Biostatistics 209 Homework #2 Solutions

1. [4 points] The follow `Stata` output gives the hazard ratio of a **1-year increase** in age (variable named `agedx`).

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       agedx |   .9144948   .0264917    -3.09   0.002     .8640186    .9679198
------------------------------------------------------------------------------
```

   *(a) [1 point] The Hazard ratio for a **1-year decrease** in age is 1.09=(1/.9145).*

   (b) [2 points] The corresponding Wald test statistic Z=+3.09 reverses sign and the p-value =0.002 is unchanged.

   (c) [1 point] The corresponding 95% CI is (1.03, 1.16)=(1/.968, 1/.864).

2. The variable `agedx` was grouped as follows to create the variable "`ages`"

   agedx: 0-2: ages =1, agedx: 2-6: ages =2, agedx: 6-12: ages =3, agedx: > 12: ages =4

   This leads the following output from a Cox model

```
No. of subjects =          103                     Number of obs   =       103
No. of failures =           74
Time at risk    =  1413.050005
                                                   LR chi2(3)      =     28.55
Log likelihood  =   -266.66906                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agedx  |
        2    |   .2684887   .1685833    -2.09   0.036     .0784258    .9191639
        3    |   .0989741   .0632982    -3.62   0.000      .028258     .346658
        4    |   .0566843   .0391472    -4.16   0.000     .0146422    .2194414
------------------------------------------------------------------------------
```

   (a) [1 point] What is the hazard ratio for someone aged 0-2 (`ages` =1) compared to someone aged >12 (`ages` =4)? *i.e., if* (`ages` =4) *is the reference group here.*

   The output gives HR(`ages` =4: `ages` =1)=0.0567. This is just a case of flipping the reference group and so the HR(`ages` =1: `ages` =4) is the reciprocal of HR(`ages` =4: `ages` =1). HR(`ages` =1: `ages` =4)=17.64=1/0.0567.

   (b) [2 points] What are the Wald test statistic, p-value and 95% CI associated with the hazard ratio in (a).

   The corresponding Wald test statistic changes sign (+4.16) and the p-value is unchanged <0.001 since if category 4 is significantly differently from category 1, then category 1 is significantly differently from category 4. The corresponding 95% CI is (4.57, 68.5)=(1/0.219, 1/.0146).

(c) [1 point] What is the hazard ratio for someone aged 2-6 (`ages` =2) compared with someone aged >12 (`ages` =4)?

Recall definition Hazard Ratio (HR) of state A to state B = HR(A:B) = hazard group A / hazard group B

You want HR(2:4), but that is not given in the output.

$$HR(2:4) = \frac{\text{hazard group 2}}{\text{hazard group 4}} = \frac{\text{hazard group 2 / hazard group 1}}{\text{hazard group 4 / hazard group 1}} = \frac{HR(2:1)}{HR(4:1)}$$

$$= .268/0.0566 = 4.73$$

Note that you cannot get p-value this way. For that you would have to perform the appropriate `lincom` command

3. For the variable `age` in the pbc dataset, check the proportional hazards assumption.

(a) [1 point] Using `stphplot` (log-log plot) or `stcoxkm`. *Note, for this question, you will need to create a grouped version of* `age`. *Use the* `Stata` *commands* `xtile` *or* `cut`:

```
use "pbc.dta"  // load dataset
stset years, failure(status) // declare data as survival

centile age, centile(0(25)100) // shows cut-off points for generating quartiles

                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
         age |       312           0     26.27789        26.27789    26.27789*
             |                    25     42.04586        40.67434    43.93096
             |                    50     49.79466        48.48598    51.84021
             |                    75     56.75291        55.96311    59.00011
             |                   100     78.43942        78.43942    78.43942*
* Lower (upper) confidence limit held at minimum (maximum) of sample
```

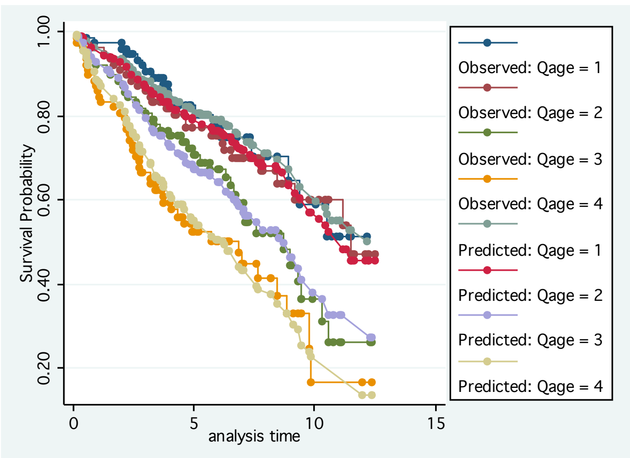
Note the quartiles of age are 42, 50 and 57 years with range from 26 to 78. It may be more sensible to categorize those aged 40-42 into the "second" group than into the youngest group, and likewise to categorize those aged 57-60 into the "third" group than into the oldest group.

