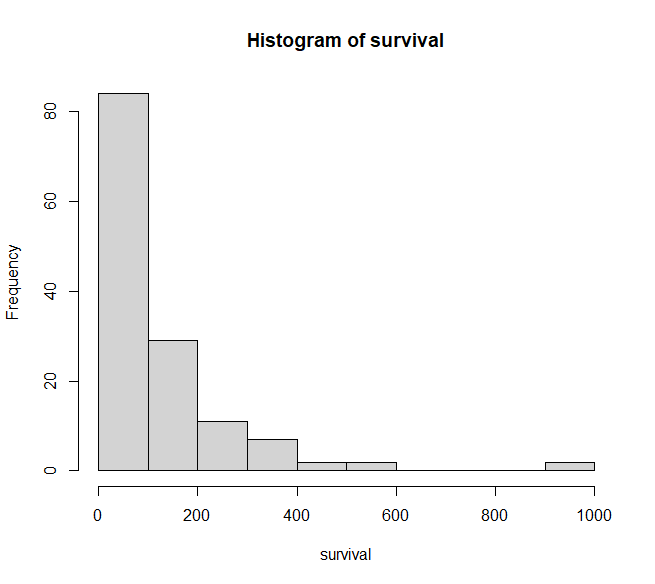
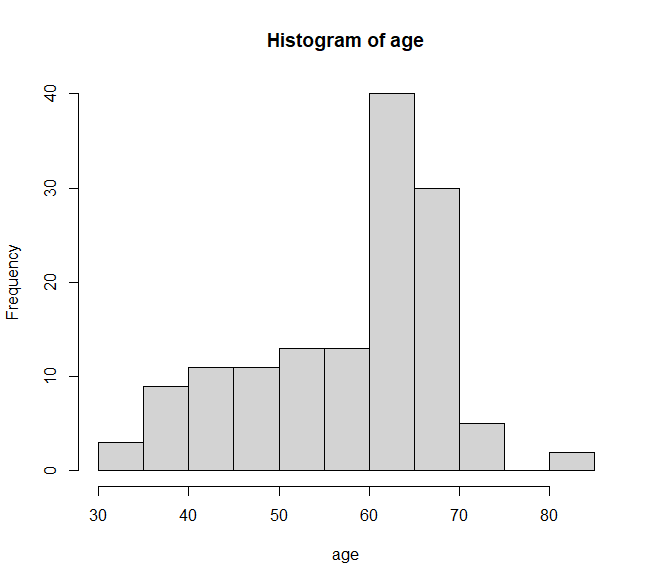
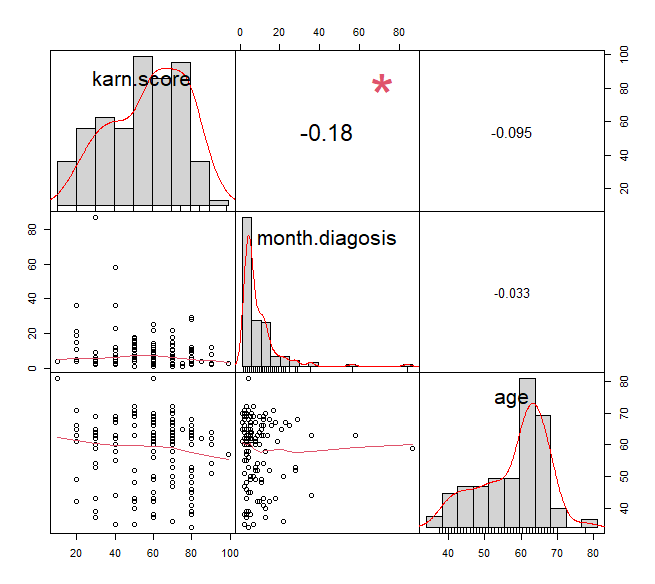
**Assignment 6: Lung Cancer Survival Analysis**

