Survival Data Analysis & Lab.
Assignment #4 SOLUTION

The dataset "vets.dat" considers survival times in days for 137 patients from the Veteran's Administration Lung Cancer Trial. The exposure variable of interest is treatment status (standard=1, test=2). Other variables of interest as control variables are cell type (four types, defined by dummy variables), performance status, disease duration, age, and prior therapy status. Failure status is defined by the status variable (0 if censored, 1 if died). A complete list of the variables is given below.

Column 1:	Treatment (standard=1, test=2),
Column 2:	Cell type 1 (large=1, other=0)
Column 3:	Cell type 2 (adeno=1, other=0)
Column 4:	Cell type 3 (small=1, other=0)
Column 5:	Cell type 4 (squamous=1, other=0)
Column 6:	Survival time (days)
Column 7:	Performance status (0=worst, $\cdots$, 100=best)
Column 8:	Disease duration (months)
Column 9:	Age
Column 10:	Prior therapy (none=0, some=10)
Column 11:	Status (0=censored, 1=died)

1. State the hazard function form of the Cox PH model that describes the effect of the treatment variable and controls for the variables, cell type, performance status, disease duration, age, and prior therapy. In stating this model, make sure to incorporate the cell type variable using dummy variables, but do not consider possible interaction variables in your model.
   $h(t, \mathbf{X}) = h_0(t) \exp\{\beta_1(\text{treatment}) + \beta_2(\text{CT}_1) + \beta_3(\text{CT}_2) + \beta_4(\text{CT}_3) + \beta_5(\text{PS}) + \beta_6(\text{DD}) + \beta_7(\text{Age}) + \beta_8(\text{PT})\}$, where CT_i denotes the cell type i dummy variable, PS denotes the performance status variable, DD denotes the disease duration variable, and PT denotes the prior therapy variable.

2. State three general approaches that can be used to evaluate whether the PH assumption is satisfied for the variables included in the model you have given in question 1. (Do not provide R analysis for this question. Just state them in general)
   The three general approaches for assessing the PH model for the above model are:
   (a) graphical, using either log-log plots or observed versus expected plots;
   (b) statistical test;
   (c) an extended Cox model containing product terms involves the variables being assessed with some function(s) of time.

3. The following printout is obtained from fitting a Cox PH model to these data. Using the

information provided, what can you conclude about whether the PH assumption is satisfied for the variables used in the model? Explain briefly.

Variables	Coef.	Std.Err.	p-val	Haz.Ratio	95% C.I.		P(PH)
Treatment	0.290	0.207	0.162	1.336	0.890	2.006	0.628
Large cell	0.400	0.283	0.157	1.491	0.857	2.594	0.033
Adeno cell	1.188	0.301	0.000	3.281	1.820	5.915	0.081
Small cell	0.856	0.275	0.002	2.355	1.374	4.037	0.078
Performance status	−0.033	0.006	0.000	0.968	0.958	0.978	0.000
Disease duration	0.000	0.009	0.992	1.000	0.982	1.018	0.919
Age	−0.009	0.009	0.358	0.991	0.974	1.010	0.198
Prior therapy	0.007	0.023	0.755	1.007	0.962	1.054	0.145

The P(PH) values given in the printout provide goodness-of-fit tests for each variable in the fitted model adjusted for the other variables in the model. The P(PH) values shown indicate that the large cell type variables and the performance status variable do not satisfy the PH assumption, whereas the treatment, age, disease duration, and prior therapy variables satisfy the PH assumption, and the adeno and small cell type variables are of borderline significant.

4. For the variables used in the PH model in question 3, describe a strategy for evaluating the PH assumption using log-log survival curves for variables considered one-at-a-time.
   A strategy for evaluating the PH assumption using log-log survival curves for variables considered one-at-a-time is given as follows:
   For each variable separately, obtain a plot of log-log Kaplan-Meier curves for the different categories of that variable. For the cell type variable, this requires obtaining a plot of four log-log KM curves, one for each cell type. (Note that this is not the same as obtaining four separate plots of two log-log curves, where each plot corresponds to one of the dummy variables used in the model.) For the variables PS, DD, and Age, which are quantitative variables, each variable must be separately categorized into two or more groups–say, low versus high values–and KM Curves are obtained for each group. For the variable PT, which is a dichotomous variable, two log-log curves are obtained which compare the "none" versus "some" groups.
   For each plot (i.e., one for each variable), those plots that are noticeably nonparallel indicate variables which do not satisfy the PH assumption. The remaining variables are assumed to satisfy the PH assumption.

