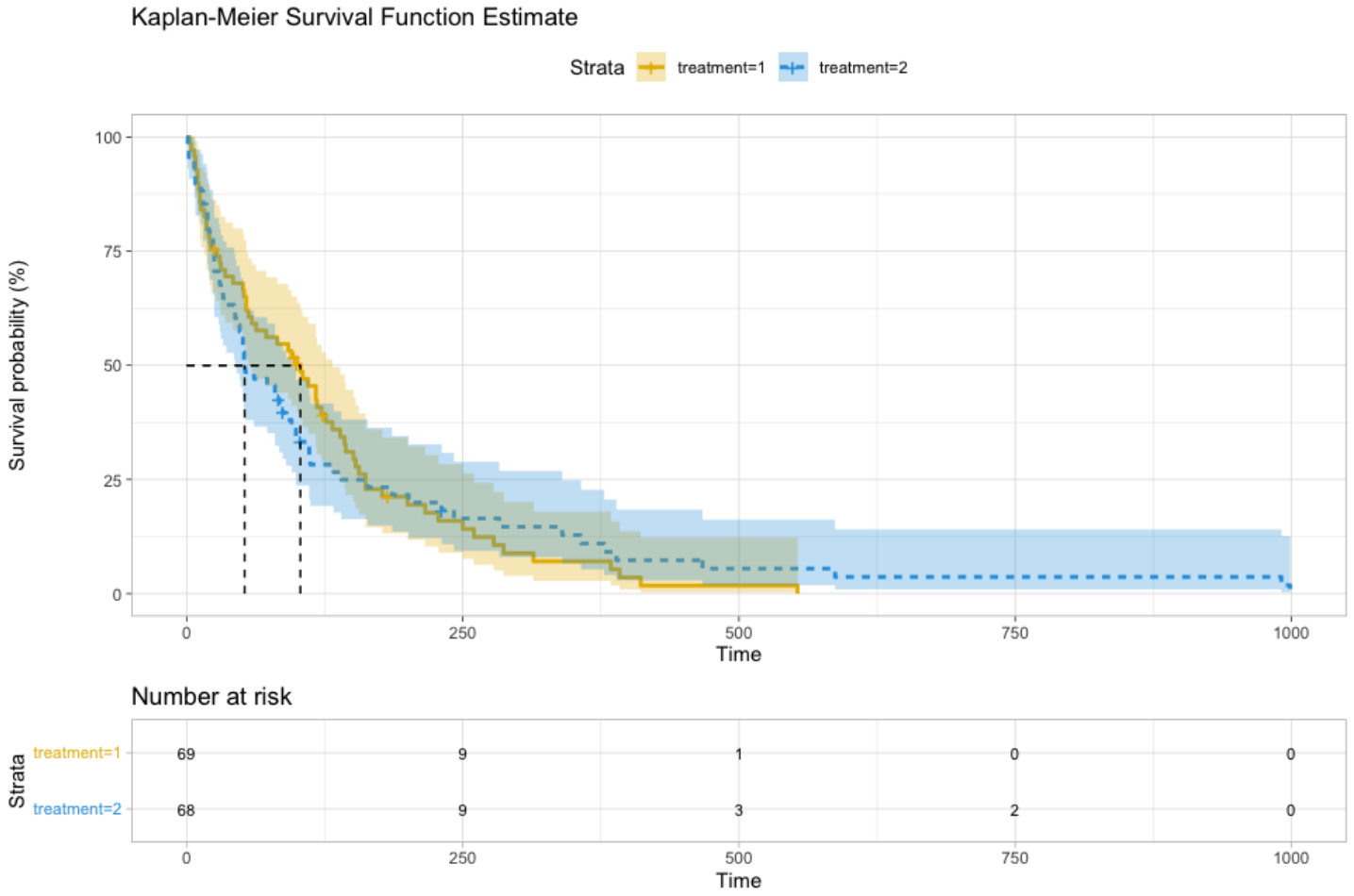
Statistical Data Mining | Survival Analysis

Pablo X Zumba

1. **Kaplan-Meier survival graphs for patients with the test vs standard treatment.**



Use this data to assess:

- **What is the probability that the patient will survive for 1 year (365 days) and 6 months (183 days) on the standard treatment vs the test treatment?**

```
treatment=1
    time      n.risk     n.event     survival     std.err lower 95% CI upper 95% CI
  365.0000     4.0000    60.0000      0.0708      0.0336      0.0279      0.1795

treatment=2
      time      n.risk     n.event     survival     std.err lower 95% CI upper 95% CI
  365.0000     6.0000    58.0000      0.1098      0.0407      0.0530      0.2272
```

Interpretation: *The standard treatment has a 0.07 - 7% chance of surviving for 1 year, while the test treatment has a 0.10 - 10% chance.*

```
treatment=1
      time      n.risk     n.event     survival     std.err lower 95% CI upper 95% CI
  183.0000    12.0000    52.0000      0.2124      0.0514      0.1322      0.3414

treatment=2
      time      n.risk     n.event     survival     std.err lower 95% CI upper 95% CI
  183.0000    14.0000    51.0000      0.2329      0.0529      0.1492      0.3634
```