* Feature Engineering-
* Converted treatment, cell.type and prior.chemo to factors.
* Created a new variable called cancer to differentiate between SCLC (Small cell lung cancer) and NSCLC (Non-small cell lung cancer) by treating “small-cell” as “SCLC” and rest as “NSCLC”.
* Data Visualizations-



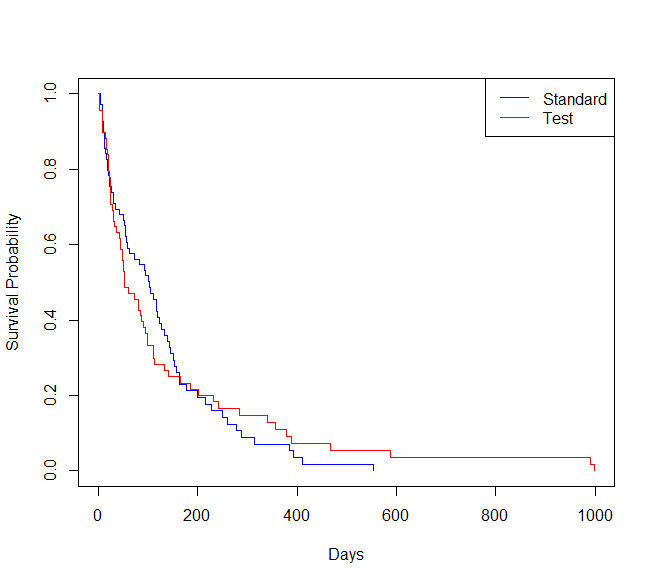
* The median survival time is 80 days, with a mean survival time of 121.6 days. The range of survival times is quite large, with the minimum survival time being only 1 day and the maximum being 999 days.
* The minimum age in the data is 34 years, while the maximum age is 81 years. The median age is 62 years, and the mean age is 58.31 years.



* No significant correlation between the variables.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Rationale** |
| *DV: y <- Surv(survival, status) = survival probability* | | |
| Treatment | +/- | The treatment type (Test or Standard) Could have either effect on days of survival. |
| cancer | +/- | SCLC is typically more aggressive than NSCLC. |
| Karnofsky score | + | A higher Karnofsky score is associated with better overall health and functional status, which could potentially contribute to longer survival times. |
| Months from diagnosis | +/- | A longer time from diagnosis may suggest a slower-growing or less aggressive cancer, which could potentially contribute to longer survival times. |
| Age | +/- | Lower age might have higher survival probability than Higher age. |
| Prior chemo | +/- | In general, patients who have received prior chemotherapy may have a better chance of survival compared to those who have not received chemotherapy. |
| Excluded |  |  |
| Cell type | +/- | Captured by cancer variable. |

* Added interaction Term combing treatment and age because certain treatment might work better or worse based on the age of a person.

Kaplan-Meier Graphs with Standard and Test Treatment

Based on the graphs, the 'Standard treatment' group has a slightly higher survival probability at all time points compared to the ‘Test treatment’ group indicating that the chemo treatment alone has higher survival probability compared to chemo + drug treatment.

Q1a. What is the probability that the patient will survive for 6 months (183 days) and 1 year (365 days) on the standard treatment vs the test treatment?

**For 183 days:** The probability that the patient will survive for 6 months (183 days) is 21.24 % on the standard treatment and 23.29 % for Test treatment.

**For 365 days:** The probability that the patient will survive for 12 months (365 days) on the standard treatment is 7 % and on test treatment is 10.98 %.

Q1b. What is the median number of days where a patient can be expected to survive if they are on the standard vs the test treatment?

The median number of days on Standard treatment is 100 days and 52 days on Test treatment.

Q2. Create three semi-parametric and parametric models to estimate the marginal effects of relevant predictors on survival outcomes. Interpret the precise effects of standard vs test treatment and other model predictors.

