Patient Survival Analysis:

I. We would like to see Kaplan-Meier survival graphs for patients with the test vs standard treatment. Use this data to assess.

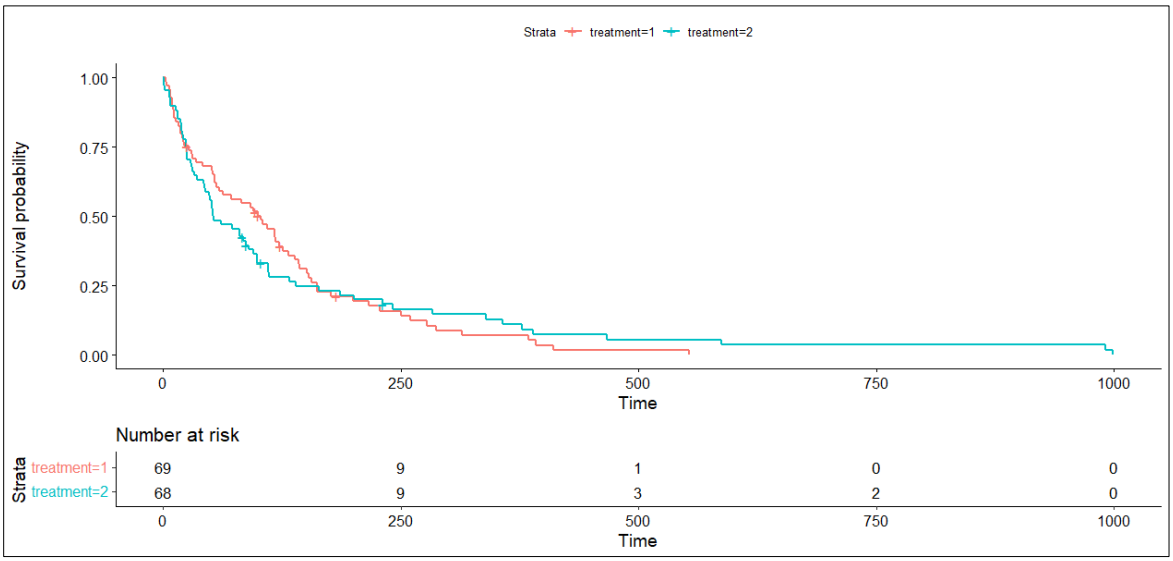
- **What is the probability that the patient will survive for 1 year (365 days) and 6 months (183 days) on the standard treatment vs the test treatment?**

Dependent Variables: Dependent variable for the survival analysis problems is the waiting time until the occurrence of an event. Observations are censored, in the sense that for some cases the event has not yet occurred at the time of analysis. These are predictors whose effect on the waiting time we wish to assess. Which is the survival_days (Time) and status (Event).

	Standard Treatment	Test Treatment
Probability that the patient will survive for 1 year (365 days)	Approx. **7%**	Approx. **11%**
Probability that the patient will survive for 6 months (183 days)	Approx. **21%**	Approx. **23%**

Estimated probability could vary within a range of 3-18% (95% CI) for the patients that received the standard treatment and a range of 5-23% (95% CI) for the patients that received the test treatment.

Estimated probability could vary within a range of 13-34% (95% CI) for the patients that received the standard treatment and a range of 15-36% (95% CI) for the patients that received the test treatment.



From our KM analysis, the survival probability of the test treatment group is higher than the standard group for both time periods. But from the plot, both treatments show the same survival probability until the first 15-20 days. Then the standard treatment group has higher survival probability until about 160-175 days. Then the test treatment group has higher survival probability. This means it takes about 175 days to see the effects of the new drug.

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

	Standard Treatment	Test Treatment
Mean number of days where a patient can be expected to survive	Approx. **103 days**	Approx. **52.5 days**

II. Create three semi-parametric and parametric models 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 months of diagnosis on survival probabilities of patients with standard and test treatments.

Factor Effects:

Positive Effect: If there is a direct proportionality (+X then +y / -X then -y) between the predictor variable (X) and the response variable (y), we can say there is a positive effect.

Negative Effect: If there is an inverse proportionality (+X then -y / -X then +y) between the predictor variable (X) and the response variable (y), we can say there is a negative effect.

