Meenu Ravi

MSCA31010: Linear & Non-Linear Models

Winter 2021 Assignment 4

Questions 1 and 2

Krall, Uthoff, and Harley (1975) analyzed data from a study on multiple myeloma in which researchers treated sixty-five patients with alkylating agents. Of those patients, forty-eight died during the study, and seventeen survived.

The data set is in the myeloma.csv. The variable Time represents the survival time in months from diagnosis. The variable VStatus consists of two values, 0 and 1, indicating whether the patient was alive or dead, respectively, at the end of the study. If the value of VStatus is 1, the patient died during the study. If the value of VStatus is 0, the patient was still alive at the end of the study and the corresponding value of Time is censored.

The following nine variables thought to be related to survival are

1. LogBUN: logarithm of blood urea nitrogen at diagnosis,
2. HGB: hemoglobin at diagnosis,
3. Platelet: platelets at diagnosis: 0=abnormal, 1=normal,
4. Age: age at diagnosis, in years,
5. LogWBC: logarithm of the number of white blood cells at diagnosis,
6. Frac: fractures at diagnosis: 0=none, 1=present,
7. LogPBM: logarithm of the percentage of plasma cells in bone marrow,
8. Protein: proteinuria at diagnosis, and
9. SCalc: serum calcium at diagnosis.

Our interest lies in identifying important prognostic factors from these nine explanatory variables.

Reference: John M. Krall, Vincent A. Uthoff, and John B. Harley (1975). “A Step-Up Procedure for Selecting Variables Associated with Survival.” *Biometrics*, volume 31, number 1, pages 49 – 57.

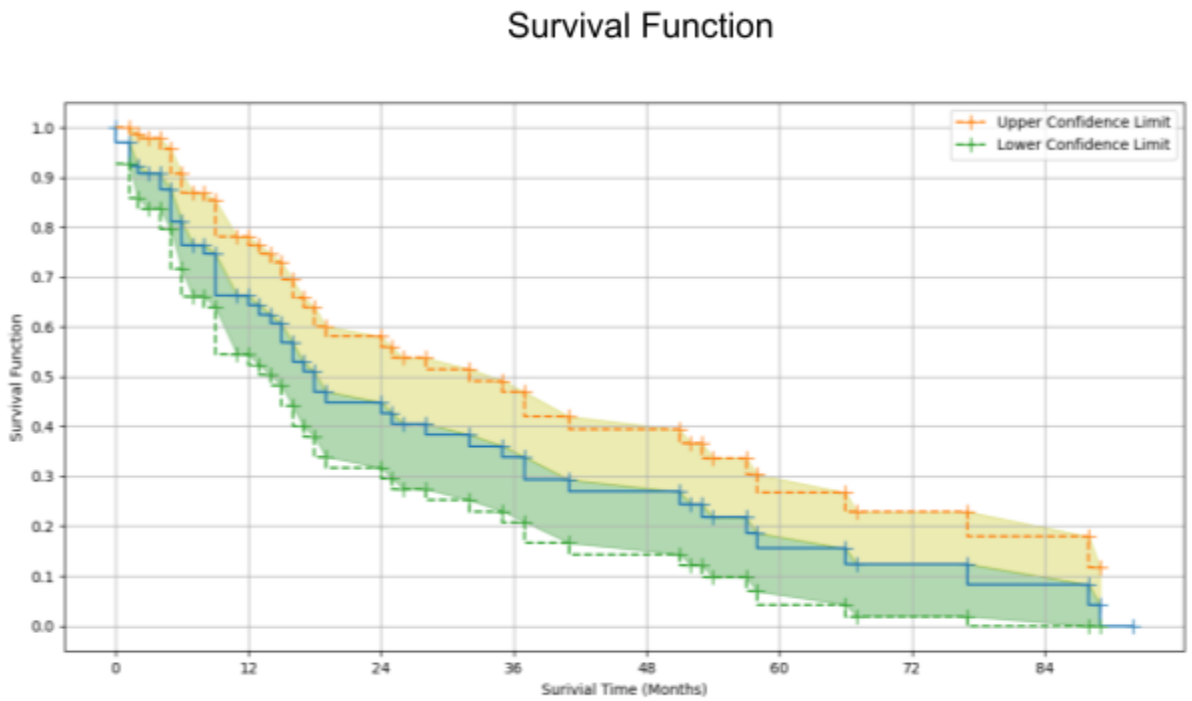
Question 1 (50 points)

(a) (30 points). Please use the Kaplan-Meier Product Limit Estimator to create the life table. The life table should include these columns: *Survival Time*, *Number Remaining*, *Number of Deaths*, *Number of Censored*, *Number at Risk*, *Probability of Survival*, *Probability of Failure*, *Standard Error for Probability of Survival*, *Lower 95% Confidence Limit for Probability of Survival*, *Upper 95% Confidence Limit for Probability of Survival*, and *Cumulative Hazard*. Please limit the number of decimal places to four.