==================================================================================

Dependent variable:

-------------------------------------------------------------

y

Cox exponential Weibull survreg: loglogistic

prop. hazards

(1) (2) (3) (4)

----------------------------------------------------------------------------------

treatment2 0.089 -0.005 -0.008 -1.366

(1.131) (1.136) (1.142) (1.075)

karn.score -0.033\*\*\* 0.033\*\*\* 0.033\*\*\* 0.037\*\*\*

(0.005) (0.005) (0.005) (0.005)

month.diagosis -0.001 0.0001 0.0001 0.005

(0.009) (0.009) (0.009) (0.010)

age -0.008 0.005 0.005 0.002

(0.012) (0.012) (0.012) (0.012)

prior.chemo10 -0.022 0.066 0.066 0.005

(0.235) (0.234) (0.235) (0.222)

cancerSCLS 0.400\* -0.412\*\* -0.413\*\* -0.470\*\*

(0.211) (0.209) (0.210) (0.197)

treatment2:age 0.004 -0.004 -0.004 0.021

(0.019) (0.019) (0.019) (0.018)

Constant 2.728\*\*\* 2.723\*\*\* 2.066\*\*\*

(0.807) (0.812) (0.793)

----------------------------------------------------------------------------------

Observations 137 137 137 137

R2 0.289

Max. Possible R2 0.999

Log Likelihood -482.058 -723.756 -723.753 -716.363

chi2 (df = 7) 54.931\*\*\* 48.677\*\*\* 67.805\*\*\*

Wald Test 47.600\*\*\* (df = 7)

LR Test 46.781\*\*\* (df = 7)

Score (Logrank) Test 50.634\*\*\* (df = 7)

==================================================================================

We can see that the coefficients of the variables are almost similar in magnitude, but the signs are flipped for parametric models compared to cox ph model which is a semi-parametric model that does not require assumptions about the underlying distribution of the survival times and is more flexible in terms of modeling the effect of covariates on the hazard function. Additionally, the Cox model allows for the estimation of hazard ratios, which provides a more interpretable measure of the effect of covariates on survival outcomes.

**Interpreting Cox PH Model-**

* Beta(treatment2) > 0. The positive coefficient for treatment 2 suggests that patients on test treatment have a higher risk of death compared to patients on standard treatment. In other words, patients on treatment 2 (test treatment) have a 9.3% higher risk of dying compared to patients on treatment 1(standard treatment), all else being equal.
* Beta(karn.score) < 0: The negative coefficient for Karnofsky performance score suggests that patients with a higher score (i.e., better functional status) have a lower risk of death. Specifically, a one order (1 to 10, or 10 to 100 ) increase in Karnofsky score is associated with a 4% decrease in the risk of death, all else being equal.
* Beta(month.diagnosis) < 0: The negative coefficient for months since diagnosis suggests that patients with a longer time since diagnosis have a lower risk of death. Specifically, a one-month increase in time since diagnosis is associated with a 0.12% decrease in the risk of death, all else being equal.
* Beta(age) < 0: The negative coefficient for age suggests that older patients have a slightly lower risk of death, although the effect is not statistically significant. Specifically, a one-year increase in age is associated with a 0.78% decrease in the risk of death, all else being equal.
* Beta(prior.chemo10) < 0: The negative coefficient for prior chemotherapy suggests that patients who had prior chemotherapy have a slightly lower risk of death. Specifically, being on prior chemotherapy is associated with a 2.2% decrease in the risk of death, all else being equal.
* Beta(cancerSCLS) > 0: The positive coefficient for cancer cell type: SCLS suggests that patients with small-cell type have a higher risk of death compared to patients with NSCLC (adeno, squamous and large). Specifically, having SCLC type cell is associated with a 49.1% increase in the risk of death, all else being equal.
* Beta(treatment2:age) > 0: The positive coefficient for the interaction between test treatment and age suggests that the effect of test treatment on the risk of death varies with age. Specifically, for each one-year increase in age, the effect of test treatment on the risk of death increases by 0.36%, all else being equal.