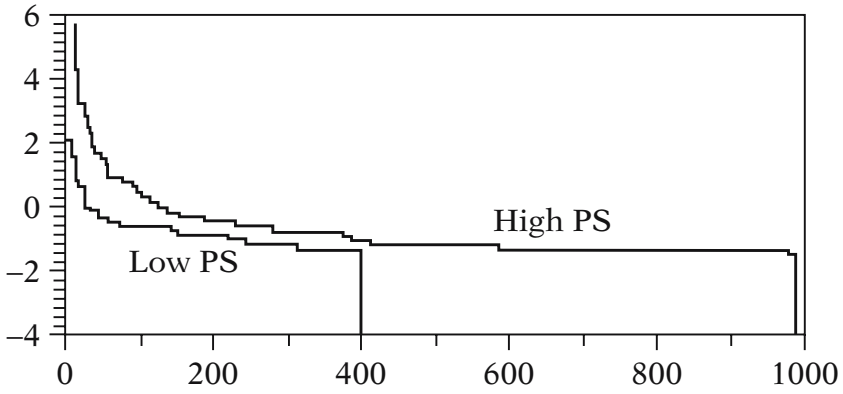
5. Again considering the variables used in question 3, describe a strategy for evaluating the PH assumption using log-log survival curves that are adjusted for other variables in the model.
   One strategy for evaluating the PH assumption for each variable adjusted for the others is to use adjusted log-log survival curves instead of KM curves separately for each of the variables in the model. That is, for each variable separately, a stratified Cox model is fit satisfying on the given variable while adjusting for the other variables. Those variables that yield adjusted log-log plots that are noticeably nonparallel are then to be considered as not satisfying the PH assumption. The remaining variables are assumed to satisfy the PH assumption.

6. For the variable "performance status," describe how you would evaluate the PH assumption using observed versus expected survival plots?

   For the performance status (PS) variable, observed plots are obtained by categorizing the variable into strata (say, two strata; low versus high) and then obtaining KM survival plots for each stratum. Expected plots can be obtained by fitting a Cox model containing the (continuous) PS variable and then obtaining estimated survival curves for values of the performance status (PS) variable that represent summary descriptive statistics for the strata previously identified. For example, if there are two strata, say, high (PS>50) and low (PS≤ 50), then the values of PS to be used could be the mean or median PS score for persons in the high stratum and the mean or median PS score for persons in the low stratum.

   An alternative method for obtaining expected plots involves first dichotomizing the PS variable–say, into high and low groups–and then fitting a Cox model containing the dichotomized PS variable instead of the original continuous variable. The expected survival plots for each group are estimated survival curves obtained for each value of the dichotomized PS variable. Once observed and expected plots are obtained for each stratum of the PS variable, they are then compared on the same graph to determine whether or not corresponding observed and expected plots are "close." If it is determined that, overall, comparisons for each stratum are close, then it is concluded that the PH assumption is satisfied for the PH variable. In determining how close is close, the researcher should look for noticeably discrepant observed versus expected plots.

7. For the variable "performance status," log-log plots which compare high (≥ 50) with low (< 50) are given by the following graph. Based on this graph, what do you conclude about the PH assumption with regard to this variable?



   The log-log plots that compare high versus low PS groups (ignoring other variables) are arguably parallel early in follow-up, and are not comparable later because survival times for the two groups do not overlap after 400 days. These plots do not strongly indicate that the PH assumption is violated for the variable PS. This contradicts the conclusion previously obtained for the PS variable using the P(PH) results.