```
xtile Qage=age, nq(4) // create variable indicating subject's quartile (labeled 1-4)
egen  Cage=cut(age), group(4) // new column label quartiles (4 equal groups) (this time 0-3)
egen  Gage=cut(age), at(25,40,50,60,80) label // using rounded numbers – label invokes icodes
label var Gage "Age Group" // give the variable Gage the name "Age Group"

tab status Qage // generates a count table of age quantile by status

           |                  4 quantiles of age
Alive/Dead |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
  Censored |        59         52         41         35 |       187
      Dead |        19         26         37         43 |       125
-----------+--------------------------------------------+----------
     Total |        78         78         78         78 |       312
```

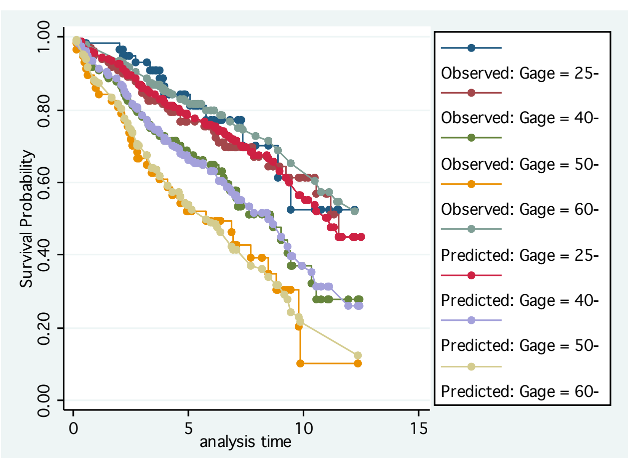
`tab status Gage` // generates a count table of (rounded) age group by status

Alive/Dead	25-	40-	50-	60-	Total
	Age Group				
Censored	45	67	52	23	187
Dead	13	33	45	34	125
Total	58	100	97	57	312

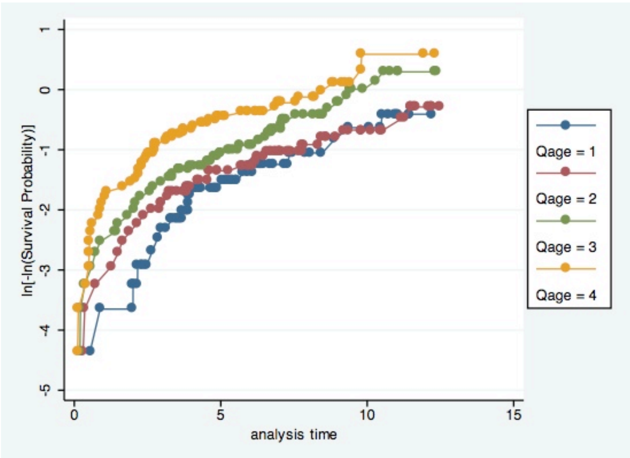
`stcoxkm, by(Qage)` // generates Kaplan-Meier and Cox PH predicted curves overlaid (stratified by quartiles)



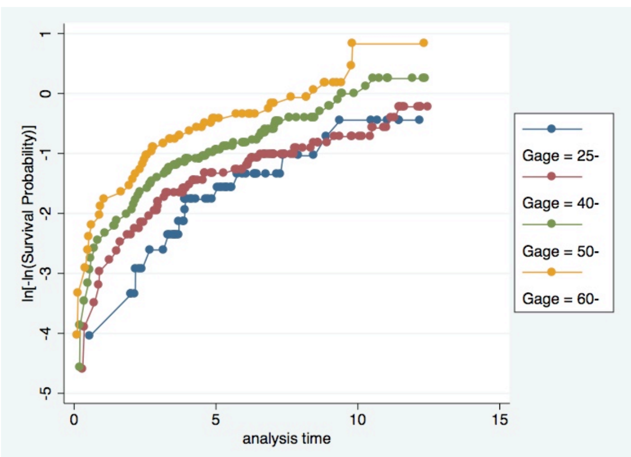
`stcoxkm, by(Gage)` // generates Kaplan-Meier and Cox PH curves overlaid (stratified by rounded age group)



`stphplot, by(Qage) nonegative nolntime` // generates log(-log(Survival function)) stratified by quartiles



`stphplot, by(Gage) nonegative nolntime` // generates log(-log(Survival function)) stratified by age group



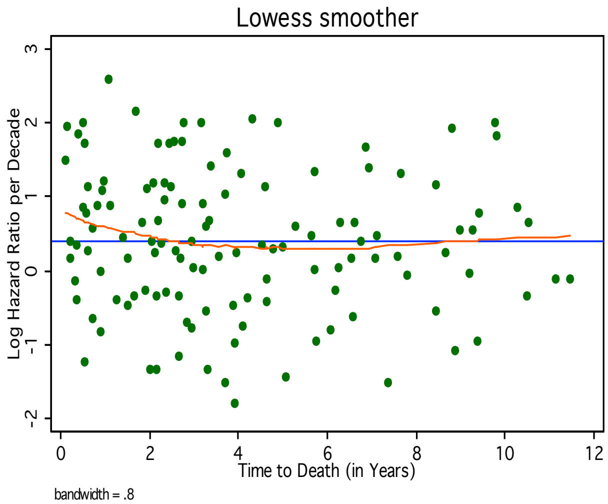
From the stcoxkm plots and log-log plots above, the Cox model appears to fit well (km and cox curves are close and log-log curves are parallel, i.e. more or less equal distances) with this grouping.

(b) [1 point] Smoothed hazard ratios for the variable `age`. ***Recall from the lecture notes: the Smoothed hazard ratios approach is NOT ideal for multilevel categorical variables and it would need a plot for each level of category.*** So we revert back to continuous age.

`gen age10=age/10` // generates age variable (age10) in ten year units

`stcox age10, scaledsch(junk)` // fits Cox PH model with age10 as predictor and generates scaled Schoenfeld residuals (junk)

`lowess junk years, xlabel(0(2)12) ytitle("Log Hazard Ratio") yline(.40, lc(blue)) scheme(s1color)` // generates lowess smoothed curve with default bandwidth (0.8). xlabel command generates time labeled every 2 years from 0 to 12; yline generates a horizontal blue line at a log hazard ratio of 0.4.



The smooth line is relatively flat, suggesting constant hazard ratios is reasonable.

(c) [1 point] The Schoenfeld test for the variable `age`.

`estat phtest, detail` // performs the test of proportional hazards (significant means statistically significant deviation from proportional hazards). The detail flag requests that the significance of the proportional hazard assumption of each individual variable in the model be displayed in addition to the global test. Since there is only one predictor in our model the specific variable test is the same as the global test.