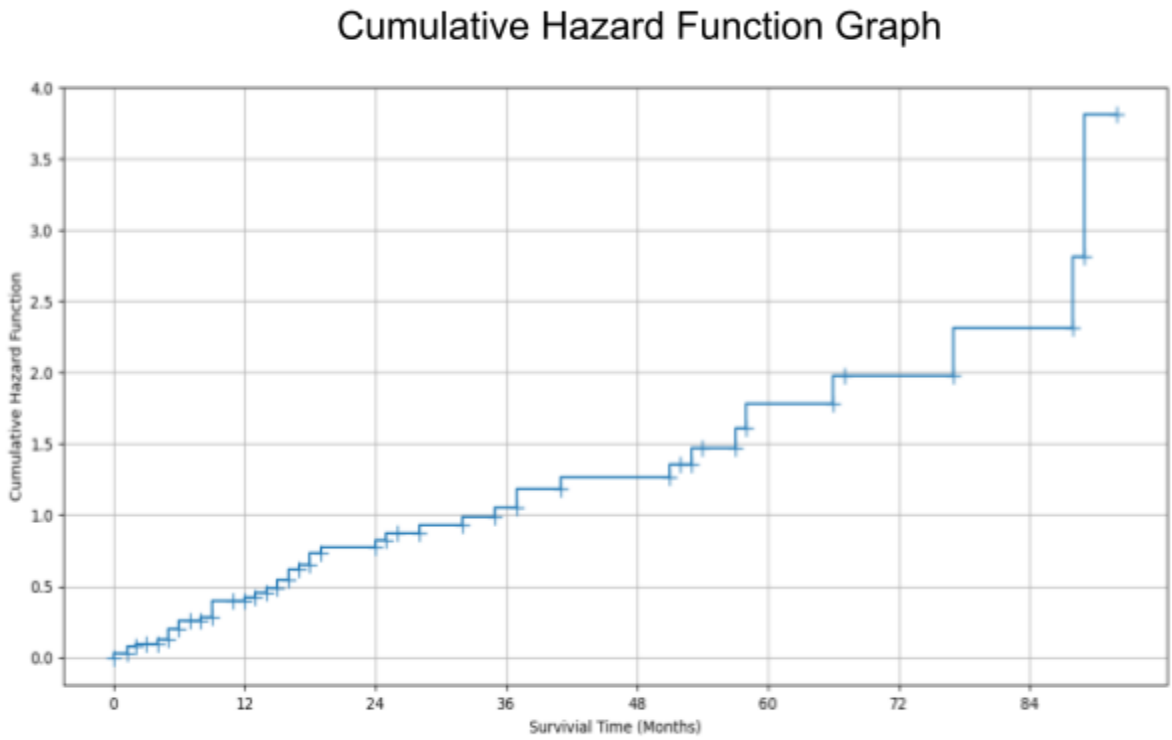
Survival Time	Number Left	Number of Deaths	Number Censored	Number at Risk	Prob Survival	Prob Failure	SE Prob Survival	Lower CI Prob Survival	Upper CI Prob Survival	Cumulative Hazard
0	65	0	0	65	1	0				0
1.25	63	2	0	65	0.9692	0.0308	0.0214	0.9272	1	0.0308
2	60	3	0	63	0.9231	0.0769	0.0331	0.8583	0.9879	0.0784
3	59	1	0	60	0.9077	0.0923	0.0359	0.8373	0.9781	0.0951
4	59	0	2	59	0.9077	0.0923	0.0359	0.8373	0.9781	0.0951
5	55	2	0	57	0.8758	0.1242	0.0411	0.7953	0.9564	0.1301
6	51	4	0	55	0.8121	0.1879	0.0489	0.7163	0.9080	0.2029
7	48	3	2	51	0.7644	0.2356	0.0533	0.6600	0.8687	0.2617
8	46	0	1	46	0.7644	0.2356	0.0533	0.6600	0.8687	0.2617
9	44	1	0	45	0.7474	0.2526	0.0547	0.6402	0.8546	0.2839
11	39	5	1	44	0.6625	0.3375	0.0603	0.5444	0.7806	0.3976
12	38	0	2	38	0.6625	0.3375	0.0603	0.5444	0.7806	0.3976
13	35	1	1	36	0.6441	0.3559	0.0613	0.5239	0.7643	0.4253
14	33	1	0	34	0.6251	0.3749	0.0624	0.5029	0.7474	0.4547
15	32	1	0	33	0.6062	0.3938	0.0633	0.4821	0.7302	0.4850
16	30	2	1	32	0.5683	0.4317	0.0648	0.4413	0.6952	0.5475
17	27	2	0	29	0.5291	0.4709	0.0660	0.3998	0.6584	0.6165
18	26	1	0	27	0.5095	0.4905	0.0664	0.3794	0.6396	0.6535
19	24	2	2	26	0.4703	0.5297	0.0668	0.3394	0.6012	0.7305
24	21	1	0	22	0.4489	0.5511	0.0671	0.3174	0.5804	0.7759
25	20	1	0	21	0.4275	0.5725	0.0672	0.2958	0.5593	0.8235
26	19	1	0	20	0.4062	0.5938	0.0672	0.2745	0.5378	0.8735