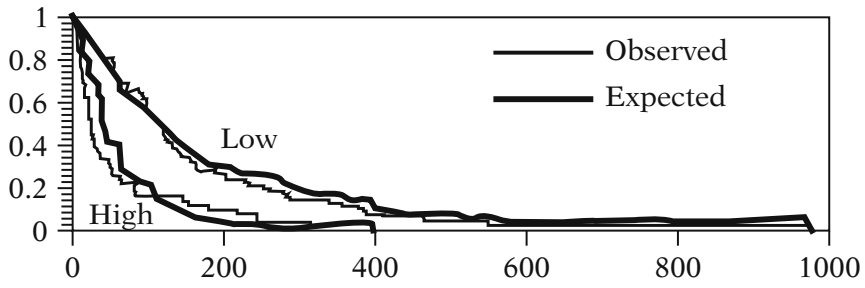
8. What are some of the drawbacks of using the log-log approach for assessing the PH assumption and what do you recommend to deal with these drawbacks?
Drawbacks of the log-log approach are:

   - How parallel is parallel?
   - How to categorize a continuous variable?
   - How to evaluate several variables simultaneously?

   Recommendations about problems:

   - Look for noticeable non parallelism; otherwise PH assumption is OK.
   - For continuous variables, use a small number of categories, a meaningful choice of categories, and a reasonable balance in sample size for categories.
   - With several variables, there are two options:
     (a) Compare log-log curves from combinations of categories.
     (b) Adjust for predictors already satisfying the PH assumption.

9. For the variable "performance status," observed versus expected plots that compare high (≥ 50) with low (< 50) are given by the following graph. Based on this graph, what do you conclude about the PH assumption with regard to this variable?



   The observed and expected plots are relatively close for low and high groups separately, although there is somewhat more discrepancy for the high group than for the low group. Deciding how close is close is quite subjective for this plots. Nevertheless, because there are no major discrepancies for either low or high groups, we consider the PH assumption satisfied for this variable.

10. State the form of an extended Cox model that allows for the one-at-a-time assessment of the PH assumption for the variable "performance status," and describe how you would carry out a statistical test of the assumption for this variable?
$h(t, \mathbf{X}) = h_0(t) \exp\{\beta_1(\text{PS}) + \delta(\text{PS})g(t)\}$, where $g(t)$ is a function of t, such as $g(t) = t$, or $g(t) = \log t$, or a heavyside function. The PH assumption is tested using a 1 df Wald or LR statistic for $H_0 : \delta = 0$.

11. State the form of an extended Cox model that allows for the simultaneous assessment of the PH assumption for the variables, treatment, cell type, performance status, disease duration, age, and prior therapy. For this model, describe how you would carry out a statistical test of the PH assumption for these variables. Also, prove a strategy for assessing which of these variables satisfy the PH assumption and which do not using the extended Cox model approach.
$h(t, \mathbf{X}) = h_0(t) \exp\{\beta_1(\text{treatment}) + \beta_2(\text{CT}_1) + \beta_3(\text{CT}_2) + \beta_4(\text{CT}_3) + \beta_5(\text{PS}) + \beta_6(\text{DD}) + \beta_7(\text{Age}) + \beta_8(\text{PT}) + \delta_1(\text{treatment} \times g(t)) + \delta_2(\text{CT}_1 \times g(t)) + \delta_3(\text{CT}_2 \times g(t)) + \delta_4(\text{CT}_3 \times g(t)) + \delta_5(\text{PS} \times g(t)) + \delta_6(\text{DD} \times g(t)) + \delta_7(\text{Age} \times g(t)) + \delta_8(\text{PT} \times g(t))\}$, where $g(t)$ is some function of time, such as $g(t) = t$, or $g(t) = \log t$, or a heavyside function. To test the PH assumption simultaneously for all variables, the null hypothesis is stated as $H_0 : \delta_1 = \delta_2 = \cdots = \delta_8 = 0$. The test statistic is a likelihood-ratio statistic of the form

$$LR = -2 \ln L_R - (-2 \ln L_F),$$

where R denotes the reduced (PH) model obtained when all δ's are 0, and F denotes the full model given above. Under H_0, the LR statistic is approximately chi-square with 8 df.

12. Using any of the information provided above and any additional analyses that you perform with this dataset, what do you conclude about which variables satisfy the PH assumption and which variables do not? In answering this question, summarize any additional analyses performed.
The question here is somewhat open-ended, leaving the students the option to explore additional graphical, GOF, or extended Cox model approaches for evaluating the PH assumption for the variables in the model. The conclusions from the GOF statistics provided in question 3 are likely to hold up under further scrutiny, so that a reasonable conclusion is that cell type and performance status variables do not satisfy the PH assumption, with the remanning variables satisfying the assumptions.