Relevant Factors for y <- [Time (survival_days), Event (status)]		
Predictor	**Effect**	**Rationale**
treatment	+	We have two treatments, which could have an effect on the patients survival.
cell_type	+/-	Lung cancer has two main types, small cell and non-small cell. We need to refactor cell_types based on the lung cancer types with two factor levels. This refactoring could have varying effects on the treatments.
months_from_diagnosis	+	More time from diagnosis can mean higher risk to the patient.
age_in_years	-	The effect of the new drug could decrease as the age of the patient increases.
prior_chemotherapy	+/-	Prior chemo session could have varying effects on the patient's survival based on the treatments.
Irrelevant Factors		
Predictor	**Effect**	**Rationale**
karnofsky_score	No Effect	This is a performance score used for pretreatments and is associated with better survival compared with patients with lung cancer. So, it is not relevant for our analysis.

```
#Stargazer
stargazer(cox1, exp1, wbl1, llg1, type='text', single.row = TRUE)
```

```
##
## ===============================================================================================
##                                              Dependent variable:
##                         ------------------------------------------------------------------
##                                                      y
##                               Cox           exponential         Weibull      survreg: loglogistic
##                          prop. hazards
##                               (1)               (2)               (3)                (4)
## -----------------------------------------------------------------------------------------------
## treatment2               -0.057 (1.251)    0.317 (1.287)     0.233 (1.388)      -1.678 (1.303)
## cell_typesmall cell    0.610*** (0.207) -0.688*** (0.205) -0.698*** (0.223)  -0.852*** (0.240)
## months_diag               0.010 (0.017)    -0.009 (0.017)    -0.010 (0.019)      -0.018 (0.021)
## age_years                 0.002 (0.012)    -0.005 (0.012)    -0.004 (0.013)      -0.004 (0.014)
## pri_chemo10               -0.171 (0.232)    0.247 (0.232)     0.229 (0.253)      -0.070 (0.278)
## treatment2:age_years      0.004 (0.021)    -0.007 (0.021)    -0.006 (0.023)      0.020 (0.021)
## treatment2:months_diag  -0.002 (0.019)   -0.002 (0.019)    0.0001 (0.021)      0.017 (0.024)
## Constant                                 5.409*** (0.703)  5.357*** (0.775)    5.113*** (0.847)
## -----------------------------------------------------------------------------------------------
## Observations                  137               137               137                137
## R2                          0.079
## Max. Possible R2            0.999
## Log Likelihood             -500.222         -742.473          -741.405           -743.262
## chi2 (df = 7)                                17.496**          13.373*            14.007*
## Wald Test                12.000 (df = 7)
## LR Test                  11.324 (df = 7)
## Score (Logrank) Test     12.329* (df = 7)
## ===============================================================================================
## Note:                                                                *p<0.1; **p<0.05; ***p<0.01
```

[Selected Model = Cox PH] – [Reason for choosing Semi-Parametric Model Over Parametric Models is that these models assume that there is a constant hazard function which in most cases is unrealistic. So, we decided to go with the Cox PH model as it does not make any assumptions in the hazard (Risk of event occurring) function.]

1. Semi-Parametric Model.

- Cox proportional hazard model provides an estimate of the hazard ratio, which is an estimate of the relative risk in the treated group (Standard Treatment) vs the control group (Test Treatment). In other words, we can estimate the hazard rate which is the probability that if the event (status – Patient dead or not-dead) has not already occurred, it will occur in the next time interval (survival_days).

	Hazard Rate: Probability that a patient dies
Patients in test treatment have decreased risk of death by	**6%** than those in the standard treatment
Patients with small cell cancer have increased risk of death by	**84%** than those with non-small cell cancer
Patients with higher months from diagnosis have increased risk of death by	**1%** than those in the lower months from diagnosis
Older patients have increased risk of death by	**0.1%** than the younger patients
Patients with prior chemotherapy have decreased risk of death by	**16%** than those with no prior chemotherapy
Older patients in test treatment have increased risk of death by	**0.4%** than those in the standard treatment
Patients in test treatment with higher months from diagnosis have decreased risk of death by	**0.3%** than those in the standard treatment

From the analysis, age and months from diagnosis does not seem to have any significant impact on the hazard rate. However, the type of cancer seems to have a significant effect and could be worth a deeper dive.

2. Parametric Models.

- Exponential, Weibull, Log-logistic models etc. provide an estimate of the survival rate. As hazard rate and survival rate have opposite slopes, this is reflected in the sign of the estimates in most cases. We get the following insights:

	Survival Rate: Time until the patient dies		
	Exponential	**Weibull**	**Log-logistic**
Patients in test treatment have	**32% increase** in time until death than patients in standard treatment	**23% increase** in time until death than patients in standard treatment	**168% decrease** in time until death than patients in standard treatment
Patients with small cell cancer have	**69% decrease** in time until death than patients with non-small cell cancer	**70% decrease** in time until death than patients with non-small cell cancer	**85% decrease** in time until death than patients with non-small cell cancer
Patients with unit increase in months from diagnosis have	**1% decrease** in time until death	**1% decrease** in time until death	**2% decrease** in time until death
Unit increase in patients age has	**0.5% decrease** in time until death	**0.4% decrease** in time until death	**0.4% decrease** in time until death
Patients with prior chemotherapy have	**25% increase** in time until death	**23% increase** in time until death	**7% decrease** in time until death
Unit increase in patients age in test treatment group has	**0.7% decrease** in time until death than patients in standard treatment group	**0.6% decrease** in time until death than patients in standard treatment group	**2% increase** in time until death than patients in standard treatment group
Patients with unit increase in months from diagnosis from test treatment group have	**0.2% decrease** in time until death than patients in standard treatment group	**0.01% increase** in time until death than patients in standard treatment group	**2% increase** in time until death than patients in standard treatment group

Like before, age and months from diagnosis does not seem to have any significant impact on the survival rate but the type of cancer does.