28	19	0	1	19	0.4062	0.5938	0.0672	0.2745	0.5378	0.8735
32	17	1	0	18	0.3836	0.6164	0.0671	0.2520	0.5152	0.9291
35	16	1	0	17	0.3610	0.6390	0.0669	0.2300	0.4921	0.9879
37	15	1	0	16	0.3385	0.6615	0.0664	0.2084	0.4686	1.0504
41	13	2	1	15	0.2933	0.7067	0.0647	0.1664	0.4202	1.1838
51	11	1	0	12	0.2689	0.7311	0.0638	0.1439	0.3939	1.2671
52	10	1	0	11	0.2445	0.7555	0.0625	0.1219	0.3670	1.3580
53	10	0	1	10	0.2445	0.7555	0.0625	0.1219	0.3670	1.3580
54	8	1	0	9	0.2173	0.7827	0.0612	0.0974	0.3372	1.4691
57	8	0	1	8	0.2173	0.7827	0.0612	0.0974	0.3372	1.4691
58	6	1	0	7	0.1863	0.8137	0.0598	0.0690	0.3035	1.6120
66	5	1	0	6	0.1552	0.8448	0.0573	0.0429	0.2676	1.7786
67	4	1	0	5	0.1242	0.8758	0.0536	0.0191	0.2292	1.9786
77	4	0	1	4	0.1242	0.8758	0.0536	0.0191	0.2292	1.9786
88	2	1	0	3	0.0828	0.9172	0.0492	0	0.1792	2.3120
89	1	1	0	2	0.0414	0.9586	0.0382	0	0.1163	2.8120
92	0	1	0	1	0	1				3.8120

(b) (10 points). Please generate the Survival Function graph. Since we measure the Time variable in the number of months, we will specify the x-axis ticks from 0 in an increment of 12. Besides plotting the Survival Function versus Time, please also add the 95% Confidence Band. Please consider using the fill_between() function in matplotlib to generate the Confidence Band as a band around the Survival Function. To receive the full credits, you must label the chart elements properly.



(c) (10 points). Please generate the Cumulative Hazard Function graph. To receive the full credits, you must label the chart elements properly.



Question 2 (50 points)

We will train a Proportional Hazard model to identify important prognostic factors from the nine explanatory variables. We will consider all explanatory variables as interval predictors. We will use the Backward Selection method to exclude non-significant predictors.

We first include all nine explanatory variables in the model. After we have trained a Proportional Hazard model, we will retrieve the summary object to obtain the test significance values (i.e., the p-value) of the explanatory variables. Next, we will look among the explanatory variables whose p-values are more than 0.15. Then, remove the explanatory variable, if any, that has the highest p-value from the model. We will repeat the Backward steps until there are no more explanatory variables that we can remove.

(a) (30 points). Please provide a Step Summary table. The table will show the explanatory variable removed from each step and the variable's test significance value.

Step	Variables	coef	exp(coef)	se(coef)	coeflower95%	coef upper95%	exp(coef) lower95%	exp(coef) upper95%	z	p	-log2(p)
0	LogWBC	0.35	1.42	0.72	-1.05	1.76	0.35	5.79	0.49	0.62	0.69
1	Platelet	-0.2	0.82	0.5	-1.18	0.78	0.31	2.19	-0.4	0.69	0.54
2	Protein	0.01	0.01	0.02	-0.04	0.06	0.96	1.06	0.44	0.66	0.6
3	LogPBM	0.35	1.43	0.47	-0.57	1.28	0.57	3.6	0.75	0.45	1.15
4	Frac	0.32	1.37	0.4	-0.46	1.09	0.63	2.98	0.8	0.43	1.23
5	Age	-0.02	0.98	0.02	-0.05	0.01	0.95	1.01	-1.34	0.18	2.47
6	SCalc	0.14	1.14	0.1	-0.06	0.33	0.94	1.39	1.36	0.17	2.52

(b) (10 points). Please provide a Parameter Estimates table that shows the explanatory variables included in the final model. Besides, show these statistics: parameter estimates, standard errors, p-values, Hazard Ratios, and 95% confidence interval for the Hazard Ratios.

Variables	coef Param Est	exp(coef) (Hazard Ratio)	se (coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
LogBUN	1.72	5.56	0.62	0.5	2.93	1.65	18.7	2.77	0.01	7.5
HGB	-0.12	0.89	0.06	-0.23	-0.01	0.79	0.99	-2.08	0.04	4.75

(c) (10 points). Please plot the estimated baseline cumulative hazard from the Proportional Hazard model versus the observed times. We will also overlay the cumulative hazard function from the Kaplan-Meier estimator for comparison. To receive the full credits, you must label the chart elements properly.