0.397	.080	0.500	0.230	1.088
	Likelihood ratio test = 2.934					

(a) For model 1, give an expression for the hazard ratio for the effect of the platelet variable adjusted for age and sex.
$\widehat{HR} = \exp[0.470 + (-0.008)\text{age} + (-0.503)\text{sex}]$

(b) Using your answer to part 3(a), compute the estimated hazard ratio for 40-year-old male. Also compute the estimated hazard ratio for a 50-year-old female.
40-year-old male:
$\widehat{HR} = \exp[0.470 + (-0.008)40 + (-0.503)1] = 0.70$
50-year-old female:
$\widehat{HR} = \exp[0.470 + (-0.008)50 + (-0.503)2] = 0.39$

(c) Carry out an appropriate test of hypothesis to evaluate whether there is any significant interaction in model 1. What is your conclusion?
The LR test for the significance of both interaction terms simultaneously yields the following likelihood ratio statistics which compares model 1 and 2: $LR = 3.920 - 3.494 = 0.426$
This statistic is approximately chi-square with 2 degrees of freedom under the null hypothesis of no interaction. This LR statistic is highly nonsignificant. Thus, we conclude that there is no significant interaction in the model 1.

(d) Considering models 2–5, evaluate whether age and sex need to be controlled as confounders?
The gold-standard hazard ratio is 0.484, which is obtained for model 2. Note that model 2 contains no interaction terms and controls for both covariates of interest. When either age or sex or both are dropped from the model, the hazard ratio (for platelets) does not change appreciably. Therefore, it appears that neither age nor sex need to be controlled for confounding.

(e) Which of the five models do you think is the best model and why?
Models 2–5 are all more or less equivalent, since they all gives essentially the same hazards ratio and confidence interval for the effect of the platelet variable. A political choice for the best model would be the gold-standard model 2, because the critical reviewer can see both age and sex being controlled in model 2. A statistical choice for the best model must be the model 5, because model 2 can be reduced to model 5 using LR test and the statisticians prefer the simplest model.

(f) Based on your answer to part 3(c), summarize the results that describe the effect of the platelet variable on survival adjusted for age and sex.

- The point estimate of the hazard ratio for normal versus abnormal platelet count is $0.484 = 1/2.07$, so that the hazard for an abnormal count is twice that for a normal count.
- There is a borderline significant effect of platelet count on survival adjusted for age and sex ($P = .071$).
- The 95% CI for the hazard ratio is given by $0.221 < HR < 1.063$, which is quite wide and therefore shows a very imprecise estimate.