Name: Himanshu Chopra

Email: himanshu.chopra@edu.dsti.institute

Github Repository: https://github.com/himchopra/survivalanalysis

Survival Analysis Project: Students will need to perform a statistical analysis of a dataset of their choice, using any of the methods seen during the class, i.e.:

- nonparametric estimation of survival for one or more groups
- nonparametric comparison of 2 or more groups
- semi-parametric Cox regression

Dataset Chosen: The Veterans' Administration Lung Cancer Trial. Source of the data is scikit library.

Dataset consists of 137 samples and 6 features

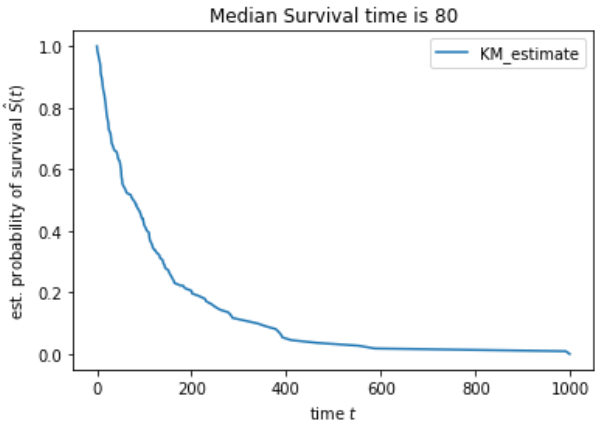
6 features are

a) Age_in_years : Datatype float : Min age 31, Max age 81, no null values
b) Celltype : Datatype category : 4 unique values ['squamous', 'smallcell', 'adeno', 'large']
c) Karnofsky_score : Datatype float
d) Months_from_Diagnosis : Datatype float
e) Prior_therapy : Datatype category: 2 unique values ['no','yes']
f) Treatment : Datatype category: 2 unique values ['standard','test']

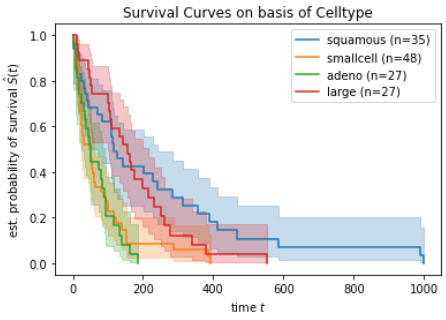
Death was observed for 128 out of 137 patients (93.43%)

Observations

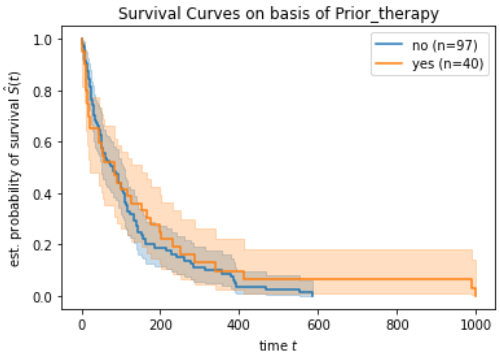
- **Median Survival days is 80 days**



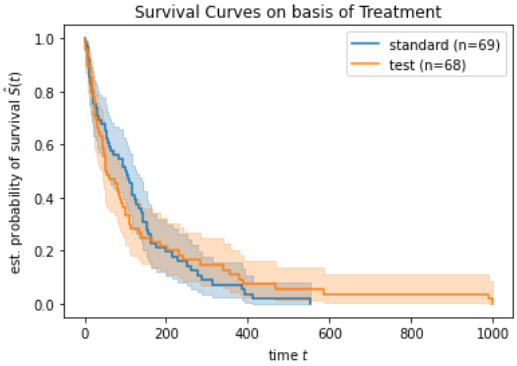
- **Median Survival days differ on basis of cell type**
    - Median of squamous is 118 days
    - Median of smallcell is 51 days
    - Median of adeno is 51 days
    - Median of large is 156 days



- Median Survival days doesn't differ on basis of Prior therapy
    - Median of no is 80
    - Median of yes is 82



- Median Survival days differ on basis of Treatment
    - Median of standard is 103
    - Median of test is 52



Performing Statistical Analysis for various groups

Logrank test : Measures and reports that whether for two event series data generating processes are statistically different. This test-statistic is chi-squared under the null hypothesis. If survival functions cross, the logrank test will give inaccurate assessment.