```
Test of proportional-hazards assumption

Time:  Time
----------------------------------------------------------------
             |       rho            chi2       df       Prob>chi2
-------------+--------------------------------------------------
age10        |    -0.08112          0.79        1         0.3748
-------------+--------------------------------------------------
global test  |                      0.79        1         0.3748
----------------------------------------------------------------
```

The p-value (0.37) is not significant, suggesting that the assumption of proportional hazards for age effect is reasonable in this dataset.

Alternatively, you can report p=0.24 when using `estat phtest` *with* `rank` *option, p=0.36 with* `km` *option, or even p=0.13 with* `log` *option.*

(d) [2 points] Based on (a)-(c), what do you conclude? Are hazards for the variable `age` proportional? How would you report the effect of `age` in a paper?

You can obtain the estimated HRs of a 10-year increase in age of about 1.96, 1.81, 1.59, 1.43, and 1.36 at 6 months, 1 year, 2 years, 3 years, and 5 years, respectively. For reference, I include a flat line which corresponds to the HR of 1.49 that is what it would be if the proportional hazards assumption is true. Almost all the potential departure from proportional hazards happens within the first year and then the age effect seems constant over the reminding more than 10 years of follow-up. Based on diagnostic plots and the Schoenfeld test, the assumption of proportional hazards for age effect is reasonable in this dataset, and the HR for a 10-year increase in age is about 1.49 with 95% CI (1.26, 1.77).