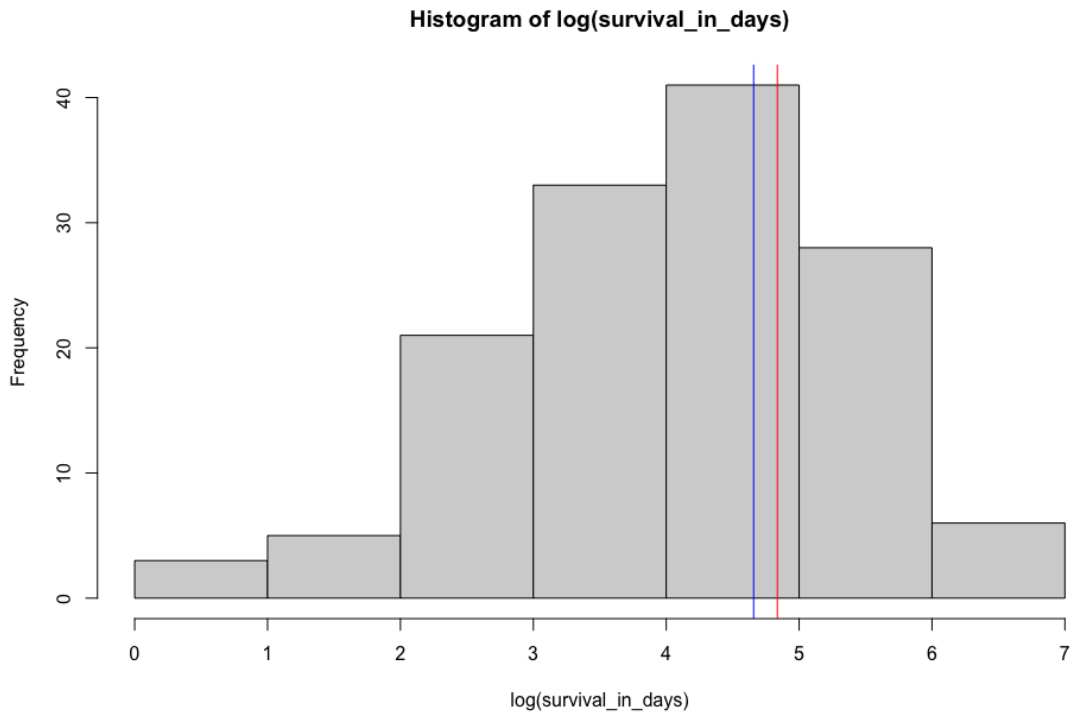
Interpretation: *The standard treatment has a 0.21 - 21% chance of surviving for 6 months, while the test treatment has a 0.23 - 23% chance.*

- **What is the average number of days where a patient can be expected to survive if they are on the standard vs the test treatment?**

105 is the average number of days a patient can be expected to survive under *standard* treatment.

126 is the average number of days a patient can be expected to survive under the *test* treatment.

In the picture below, a logarithmic transformation shows that the test treatment (red line) has a higher mean number of days than the standard treatment (blue line).



2. **Create at least three semi-parametric and parametric models (either changing underlying parametric distributions or sets of variables) to estimate the marginal effects of relevant predictors on survival outcomes. Interpret the coefficients of these models to explain the precise effects of age and diagnosis months on patients' survival probabilities with standard and test treatments. Stargazer tables should work for the nonparametric and semiparametric models you are using.**

   Interpretation:

   - The most significant independent variable in all survival models is the Karnofsky score, which measures how well cancer patients can perform daily tasks. A higher score indicates better performance. Overall, the models show that events are less likely to occur if the Karnofsky score is higher. In theory, the Karnofsky score could be correlated with patients' ages since an aging body is less able to perform daily functions. There needs to be a further correlation analysis between Karnofsky and the age of the patient.
   - Those with a positive coefficient for treatment, months from diagnosis, and cell type are more likely to experience the event.
   - Diagnosis month of patients does not seem to add significance to all the survival models since coefficients are very low.
   - Older individuals have a higher probability of experiencing an event, according to the age coefficient.

Comparison of marginal effects using different survival models

	Dependent variable:			
	survival_in_days			
	Cox prop. hazards	*exponential*	*Weibull*	*survreg: loglogistic*
	(1)	(2)	(3)	(4)
treatment	0.223	-0.184	-0.185	-0.056
	(0.188)	(0.182)	(0.182)	(0.186)
months_from_diagnosis	0.002	-0.003	-0.003	0.005
	(0.009)	(0.009)	(0.009)	(0.010)
age_in_years	-0.004	0.001	0.001	0.009
	(0.009)	(0.009)	(0.009)	(0.009)
prior_chemotherapy	-0.077	0.112	0.112	0.029
	(0.223)	(0.218)	(0.218)	(0.226)
karnofsky_score	-0.035***	0.035***	0.035***	0.040***
	(0.005)	(0.005)	(0.005)	(0.005)
cell_type	0.129*	-0.136*	-0.137*	0.021
	(0.078)	(0.075)	(0.075)	(0.090)
Constant		3.149***	3.154***	1.279*
		(0.710)	(0.713)	(0.753)
Observations	137	137	137	137
R^2	0.283			
Max. Possible R^2	0.999			
Log Likelihood	-483.111	-724.014	-724.012	-719.581
chi^2 (df = 6)		54.415***	48.158***	61.370***
Wald Test	46.860*** (df = 6)			
LR Test	45.545*** (df = 6)			
Score (Logrank) Test	49.323*** (df = 6)			
Note:				*p<0.1; **p<0.05; ***p<0.01