1. **First log rank test is done for Prior Therapy:** As we have observed earlier, the survival functions for “prior_treatment = Yes” and “prior_treatment = No” cross and hence logrank test is not the best evaluator. Nonetheless, trying it

   p-value = 0.48 > 0.05 hence we fail to reject the null hypothesis. Hence, survival function of prior_treatment = Yes / No are not different.

   As an alternative to logrank test, trying cox regression with only 1 variable, again we got p=0.48 which is revalidation of output from logrank test

   Hence, Prior Therapy has not shown ability to predict cancer free survival time

```
index = (df_x["Prior_therapy"] == 'yes')
results = logrank_test(df_y["Survival_in_days"][index],df_y["Survival_in_days"][~index],df_y["Status"][index],df_y["Status"][~index],alpha=0.95)
results.print_summary()
```

t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
alpha	0.95
test_name	logrank_test

	test_statistic	p	-log2(p)
0	0.50	0.48	1.06

```
cph = CoxPHFitter()
cph.fit(df_model[['Prior_therapy=yes','Status','Survival_in_days']],'Survival_in_days',event_col = 'Status')
cph.print_summary()
```

model	lifelines.CoxPHFitter
duration col	'Survival_in_days'
event col	'Status'
baseline estimation	breslow
number of observations	137
number of events observed	128
partial log-likelihood	-505.19
time fit was run	2023-01-05 17:59:53 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Prior_therapy=yes	-0.14	0.87	0.20	-0.54	0.25	0.59	1.28	0.00	-0.71	0.48	1.07

Concordance	0.49
Partial AIC	1012.38

2. **Log rank test for basis for Treatment:** p-value of 0.93, hence groups are not statistically different at 95% confidence level. Same results are observed using both logrank_test and CoxPHFitter. Hence, treatment has not shown ability to predict cancer free survival time

```
index = (df_x["Treatment"] == 'test')
results = logrank_test(df_y["Survival_in_days"][index],df_y["Survival_in_days"][~index],
                       df_y["Status"][index],df_y["Status"][~index],
                       alpha=0.95)
results.print_summary()
print(results.p_value)
print(results.test_statistic)
```

t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
alpha	0.95
test_name	logrank_test

	test_statistic	p	-log2(p)
0	0.01	0.93	0.11

```
0.9277272333400758
0.008227343202350305
```

```
cph = CoxPHFitter()
cph.fit(df_model[['Treatment=test','Status','Survival_in_days']],'Survival_in_days',event_col = 'Status')
cph.print_summary()
```

model	lifelines.CoxPHFitter
duration col	'Survival_in_days'
event col	'Status'
baseline estimation	breslow
number of observations	137
number of events observed	128
partial log-likelihood	-505.44
time fit was run	2023-01-05 18:05:11 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Treatment=test	0.02	1.02	0.18	-0.34	0.37	0.71	1.45	0.00	0.10	0.92	0.12

Concordance	0.53
Partial AIC	1012.89
log-likelihood ratio test	0.01 on 1 df
-log2(p) of ll-ratio test	0.12

3. **CoxPHFitter for Celltype :** Following are the p-values for large, smallcell and squamous respectively <0.005, 0.56, 0.005. Hence, not all groups are statistically different. Same is observed through multivariate_logrank_test. Hence, Celltype has not shown ability to predict cancer free survival time

4. **Age_in_years :** Again not significant. Hence, Age has not shown ability to predict cancer free survival time
5. **Karnofsky_score :** Significant. Hence, karnofsky score has **shown** ability to predict cancer free survival time

```
cph = CoxPHFitter()
cph.fit(df_model[['Karnofsky_score','Status','Survival_in_days']],'Survival_in_days',event_col = 'Status')
cph.print_summary()
```

model	lifelines.CoxPHFitter
duration col	'Survival_in_days'
event col	'Status'
baseline estimation	breslow
number of observations	137
number of events observed	128
partial log-likelihood	-484.43
time fit was run	2023-01-05 18:14:19 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Karnofsky_score	-0.03	0.97	0.01	-0.04	-0.02	0.96	0.98	0.00	-6.59	<0.005	34.37

Concordance	0.71
Partial AIC	970.87
log-likelihood ratio test	42.03 on 1 df
-log2(p) of ll-ratio test	33.37

References:

1. https://lifelines.readthedocs.io/en/latest/index.html
2. https://scikit-survival.readthedocs.io/en/stable/